7-26-2020

Perceptual Learning of Noise-Vocoded Speech

Julia R. Drouin
University of Connecticut, julia.drouin@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
https://opencommons.uconn.edu/dissertations/2588
Perceptual Learning of Noise-Vocoded Speech
Julia R. Drouin, Ph.D.
University of Connecticut, 2020

A challenge in the cochlear implant adaptation literature is accounting for wide variability in speech perception outcomes following implantation. While many preimplantation factors contribute to the variance observed in outcomes, formal auditory training has been proposed as a way to maximize speech comprehension benefits for cochlear implant users. However, there is currently no standardized set of recommendations as to how training should be implemented in the patient population. The goal of this dissertation was to examine two potential training variables of interest: time of day in which the training occurs and the task structure of training. This dissertation examined these factors in the context of speech learning for normal hearing listeners adapting to degraded speech input as a first step towards identifying training parameters that could hold clinical utility.

In chapter one, an interdisciplinary perspective is used to bridge the clinical rehabilitation literature with basic research examining perceptual learning of speech. Findings demonstrating a facilitative role of auditory training on adaptation to degraded speech signals in normal hearing listeners are reviewed in the context of rehabilitation for cochlear implant users. In chapter two, we examine the influence of sleep-based memory consolidation on perceptual learning of degraded speech, which has been previously shown to improve perception of other novel speech inputs. For listeners trained in close proximity to sleep, a significant stabilization effect was observed for trained and novel items that persisted at a one-week follow-up interval. This finding demonstrates a significant influence of sleep-based learning paradigms on perceptual learning for acoustically degraded speech signals. Finally, in chapter three, we examined the role of training
task on perceptual learning outcomes both in the immediate and a later time interval. Across training groups, listeners demonstrated robust perceptual learning, regardless of whether the task was linguistic or non-linguistic and in the absence of linguistic feedback. This finding suggests alternate routes to degraded speech learning that may be customized based on the needs of the listener. Each section includes recommendations for future research that could foster translation of principles of speech learning in normal hearing listeners to aural rehabilitation protocols for cochlear implant patients.
Perceptual Learning of Noise-Vocoded Speech

Julia R. Drouin

B.A., University of Connecticut, 2014
Au.D., University of Connecticut, 2020

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2020
Copyright by

Julia R. Drouin

2020

ii
APPROVAL PAGE

Doctor of Philosophy Dissertation

Perceptual Learning of Noise-Vocoded Speech

Presented by
Julia R. Drouin, B.A., Au.D.

Major Advisor
__________________________________
Rachel M. Theodore

Associate Advisor
__________________________________
Emily B. Myers

Associate Advisor
__________________________________
Erika Skoe

Associate Advisor
__________________________________
James S. Magnuson

University of Connecticut
2020

iii
Acknowledgements

My graduate school career was truly made possible because of the unwavering support and patience from my advisor, Dr. Rachel Theodore. As a new student who really struggled to find my confidence and voice early on in graduate school, Dr. Theodore was my biggest advocate every step of the way. Every time I wanted to push my research in a new direction, she was there to encourage me to go there – reminding me that I would not be alone in my journey, but rather that we would tackle the new challenges as a team. One of the most impressive attributes of Dr. Theodore is her willingness to not just learn new things, but master them (and very quickly for that matter) – this is a quality not often seen in tenured faculty members and a model I hope to follow in my academic career. Through the many hours we’ve put in together on various projects, I’ve been reminded that good science is hard, but a sense of humor and surrounding yourself with a great team makes the process a little easier and fun. I’m grateful to have had the opportunity to complete my training with Dr. Theodore and look to her as the standard as I embark on my own academic career.

I am grateful to have completed my training at an institution with an infrastructure set up to support students pursuing language research. As a consequence, I’ve had the opportunity to work with faculty and students from various departments and backgrounds and engage in a community where I felt comfortable communicating across disciplines. I believe this training will help me as I work towards translational research in the audiology field. I thank my committee members and readers Dr. Emily Myers, Dr. Erika Skoe, Dr. James Magnuson, and Dr. Lendra Friesen for their useful feedback and comments on this dissertation and tremendous support throughout my graduate school career. I also thank the UConn Neurobiology of Language community which is where I’ve been able to meet friends and colleagues that will
extend well beyond our UConn years. Thank you to my lab mates, Nick Monto, Nikole Giovannone, and Lee Drown for their friendship, humor, and support – I’ve been so impressed and inspired to see their progress as they’ve carved out their own research trajectories. I am also deeply grateful to have had the support and friendship of Pam Fuhrmeister throughout graduate school. Tackling all of our milestones together, Pam has made the Ph.D. process so much more enjoyable and I’m excited to see where our Ph.D. training takes us in the future.

Finally, on a personal note, I’m thankful to have supportive and caring parents and brothers who don’t always know what I do (and keep asking when I’ll finally be done with school), but nonetheless are my biggest cheerleaders to keep going. My mom in particular was the person who inspired me to pursue a career in science and my dad was the motivation to never doubt myself and my accomplishments. My three brothers were always there to remind me of the real world and my life outside of school. I feel incredibly lucky to have had the overwhelming support and love from my family every step of the way in graduate school. Finally, I want to thank my significant other (and soon to be husband), Charles, for his constant encouragement throughout graduate school. Meeting him has without doubt been the best thing to come out of my graduate school career and I’m so excited for our next chapter.
Foreword

This dissertation reviews aspects of perceptual learning for degraded speech signals in chapter one, and examines two potential training variables of interest in chapter two and chapter three. The review outlined in chapter one and the empirical work outlined in chapter two and chapter three were written as separate manuscripts for publication. As such, each chapter is intended to stand on its own.
# Table of Contents

## I. Leveraging interdisciplinary perspectives to optimize auditory training for cochlear implant users

1. Introduction ........................................................................................................................................... 1
2. Current forms of active auditory training in aural rehabilitation ......................................................... 6
   2.1. Bottom-up focused auditory training ................................................................................................. 8
   2.2. Top-down focused auditory training ................................................................................................ 11
3. Lexical influences on speech perception ................................................................................................. 13
   3.1. Lexical effects in NH listeners ........................................................................................................... 13
   3.2. Lexical effects in CI users ............................................................................................................... 15
4. Factors that promote perceptual learning for degraded speech signals ..................................................... 16
   4.1. A role for lexical information ............................................................................................................ 17
   4.2. Alternate routes to perceptual learning .............................................................................................. 19
5. Translating principles of learning to auditory training for CI users ......................................................... 22

## II. Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech

1. Introduction ............................................................................................................................................ 25
   1.1. Role of sleep consolidation on perceptual learning of speech ......................................................... 27
   1.2. Maintenance of gains following perceptual learning ........................................................................ 30
   1.3. The current study ............................................................................................................................ 32
2. Methods .................................................................................................................................................. 34
   2.1. Participants ........................................................................................................................................ 34
   2.2. Stimuli .............................................................................................................................................. 34
   2.3. Procedure ......................................................................................................................................... 35
3. Results ................................................................................................................................................... 38
   3.1. Training ........................................................................................................................................... 38
   3.2. Test .................................................................................................................................................. 39
4. Discussion .............................................................................................................................................. 48
   4.1. Effects of training on immediate perceptual learning ....................................................................... 48
   4.2. Role of sleep-based consolidation on learning ............................................................................... 50
   4.3. Maintenance of learning at a later time interval ............................................................................. 52
   4.4. Clinical implications and future directions ..................................................................................... 54

## III. Many tasks, same outcome: Role of training task on learning and maintenance of noise-vocoded speech

1. Introduction ............................................................................................................................................ 56
   1.1. Role of lexical context on perceptual learning for noise-vocoded speech ....................................... 57
   1.2. Role of non-linguistic training tasks on perceptual learning for speech ....................................... 61
   1.3. The current study ............................................................................................................................ 66
2. Methods .................................................................................................................................................. 70

vii
Supplemental Material ................................................................. 96
References .................................................................................. 102
Addendum ................................................................................... 118
I. Leveraging interdisciplinary perspectives to optimize auditory training for cochlear implant users

1. Introduction

Over the past 50 years, significant advancements have been made in the clinical management of severe to profound sensorineural hearing loss. Cochlear implants (CIs) are now widely considered to be an effective treatment option for both pediatric and adult patients. CIs are surgically implanted medical devices that use electrical stimulation of the auditory nerve to create the sensation of hearing. In the CI system, acoustic energy is detected by an external microphone and sent to a speech processor that is worn on the head. The speech processor is involved in extracting relevant acoustic features and determining electrode stimulation patterns. The signal is then transmitted via radio waves to an internal component, which transforms the input into electrical pulses to stimulate the auditory nerve (e.g., Loizou, 1998). Current electrode arrays are designed to provide stimulation for up to 22 intra-cochlear electrodes. The way in which electrodes are stimulated varies depending on the selected processing strategy. For example, in a continuous interleaved (CIS) processing strategy, acoustic information is bandpass-filtered based on the number of electrodes, and electrical pulses are generated using amplitudes proportional to the amount of energy contained in each bandpass channel. Electrodes are stimulated in a non-simultaneous fashion to minimize channel interaction (e.g., Loizou, 1998; Loizou, 2006). Conversely, an advanced combination encoder (ACE) processing strategy uses a subset of the available electrodes with stimulation occurring for frequency bands that contain the largest amplitudes (e.g., Hu & Loizou, 2008).
Listening through an implant is a markedly different experience than acoustic hearing. In the normal-hearing (NH) system, the cochlea contains thousands of hair cells that code fine-grained differences in frequency, which supports a high-resolution representation of auditory information in the environment. Conversely, most CI users only have up to 22 electrodes available to represent the auditory signal, which results in a coarser signal representation. As a consequence, the implant does not afford access to the fine-grained spectral resolution that is available to NH listeners. Paradoxically, studies examining speech perception outcomes in CI users have demonstrated that listeners do not necessarily need access to the richness of spectral information present in the natural speech signal in order to achieve reliable speech comprehension – many CI users recognize speech transmitted through the implant, as do NH participants listening to spectrally degraded speech (e.g., Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). However, there is tremendous individual variability in speech perception outcomes among CI users. While some users show performance at or near the typical range, many others do not (e.g., Busby, Roberts, Tong, & Clark 1991; Dawson & Clark, 1997; Niparko et al., 2010). A variety of patient-related factors have been examined to determine their relevant contributions to speech perception outcomes. Research has found that many factors contribute to patient performance including age of implantation (e.g., Gantz et al., 1988; Parkin, Stewart, Dankowski, & Haas, 1990; Shea, Domio, & Orchik, 1990), pre- vs. post-lingual language status at time of implantation (e.g., Dawson et al., 1992), etiology of deafness (e.g., Battmer et al., 1995), and residual hearing following implantation (e.g., Gantz et al., 1988; van Dijk, van Olphen, Langereis, Mens, Brokx, & Smoorenburg, 1999; Friedland, Venick, & Niparko, 2003). In addition, other work has sought to link parameters of the device itself to outcome performance including number of active electrodes (e.g., Friesen, Shannon,
Baskent, & Wang, 2001), coding strategy (e.g., Skinner, Holden, Whitford, Plant, Psarros, & Holden, 2002), and insertion depth (e.g., Finley & Skinner, 2009). Both patient- and device-related factors have been shown to partially account for the variance observed in patient outcomes following implantation. However, to date there is no comprehensive set of predictors that can reliably account for differences in outcome performance observed with the device. That is, while a substantial portion our current knowledge base has justified the efficacy of the CI in order to demonstrate a benefit of the device at an implementation level (e.g., Pisoni, Kronenberger, Harris & Moberly, 2017), we have a more limited understanding as to why a CI works well for some patients, yet poorly for others. As a result, the current evidence base does not account for why patients with comparable language status, medical history, demographic background, and device may perform differently with the implant (e.g., Moberly, Bates, Harris, & Pisoni, 2016; Pisoni, Cleary, Geers, & Tobey, 1999).

One account as to why some CI users show poor performance outcomes is the lack of availability of acoustic cues. However, research examining NH participants has found that individuals can recognize speech in the absence of acoustic cues that they routinely have access to. For example, listeners readily demonstrate comprehension of compressed speech (Licklider & Pollack, 1948), sine-wave speech (e.g., Remez, Rubin, Pisoni, & Carrell, 1981), and noise-vocoded speech (e.g., Shannon et al., 1995), suggesting that availability of all acoustic cues is not necessary to achieve comprehension – instead listeners dynamically learn to adapt to the input to which they are exposed. This is an important consideration in the context of CI rehabilitation as it demonstrates that listeners do not necessarily need access to the full spectrum of acoustic cues in natural speech because adaptation mechanisms can help facilitate comprehension for acoustically degraded input. Thus, an alternate approach to understanding
performance outcomes in CI users might not be which acoustic cues are available per se, but rather how the individual learns to adapt to the input transmitted through the CI.

One proposed technique to improve listening through a CI is auditory training. In general, auditory training can be characterized as passive (i.e., incidental) or active training. In passive auditory training, listeners learn to adapt to the input transmitted through the CI in an unsupervised and unstructured manner. Passive training is a relatively common recommendation in the clinical domain, particularly for adult patients, as it allows the patient to integrate use of the CI into their everyday life. However, passive adaptation may be a time-consuming process, requiring frequent exposure to routine sounds in order to accurately pair an auditory stimulus to meaning. One study illustrated this point by examining the role of passive learning on adaptation for CI users undergoing an update to their processing strategy – a manipulation that requires even experienced users to adjust the mappings between auditory input and linguistic representations (Fu, Shannon, & Galvin, 2002). In this study, the frequency mapping was modified for a group of experienced CI users, who were allowed a three-month time period to passively adapt to the new parameters. They found that the CI users showed only minimum adaptation (i.e., no or modest improvements in speech recognition compared to baseline measures) in the time period of passive adjustment, though performance did improve with continued experience. These results suggest that unsupervised learning may not promote peak performance, particularly in the case where the CI map is novel or is not optimally configured for the user.

In contrast, listeners who engage in active auditory training engage in listening tasks that target a specific goal, such as phoneme, word, or sentence comprehension. As such, learning can be facilitated using rich contexts and feedback to meet the goal. Active auditory training may be
beneficial for all CI users; it can help poor performers by providing structured opportunities to learn meaningful linguistic information, and can aid high performers by providing learning opportunities under difficult listening conditions (e.g., poor signal-to-noise ratio, novel talkers; Fu & Galvin, 2008). Moreover, it is common for CI manufacturers to update processing strategies and technology regularly, requiring even experienced CI users to adapt to novel input. However, few CI users participate in active auditory training as part of rehabilitation, with even fewer programs targeted at post-lingual deafened adults (Fu & Galvin, 2007; Prendergast & Kelley, 2002). As a consequence, most CI users adapt to the implant in a passive fashion, though growing research suggests that this form of adaptation may not be sufficient to maximize listeners’ performance (e.g., Fu & Galvin, 2008).

Many critical challenges exist for implementing auditory training as part of standardized aural rehabilitation. First, there is significant variability in speech perception outcomes following training; some individuals demonstrate a benefit, while others do not (Fu et al., 2005; Wu, Yang, Lin, & Fu, 2007; Busby et al., 1991). As a consequence, it is unclear how recommendations might differ across patients, and whether training may only be beneficial for patients demonstrating poor performance relative to patients who perform at a high level. In addition, because there are no concrete clinical recommendations on what the structure of active training should be (i.e., tasks, training duration, outcome measures), it is likely that few clinicians prioritize training as part of aural rehabilitation given the lack of evidence-based recommendations. Another challenge is that gains are often only measured immediately following training, with few studies examining whether performance is maintained over time, which is a critical factor to establish for any rehabilitation program. Finally, even in cases where auditory training appears to promote improvements in speech understanding, the outcome
variables (e.g., task accuracy) may not explain why the improvement occurred because the measures do not index mechanistic processing changes. Standard clinical assessments of speech perception outcomes include word (e.g., Maryland CNC lists) and sentence (e.g., AzBio lists) comprehension in quiet and noise. Using these metrics alone, it is difficult to determine why one patient shows a greater benefit from training relative to a patient who has a similar profile yet gleans little benefit from training. The use of other physiologic measures may better assess changes at lower processing levels that precede changes in behavioral responses, though the relationship between physiologic changes and mechanistic changes is an area of continued research.

Here we review current forms of active auditory training including bottom-up focused training, in which attention is directed towards fine-grained acoustic differences in the signal, and top-down focused training paradigms, in which attention is directed towards the global structure of a sentence and/or contextual cues. For the nature of this review, our focus is on examining task-specific differences across training types with the population of interest being post-lingual adults, unless otherwise stated. We adopt an interdisciplinary perspective to bridge findings from the clinical rehabilitation literature with findings from the psycholinguistics domain, highlighting the role of top-down lexical feedback on adaptation to acoustically degraded speech in NH listeners. In this review, we argue that active auditory training may be necessary to optimally target the plasticity mechanisms underlying improvements in speech perception for listeners adapting to a CI.

2. Current forms of active auditory training in aural rehabilitation

Formal auditory training is aimed at enhancing auditory skills and improving speech understanding through a variety of listening exercises (Sweetow & Palmer, 2005; Sweetow &
Sabes, 2007; Rayes, Malky, & Vickers, 2019). It has been argued that auditory training should be a standard part of aural rehabilitation as a way to structure how the listener learns to understand the transmitted signal. However, few CI users engage in formal training, with even fewer programs targeted at post-lingual deafened adults, which may be due to a failure to prioritize training as part of the aural rehabilitation process for these users (Fu & Galvin, 2007; Prendergast & Kelly, 2002). Thus, current training protocols are often targeted at the pediatric population who demonstrate greater degrees of plasticity. While neural structure and function is maximally plastic during the first few years of life, plasticity mechanisms are present throughout the lifespan (e.g., Schramm, Fitzpatrick, & Seguin, 2002; Tong, Busby, & Clark, 1988; Sharma, Gilley, Dorman, & Baldwin, 2007; Most, Shrem, & Duvdevani, 2010). Adults continue to learn new skills, but often require more exposure or practice to do so. This principle underlies flexibility in the CI candidacy criterion, which now allows pre-lingual or post-lingual deafened adults the opportunity to receive an implant. Theories of residual plasticity would predict that implanted adults could reach the performance of younger implanted individuals with sufficient and prolonged experience (Kral & Sharma, 2012). However, research has shown that late-implanted individuals, in this case pre-lingual deafened participants over the age of 12, who used their CI for an extended period of time (i.e., at least six months post-implantation) showed tremendous variability in open-set sentence recognition (Schramm, Fitzpatrick, & Segun, 2002). Of the 15 participants in their sample, 10 participants showed an improvement from pre- to post-implantation performance on open set word recognition. However, nine participants performed at or below 40% performance, a level considered to represent minimal benefit. Thus, not all of the patients showed an improvement post-implantation, and of those patients who did demonstrate improvement, most of the sample did not perform at a level consistent with an acceptable
benefit. Thus, the results of Schramm et al. (2002) indicate that extended passive listening through a CI may not maximize speech understanding. Studies following newly implanted individuals suggest that most gains occur within the first 3–6 months following implantation (Kessler, Loeb, & Barker, 1995; Spivak & Waltzman, 1990; Waltzman, Cohen, & Shapiro, 1986; Fu & Galvin, 2008), with less improvement observed after this time point (e.g., Dorman, Loizou, & Rainey, 1997; Pelizzone & Cosendai, & Tinembart, 1999; Helms et al., 2004). This has prompted discussion in the audiological field that CI users may need other rehabilitative support to fully acclimate to their device (e.g., Fu & Galvin, 2008).

There are two primary approaches to active auditory training for rehabilitation: bottom-up focused training and top-down focused training. Both approaches have the same end goal of improving speech understanding for CI users, though they use different strategies to meet that goal. Bottom-up approaches, also called analytic training, focus on the building blocks of speech perception, including sensitivity to fine-grained acoustic details in the speech signal. Top-down approaches, also called synthetic training, focus on training the listener to use lexical and contextual cues to fill in perceptual gaps while processing speech (Fu & Galvin, 2007).

2.1. Bottom-up focused auditory training

The bottom-up approach encourages listeners to focus on the acoustic signal itself. True bottom-up training is designed to promote improvements in processing efficiency of the sensory signal (Fu & Galvin, 2007). This is particularly prominent for CI users as they experience a spectral mismatch in perception because the auditory input is filtered through a limited number of electrodes. While the amplitude envelope within each channel is relatively preserved, the spectral structure of the channel is poorly conveyed (e.g., Shannon et al., 1995; Nelson, Jin, Carney and Nelson, 2003; Moberly et al., 2014). Previous research has demonstrated that NH
listeners rely on a variety of spectral and temporal cues for phonemic decisions including formants (e.g., Hillenbrand, Getty, Clark, Wheeler, 1995; Neel, 2008), formant transitions (e.g., Delattre, Liberman, Cooper, 1955; Stevens & Klatt, 1974; Stevens & Blumstein, 1978; Walley & Carrell, 1983), spectral center (e.g., van Son & Pols, 1997), and periodicity (e.g., Cole & Cooper, 1975). A critical question is the degree to which CI users rely on the same acoustic cues that NH listeners use to achieve comprehension. Research has found evidence that some CI users demonstrate similar acoustic cue weighting strategies to that of NH listeners. For example, Iverson, Smith, and Evans (2006) found that vowel recognition in CI users was adversely affected when the speech signal was processed to remove formant movement or equate vowel duration, suggesting that these cues, as with NH listeners, are strongly weighted in perception for CI users. However, there is also evidence suggesting that CI users may weight acoustic cues differently than NH listeners. For example, Hedrick and Carney (1997) studied four post-lingual deafened CI users to determine the relative contributions of amplitude and formant transition information on consonant-vowel identification. They found that compared to a NH control group, the CI users consistently relied more on the relative amplitude cues while the NH listeners used both formant structure and amplitude information, suggesting a difference in perceptual weighting strategies between the groups. Similar findings were demonstrated by Moberly et al. (2014), who examined labeling patterns of the /ba/-/wa/ contrast in a sample of 20 post-lingual CI users. While NH listeners weighted the spectral structure over amplitude structure exclusively, CI users showed variability in which cues were prioritized. Some users relied on amplitude structure to a greater degree, in line with a theory of shifted perceptual strategies in which CI users weight the acoustic cues that are most saliently delivered. However, it was found
that the CI users who adopted a strategy closer to NH listeners, in which spectral structure was weighted over amplitude cues, demonstrated the best word recognition performance overall.

It is thought that bottom-up training approaches promote changes in how acoustic information is weighted by providing structured training in which listeners engage in tasks that direct attention towards fine-grained, meaningful acoustic cues. A limited number of studies have examined bottom-up auditory training in CI users and results have been equivocal. For example, Busby, Roberts, Tong, and Clark (1991) did not observe significant improvements in vowel perception for a small sample of pre-lingual deafened CI patients who completed 10 one-hour training sessions. Conversely, Dawson and Clark (1997) examined five CI patients following 10 weeks of bottom-up focused training sessions using explicit vowel training tasks that focused on rhyme generation, discrimination, and identification. They found that four out of the five participants showed some improvement in at least one of the trained areas, and one patient showed generalization to novel contexts. Similar results were reported by Fu, Galvin, Wang, and Nogaki (2005), who examined 10 CI patients who completed at-home training that was customized based on their baseline performance. Participants completed a three-alternative forced-choice discrimination task. On each trial of this task, participants heard two identical sounds and one outlier sound that differed maximally with respect to acoustic features. Over the course of the experiment, the outlier sound became increasingly similar to the identical sounds, making the experiment more difficult as it progressed. Participants also completed three-alternative forced-choice vowel, consonant, and nonsense syllable identification tasks that also increased in difficulty with improved performance. They found improvements in vowel and consonant recognition, with some participants also showing generalization to sentence recognition tasks.
Overall, it is difficult to conclude whether strict bottom-up focused training promotes significant improvements in speech comprehension as these studies suffer from limited sample sizes, document wide variability in benefits across patients, and show limited generalization. However, the remaining low-level acoustic cues available to CI users may be more amenable to auditory training and may play a greater role for attending to speech in acoustically simple environments, like listening in quiet (e.g., Fu & Galvin, 2007). A key limitation with the bottom-up approach is that the tasks used during training may be considered less functionally relevant than top-down approaches. Specifically, training focused on recognition of single phonemes is not typically a task required to achieve comprehension under everyday listening conditions. Instead, most listeners require comprehension of a continuous stream of sentences in everyday environments and have access to broad contextual cues. Spoken language comprehension occurs through interactive processes at both the low-level acoustic level and higher-order linguistic level. Future research is needed to explicate the degree to which bottom-up focused training generalizes to more functionally relevant communication tasks, such as single words or sentences.

2.2. Top-down focused auditory training

Top-down based approaches to auditory training emphasize the use of lexical and contextual cues to decode the speech signal. This type of training targets skills that support auditory processing and language skills. While bottom-up approaches first direct attention towards the fine-grained structure of the input, top-down approaches first direct attention towards the global structure of the signal, which is thought to closely mirror how NH listeners learn language (Nittrouer & Caldwell-Tarr, 2016). Unlike analytic approaches, the focus of the top-down approach is on using contextual cues to facilitate comprehension. Materials and
training tasks for top-down based auditory training include transcribing words or sentences under clear or challenging listening conditions, such as in background noise or with multiple talkers. One common form of top-down auditory training is connected discourse tracking (DeFilippo & Scott, 1978). This type of training requires a talker (sender) and listener (receiver) who create a set of speech materials for the talker to read. The talker reads the prepared text and the receiver must repeat back exactly what was said, with access to visual and/or auditory information. If the listener makes an error, then the talker may paraphrase the materials or prompt the listener with additional cues (Levitt, Waltzman, Shapiro, Cohen, 1986). Clinicians assess the number of words correct per minute and can compare performance in auditory only, visual only, or audiovisual conditions to determine individual contributions of each domain on comprehension. Research utilizing connected discourse tracking has found variable results on its efficacy. Levitt et al. (1986) examined performance of five CI patients using Connected Discourse Tracking over a 10-week training period. They report that all five subjects showed significant improvements in tracking rate, though the magnitude of learning differed considerably, and none of the participants reached the threshold considered within normal listening levels. The use of connected discourse tracking has been a topic of debate in the audiology domain because the paradigm has many uncontrolled variables (e.g., talker differences, listener differences, text materials, repeated presentations) and the guidelines to compare performance within subjects are not well established (Tye-Murray & Tyler, 1988). However, use of top-down focused training paradigms may hold promise for aural rehabilitation, as they are commonly used in the psycholinguistics literature for adaptation to degraded speech input for NH listeners. In the following section, we review the role of lexical information on speech perception for NH
listeners, and highlight findings demonstrating that contextual cues play an influential role for adaptation to atypical speech input.

3. Lexical influences on speech perception

3.1. Lexical effects in NH users

Everyday listening environments present a challenge for even the most skilled NH individuals because listening often occurs in suboptimal environments and talkers show a high degree of variability in the acoustic realization of speech sounds (e.g., Delattre, Liberman, Cooper, 1955; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Peterson & Barney, 1952; Jusczyk, 1997; Theodore, Miller, & DeSteno, 2009; Newman et al., 2001). These findings have been formalized as the “lack of invariance” problem for speech perception, which reflects the fact that there is no one-to-one mapping between the acoustic signal and a given speech sound. However, structured variability is not inherently negative as it can give listeners information about the talker, for example, which listeners can use to dynamically modify the structure of phonetic category representations (e.g., Allen, Miller, DeSteno, 2003; Theodore, Miller, & DeSteno, 2009; Clayards, Tannenhaus, Aslin, & Jacobs, 2008; Miller, 1994; Drouin, Theodore, & Myers, 2016).

Seminal research has shown that listeners use lexical information to guide speech perception, which is particularly useful when the speech signal is variable, ambiguous, or degraded. Phonemic restoration is one demonstration of a lexical effect on perception, in which listeners fill in missing or degraded input using contextual knowledge. Warren (1970) first demonstrated that when a cough or tone replaced the initial /s/ in a word like legislatures, listeners could not reliably identify when the gap occurred and many participants reported that the word was fully intact, suggesting an apparent restoration of the missing phoneme, which has
also been observed in listeners with hearing loss (Başkent, Eiler, & Edwards, 2010). The lexical effect has also been observed with perception of ambiguous speech input. For example, Ganong (1980) exposed listeners to stimuli varying along a voice-onset-time continuum where one end of the continuum yielded perception of a lexical item (e.g., *kiss* or *gift*), while the other end did not (e.g., *giss* or *kift*). He found that listeners were biased to perceive more of the speech continuum as a /g/ when in the context of /ɪft/, where only the /g/ interpretation yields a real English word. However, when the same voice-onset-time variants were placed in the context of /ɪs/, listeners perceived more of the continuum as /k/, as only the /k/ interpretation yields a lexical interpretation.

Collectively, these findings demonstrate how speech perception for NH listeners reflects interactions at both the phonemic and lexical levels – all of which occurs with relative ease and within milliseconds as the speech signal unfolds in time. Indeed, many studies on the time course of lexical access have shown that access to even a single phoneme creates lexical competition, supporting activation of lexical candidates consistent with the phonemic input (e.g., Allopenna, Magnuson, & Tannehaus, 1998). For example, Allopenna et al. (1998) used a visual world eye-tracking paradigm in which listeners heard a target word (e.g., *beaker*) while simultaneously viewing an array of pictures containing a cohort competitor (e.g., *beetle*), a rhyme competitor (e.g., *speaker*), and an unrelated word (e.g., *dolphin*). As the first few hundred milliseconds of the target word unfolded over time, the cohort competitor showed the greatest initial competition. As the end of the target word was processed, the rhyme competitor emerged as the dominant competitor. Such patterns have been modeled using interactive accounts of spoken word recognition, such as TRACE, in which lower-level perceptual processing is altered based
on top-down lexical feedback that unfolds over time (e.g., McClelland & Elman, 1986; Allopenna et al., 1998).

3.2. Lexical effects in CI users

An important consideration is how lexical information may be weighted differently depending on the specific listening context. When the quality of the signal is good (e.g., clear speech, minimal background noise), bottom-up information may contribute more to processing. However, when the stimulus quality is poor (e.g., degraded input, background noise), top-down feedback may play a more dominant role in speech processing (e.g., Norris et al., 2003; Fu & Galvin, 2007). Under this view, top-down feedback might contribute more to speech processing for CI users because the transmitted signal contains limited spectral information. Strikingly, previous research has shown atypical patterns of lexical influences on speech perception for CI users. For example, recent work on lexical access in CI users using a visual world paradigm has demonstrated significant delays in lexical activation which may be more pronounced for pre-lingual CI users (Farris-Trimble, McMurray, Cigrand, & Tomblin, 2014; McMurray, Farris-Trimble, & Rigler, 2017). Delays in lexical access at the word level may have detrimental, cascading effects at the sentence level because the opportunity to build sentence structure is impaired as the delay lengthens. Indeed, one study found that pediatric CI users do not use sentence context to facilitate word recognition; instead, they appear to process sentences as a string of unrelated words (Conway, Deocampo, Walk, Anaya, & Pisoni, 2014). While the current evidence base suggests atypical patterns of lexical access among CI users (Conway et al., 2014; McMurray et al., 2017), the existing literature characterizing how lexical efficiency changes with experience and training in this population is extremely limited. Future research is needed to examine how auditory training for CI users may be used to foster a perceptual system that uses
lexical cues to improve processing efficiency. In the next section, we review the role of both lexically-oriented and non-linguistic training paradigms on improving perception of degraded speech signals for naïve NH listeners, which lay the foundation to promote optimal translation to the clinical population.

4. Factors that promote perceptual learning for degraded speech signals

A common occurrence in everyday listening conditions is the experience of poor speech understanding when initially parsing an acoustically degraded speech signal, which improves with sufficient exposure or with sufficient context cues. This occurs regularly in the initial activation period for CI patients, who often may not perceive speech during activation, yet show perceptual improvements with extended use. Audiologists counsel users on the importance of wearing the device regularly to improve how the brain learns the new sensory signals transmitted through the CI. This also occurs for NH listeners who may initially encounter a novel talker with an unfamiliar accent or dialect. Despite initial difficulty, listeners report a rapid improvement in their speech understanding with exposure. This phenomenon is referred to in the psychology literature as perceptual learning. Perceptual learning can be defined as adaptive changes in an organism’s perceptual system that enhance the system for future interactions (Goldstone, 1998). In the speech domain, perceptual learning can result in long-lasting changes to the mapping process between the speech signal and phonetic categories (Goldstone, 1998; Norris, Cutler, & McQueen, 2003). A body of research has studied perceptual learning in lab-based settings in NH listeners to examine the factors that underlie improvements in speech understanding. Across many studies and different forms of acoustic degradations, research has demonstrated that NH listeners consistently rely on lexical information to promote improvements in comprehension of
challenging speech signals, and that improvements can be leveraged through a brief training period.

4.1. A role for lexical information

One example of the influence of lexical information on speech learning occurs in the lexically guided perceptual learning paradigm. In this paradigm, listeners complete a brief training exposure phase followed by a test phase. During the training phase, participants listen to words and nonwords and complete a lexical decision task (Norris et al., 2003; Kraljic & Samuel, 2005; Drouin et al., 2016; Drouin et al., 2018). Critically, listeners are exposed to an atypical production during the training phase, such as a fricative that is spectrally ambiguous between /s/ and /ʃ/. For one group of listeners (e.g., /s/-bias group), the ambiguous replaces the medial /s/ in words that normally contain an /s/ (e.g., *dinosaur*). For a different group of listeners (e.g., /ʃ/-bias group), the same ambiguous sound replaced the medial /ʃ/ in words that normally containing an /ʃ/ (e.g., *publisher*). Accordingly, lexical context can be used to resolve the ambiguity in the atypical productions. In the test phase, listeners categorize items from a nonword continuum (e.g., */æʃi/-*/æsi*) to assess changes in the mapping between speech acoustics and the phonetic categories. The standard result in this paradigm is that performance differs between the two exposure groups at test in line with their exposure during training; specifically, listeners in the /s/-bias training group categorize more test items as /s/ compared to listeners in the /ʃ/-bias group. This finding demonstrates a persistent influence of lexical context on speech perception, even when disambiguating lexical context is removed. It has been confirmed that this effect is explicitly driven by lexical context as learning is absent when the ambiguous sound is placed in a nonword context during exposure (e.g., Norris et al., 2003).
Lexically guided perceptual learning is also observed for listeners learning to adapt to noise-vocoded speech. Noise-vocoding is a digital signal manipulation that is frequently used in the research domain to approximate the auditory experience of an individual using a CI. Noise-vocoding consists of dividing the natural speech signal into frequency bands, extracting the amplitude envelope, and replacing the fine structure in each band with noise (e.g., Faulkner, Rosen, & Smith, 2000; Loizou, Dorman, & Tu, 1999; Shannon et al., 1995; Davis, et al., 2005). This manipulation has been useful in research studies to determine the spectral resolution that is necessary for proficient speech comprehension. Standard noise-vocoded learning paradigms consist of a pre-test, or baseline measure of speech understanding, followed by a brief period of linguistic training with feedback. A post-test transcription measure is then compared to baseline to gauge improvement as a function of training. Numerous training studies have found that while perception of noise-vocoded speech is poor at initial exposure, listeners show robust perceptual learning with only minutes of training exposure, which generalizes across sentences (e.g., Loebach, Bent, & Pisoni, 2008), talkers (Huyck, Smith, Hawkins, & Johnsrude, 2017), and stimulus types (e.g., Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Loebach, Pisoni, & Svirsky, 2009).

Previous research suggests that perceptual learning of noise-vocoded speech, like other atypical input, is strongly influenced by stimulus lexicality and feedback. In a series of experiments, Davis et al. (2005) trained NH listeners on noise-vocoded sentences composed of either words or nonwords. During training, listeners were provided with interleaved feedback in which they heard the noise-vocoded sentence, then heard the clear version, and finally heard the same noise-vocoded sentence repeated again. At test, participants were asked to report as many words as possible and no feedback was provided. They found that listeners trained with vocoded
sentences composed of words performed significantly better than those trained with nonword sentences. In fact, the participants trained with nonword sentences were indistinguishable from naïve listeners with no training, highlighting the role of lexical context in facilitating comprehension of degraded input.

In addition to stimulus lexicality, research has also examined the role of lexical feedback in perceptual learning of noise-vocoded speech. Listeners do not necessarily require explicit feedback to learn noise-vocoded speech (Davis et al., 2005); however, feedback may allow learning to occur more efficiently. Studies have examined the role of both the type of feedback and the time course of feedback on learning noise-vocoded speech. Top-down approaches to adaptation posit that learning occurs via a comparison process between a given input and its target representation (e.g., Norris et al., 2003; Mirman, McClelland, & Holt, 2006). Providing explicit feedback as to what the target item is can allow for ambiguous representations to be adjusted to reflect the intended item (Hervais-Adelman et al., 2008). Research has shown that both written and auditory feedback are effective during training for impoverished input (Davis et al., 2005; Schwab, Nusbaum, Pisoni, 1985; Greenspan, Nusbaum, Pisoni, 1988) and that perceptual learning occurs more rapidly for listeners who receive feedback prior to hearing the stimulus, suggesting that knowing the identity of the target can allow learning to occur more rapidly (Davis et al., 2005; Hervais-Adelman et al., 2005).

4.2. Alternate routes to perceptual learning

The architecture of the TRACE model of spoken word recognition (McClelland & Elman, 1986) is consistent with a facilitative role for top-down auditory training on speech perception. In this framework, engaging the lexicon during training allows for an improved tuning in the mapping of the sensory signal to prelexical representations through a top-down
feedback mechanism. Top-down approaches to adaptation posit that learning occurs via a comparison process between a given input and a target representation (e.g., Norris et al., 2003). Providing explicit lexical feedback as to what the target item is can allow for potentially ambiguous representations to be adjusted to reflect the intended item – leading to a refinement in the mapping process given experience (Hervais-Adelman et al., 2005). Support for this framework comes from the finding that perceptual learning of noise-vocoded speech is absent for listeners who are trained with nonwords sentences (Davis et al., 2005), in line with other findings from the psycholinguistic domain (e.g., Norris et al., 2003). However, this finding was not replicated by listeners trained on single nonwords, suggesting that when the memory load is sufficiently reduced, listeners can utilize non-linguistic information for speech adaptation. Nevertheless, lexical context may maximize perceptual learning for vocoded speech as the magnitude of learning was smaller for single word training (e.g., Hervais-Adelman et al., 2008), compared to training with sentences (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Research has shown that both written and auditory lexical feedback are effective for training of impoverished input (Davis et al., 2005; Schwab, Nusbaum, Pisoni, 1985; Greenspan, Nusbaum, Pisoni, 1988) and that perceptual learning occurs more rapidly for listeners who receive feedback prior to hearing the stimulus, which suggests that knowing the identity of the target can allow learning to occur more rapidly (Davis et al., 2005; Hervais-Adelman et al., 2005). This finding is in line with the TRACE supervised learning framework, which allows backpropagation between processing layers to integrate available contextual information into the percept (McClelland & Elman, 1986; Davis et al., 2005; Hervais-Adelman et al., 2005). Under this framework, lexically rich contexts during training promotes an optimal environment in which to learn degraded input.
An alternative account posits that learning may be mediated by the similarity of the tasks used during training and test. In a transfer-appropriate processing (TAP) framework (e.g., Franks et al., 2000), performance is maximized when the task used to assess learning at test is identical to the task used during training. Under this theoretical framework, learning occurs because of a match between training and test tasks. Support for this view comes from studies demonstrating perceptual learning on tasks that mirror the test task (i.e., Davis et al., 2005). Thus, advocates for this view argue that linguistic focused training tasks are necessary to observe a linguistic benefit. This view has been challenged by findings that demonstrate a linguistic benefit using training tasks that direct attention towards non-linguistic characteristics of the signal. For example, recent work has demonstrated perceptual learning of noise-vocoded speech using non-linguistic training tasks that differ from the test task (Loebach, Bent, & Pisoni, 2008). In this study, three groups of NH listeners heard noise-vocoded sentences produced by multiple talkers during a training period. During training, one group was asked to transcribe the sentence, one group was asked to identify the talker, and another group was asked to identify the gender of the talker. All listeners received feedback in line with the assigned task. At test, all participants, regardless of training group, were asked to complete a transcription task to measure language comprehension. The results showed that perceptual learning was equivalent for listeners trained with either the linguistic transcription task or the non-linguistic talker identification task. This finding suggests that a linguistic benefit can be obtained regardless of whether the training phase mirrors the task used during test, and is in line with other forms of perceptual learning for adaptation to ambiguous speech sounds (e.g., Drouin et al., 2018).

The findings from Loebach et al. (2008) are considered with respect to a depth of processing framework (e.g., Craik & Lockhart, 1972). In this framework, perceptual learning
occurs if the task used during training is sufficiently challenging so that the listener needs to attend to fine-grained acoustic information in order to accurately perform the task. Under this view, perceptual learning is directly mediated by the attentional requirements of the task and thus can be achieved using a variety of top-down or bottom-up focused training strategies, so long as the listener is sufficiently challenged to engage with the stimuli. Support from this view comes from the Loebach et al. (2008) finding showing that perceptual learning, as measured through a linguistic transcription task, was equally maximized for two groups of listeners who completed a challenging task during training (i.e., lexical transcription and talker identification), but was less robust for listeners who completed an easier gender identification task during training. They argue that this finding reflects the attentional requirements of the training task. Specifically, the talker identification and transcription tasks were more cognitively challenging tasks relative to the gender identification training group, which resulted in a disparity in how listeners encoded and learned the novel input. Together, these findings suggest that similarity between training and test asks in addition to the attentional requirements of the task are both important factors to consider when designing adaptation studies where the outcome goal is comprehension.

5. Translating principles of learning to auditory training for CI users

Predicting variability in speech perception outcomes following implantation remains a challenge in the CI literature. Here we suggest that active auditory training may offer a means to maximize adaptation for both low and high performers. In this review, we have focused on findings from the psycholinguistics literature demonstrating that robust perceptual learning can occur even in the face of acoustic degradation. Human speech perception is a highly plastic and adaptive process that allows for the compensation of a variety of listening conditions with relative ease. The conditions that promote ease of understanding for atypical speech input has
been the primary focus of this review. We have examined the influential effects of lexical knowledge on processing impoverished input for NH listeners, with a focus on how active top-down feedback can propagate to improve the mapping between a sensory signal and prelexical representations. This is an important consideration with respect to the rehabilitation literature because it suggests that passive listening through the implant, without engagement in active training, may not sufficiently maximize opportunities to improve perception. Translational research characterizing how CI users benefit from lexical context and how lexically driven training approaches promote improvements in perception is needed to begin to formalize concrete rehabilitation recommendations. While many outstanding questions still exist, the literature reviewed here for NH listeners lays the foundation for establishing a benefit of lexical context and feedback for degraded speech signals – raising the possibility that modeling training paradigms in this way for CI users may also promote similar benefits. If lexical context is the putative factor for driving perceptual learning of degraded speech input transmitted through a CI, then listeners trained using lexically oriented paradigms may show the greatest long-term benefit from their device.

We have also reviewed alternate routes to perceptual learning under different theoretical frameworks and consider how task-based listening strategies may mediate learning outcomes. This is particularly important when considering how to design optimal training paradigms. If training using cognitively easier tasks (i.e., talker identification instead of lexical transcription) promotes equivalent learning outcomes on linguistic tasks, then it opens the opportunity to better customize training to patient needs. An understudied area in the current research domain with respect to bottom-up or top-down focused listening strategies is the degree to which gains are maintained over long-term time periods. Only through a longitudinal design may we better
document potential differences among training groups. Namely, would two groups of listeners trained using different tasks show equivalent performance in the short-term, but differ with respect to long-term benefit of gains? Currently, we do not have the evidence base to definitively answer this question, which contributes to hesitations about recommending a specific training protocol.

Finally, there is a significant need to better understand how individual patient factors interact with training outcomes and to characterize differences in plasticity across patient profiles. Even within the NH population there remains significant variation in how individuals adapt to acoustically poor input. While most training studies have focused exclusively on outcomes at the group level, characterizing individual differences in learning is emerging as an active area of research. As reviewed, both bottom-up and top-down processes drive perceptual adaptation to acoustically poor input, therefore differences may emerge from variation in how listeners code the low-level sensory signal or how listeners integrate higher-order contextual cues in the mapping process. Could customized training programs be used to specifically target weaknesses in either domain to close the gap between high and low performers? We propose that the next phase of auditory training research should be aimed at examining the use of novel outcome measures, long-term assessments, training variables (e.g., task, duration, time of intervention), and individual difference metrics. By establishing these parameters, scientists and clinicians will be able to facilitate recommendations for the conditions that promote general speech learning as well as how training paradigms might be customized to the needs of individual CI users.
II. Sleep-based memory consolidation stabilizes perceptual learning of noise-vocoded speech

1. Introduction

Language comprehension requires mapping the highly variable acoustic speech signal to stable linguistic representations. Perceptual learning mechanisms allow the listener to dynamically “tune in” to idiosyncratic variation in the speech signal for language comprehension. One example of perceptual learning can be observed when listeners face extreme acoustic distortion, as in the case of noise-vocoded speech. Noise-vocoded speech is often used to simulate the auditory experience of a cochlear implant (CI) user, and is a challenging auditory stimulus for speech perception because it limits the amount of spectral information available in the speech signal (Shannon et al., 1995). Inexperienced listeners of noise-vocoded speech initially show poor perception of the signal, but performance rapidly improves following a short period of focused training (Davis et al., 2005; Hervais-Adelman et al., 2008; Loebach et al., 2008) Because noise-vocoding is a novel form of speech for normal-hearing listeners, it provides a useful tool for assessing perceptual learning.

In a standard perceptual learning for noise-vocoded speech paradigm, listeners complete a baseline recognition task prior to training. Then, listeners complete a training phase in which a transcription task is performed for noise-vocoded signals, most often with feedback provided on each trial. Immediately following training, listeners again complete a recognition task, which is compared against the baseline measure in order to assess the magnitude of perceptual learning. Previous research suggests that top-down focused training paradigms in which a patient hears a speech stimulus and receives immediate lexical feedback promotes heightened perceptual
learning as compared to individuals who engage only in passive listening with no task or feedback (Davis et al., 2005). In addition, a variety of training tasks may be appropriate for optimizing perceptual learning of speech including word/sentence transcription and talker identification (e.g., Loebach et al., 2008).

A critical marker of perceptual learning is transfer of learning to new stimuli. Previous research has shown that listeners trained to comprehend noise-vocoded speech show generalization to novel words (Hervais-Adelman et al., 2008), sentences (e.g., Loebach et al., 2008), talkers (Dahan & Mead, 2010; Huyck et al., 2017), non-words (Hervais-Adelman et al., 2008), frequency regions (Hervais-Adelman et al., 2011) and environmental non-speech stimuli (Loebach et al., 2009). Furthermore, training on an eight-channel manipulation generalizes to fewer channel manipulations, including six- and four-channel manipulations (e.g., Loebach et al., 2008). However, generalization for noise-vocoded speech is not ubiquitous across all perceptual learning studies. For example, Dahan and Mead (2010) carefully controlled phonetic contexts for test and training items for listeners trained on noise-vocoded speech. They found that the degree of generalization observed was closely linked to the degree of similarity of the training items to the test items. Specifically, generalization was limited to test items that shared similar onset, nucleus, and rhyme contexts to training items. Moreover, their findings demonstrating generalization across talkers were equivocal. In one experiment, listeners who were trained and tested on different talkers were more accurate at identifying test items produced by the same talker during training and that appeared in a similar phonetic context (Dahan & Mead, 2010). In a follow-up experiment, Dahan and Mead (2010) found comparable performance on test item identification, regardless of whether the test item had been produced by the same speaker or different speaker from training. Given that generalization is observed in many studies, but not all
(e.g., Dahan & Mead, 2010), there is a need to carefully examine factors that promote
generalization within the perceptual learning literature.

1.1. Role of sleep consolidation on perceptual learning of speech

Rapid perceptual learning for noise-vocoded speech has been well documented in the
literature (e.g., Shannon et al., 1995; Davis et al., 2005), with some studies noting generalization
(e.g., Davis et al., 2005; Loebach et al., 2008). Currently, there is limited research examining the
factors that promote stability and generalization of perceptual learning for noise-vocoded stimuli.
One area of active research has implicated sleep as an important factor for perceptual learning of
other types of speech stimuli. Specifically, perceptual learning has been shown to benefit from a
period of offline memory consolidation via sleep (e.g., Gomez et al., 2011; Earle & Myers, 2014
for review). Effective speech learning engages various memory processes, with many of these
memory processes inherent in our explanations on why perceptual learning occurs. Acquisition
of novel speech signals is considered a procedural skill as it involves implicit skill learning
instead of explicit recall on the part of the perceiver. Memory consolidation, then, is considered
to be the process by which skill learning stabilizes and moves to long-term neural systems.
Previous research has suggested that consolidated memories seem to be more resistant to
interference effects (e.g., Earle & Myers, 2014; Earle & Myers, 2015) and may promote better
long-term recall (e.g., Holz et al., 2012). Memory consolidation is thought to be a sleep-
dependent process as rapid eye-movement (REM) sleep has been implicated in the consolidation
of procedural memories (Marshall & Born, 2007).

Several studies have demonstrated a role for sleep in speech learning. For example, Fenn,
Margoliash, and Nusbaum (2003) trained and tested two groups of participants on synthetic
“text-to-speech” single words. One group of participants completed their first training and test
interval at 9am and then returned approximately 12 hours (9pm on the same day) and 24 hours (9am the next day) following the initial training. Another group of participants received identical training structure presented at different time intervals – with the initial training/test occurring at 9 PM, and then returning approximately 12 hours (9 AM the next day, following a period of sleep) and 24 hours (9 PM the second day) after the initial training. Listeners were asked to transcribe a synthetic word on each trial with each word being presented only one time across any session, thus requiring listeners to generalize the mapping from synthetic speech to prelexical representations. They found that while both groups of participants received an identical training format, there was a significant difference in performance following a period of sleep. Namely, performance 12 hours following the initial session seemed to stabilize for listeners trained in the evening, with that performance remaining stable 24 hours later. Conversely, listeners trained in the morning seemed to show a degradation in performance 12 hours later, with performance restored only after a period of offline consolidation 24 hours later. Thus, only at the 12-hour follow-up test – when a period of sleep had occurred for the evening group and a period of wakefulness had occurred for the morning group – did performance between the two groups differ. These results have been replicated and extended to show that sleep supports generalization for synthetic speech, not just learning involved in rote memorization (Fenn et al., 2013). Together, these findings highlight the role of sleep in adaptation to atypical speech input, as well as the importance of using assessments at time intervals displaced from initial training to observe changes over time.

Other work has highlighted the role of sleep-consolidated learning for other forms of speech input including adaptation to non-native speech sounds and accented speech. Earle and Myers (2015) trained two groups of monolingual English participants on a non-native contrast
using a discrimination task. The morning group completed their initial training session between 8 – 10 AM and returned approximately 12 hours later (between 6 – 9 PM on the same day) to assess maintenance of learning. The evening group completed the same tasks, but the initial training session was completed between 6 – 9 PM and they returned approximately 12 hours later (between 8 – 10 AM) on the next day. They reported that only participants who completed training in the evening and subsequently engaged in a period of sleep demonstrated significant improvement in non-native contrast discrimination. Earle and Myers (2015) suggest that both sleep-consolidated memory and hearing interfering stimuli throughout the day play a role in acquiring non-native speech sounds, which allows those trained at night to benefit to a greater degree from training compared to those trained in the morning.

Training in close proximity to sleep also promotes generalization for accent adaptation. Research has shown that after exposure to a single talker who produces an ambiguous speech sound, listeners generalize to other phonetically similar speakers not previously heard during training (Xie & Myers, 2017). However, when accented talkers differ in an acoustic cue dimension not routinely utilized by native speakers, in this case only the acoustic parameter of burst duration – not vowel duration – reliably cued a stop consonant distinction, only participants who received training in the evening maintained generalization to an untrained talker (Xie et al., 2018). It has been proposed that sleep preferentially promotes speech learning that requires greater perceptual adjustment on the part of the listener, such as learning to utilize an acoustic cue dimension not routinely used (Xie et al., 2018) or learning to perceive an initially challenging stimulus (Fenn et al., 2003). Sleep-consolidated memory appears to be less critical when the learning process requires less comprehensive adjustments on the part of the listener. For example, Eisner and McQueen (2006) used a lexically-guided perceptual learning paradigm
to train two groups of listeners to perceive an ambiguous speech sound as a member of an existing phonetic category: the morning group received training in the morning and was tested later the same day, while the evening group trained in the evening and was tested the next morning following a period of sleep. During training, all participants were exposed to an ambiguous fricative (midway between /f/ and /s/) in lexical items that would normally contain those fricatives; thus, listeners were differentially biased to perceive the ambiguous sound as either an /s/ or /f/. At test, listeners categorized nonwords items along an /s/-/f/ continuum. At test, the results showed robust phonetic recalibration that reflected the lexical context in which the ambiguous fricative was trained. Critically, they found no difference in phonetic recalibration for participants who trained in the morning relative to those who trained in the evening, though the learning effect was numerically more robust in the evening group. Of note, they did not find evidence of heightened or decayed adaptation following the sleep period, in contrast to previous findings (Fenn et al., 2003). Eisner and McQueen (2006) hypothesized that this was because lexically guided perceptual learning involves relatively minimal adjustment to phonetic categories on the part of the listener, given that listeners regularly make slight adjustments to phonetic categories to accommodate to talker-specific idiosyncrasies. Sleep consolidation has thus been implicated for speech adaptation that requires a more comprehensive retuning for novel forms of speech that the perceiver does not regularly encounter, and thus may also be important for learning of degraded speech signals.

1.2. Maintenance of gains following perceptual training

An understudied area in the perceptual learning literature concerns the long-term outcomes of learning. Specifically, perceptual learning for a variety of speech signals often only measures outcomes of training in the immediate – relatively fewer studies examine the outcomes
of learning after the initial training and test session. It is therefore unknown how gains observed after a training session are maintained over time. Given that perceptual learning inherently refers to long-lasting changes in an individual’s perceptual system (Goldstone, 1998), it is theoretically important to characterize performance observed at time intervals that are displaced from training. Moreover, from a clinical standpoint, the overarching goal of many adaptation studies for noise-vocoded speech is to identify training variables of interest that may hold value for rehabilitation protocols for patients adapting to poor acoustic input, including CI users. Auditory training has been examined as a possible rehabilitation tool for patients adapting to the acoustic input transmitted through the CI (e.g., Fu & Galvin, 2007), but there is a critical need to establish how gains are maintained over time using basic paradigms of perceptual learning for degraded speech signals.

With respect to perceptual learning for noise-vocoded signals, McGettigan, Rosen, and Scott (2014) exposed participants to multiple levels of noise-vocoded speech in sentence, word, consonant, and vowel contexts. Learning was assessed immediately and approximately one-week later, with no additional intervening exposure. Across various tasks, significant improvement in performance was reported one-week later relative to the immediate assessment. However, because the time of day at which the tasks were administered was not controlled in McGettigan et al. (2014), it is difficult to determine the role of sleep-mediated memory consolidation on learning from this study.

Performance at even later time intervals from training has been reported in the non-native speech learning literature. In one seminal study, Lively et al., 1994 trained monolingual Japanese participants on the English /r/-/l/ phonetic contrast. An improvement in performance was observed immediately following training, and those gains were maintained three months later
without any intervening additional training. Moreover, at a 6-month follow-up, benefit from training continued to be observed, though it was less robust compared to earlier tests. Thus, the performance reported here aligns with Goldstone’s (1998) standards for perceptual learning, as gains from initial training remained robust many months after the initial training session occurred. Thus, the addition of later test intervals allows the time course of performance to be characterized, enabling identification of improvement or decay in performance over time that may be missed when testing only in the immediate.

1.3. The current study

In the current study, we trained two groups of normal hearing listeners to improve their perception of noise-vocoded speech. One group received training in the morning and the other group received training in the evening. In addition to training, listeners completed four tests. The first test was administered immediately prior to training (to establish baseline performance), the second test was administered directly after training to assess immediate learning, the third test was administered 12 hours post-training to examine potential effects of sleep consolidation, and the fourth test was administered seven days following training to assess maintenance of learning. We included both trained and novel items at each test to assess generalization of learning.

The scientific goal of the current study was twofold. First, we aimed to identify the role of sleep-mediated memory consolidation on perceptual learning for noise-vocoded speech. As previously reviewed, sleep consolidation has been shown to play an important role in stabilizing procedural learning for difficult to learn stimuli such as non-native speech sounds (e.g., Earle & Myers, 2015), accented speech (e.g., Xie et al., 2018), and synthetic speech (Fenn et al., 2003). As the noise-vocoding manipulation requires extensive restructuring of the mapping between the acoustic signal and prelexical representations to achieve comprehension, sleep consolidation may
be critical for this type of adaptation. We hypothesized that while performance immediately following training would be comparable between the morning and evening groups, listeners trained in the evening would show stabilization of learning at the 12-hour follow-up test. We predicted that the morning group on the other hand would demonstrate a decay in performance at the 12-hour follow-up test because the period of wakefulness – as opposed to sleep – would interfere with learning stabilization. Moreover, we predicted if learning and generalization benefits from a period of offline consolidation, listeners trained in the evening would show stable performance for trained and novel items at the 12-hour follow-up, while participants trained in the morning would show a degradation for both item types.

The second goal of the current study was to characterize the time course of perceptual learning for degraded speech signals. Namely, the one-week follow-up was included to determine whether learning is maintained or diminished at a later interval displaced from initial training. We predicted that if the initial encoding experience facilitates maintenance of learning, then participants trained in the evening will demonstrate stable performance at the one-week follow-up, while participants trained in the morning will show a decline in performance. In particular, we hypothesized that sleep-consolidated learning will preferentially stabilize performance for novel items in the evening group relative to listeners trained in the morning. This result would be consistent with notion that offline memory consolidation preferentially maintains long-term generalization for degraded speech signals, extending findings demonstrating consolidation as a critical factor for generalization at time intervals more immediate to training (e.g., Fenn et al., 2013; Earle & Myers, 2015; Xie et al., 2018).
2. Methods

2.1. Participants

Participants (n = 64) were recruited from the University of Connecticut community and received either course credit or monetary compensation for their participation. The participants (40 women, 24 men) were between 18 and 35 years of age (mean = 20, SD = 2). All participants were monolingual speakers of American English with no self-reported history of speech, language, or hearing problems. All participants passed a pure tone hearing screening on the day of testing, administered at 25 dB HL at 500, 1000, 2000, and 4000 Hz. We tested but excluded four additional participants because they failed to return to at least one test session.

2.2. Stimuli

The stimuli were obtained from the University of Washington/Northwestern University (UW/NU) Corpus 1.0 (McCloy, Souza, Wright, Haywood, Gehani, & Rudolph, 2013). The stimuli consisted of 150 sentences drawn from a subset of the phonetically balanced Harvard/IEEE sentences (Egan, 1944; Egan, 1948; Rothauser, 1969; for full list of sentences used see appendix). Each sentence contained five or six keywords embedded into a semantically rich sentence context (e.g., A pot of tea helps to pass the evening contains five keywords; The ship was torn apart on the sharp reef contains six keywords). Each sentence was produced by five different talkers (three female; two male) from the northern cities geographic region. Sentences were recorded with a 44.1 kHz sampling rate and 16-bit depth. Audio files were for peak intensity across recordings.

All stimuli were processed in Praat (Boersma, 2001) using the band filtering and vocoding resynthesis scripts in the GSU Praat Tools suite (Owren, 2008). First, sentences were bandpass filtered with a lower frequency of 50 Hz and an upper frequency of 8000 Hz with a 100
Hz smoothing bandwidth. Then, sentences were noise-vocoded using default lower frequency limits to create six nonlinearly spaced bands. A six-channel vocoding synthesis was implemented given that previous research has demonstrated that normal hearing listeners demonstrate comparable performance to cochlear implant users with this number of bands (e.g., Dorman & Loizou, 1998), and is in line with previous noise-vocoding training studies (e.g., Davis, et al., 2005).

2.3. Procedure

The methods and procedure were approved by the Institutional Review Board at the University of Connecticut. All participants gave informed consent before beginning the experiment. All participants completed a single training phase and four test phases. Participants completed testing individually in a sound-attenuated booth. Stimuli were presented over headphones at a comfortable amplitude that was held constant across participants. The SuperLab software (version 4.5) on a Mac OS X desktop or Mac OS X laptop was used for stimulus presentation and data collection. Participants typed their responses using the keyboard on the computer.

Participants were randomly assigned to either the morning or evening group. For participants in the morning group, session 1 occurred between 8 – 10 AM on day one, session 2 occurred between 6 – 8 PM on day one, and session 3 occurred at any time of day exactly one-week after session 1. For participants in the evening group, session 1 occurred between 6 – 8 PM on day one, session 2 occurred between 8 – 10 AM on day two, and session 3 occurred at any time of day exactly one-week after session 1. The session structure is illustrated in Figure 1 and was identical between training groups except for the time of day at which the first two sessions occurred. Session 1 consisted of a baseline pretest to assess performance prior to training, a
training phase, and a test phase completed immediately following training (posttest1). During session 2 (posttest2) and session 3 (posttest3), participants completed a single test phase only.

**Figure 1.** Graphic illustrating session structure across days for participants in the morning (tan) and evening group (blue). While the number of hours between session 1 and session 2 is identical between the morning and evening group, the dashed line indicates the period of sleep that occurs between session 1 and session 2 for the evening group. Procedural details for the training and test phases are described below.

**Training.** Training occurred during session 1. During training, participants heard 30 unique sentences, each produced once by each of the five talkers, yielding 150 trials total. Participants were randomly assigned to one of four training randomizations, resulting in four unique presentations of the 150 trials. On each trial, listeners first heard a sentence produced by one of the talkers. Then the prompt “What did the talker say?” appeared on the screen with a text box below it. The participant was instructed to type in any words that they heard in the sentence and were encouraged make their best guess if they were unsure. After typing in their response,
the participant pressed the “Enter” key to record the response. After the response was recorded, feedback was provided as the correct transcription of the sentence, which was displayed on the screen for 2000 milliseconds. The interstimulus interval was 2000 milliseconds, timed from the offset of feedback. If a participant failed to respond within 5000 milliseconds after the stimulus onset, then no response was recorded, and the experiment advanced to the next trial. The entire training phase lasted approximately 30 minutes.

**Test.** Four test sessions were completed during the experiment. A baseline pretest was performed at the start of the experiment to assess performance prior to training. Posttest1 was completed immediately following training. Posttest2 was completed during session 2, which occurred 10 – 12 hours following session 1. Finally, posttest3 was completed during session 3, which occurred 7 days after session 1. Each test phase consisted of 60 sentences produced by one of five talkers, with equal number of productions across the five talkers at each test phase. Of the 60 sentences heard during any given test phase, 30 sentences were also heard during training (trained items), while 30 sentences were never heard during training and only presented at test phases (novel items). Trials were unique across test sessions – while a given sentence could be repeated across sessions, there was no repeat of a given talker and sentence pairing at any of the test sessions. The order in which test items were presented was randomized across participants in each test session.

On each trial, listeners heard a sentence produced by one of five talkers and then the prompt “What did the talker say?” appeared on the screen with a text box below it. Identical to training, the participant was instructed to type in any words that they heard in the sentence and were encouraged to make their best guess if unsure. The participant was informed that they may hear the same sentence more than once during test sessions. After typing in their response, the
participant pressed the “Enter” button to record their response and move onto the next trial.

Unlike training, participants did not receive feedback in any form during test. The interstimulus interval was 2000 milliseconds, timed from the offset of the participants’ response. If a participant failed to respond within 5000 milliseconds after the stimulus onset, then no response was recorded, and the experiment advanced to the next trial. Each test phase lasted approximately 15 minutes.

3. Results

3.1. Training

Each sentence was scored for accuracy by a trained research assistant. A response was considered correct if the transcription contained two more correctly transcribed keywords, and a response was considered incorrect if the transcription contained fewer than two correctly transcribed keywords. Credit was given for common misspellings, typos, and alternate spellings with the constraint that the misspelling did not yield a real word and/or represent a change in morpheme status. For example, if the target sentence was *the box was thrown beside the parked truck*, then all of the following are examples of acceptable responses for the keyword *truck*: *truej, truc, truckj, and truk*. The following are examples of unacceptable responses for the keyword *truck*: *trucks, trucker, dumptruck, trick*, and *tri*. Training and test data in addition to an analysis script (in R) are available at: https://osf.io/nja9w/?view_only=18e9672861fc4f568f1eb9d61543479e. The analysis script operates on trial-level data to reproduce all statistical analyses presented in this manuscript in addition to generating the figures.

Figure 2 shows the accuracy distributions during training for the two training groups, which were determined by first calculating mean accuracy across the 150 training items for each
Consistent with past research, both groups showed high accuracy during the training phase, indicating that listeners learned to perceive the noise-vocoded speech. In addition, visual inspection of this figure suggests that accuracy during training was equivalent between the two training groups.

**Figure 2.** Boxplots showing the distribution of mean proportion correct responses during the training phase for the morning and night training groups.

To examine this pattern statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear mixed effects model (GLMM) with the binomial response family as implemented in lme4; the Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using lmerTest (Kuznetsova et al., 2017). The fixed effect of the model was group (sum coded; morning = -1, evening = 1). The random effects structure consisted of random intercepts by subject and random intercepts by sentence. The main effect of group was not significant ($\hat{\beta} = -0.069$, $SE = 0.066$, $z = -1.044$, $p = 0.297$), providing evidence that accuracy between the two groups did not differ during the training phase.

### 3.2. Test

Each test sentence was scored for accuracy following the procedure outlined for the training phase. Figure 3 shows the accuracy distributions at each test for the two training groups...
separately for trained and novel items, which was determined by calculating mean accuracy across the appropriate test items for each participant. Consider first performance for the trained items. Visual inspection suggests that both groups improved from pretest to the first posttest, indicating that learning occurred following training. In addition, inspection of the group difference between posttest1 and posttest2 (which took place 12 hours after posttest1) suggests the presence of a sleep-based consolidation effect. Specifically, the morning group shows a decline in performance between posttest1 and posttest2, whereas the evening group does not. Moreover, visual inspection suggests that learning is maintained over time, in that performance for posttest3 is equivalent or improved relative to performance at posttest2 for both groups.

Now consider performance for the novel items. As for the trained items, learning is observed in that performance for posttest1 is improved relative to the pretest. A sleep-based consolidation effect is also suggested, in that performance at posttest2 remains stable relative to posttest1 for the evening group, but declines for the morning group. Performance between posttest3 and posttest2 again suggests maintenance of learning, though in contrast to the trained items, the evening group appears to show higher accuracy compared to the morning group for novel items at posttest3.
To examine these patterns statistically, three sets of analyses were performed: the learning analysis examined performance between pretest and posttest1, the consolidation analysis examined performance between posttest1 and posttest2, and the maintenance analysis examined performance between posttest2 and posttest3. Each is addressed in turn.

**Learning.** The learning analysis examined performance between pretest and posttest1. Recall that pretest took place immediately before training and posttest1 took place immediately after training. As can be seen in Figure 4, both groups improved from pretest to posttest1, though it appears that they did so to a larger degree for trained compared to novel items. To examine these patterns statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a GLMM with the fixed effects of group (morning = -1, evening = 1), test (pretest = -1, posttest1 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject, random intercepts by sentence, and random slopes by subject for test and type.
Figure 4. Boxplots showing the distribution of mean proportion correct responses in the learning analysis (pretest vs. posttest1) for the morning and night training groups separately for trained and novel items. Lines show performance for the individual participants in each group.

The model revealed a main effect of test ($\hat{\beta} = 1.403$, $SE = 0.119$, $z = 11.725$, $p < 0.001$), with higher accuracy at posttest1 compared to pretest, and a main effect of type ($\hat{\beta} = -0.827$, $SE = 0.186$, $z = -4.437$, $p < 0.001$), reflecting higher accuracy for trained compared to novel items. There was also an interaction between test and type ($\hat{\beta} = -0.605$, $SE = 0.113$, $z = -5.379$, $p < 0.001$). Post-hoc paired comparisons (here and throughout) were performed using the emmeans package to explicate the nature of the interaction, applying the Tukey method to adjust for multiple comparisons. The paired comparisons showed no difference between trained and novel items at pretest ($\hat{\beta} = 0.443$, $SE = 0.435$, $z = 1.017$, $p = 0.739$), but significantly higher accuracy for trained compared to novel items at posttest1 ($\hat{\beta} = 2.865$, $SE = 0.436$, $z = 6.578$, $p < 0.001$). Moreover, accuracy improved from pretest to posttest1 for both trained ($\hat{\beta} = -4.017$, $SE = 0.150$, $z = -26.747$, $p < 0.039$) and novel ($\hat{\beta} = -1.595$, $SE = 0.440$, $z = -3.627$, $p = 0.002$) items. Thus, the
interaction in the omnibus model reflects increased improvement between pretest and posttest1 for trained compared to novel items, and not a failure to improve for novel items.

The omnibus model showed no main effect of group ($\hat{\beta} = 0.007, SE = 0.080, z = 0.092, p = 0.927$), but there was an interaction between group and test ($\hat{\beta} = -0.139, SE = 0.051, z = -2.707, p = 0.007$). Paired comparisons showed an improvement from pretest to posttest1 for both the morning group ($\hat{\beta} = -3.084, SE = 0.262, z = -11.760, p < 0.001$) and the evening group ($\hat{\beta} = -2.527, SE = 0.259, z = -9.771, p < 0.001$). The beta estimate is larger in the former compared to the latter, which may suggest that the interaction between group and test in the omnibus model reflects increased improvement for the morning compared to the evening group. However, no reliable difference was observed between the two groups at either pretest ($\hat{\beta} = -0.293, SE = 0.169, z = -1.739, p = 0.304$) or posttest1 ($\hat{\beta} = 0.264, SE = 0.210, z = 1.257, p = 0.590$).

The omnibus model also showed an interaction between group and type ($\hat{\beta} = 0.084, SE = 0.041, z = 2.061, p = 0.039$). Paired comparisons showed that accuracy for trained items was higher than accuracy for novel items for both the morning group ($\hat{\beta} = 1.823, SE = 0.383, z = 4.765, p < 0.001$) and the evening group ($\hat{\beta} = 1.485, SE = 0.381, z = 3.899, p = 0.006$). Given the smaller beta estimate in the latter, the interaction may reflect a weaker item effect in the evening compared to the morning group. However, paired comparisons showed no difference between the two groups for either trained ($\hat{\beta} = 0.155, SE = 0.174, z = 0.886, p = 0.812$) or novel items ($\hat{\beta} = -0.184, SE = 0.185, z = -0.993, p = 0.753$). The three-way interaction between group, test, and type was not significant ($\hat{\beta} = 0.046, SE = 0.033, z = 1.416, p = 0.157$). Collectively, the learning analysis demonstrates that the morning and night groups both showed improved perception of noise-vocoded signals following training. Training generalized to novel items, though the learning effect was stronger for trained compared to novel items.
Consolidation. The consolidation analysis examined performance between posttest1 and posttest2. Recall that posttest2 took place 12 hours after posttest1, a period of time that included sleep for the evening group but not for the morning group. As can be seen in Figure 5, the evening group appears to show increased accuracy at posttest2 compared to the morning group for both trained and novel items.

Figure 5. Boxplots showing the distribution of mean proportion correct responses in the consolidation analysis (posttest1 vs. posttest2) for the morning and night training groups separately for trained and novel items. Lines show performance for the individual participants in each group.

To examine these patterns statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a GLMM with the fixed effects of group (morning = -1, evening = 1), test (posttest1 = -1, posttest2 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject and by sentence, and random slopes by subject for test and type. The results of the model showed a main effect of test ($\hat{\beta} = -0.393$, $SE = 0.118$, $z = -3.311$, $p < 0.001$), indicating higher accuracy at posttest1 compared to posttest2, and a main effect of type ($\hat{\beta} = -1.445$, $SE = 0.207$, $z = -6.993$, $p < 0.001$), indicating higher accuracy for trained compared
to novel items. There was no main effect of group ($\hat{\beta} = 0.119, SE = 0.098, z = 1.217, p = 0.223$); however, a robust group by test interaction was observed ($\hat{\beta} = 0.228, SE = 0.032, z = 7.009, p < 0.001$). Post-hoc paired comparisons showed that accuracy for the morning group declined between posttest1 and posttest2 ($\hat{\beta} = 1.243, SE = 0.246, z = 5.045, p < 0.001$), whereas there was no difference in accuracy between the two test sessions for the evening group ($\hat{\beta} = 0.327, SE = 0.246, z = 1.333, p = 0.542$). Moreover, accuracy was higher for the night compared to the morning group at posttest2 ($\hat{\beta} = -0.697, SE = 0.201, z = -3.473, p = 0.003$), but no such difference was observed between the two groups at posttest1 ($\hat{\beta} = 0.219, SE = 0.213, z = 1.026, p = 0.734$).

In the omnibus model, there was no interaction between group and type ($\hat{\beta} = 0.005, SE = 0.061, z = 0.089, p = 0.929$) or between test and type ($\hat{\beta} = 0.019, SE = 0.119, z = 0.163, p = 0.870$). However, the three-way interaction between group, test, and type was reliable ($\hat{\beta} = -0.105, SE = 0.032, z = -3.247, p = 0.001$). To explicate the interaction, separate GLMMs were constructed for trained and novel items with the same fixed and random effects structure of the omnibus model except for the removal of type as a fixed effect. The group by test interaction was significant for both trained ($\hat{\beta} = 0.334, SE = 0.049, z = 6.779, p < 0.001$) and novel items ($\hat{\beta} = 0.124, SE = 0.042, z = 2.980, p = 0.003$). The beta estimates for the interaction terms suggest a weaker consolidation effect for novel compared to trained items. Collectively, the results of the consolidation analysis are consistent with a facilitative effect of sleep-based consolidation on adaptation to noise-vocoded speech. Compared to listeners who were trained in the morning, listeners who were trained prior to sleep showed improved comprehension of noise-vocoded signals 12 hours after training, which was the consequence of learning retention in the night group and a relative loss of learning in the morning group.
**Maintenance.** The maintenance analysis examined performance between posttest2 and posttest3. Recall that posttest3 took place one-week after posttest2. As can be seen in Figure 6, the evening group appears to show increased accuracy at posttest3 compared to the morning group for novel items, though the two groups appear equivalent for trained items.

**Figure 6.** Boxplots showing the distribution of mean proportion correct responses in the maintenance analysis (posttest2 vs. posttest3) for the morning and night training groups separately for trained and novel items. Lines show performance for the individual participants in each group.

Trial-level responses (0 = incorrect, 1 = correct) were submitted to a GLMM with the fixed effects of group (morning = -1, evening = 1), test (posttest2 = -1, posttest3 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject and by sentence, and random slopes by subject for test and type. The results of the model showed a main effect of group ($\hat{\beta} = 0.246, SE = 0.096, z = 2.650, p < 0.010$), indicating higher accuracy for the night compared to the morning group, and a main effect of type ($\hat{\beta} = -1.446, SE = 0.210, z = 6.882, p < 0.001$), indicating higher accuracy for trained compared to novel items. There was no main effect of test ($\hat{\beta} = 0.147, SE = 0.121, z = 1.217, p = 0.224$), no interaction between group and
type ($\hat{\beta} = -0.029, SE = 0.059, z = -0.496, p = 0.619$), and no interaction between test and type ($\hat{\beta} = -0.048, SE = 0.121, z = -0.395, p = 0.692$).

The omnibus model did reveal a significant interaction between group and test ($\hat{\beta} = -0.104, SE = 0.032, z = -3.226, p = 0.001$), which was further mediated by type as indicated by a significant three-way interaction between group, test, and type ($\hat{\beta} = 0.074, SE = 0.031, z = 2.340, p = 0.019$). To explicate the interaction, separate GLMMs were constructed for trained and novel items following the same fixed and random effects structure as the omnibus model except for removing the fixed effect of type. For the trained items, a significant group by test interaction was observed ($\hat{\beta} = -0.177, SE = 0.046, z = -3.819, p = 0.001$). Post-hoc paired comparisons showed no difference for trained items between the two groups at posttest3 ($\hat{\beta} = -0.193, SE = 0.261, z = 0.741, p = 0.881$), though the night grouped showed higher accuracy than the morning group at posttest2 ($\hat{\beta} = -0.903, SE = 0.264, z = -3.420, p = 0.003$). These between-subjects comparisons reflect improvement on trained items between posttest2 and posttest3 for listeners in the morning group ($\hat{\beta} = -0.736, SE = 0.127, z = -5.795, p < 0.001$), and equivalent performance on trained items between posttest2 and posttest3 for listeners in the night group ($\hat{\beta} = -0.026, SE = 0.144, z = 0.179, p = 0.998$). For novel items, there was a main effect of group ($\hat{\beta} = -1.011, SE = 0.282, z = -3.583, p < 0.001$), indicating increased accuracy for the night compared to the morning group. However, there was no interaction between group and test ($\hat{\beta} = -0.031, SE = 0.043, z = -0.724, p = 0.469$), suggesting that listeners in the night group showed increased accuracy for novel items compared to the morning group at both posttest2 and posttest3. Collectively, the maintenance analysis provides evidence that both groups of listeners
maintained learning one-week after training, though those in the night group showed modestly higher accuracy compared to the morning group for novel items.¹

4. Discussion

In the current study, we examined the role of sleep-based memory consolidation on perceptual learning for noise-vocoded speech at both immediate and later time intervals. Listeners were trained to perceive noise-vocoded sentences with feedback and completed testing on trained and novel items before training, immediately after training, 12-hours later, and one-week later. Critically, one group of participants received training in the morning, while a different group of participants completed training in the evening. The goal of this design was to (1) assess the role of training on immediate learning, (2) examine how sleep-based memory consolidation promotes learning and generalization and (3) characterize the maintenance of gains over time. Overall, we found that participants trained in both the morning and evening groups showed robust and equivalent perceptual learning immediately after training. Following a 12-hour period of wakefulness for the morning group and sleep for the evening group, participants who received training in the evening showed stabilization of gains while participants who trained in the morning showed a degradation in performance. When participants returned one-week later, performance was equivalent for both the morning and evening groups for trained items, but generalization performance was modestly better for participants who received their initial training in the evening. We discuss each finding in detail in the sections below.

4.1. Effect of training on immediate perceptual learning

¹ See addendum for an additional set of analyses using proportion correct keywords on trial-level data.
We observed a robust perceptual learning effect for listeners trained to perceive noise-vocoded sentences. Relative to pretest performance, listeners in both the morning and evening training group demonstrated a significant improvement at posttest1. Our results replicate previous findings demonstrating perceptual learning of degraded speech following a period of lexically focused training with feedback (e.g., Davis et al., 2005; Hervais-Adelman et al., 2008; Loebach et al., 2008). Namely, Davis et al. (2005) demonstrated written feedback in the form of correct transcription bootstraps learning for noise-vocoded sentences. In our paradigm, listeners received the correct written transcription immediately after logging their response allowing for top-down feedback to update the mapping between the degraded speech signal and prelexical representations for future trials. The remarkable plasticity of the perceptual system is demonstrated by the poor performance at pretest with rapid improvement at posttest1 despite the relatively limited training session. We found no evidence that the time of day in which training occurs influences immediate perceptual learning.

In line with previous findings, accuracy for trained items was significantly higher relative to novel items. This finding replicates various studies in the noise-vocoded (e.g., Loebach et al., 2008; Hervais-Adelman et al., 2008; Dahan & Mead, 2010; Huyck et al., 2017). Listeners also showed greater improvement for trained items at posttest1 compared to novel items, which likely reflects familiarity and recency effects for the sentences heard during test. The fact that we observed generalization to untrained sentences demonstrates that listeners show some abstraction with even a brief training period and is in line with previous generalization findings for perceptual learning (e.g., Kraljic & Samuel, 2007; Kraljic & Samuel, 2006). We take this finding as evidence that brief, lexically-oriented training paradigms are appropriate for promoting learning and generalization to noise-vocoded sentences.
4.2. Role of sleep-based consolidation on learning

In our 12-hour follow-up test, we explicitly examined the role of sleep consolidation on perceptual learning for noise-vocoded speech. Recall, both the morning trained participants and evening trained participants underwent an identical training structure in the first session and returned for posttest2 approximately 12-hours later. For the morning trained participants, the 12-hour follow-up occurred in the evening of the same day they received training, thus an interval of wakefulness occurred between the initial session and posttest2. Conversely, the evening trained participants returned the following morning, resulting in a period of sleep in between the initial session and posttest2. A sleep-mediated consolidation effect would manifest as a difference in performance between the morning trained participants and evening trained participants at posttest2. Strikingly, we observed clear evidence that sleep-consolidation stabilized perceptual learning for noise-vocoded speech. Specifically, listeners trained in the evening showed equivalent performance for trained and novel items at a 12-hour time interval relative to performance immediately after training. Conversely, participants trained in the morning demonstrated a deterioration of both learning and generalization at the 12-hour follow-up compared to performance immediately following training. As performance at posttest1 was comparable between the two training groups, the change observed at posttest2 cannot be attributed to performance differences at an earlier time point. Rather, the unchanged performance for the evening group and relative decline in performance for the morning group suggests that sleep-consolidation stabilized gains obtained from training.

Our finding of unchanged performance after sleep for the evening group and deterioration in performance for the morning group is in line with previous findings demonstrating a stabilization effect of sleep on learning for other types of auditory signals. For example, using
synthetic speech items, Fenn et al. (2003) observed stabilized performance for participants trained in the evening and a decrease in performance over the course of the day for participants trained in the morning, that recovered only after a period sleep. Further, work on non-native speech learning has shown enhanced discrimination performance and stable identification performance for listeners trained in the evening relative to participants trained in a morning interval (Earle & Myers, 2015). Earle and Myers (2015) suggest that both sleep-consolidated memory and hearing interfering stimuli throughout the day play a role in acquiring non-native speech sounds, which allows those trained at night to benefit to a greater degree from training compared to those trained in the morning. Our results extend previous findings and indicate sleep facilitates training benefits for a different type of acoustic signal, that of noise-vocoded speech. Noise-vocoded speech learning differs from the attentional requirements of previous speech learning studies in that the quality of the signal requires a comprehensive adjustment for all prelexical representations – as opposed to more focused adjustment as in the case of non-native speech sound learning (e.g., Earle & Myers, 2014). Thus, we take this finding as evidence that sleep-based consolidation paradigms provide benefit for not just clear, but also degraded speech signals.

For our evening trained participants, we did not observe an enhancement effect following sleep – rather listeners showed stable performance after the 12-hour interval compared to a decline in performance for the morning trained listeners. This finding contrasts with predictions made by the “two-stage model” of sleep consolidation (Walker, 2005) and diverges from other work reporting enhancement effects for novel speech items following a period of sleep. According to this theoretical framework, acquisition of a new procedural skill occurs in two stages; first, a period of wakefulness fosters a time-sensitive stabilization process, followed by a
period of sleep that results in skill enhancement, or performance at a higher level than the level that was obtained prior to sleep. It is hypothesized that the enhancement observed following sleep, in the absence of other interfering stimuli, is related to synaptic strengthening that occurs during REM sleep (e.g., (Walker et al., 2003; Diekelmann & Born, 2007). We found no evidence that sleep enhanced performance, as the evening group never exceeded a level observed immediately after training. Instead, we report a decline in performance for the morning trained group only. Our results are consistent with findings in other domains demonstrating sleep-based stabilization of performance for motor sequence learning (e.g., Nettersheim et al., 2015), but diverge from studies demonstrating an enhancement role of sleep (e.g., Earle & Myers, 2015) who reported improved in non-native speech sound discrimination in their night trained participants, relative to their morning trained participants. Instead, our results are consistent with a model in which procedural skill learning transfers from short-term neural storage regions and stabilizes in long-term memory (Marshall & Born, 2007).

4.3. Maintenance of learning at a later time interval

A critical question we sought to examine in the current study was the role of sleep-based memory consolidation on long-term learning outcomes. As previously reviewed, many studies have implicated sleep-based consolidation in learning novel speech input (e.g., Fenn et al., 2003; Earle & Myers, 2015), but few studies have examined performance at a time interval significantly displaced from the initial training session. The goal with our design was to characterize performance for our morning and evening trained groups one-week after the initial training session. Given that our night training group showed stabilized performance after a period of sleep, while the morning trained group demonstrated a degradation in performance following a period of wakefulness, the one-week follow-up was used to observe whether the
night training benefit persisted in the long-term. We found that performance for the morning and evening trained groups was equivalent one-week later for trained items. This finding reflected unchanged performance for the evening group from posttest2 to posttest3, while the morning group showed an improvement from posttest2 to posttest3 that reflected a return to immediate post-training performance. Thus, it appears that after a week elapsed, both groups, regardless of the time of day in which training initially occurred, showed stabilization in performance for trained items. Our one-week follow-up finding is in line with other work demonstrating a recovery effect for participants trained in the morning following a period of sleep (e.g., Fenn et al., 2003; Qin & Zhang, 2019), though the recovery performance was only measured 24 hours following the initial training.

Notably, a difference was observed between the morning trained group and the evening trained group for novel items. While the evening trained group maintained their performance between posttest2 and posttest3 for novel items, the morning trained group performed poorer—performance was maintained for the morning group at posttest2 to posttest3 thus resulting in poorer accuracy relative to the evening group. The current study provides evidence that the performance level achieved prior to sleep is maintained in the long-term, as both groups showed no difference in performance when compared to their pre-sleep posttest. That the group level training difference in performance only emerged for novel items at the one-week follow-up is evidence that sleep might preferentially support the abstraction process required for generalization. Our findings converge with a growing evidence base demonstrating a facilitative role of sleep specifically on generalization for novel talkers (e.g., Xie et al., 2018; Qin & Zhang, 2019), words (Fenn et al., 2003; Fenn et al., 2013), and tasks (e.g., Earle & Myers, 2015). We extend previous findings on generalization to demonstrate that the sleep-based learning effect is
present for novel items well beyond the post-sleep testing period and support a theoretical framework in which a dissociation between rote-learning and abstraction processes are independently supported by sleep-related processes (e.g., Fenn et al., 2013).

4.4. Clinical implications and future directions

Auditory training is becoming an increasingly more integrated part of clinical audiology. Our results support the notion that listeners learning to adapt to degraded speech input, as is in the case of a CI user adapting to the acoustic input transmitted through the implant, benefit from a focused period of auditory training. There are currently no clinically agreed upon set of recommendations for translating auditory training to the patient population. To our knowledge, this is the first study to establish a role of sleep-consolidated learning for adaptation to the degraded speech signal of noise-vocoded speech. Our work suggests that the time of day in which a training session occurs may be an important factor to consider when designing training protocol. Namely, training in the evening immediately prior to sleep stabilized learning and generalization gains relative to training in the morning. As maintaining gains and promoting generalization is the focus of any rehabilitation program, training in close proximity to sleep may hold promise for optimizing future rehabilitation programs. We view the current study as a foundational step in examining sleep-related learning in CI users. Future research is aimed at explicating the relationship between speech learning and sleep in the patient population.

A growing body of research has implicated a role of interference in learning for other types of speech input (e.g., Earle & Myers, 2015). Training to a maximal criterion may be a means to overcome the stabilization effect of sleep alone and future research is aimed at examining how training parameters interact with facilitatory effects of sleep for degraded speech input (e.g., Fuhrmeister et al., 2020). A limitation of the current work is that our methods did not
allow a way to track sleep activity and as such draw direct conclusions about how sleep itself
influenced learning. For example, recent work has reported that sleep duration is closely related
to non-native speech sound learning (e.g., Earle et al., 2017), which raises the question of similar
effects for degraded speech. The use of sleep tracking methods will enhance conclusions for
future research. A strength of the current design was the inclusion of a one-week follow-up test
to assess maintenance of learning – however, the inclusion of even later time intervals will more
concretely contribute to questions about the long-term retention of perceptual learning.
III. Many tasks, same outcome: Role of training task on learning and maintenance of noise-vocoded speech

1. Introduction

Normal hearing (NH) listeners adapt to atypical speech input rapidly, even with relatively limited experience (e.g., Norris et al., 2003; Clarke & Garrett, 2004; Drouin et al., 2016; Davis et al., 2005). Characterizing the mechanisms that support dynamic adaptation to novel speech input in NH listeners has been the focus of many studies in the speech learning domain, with the broad goal of translating principles of speech adaptation for aural rehabilitation in clinical populations. One such population is that of cochlear implant (CI) users. CIs restore a sense of audition to profoundly hearing-impaired individuals. However, because they process spectral information with a limited number of electrodes, as opposed to thousands of sensory cells, the auditory signal generated by the CI is a coarse representation of the speech signal. For some CI users, adaptation to their implant requires months to years of practice to show adequate speech comprehension, while a subset of users fail to perform at a comprehension level considered to demonstrate an acceptable benefit – even with prolonged use (e.g., Schramm et al., 2002). Auditory training is a tool used to help CI users adapt to their implant, and may be necessary to optimally target the plasticity mechanisms that support improvements in speech perception (e.g., Fu & Galvin, 2007). However, there are many outstanding questions as to how to effectively implement training as part of the rehabilitation process, and, as a consequence, there is no standardized recommendation for designing training protocols aimed at improving speech comprehension in this population. Moreover, because of the high degree of individual variation in CI user outcomes (e.g., Niparko et al., 2010), it can be difficult to isolate specific training-related factors
that influence performance. Therefore, one approach to better understand how training variables interact with speech learning is to implement training studies in NH listeners exposed to acoustically degraded speech signals to isolate the role of specific training variables of interest on performance and then use those optimal parameters to observe performance in the clinical population.

To assess speech learning in NH listeners, many studies have used noise-vocoding as an acoustic manipulation to approximate the auditory signal transmitted through a CI. Initially, the noise-vocoded signal is described as “robotic” or “unintelligible” by the naïve NH listener, but given sufficient experience, the listener learns to map the noise-vocoded signal to meaning.

Studies using noise-vocoded signals are designed to assess comprehension benefits that reflect training, and thus most studies include a pre-test to establish a baseline before training occurs and a post-test administered immediately after training to determine the benefit gained from training. In the following sections, we review previous research on perceptual learning of noise-vocoded speech, with a focus on how the task used during training may contribute to different learning outcomes both in the immediate and at later time intervals.

**1.1. Role of lexical context on perceptual learning for noise-vocoded speech**

Listeners make use of both bottom-up coarticulatory cues and top-down contextual cues to guide perception of the speech signal. Bottom-up coarticulatory cues constrain the likelihood of subsequent phonemes in a principled fashion based on speech patterns (e.g., Liberman, Cooper, Shankweiler, & Studert-Kennedy, 1967; Repp & Liberman, 1984; Fowler, 1986).

Because individual phonemes are not produced in isolation, but rather in sequences, the state of the vocal tract at any given time reflects the previous, current, and subsequent segment (e.g., Miller & Eimas, 1995). With respect to top-down contextual cues, in most everyday listening
conditions, the speech signal is processed in the context of meaningful words or sentences. Listeners capitalize on sentence context such that preceding words help listeners predict succeeding words in a sentence (e.g., Elliot, 1995). Consider the phrase, *The doctor prescribed the*. A limited set of words are contextually appropriate for the last word in the sentence including *drug, medication, or treatment*. Given the sentence context, it would be unlikely for the final word in this example to be *dog or airplane*. These two levels of processing are not modular; instead, bottom-up processing at the phonemic level interacts during online processing with information at higher levels including the lexical, semantic, and syntactic levels (e.g., Elman & McClelland, 1988; Marslen-Wilson & Tyler, 1980).

A central focus in the speech perception domain involves characterizing how the speech signal is processed when speech acoustics are degraded. Under these conditions, NH listeners have been shown to use lexical context to learn to map the degraded signal to lexical representations. With respect to noise-vocoded speech, seminal research examining perceptual learning of noise-vocoded speech has demonstrated that perception of this challenging signal is not static – a brief training period can be leveraged to promote comprehensive changes in how the listener maps the degraded signal (e.g., Davis et al., 2005). In a series of experiments, Davis and colleagues (2005) trained NH participants to perceive noise-vocoded sentences and measured comprehension (indexed as proportion of words correct) before training and immediately after training under a variety of training conditions. First, they investigated the role of feedback on perceptual learning for noise-vocoded sentences. Participants transcribed noise-vocoded simple sentences and received auditory feedback in the form a clear version of the sentence on each trial. For one group, the training presentation began with the distorted sentence, followed by the clear sentence, and ended with another repetition of the same distorted sentence.
In a different condition, the training presentation included two repetitions of the distorted sentence in a row, followed by the clear version. They predicted that hearing the clear version of the sentence in the middle of the two distorted versions would promote heightened perceptual learning of the noise-vocoded sentences because the identity of the words of the sentence would be known – creating what they termed a “pop-out” effect. Indeed, the results from this experiment indicated that the distorted-clear-distorted presentation order did in fact lead to greater perceptual learning relative distorted-distorted-clear presentation order. They take this finding as evidence that sentence content supports a comparison process between the distorted and clear sentences that allows changes in prelexical representations to be updated dynamically.

In a follow-up experiment, they examined the role of feedback modality on learning outcomes. Specifically, listeners were presented training items in the distorted-clear-distorted order (that maximized perceptual learning), but critically one group received the clear version in the form of the clear auditory version of the sentence, while in another condition the clear feedback was in the form of correct written transcription of the sentence. The results showed that learning was comparable across both feedback modalities, suggesting that the underlying mechanism that supported learning was amodal and thus occurred at the phonological level or higher.

The results reported from Davis et al. (2005) establish that sentence feedback supports rapid perceptual learning, whether it appears in either written or auditory form. However, a critical question is the degree to which the sentence context itself plays a role on perceptual learning as the previously described experiments used sentences in made up of real words in simple contexts. In a different set of experiments, Davis et al. (2005) sought to explicate the role of sentence context on training related changes in perception of noise-vocoded speech. In their manipulation, they exposed listeners to two noise-vocoded sentence manipulations. In one
condition, the sentences were composed of real English words, while in another condition the sentences were composed of nonwords. Feedback was given in both conditions in the form of distorted-clear transcription-distorted. A clear effect of lexicality was noted; learning was not observed when the sentences were composed strictly of nonwords. Further, using a different manipulation, Davis et al. (2005) determined that changes in performance are not necessarily linked to sentence level meaning or syntactic structure, but instead appear to be related to the lexicality of the sentence itself. It was observed that sentences composed of real words without sentence level meaning (e.g., *The effect supposed to the consumer*) and Jabberwocky sentences composed of nonwords with the same syntactic structure as English (e.g., *The tekeen gerund to the sumeeun*) promoted equivalent perceptual learning to sentences composed of semantically and syntactically appropriate real words (e.g., *The police returned to the museum*).

Collectively, this series of studies suggests that lexical context itself is a driving factor behind perceptual learning of noise-vocoded sentences, and thus paradigms designed with lexical contexts should promote maximal gains. Indeed, other work has replicated and extended these findings to demonstrate that training paradigms that use lexical transcription tasks, provide lexical feedback, and use sentences composed of real words all promote robust learning that can generalize across talkers (e.g., Huyck et al., 2017). Across these various findings, it has been proposed that lexical information is necessary for learning to occur. However, there are at least three independent components of lexical processing that could contribute to the lexical advantage observed in these studies. Because lexical context, lexical feedback, and lexical focus (as indexed as task-based attention towards the lexicon while engaging in a lexically oriented transcription task) always co-occurred in these training paradigms, the independent contributions of each component on learning cannot be explicated in this design. That is, is lexical context
itself sufficient for learning to occur, even in the absence of a lexical transcription task and/or lexical feedback? In the following section, we review other routes to perceptual learning of noise-vocoded speech, which demonstrate that lexically oriented training tasks are not necessarily required for perceptual learning to occur.

1.2. Role of non-linguistic training tasks on perceptual learning for speech

As reviewed above, standard perceptual learning paradigms use lexically oriented training tasks with feedback (e.g., Davis et al., 2005). However, an important consideration in designing training protocols is to determine whether comparable comprehension benefits can be obtained using non-linguistic training tasks. That is, what type of task engagement is both necessary and sufficient for learning to occur? Previous research has demonstrated that a linguistic benefit for processing noise-vocoded speech can be observed using the indexical task of talker identification (Loebach et al., 2008). In their study, Loebach and colleagues (2008) manipulated the task used to train three groups of NH listeners to perceive noise-vocoded sentences. The training stimuli for all three groups consisted of high-predictability sentences produced by five talkers. During training, one group of listeners completed a lexical transcription task, one group of listeners completed a talker identification task, and a different group of listeners completed a gender identification task. All training groups received feedback consisting of the correct response based on the training task. Learning was assessed before training (pre-test) and immediately after training (post-test) for anomalous (meaningless, but composed of real words) and low-predictability sentences. Generalization was assessed across sentences using novel anomalous and low-predictability items. During test, all listeners completed the same lexical transcription task on low-predictability and anomalous noise-vocoded sentences and did
not receive feedback. Accuracy was assessed as correct transcription of the final word in the sentence.

In line with previous studies on perceptual learning for noise-vocoded speech, Loebach et al. (2008) found that the lexical transcription group showed a significant improvement in performance at post-test relative to performance at pre-test. Remarkably, they also observed that the listeners trained on talker identification showed transcription performance at test equivalent to those trained on transcription. That is, even though the talker identification group never completed a transcription task at test, they still demonstrated comparable transcription performance to the group of listeners who did complete the transcription task. In contrast, the gender identification training group demonstrated a significantly poorer benefit from training relative to the two other groups. In addition, there was poorer generalization in the gender identification group relative to the talker identification and transcription training groups, who did not differ from each other. Overall, Loebach et al. (2008) conclude that perceptual learning of noise-vocoded speech can occur in the absence of a linguistic training task, so long as the task is sufficiently engaging and provides deep engagement towards the acoustic details of the stimuli.

Indeed, other studies examining task-based listening strategies on perceptual learning of noise-vocoded speech suggest that learning is dependent on task-based attentional requirements. For example, Huyck and Johnsrude (2012) trained listeners on noise-vocoded meaningful sentences. While one group attended to sentence-level meaning via a transcription task, a different set of listeners attended to either auditory or visual distractors that co-occurred with the noise-vocoded sentences and completed a target detection task. Feedback was not provided for any training group, regardless of task. Relative to their pre-test baseline performance, the groups who attended to the auditory or visual distractor during training performed significantly poorer
than the group who attended to the sentence-level meaning. The distractor training groups performed at the same level as their no training control group, suggesting that they received limited benefit from the training period at all – despite hearing the same stimuli as the transcription training group. Instead, it appeared that the way in which the different groups engaged with the stimuli through their assigned task predicted post-test performance.

Follow-up work by Wild et al. (2012) used functional magnetic resonance imaging to examine the role of task-based attention on speech processing for clear or noise-vocoded sentences. Using the same task structure (transcription, auditory distractor, or visual distractor) and meaningful sentences for their training groups as in Huyck and Johnsrude (2012), they found that while clear speech did not require direct attention to be processed, attention greatly promoted processing for noise-vocoded speech. Namely, while all training groups showed recruitment of auditory cortex regions while listening to noise-vocoded sentences, frontal regions implicated in higher-order language processing were only recruited for the group attending to the speech signal via a transcription task for noise-vocoded sentences. They suggest that this neural pattern of activation reflects engagement in effortful listening for the transcription group that was not observed for the distractor groups. That recruitment of frontal regions was only observed for the transcription training group, and not the auditory or visual distraction groups despite equal access to the auditory signal, suggests that the attentional focus of the listener plays a significant role in how the signal is learned.

Together, findings from Loebach et al. (2008) suggest that a sufficiently engaging non-linguistic training task, like talker-identification, can promote equivalent comprehension outcomes to a group of listeners who explicitly engaged in a transcription task. This finding converges with work other work demonstrating that processing of noise-vocoded of sentences
relies uniquely on the attentional engagement which can be leveraged using different training
tasks (Huyck & Johnsrude, 2012; Wild et al., 2012). Together, these findings suggest that
learning is graded relative to task difficulty, such that cognitively easier tasks, like attending to
an auditory or visual distractor or identifying the gender of the talker, do not promote the same
degree of improvement as tasks that require deep engagement, such as transcription or talker
identification. A relationship between task demands on learning is predicted by a depth of
processing framework, which posits that increased engagement as a consequence of challenging
training tasks will maximize perceptual learning (e.g., Craik & Lockhart, 1972). In this
framework, it is argued that transfer of non-linguistic training to linguistic tasks occurs when the
task is challenging enough to promote deep engagement towards the acoustic-level details of the
speech stimuli. The task of talker identification requires considerably more controlled attention
to the acoustic details of the speech signal and deeper processing compared to the task demands
required to identify the gender of the talker. This is reflected in the training accuracy data of
Loebach et al. (2008) who reported that accuracy was only around 50% for the talker
identification training group and near ceiling for the gender identification training group,
suggesting that the talker identification task was indeed harder for listeners relative to the gender
identification task.

An alternate approach to predicting perceptual learning outcomes is made by the transfer
appropriate processing (TAP) theories of learning and memory (e.g., Morris et al., 1977; Franks
et al., 2000). TAP theories posit task-specific learning and argue that performance at test is
heightened when the task used at test is identical to the task used during training. Under this
framework, training on non-linguistic tasks should not promote improvements in linguistic tasks.
This framework was not supported by the findings reported by Loebach et al. (2008). While the
transcription group improved in performance relative to pre-test, the talker identification training group improved to the same degree, suggesting that task-specific training is not required to achieve heightened perceptual learning.

To sum, many studies have implicated a privileged role of lexical information on perceptual learning for noise-vocoded speech (e.g., Davis et al., 2005), while other work has argued that the training task demands – as opposed to lexical-based attention – are the driving factor as to whether perceptual learning is observed (e.g., Loebach et al., 2008). A challenge when considering these findings together is the differences in how lexical influence is characterized. Specifically, lexical influence could be defined as lexically oriented attentional strategy through a transcription task, supporting lexical context, or lexical feedback. In the Loebach et al. (2008) design, listeners across training groups listened to high-predictability sentences and received feedback. Thus, even though the talker-identification training group did not engage in a lexical task, they still had supporting lexical and semantic context given the high-predictability sentences used during training, which itself could independently drive perceptual learning. We have evidence that adaptation to ambiguous speech sounds can occur irrespective of the training task – with equivalent perceptual learning observed for perception-focused and comprehension-focused listening tasks, so long as supporting lexical context is available (Drouin et al., 2018). Moreover, because most studies on perceptual learning of noise-vocoded speech provide lexical feedback (e.g., Davis et al., 2005), it is unknown whether feedback itself is strictly necessary to promote learning because a condition using a transcription task without feedback was not included. These discrepancies across studies motived the design of the current study, outlined below, to isolate individual contributions of lexical focus, lexical feedback, and lexical context on perceptual learning of noise-vocoded speech.
1.3. The current study

In the current study, three groups of participants completed different tasks during a training phase where they heard noise-vocoded sentences. During training, all participants heard anomalous sentences. One group of listeners completed a transcription task and received lexical feedback (transcription group), one group of listeners completed a talker identification task and received talker feedback (talker identification group), and the third group of participants completed a transcription task and received no feedback (no feedback group). Listeners were tested prior to training, immediately after training, and one-week following training in order to examine test performance over time. The test and training items were identical across participants – only the training task differed across the three groups. We examined performance at test on low-predictability and high-predictability sentences, and measured accuracy for words at sentence-onset and words at sentence-offset. Using this design, the current study aimed to extend previous findings on perceptual learning of noise-vocoded speech by (1) examining the independent contributions of lexical influence on learning outcomes, (2) characterize both the learning (immediate) and maintenance (long-term) of training gains, and (3) evaluate test performance using an understudied metric in the speech learning domain – that of sentence-initial words. We describe the motivation for each aim in turn.

The first goal was to characterize the independent contributions of lexical influence on learning outcomes. Our design allowed for examination of individual aspects of lexical influence on perceptual learning of noise-vocoded speech. As illustrated in Figure 1, all training groups heard sentences composed of real words during training, though there was no supporting sentence context because all sentences were anomalous. The inclusion of the transcription training group acted as a reference because this group completed a lexical transcription task
(lexical focus) and received lexical feedback. The no feedback group completed a lexical transcription task but did not receive feedback – allowing the individual contribution of feedback on learning to be determined by comparing performance to the transcription group with feedback. The talker identification group did not complete a lexical task, nor did they receive lexical feedback – following the protocol used in Loebach et al. (2008). However, recall that the training phase in Loebach et al. (2008) used high-predictability sentences, as opposed to anomalous sentences, which could drive learning independently of task. We hypothesized that if lexical feedback itself is necessary to maximize perceptual learning, then listeners trained on the transcription task without feedback will show decreased learning than listeners trained on transcription with feedback. However, if completing the lexical task during training is sufficient for learning to occur, then we would expect the no feedback group to perform equivalently to the transcription group. With respect to the talker identification, we aim to parse out whether task engagement, regardless of lexicality, promotes perceptual learning in the absence of useful sentence context. If perceptual learning is equivalent to performance for the transcription group it would suggest that task engagement, as opposed to sentence context, drives the perceptual learning process, as predicted by Loebach et al. (2008) and in line with predictions made by the depth of processing framework. However, if learning is diminished in the talker identification group relative to the transcription group, it would suggest that learning using a non-linguistic task is maximized when supporting sentence context is available.
The second goal of the study was to characterize both the learning (i.e., performance immediately after training) and maintenance (i.e., performance at a later time interval) of training gains. There is currently limited research examining the long-term outcomes of perceptual learning of noise-vocoded speech and as such it is unknown whether gains obtained from training persist in the long-term. This is particularly important to consider in the context of how gains obtained using different training tasks might change over time. As neither theoretical framework makes predictions as to how learning persists over time, we sought to characterize the pattern of learning immediately after training and one-week later to determine whether the same task-independent pattern of learning was observed. We hypothesized that if the training task differentially affects the maintenance of gains over time, such that maintenance is linked to task-specificity during encoding, then the talker identification group will show diminished gains at the one-week follow-up relative to the transcription and no feedback group. However, if training task does not affect either learning or maintenance of gains over time, then performance will be equivalent for all training groups both in the immediate and at the later time interval. This result
would be consistent with the hypothesis that engagement in an engaging training task is sufficient to maintain comparable comprehension outcomes.

The third, and final, goal of the study was to characterize learning and maintenance using a relatively understudied metric in the speech learning domain – that of accuracy for sentence-initial words. As reviewed above, listeners use both contextual cues and coarticulatory cues to facilitate predictions about succeeding words in a sentence. We chose to implement both low-predictability and high-predictability sentences during test to determine how training, in the absence of sentence context cues, promotes perceptual learning for two contextually different sentence types. It is well established that listeners transcribing high-predictability sentences show context effects that promote greater accuracy for sentence-final words relative to sentence-initial words (*The cabin was made of logs*) (e.g., Elliot, 1995). Conversely, accuracy differences as a function of word position are not expected for low-predictability sentences as they lack the contextual cues used to facilitate word-final identification (*Tom has been discussing the beads*). Previous studies on perceptual learning of noise-vocoded sentences typically report accuracy in terms of word-final accuracy or summed total across several keywords in a sentence. However, word-final accuracy may be already enhanced based on the sentence context (i.e., high-predictability sentences) and recency effects, or the retrieval of items still being actively maintained (e.g., Postman & Phillips, 1965). The use of low- and high-predictability sentences at test in our design allows for context to be controlled with respect to final keyword accuracy. Moreover, by examining accuracy of word-initial keyword in addition to word-final we can isolate for the potential additive effect of recency. Thus, examining word-initial accuracy for high-predictability sentences might be a more sensitive metric to index perceptual learning as we expect performance to be high for word-final accuracy even without training given the context.
Currently, it is unknown if training promotes the same degree of improvement for word-initial items as has been observed for word-final items. For low-predictability items, we hypothesized that training would promote an equivalent improvement to words at both the beginning and end of the sentence as context support was not available to boost final keyword accuracy. At baseline, we anticipated that accuracy for the final keyword would be greater than accuracy for the initial keyword for high-predictability items as sentence context would be available. However, after training, we hypothesized that we would observe a greater change in accuracy for word-initial items relative to word-final items for high-predictability sentences, such that performance at sentence onset would match performance at sentence offset. This result would suggest that training promotes a global change in sentence understanding irrespective of sentence context.

2. Methods

2.1. Participants

Participants (n = 108) were adults between 18 and 35 years of age (mean = 25 years, SD = 5 years) who were recruited from the Prolific participant pool (http://www.prolific.co). Participants were randomly assigned to one of three training groups (n = 36 in each training group). We used the Prolific filter option to limit participants who self-reported that they were a monolingual speaker of English, born in the United States and currently resided in the United States, and that they had no history of language related disorders. All participants passed a headphone screen administered prior to the start of each experimental session (Woods et al., 2017). All participants completed session one and were invited to participate in session two; 71 participants.

---

2 An additional 14 participants were tested but excluded from data analysis due to failure to pass the headphone screen (n = 13) or failure to complete the transcription task (n = 1).
participants completed both sessions (n = 24 in no feedback group; n = 25 in transcription group n = 22 in talker identification group).

2.2. Stimuli

The stimuli were obtained from the Nationwide Speech Project (NSP) corpus (Clopper & Pisoni, 2006). All recordings in this corpus were captured in a sound-attenuated booth using a 44.1 kHz sampling rate and 16-bit encoding. Five talkers from the New England region were selected (3 male; 2 female) from this corpus. For each talker, stimuli consisted of high-predictability (n = 45) and low-predictability (n = 45) sentences drawn from the Speech Perception in Noise (SPIN) test (Kalikow, Stevens, & Elliott, 1977), in addition to anomalous sentences (n = 30). The SPIN sentences consisted of five- to eight-word phonetically-balanced sentences. We identified two keywords in each sentence that reflected the beginning and end of the sentence, respectively. The high predictability sentences provide semantically rich contexts such that the final keyword in each sentence was predictable based on preceding words (e.g., *The doctor prescribed the drug*). For the low predictability sentences, the final keyword in each sentence was not predictable based on preceding words (e.g., *Mr. Smith knew about the bay*). The anomalous sentences were semantically meaningless, though syntactically correct (e.g., *Ruth’s problems are made from bark*). We presented anomalous sentences during the training period, while low-predictability and high-predictability sentences were presented during Test-00, Test-01, and Test-02. Test sentences were arranged in three lists so that novel sentences were presented at each time point and no test sentence was ever heard more than one time.

Using the Praat software (Boersma, 2001), the selected sentences from the NSP corpus were edited to remove trailing and leading silence. All stimuli were then processed using customized speech processor settings using the Tiger CIS software (http://tigerspeech.com)
following the parameters reported in Loebach et al. (2008). The noise-vocoding parameter was selected with white noise as the carrier. First, an eight-channel simulation was selected with eight non-linearly spaced channels. In line with Loebach et al. (2008), we set frequency bandwidth based on the Greenwood function (Greenwood, 1990) and used a lower frequency range of 200 Hz, an upper frequency range of 7000 Hz, and a 24 dB/octave filter slope. A low pass filter, with a cutoff frequency of 400 Hz and 24 dB/octave slope, was used to derive the fine structure in each band, which was then replaced by white noise modulated by the amplitude envelope for each band. All stimuli were then RMS normalized in Praat to ensure equal intensity across recordings.

2.3. Procedure

The procedure was approved by the Institutional Review Board at the University of Connecticut. The experiment was hosted on the Gorilla web-based platform (Anwyl-Irvine et al., 2019). All participants gave informed consent before participating in the experiment. First, all participants passed a headphone screen using the procedure outlined in Woods et al. (2017), designed to screen for compliance with headphone use in online experiments. The auditory stimuli were presented over headphones and participants typed their responses using the keyboard on their computer.

The experiment consisted of a single training phase and three test phases. Session one consisted of pretest (Test-00), training, and an immediate post-test (Test-01). Session two was completed one-week following session one and consisted of a single test (Test-02). Session one was approximately 35 minutes in duration and session two was approximately 20 minutes in duration. To promote increased retention of participants in session two, we allowed participants to complete the experiment one day before and one day after the one-week mark, resulting in a
time window for session two of 6 – 8 days post session one (mean days post-training = 6, SD = 1 day). Participants were randomly assigned to one of three training groups: transcription training, talker identification training, or no feedback training. As shown in Figure 2, the structure of the experiment was identical across participants; the only difference among training groups was the specific training task.

**Figure 2.** Graphic illustrating the task and session structure across the no feedback transcription training group (purple), transcription with feedback training group (tan), and talker identification group (blue). The only difference between the training groups was the task performed during training.

**Training.** The training phase was completed during session one. First, the training phase began with a familiarization for all participants consisting of ten trials. The stimuli used during familiarization consisted of two sentences produced by each of the five talkers. For participants in the transcription and no feedback training group, the auditory presentation of each sentence
was coupled with a written transcription of the sentence on screen. For participants in the talker identification training group, each auditory presentation of the sentence was coupled with a written identifier on screen of which talker was speaking. The goal of the familiarization trials was to promote a degree of knowledge of the talkers, sentences, and vocoding manipulation before the experiment began. The familiarization trials for every training group were not used in any analyses and no responses were recorded.

In training, participants heard 30 anomalous sentences produced by each of the five talkers, resulting in 150 trials total. Every sentence was heard five times with one production per talker. Participants were randomly assigned to one of three training groups. While the stimuli were identical between the training groups, the task structure differed. For every training group, each trial began with the auditory presentation of a noise-vocoded sentence produced by a single talker. Following this, participants in the transcription training group were asked “What do you hear?” while a textbox appeared on screen. The participant typed their response in the textbox, pressed the enter key to register their response, and then received written feedback in the form of correct transcription of the sentence that appeared on screen for 2500 milliseconds. Participants in the no feedback group were asked “What do you hear?” while a textbox appeared on screen. The participant typed in their response and pressed the enter key to register their response. Critically, the no feedback group did not receive any form of feedback. Participants in the talker identification group were asked “Who do you hear?” while a display of five response boxes with the names Jim, Rachel, Erik, Emily, and Lenny appeared on screen in a fixed order. They clicked on a response box to register their response and then received written feedback in the form of the correct talker that appeared on screen for 2500 milliseconds. Participants in the transcription and no feedback group were instructed to guess any words in the sentence, while participants in the
talker identification group were instructed to take their best guess of which talker was speaking. The interstimulus interval was 1000 milliseconds timed from the response for the no feedback training group and the offset of feedback for the transcription and talker identification training groups.

**Test.** Three test sessions were administered. A pre-test baseline measure was administered to assess performance prior to training (Test-00) and an immediate post-test (Test-01) was administered immediately after training in session one. A follow-up post-test (Test-02) was administered one-week later. We used equal proportion of high-predictability and low-predictability sentences at each test, yielding 30 trials at each test (15 high predictability trials; 15 low predictability trials). The structure of the test task was identical across participant groups. First, participants heard an auditory presentation of a noise-vocoded sentence produced by one of five talkers. Next, the question “What do you hear?” appeared on screen with a textbox. The participant typed their response in the textbox and pressed the enter key to register the response. Critically, participants did not receive any form of feedback during the test phase. Participants were instructed to guess any words in the sentence. The interstimulus interval was 1000 milliseconds, timed from the participant’s response.

3. **Results**

3.1. **Training**

For participants in the transcription training group and no feedback training group, responses for each keyword were considered correct if the participant’s response matched the keyword in question. Each trial was scored by the first author. In addition to identical matches between keyword and transcription, a keyword was considered correctly transcribed if the transcription contained a typo or misspelling, with the caveat that the typo/misspelling did not
change tense or create a different word. For example, given the target sentence \textit{maple syrup is made from sap}, all of the following would be considered as correct responses for the keyword \textit{maple: maple, maplej, mpel, miple, maid}. The following would be examples of incorrect responses for the keyword \textit{maple: maples, mavel, mpl}. For the talker identification group, a trial was considered correct if the talker was identified correctly.

Figure 3 shows the distribution of mean accuracy scores across participants for the three training groups, which were determined by first calculating mean accuracy across the 150 training items for each participant. For the no feedback and transcription training groups, a trial was marked as accurate if final keyword was correct, in line with the scoring procedure used in Loebach et al. (2008). For the talker identification training group, a trial was marked as accurate if the correct talker was identified. Visual inspection of Figure 3 suggests a training group effect in accuracy such that accuracy was highest for the transcription, followed by the no feedback, and talker identification training group.

\textbf{Figure 3.} Boxplots showing the distribution of mean proportion correct responses during the training phase for the no feedback, lexical, and talker training groups. The dashed line indicates the level of chance for the talker ID training group.
To examine this pattern statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear mixed effects model (GLMM) with the binomial response family implemented in the lme4 package in R (Bates et al., 2015). The Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using lmerTest (Kuznetsova et al., 2017). The fixed effect was training condition (treatment-coded, with the transcription group as the reference level). The random effects structure consisted of random intercepts by subject, random intercepts by sentence, and random intercepts by talker.

Compared to the transcription with lexical feedback condition, accuracy during training was lower for both the talker identification condition ($\hat{\beta} = -1.547, SE = 0.123, z = -12.536, p < 0.001$) and the transcription with no feedback condition ($\hat{\beta} = -0.748, SE = 0.123, z = -6.062, p < 0.001$). Because the condition coding structure does not support direct comparison between the talker identification and transcription with no feedback conditions, post-hoc comparisons were conducted with the pairs() function of the emmeans package (Lenth, 2020), using the Tukey method to adjust for multiple comparisons (here and throughout). In this comparison, the no feedback group and talker identification group differed significantly ($\hat{\beta} = -0.800, SE = 0.122, z = -6.543, p < 0.001$), providing evidence that accuracy for the transcription group was highest, followed by the no feedback group, followed by the talker identification group.³

³ We examined training performance in our subset of participants who returned for session two (n = 71) to confirm reliability of this sample relative to the full sample using the same GLMM model structure. Compared to the transcription with lexical feedback condition, accuracy during training was lower for both the talker identification condition ($\hat{\beta} = -1.449, SE = 0.158, z = -9.182, p < 0.001$) and the transcription with no feedback condition transcription ($\hat{\beta} = -0.749, SE = 0.154, z = -4.854, p < 0.001$). Because the condition coding structure does not support comparison between the talker identification and transcription with no feedback conditions directly, post-hoc comparisons were conducted and revealed that accuracy for the talker identification group was lower than the no feedback training group ($\hat{\beta} = -0.699, SE = 0.158, z = -4.429, p < 0.001$), replicating the finding in the full sample of a monotonic decrease in accuracy across the transcription, no feedback, and talker identification training groups.
3.2. Test

The same scoring procedure outlined for the transcription task was used to score test items. We considered test performance in two analyses: learning and maintenance. In each, low-predictability and high-predictability items were analyzed separately. The learning analysis (n = 108) assessed performance between Test-00 and Test-01 to examine performance immediately after training. The maintenance analysis assessed performance between Test-01 and Test-02 (n = 71) to examine how performance is maintained over a one-week period. Each is addressed in turn.

Learning. The learning analysis examined performance between Test-00 and Test-01. As outlined previously, Test-00 occurred before training and Test-01 occurred immediately after training. Separate learning analyses were performed on the low-predictability items and high-predictability. First, we examined the learning analysis for low-predictability items. In Figure 3, we depict the distribution of mean proportion correct responses for keyword 1 (left panel) and keyword 3 (right panel) at Test-00 and Test-01. Visual inspection of Figure 3 reveals that performance at Test-01 was higher than performance at Test-00 for all training groups across keywords for low-predictability items. Moreover, it appears that all training groups demonstrated similar improvement regardless of word position – keyword 1 and keyword 3 appear to demonstrate equivalent improvement.

To examine these patterns statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a GLMM with the fixed effects of training condition (treatment coded; transcription as the reference), test session (Test-00 = -1, Test-01 = 1), and keyword (KW1 = -1, KW3 = 1). The model also included random intercepts by subject, sentence, and talker, as well as random slopes by subject for test session and keyword.
For low-predictability items, the model revealed a main effect of test session ($\hat{\beta} = 0.590$, $SE = 0.053$, $z = 11.084$, $p < 0.001$), indicating that performance at test-01 was higher in accuracy relative to performance at test-00. As can be calculated from Table 1, this difference reflects a

\[ p(Correct) \]

\[ \text{Condition} \]

\[ \text{NoFeedback} \]

\[ \text{Transcription} \]

\[ \text{TalkerID} \]

\[ \text{Low predictability sentences} \]

\[ \text{KW1} \]

\[ \text{KW3} \]

\[ \text{Task} \]

\[ \text{Test-00} \]

\[ \text{Test-01} \]

\[ \text{Test-00} \]

\[ \text{Test-01} \]

\[ 0.00 \]

\[ 0.25 \]

\[ 0.50 \]

\[ 0.75 \]

\[ 1.00 \]

\[ \text{Figure 4.} \] Boxplots showing the distribution of mean proportion correct responses in the learning analysis (Test-00 vs. Test-01) for each training group across low-predictability items. Performance at keyword 1 is shown on the left panels and performance at keyword 3 is shown on the right panels.

We examined learning in our subset of participants who returned for session two ($n = 71$) to confirm reliability of this sample relative to the full sample using the same GLMM model structure for low-predictability items. The model revealed a main effect of test session ($\hat{\beta} = 0.606$, $SE = 0.062$, $z = 9.766$, $p < 0.001$), a two-way interaction between the no feedback group and transcription group by keyword ($\hat{\beta} = -0.195$, $SE = 0.093$, $z = -2.110$, $p = 0.035$), a two-way interaction between the test session and keyword ($\hat{\beta} = 0.154$, $SE = 0.059$, $z = 2.604$, $p = 0.009$), and a three-way interaction between the no feedback group and transcription group by test session and keyword ($\hat{\beta} = -0.200$, $SE = 0.083$, $z = -2.399$, $p = 0.016$). Post-hoc analyses revealed the two-way interaction between training session and keyword was driven by an equivalent
mean increase of 20.2% after training. With respect to training task, there was no difference between the talker identification and the transcription group \( (\hat{\beta} = -0.198, SE = 0.127, z = -1.557, p = 0.120) \), nor was there a reliable difference between the no feedback and transcription group \( (\hat{\beta} = 0.064, SE = 0.127, z = 0.504, p = 0.614) \). We also observed no effect of keyword on accuracy \( (\hat{\beta} = -0.001, SE = 0.053, z = -0.018, p = 0.985) \), as would be expected using these low-predictability items. The two-way interaction between test session and training conditions was not reliable between the no feedback and transcription training group \( (\hat{\beta} = -0.077, SE = 0.074, z = -1.030, p = 0.303) \), but was reliable between the talker identification and transcription training group \( (\hat{\beta} = -0.193, SE = 0.074, z = -2.589, p = 0.010) \). Post-hoc comparisons revealed this two-way interaction was driven by a change in performance from Test-00 to Test-01 for the transcription group \( (\hat{\beta} = -1.180, SE = 0.106, z = -11.084, p < 0.001) \) and talker identification group \( (\hat{\beta} = -0.795, SE = 0.105, z = -7.549, p = 0.001) \), and not a between-groups difference at either Test-00 \( (\hat{\beta} = 0.005, SE = 0.144, z = 0.035, p = 1.000) \) or Test-01 \( (\hat{\beta} = 0.391, SE = 0.151, z = 2.590, p = 0.100) \). Thus, it appears the learning effect for low-predictability items was equivalent across all training groups, regardless of task. The two-way interaction between test session and keyword was reliable \( (\hat{\beta} = 0.102, SE = 0.049, z = 2.092, p = 0.036) \). Post-hoc comparisons revealed this was driven by a change in keyword accuracy from Test-00 to Test-01 for keyword 1 \( (\hat{\beta} = -0.973, SE = 0.083, z = -11.664, p < 0.001) \) and keyword 3 \( (\hat{\beta} = -1.029, SE = 0.084, z = -12.316, p < 0.001) \), and not a difference in performance between keywords at Test-00.

change in keyword accuracy between sessions for keyword 1 \( (\hat{\beta} = -0.932, SE = 0.101, z = -9.263, p < 0.001) \) and keyword 3 \( (\hat{\beta} = -1.105, SE = 0.102, z = -10.831, p < 0.001) \). All other post-hoc comparisons for every main effect and interactions did not reveal any reliable differences. Thus, the learning performance of the subset of participants for low-predictability items was comparable to the learning performance observed for the full set of participants.
Finally, we observed a three-way interaction for the transcription training group relative to the no feedback group for test session and keyword, but post-hoc analyses revealed no group differences in accuracy of keyword 1 at test-00 ($\hat{\beta} = -0.0672, SE = 0.179, z = -0.375, p = 1.000$), or test-01 ($\hat{\beta} = -0.234, SE = 0.187, z = -1.251, p = 0.985$), nor was there a group accuracy difference of keyword 3 at test-00 ($\hat{\beta} = -0.214, SE = 0.172, z = -1.243, p = 0.986$), or test-01 ($\hat{\beta} = 0.259, SE = 0.177, z = 1.468, p = 0.949$). This suggests that the three-way interaction was spurious, as post-hoc analyses revealed learning was achieved equally across groups and keywords. All other two-way and three-way interactions were not reliable.

Next, we examined learning for high-predictability items. In Figure 5 we depict the distribution of mean proportion correct responses for keyword 1 (left panel) and keyword 3 (right panel) at Test-00 and Test-01 for high-predictability items. Visual inspection of Figure 4 reveals that performance at Test-01 was higher than performance at Test-00 for all training groups across keywords and sentence type. Moreover, it appears accuracy for keyword 3 was higher for accuracy of keyword 1 each time point, but a learning effect is apparent for both keywords.
**Figure 5.** Boxplots showing the distribution of mean proportion correct responses in the learning analysis (Test-00 vs. Test-01) for each training group across high-predictability items. Performance at keyword 1 is shown on the left panels and performance at keyword 3 is shown on the right panels.

The high-predictability analysis was performed using the same model structure described above. The model revealed a main effect of test session ($\hat{\beta} = 0.357$, $SE = 0.075$, $z = 4.780$, $p < 0.001$), indicating that performance at Test-01 was higher in accuracy relative to performance at Test-00. As can be calculated from Table 1, this difference is driven by a 12.9% increase in performance.

---

5 We examined learning in our subset of participants who returned for session two ($n = 71$) to confirm reliability of this sample relative to the full sample using the same GLMM model structure for high-predictability items. The model revealed a main effect of test session ($\hat{\beta} = 0.446$, $SE = 0.070$, $z = 6.385$, $p < 0.001$) and keyword ($\hat{\beta} = 0.440$, $SE = 0.069$, $z = 6.357$, $p < 0.001$). All other training group and interaction comparisons did not reveal any reliable differences. Thus, the learning performance of the subset of participants for high-predictability items was comparable to the learning performance observed for the full set of participants.
mean accuracy for high-predictability items at Test-01 relative to Test-00. With respect to training task, there was no difference between the talker identification and the transcription group ($\hat{\beta} = -0.078$, $SE = 0.166$, $z = -0.470$, $p = 0.638$), nor was there a reliable difference between the no feedback and transcription group ($\hat{\beta} = 0.101$, $SE = 0.167$, $z = 0.608$, $p = 0.543$).

Unlike the low-predictability items, we observed a significant effect of keyword on accuracy ($\hat{\beta} = 0.438$, $SE = 0.057$, $z = 7.661$, $p < 0.001$), indicating that accuracy for keyword 3 was higher relative to accuracy for keyword 1, in line with a role of context. The two-way interaction between test session and training condition were not reliable between the talker identification to transcription training group ($\hat{\beta} = 0.104$, $SE = 0.106$, $z = 0.984$, $p = 0.325$) or the no feedback to transcription training group comparison ($\hat{\beta} = 0.176$, $SE = 0.106$, $z = 1.655$, $p = 0.098$), indicating that the learning effect was equivalent across training groups. The two-way interaction between session and keyword was not reliable ($\hat{\beta} = 0.030$, $SE = 0.057$, $z = 0.524$, $p = 0.600$), indicating that training promoted similar learning across keywords and training groups, in line with results reported for the low-predictability items. All other two-way and three-way interactions were not reliable. Collectively, the learning analysis reveals that all training groups learned to the same degree and this learning effect promoted equal improvements for words at sentence-onset as they did for words at sentence-offset.
Table 1. Mean accuracy (proportion correct), standard deviation, and standard error of the mean for low-predictability and high-predictability items for the learning analysis (n = 108).

<table>
<thead>
<tr>
<th>Test session</th>
<th>Sentence type</th>
<th>Mean accuracy</th>
<th>Standard deviation</th>
<th>Standard error of the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-00</td>
<td>Low-predictability</td>
<td>0.353</td>
<td>0.148</td>
<td>0.010</td>
</tr>
<tr>
<td>Test-01</td>
<td>Low-predictability</td>
<td>0.554</td>
<td>0.163</td>
<td>0.011</td>
</tr>
<tr>
<td>Test-00</td>
<td>High-predictability</td>
<td>0.592</td>
<td>0.165</td>
<td>0.011</td>
</tr>
<tr>
<td>Test-01</td>
<td>High-predictability</td>
<td>0.721</td>
<td>0.150</td>
<td>0.010</td>
</tr>
</tbody>
</table>

**Maintenance.** The maintenance analysis examined performance between Test-01 and Test-02. As reviewed previously, Test-01 was administered immediately after training, while Test-02 was administered one-week following training. For this analysis, we examine performance for the 71 participants who completed both Test-01 and Test-02. First, we examine the learning analysis for low-predictability items. In Figure 6, we depict keyword performance of each training group at Test-01 and Test-02 for low-predictability items. Visual inspection of Figure 6 reveals that performance at Test-01 appears similar to performance at Test-02 for all training groups across keywords – suggesting that learning is maintained equally across training conditions and keyword position.
To examine these patterns statistically, trial-level responses (0 = incorrect, 1 = correct) were submitted to a GLMM with the fixed effects of training condition (treatment coded; transcription as the reference), test session (Test-01 = -1, Test-02 = 1), and keyword (KW1 = -1, KW3 = 1). The model also included random intercepts by subject, random intercepts by sentence, and random intercepts by talker, as well as random slopes by subject for test session and keyword. We used this model structure separately for both low-predictability and high-predictability items and report analyses for each.

For low-predictability items, the model revealed a main effect of test session ($\hat{\beta} = -0.156$, $SE = 0.061$, $z = -2.555$, $p = 0.011$), indicating that performance at Test-01 was higher in accuracy.
relative to performance at Test-02. As can be calculated from Table 2, this difference reflects a 4.4% decrease in mean accuracy one-week later, suggesting a modest decline in performance.

With respect to training task, there was no difference between the talker identification and the transcription group ($\hat{\beta} = -0.212$, $SE = 0.153$, $z = -1.382$, $p = 0.167$), nor was there a reliable difference between the no feedback and transcription group ($\hat{\beta} = -0.026$, $SE = 0.150$, $z = -0.176$, $p = 0.861$). We also observed no effect of keyword on accuracy ($\hat{\beta} = 0.103$, $SE = 0.067$, $z = 1.545$, $p = 0.122$), as would be expected using these low-predictability items. The two-way interaction between test session and training conditions was not reliable between the talker identification to transcription group ($\hat{\beta} = 0.088$, $SE = 0.089$, $z = 0.987$, $p = 0.324$), nor was the no feedback to transcription training group ($\hat{\beta} = 0.042$, $SE = 0.088$, $z = 0.478$, $p = 0.633$), indicating the maintenance of gains did not differ as a function of training task. We observed a two-way interaction between test session and keyword ($\hat{\beta} = -0.127$, $SE = 0.059$, $z = -2.176$, $p = 0.030$). However, post-hoc comparisons revealed this two-way interaction was not driven by a difference in keyword performance at Test-01 ($\hat{\beta} = -0.029$, $SE = 0.106$, $z = -0.277$, $p = 0.993$) or Test-02 ($\hat{\beta} = -0.031$, $SE = 0.105$, $z = -0.298$, $p = 0.991$), nor was this driven by a change in accuracy of keyword 1 ($\hat{\beta} = 0.227$, $SE = 0.100$, $z = 2.274$, $p = 0.104$) or keyword 3 ($\hat{\beta} = 0.225$, $SE = 0.101$, $z = 0.231$, $p = 0.115$) across time points. Thus, it appears this interaction was spurious, and maintenance of gains were equivalent across keywords and sessions. Finally, we observed a three-way interaction for the transcription training group relative to the no feedback group for test session and keyword ($\hat{\beta} = 0.229$, $SE = 0.084$, $z = 2.742$, $p = 0.006$). Post-hoc analyses revealed no group differences in accuracy of keyword 1 at Test-01 ($\hat{\beta} = -0.341$, $SE = 0.205$, $z = -1.666$, $p = 0.884$), or Test-02 ($\hat{\beta} = 0.034$, $SE = 0.202$, $z = 0.166$, $p = 1.000$), nor was there a group accuracy difference of keyword 3 at Test-01 ($\hat{\beta} = 0.477$, $SE = 0.222$, $z = 2.152$, $p = 0.585$),
or Test-02 ($\hat{\beta} = -0.065, SE = 0.231, z = -0.280, p = 1.000$). This suggests that learning was maintained equally across groups and keywords for low-predictability items. All other two-way and three-way interactions were not reliable.

Finally, we examined learning for high-predictability items. In Figure 7, we depict the distribution of mean proportion correct responses for keyword 1 (left panel) and keyword 3 (right panel) at Test-00 and Test-01 for high-predictability items. Visual inspection of Figure 7 reveals that performance at Test-02 was comparable to performance at Test-01 for all training groups. Moreover, accuracy for keyword 3 remains higher than accuracy for keyword 1 at both time points, but a maintenance effect is apparent for both.

*Figure 7.* Boxplots showing the distribution of mean proportion correct responses in the maintenance analysis (Test-01 vs. Test-02) for each training group across high-predictability items. Performance at keyword 1 is shown on the left panels and performance at keyword 3 is shown on the right panels.
For high-predictability items, the model revealed a main effect of test session ($\hat{\beta} = 0.166, SE = 0.078, z = 2.113, p = 0.034$), revealing that performance at Test-01 was higher in accuracy relative to performance at Test-02. As can be calculated from Table 2, this difference reflects a 3.4% decrease in mean accuracy one-week later, suggesting a modest decline in performance.

With respect to training task, there was no difference between the talker identification and the transcription group ($\hat{\beta} = 0.210, SE = 0.224, z = 0.937, p = 0.349$), nor was there a reliable difference between the no feedback and transcription group ($\hat{\beta} = 0.252, SE = 0.220, z = 1.147, p = 0.251$). We observed an effect of keyword on accuracy ($\hat{\beta} = 0.503, SE = 0.072, z = 7.014, p < 0.001$), indicating that accuracy for keyword 3 was higher relative to accuracy for keyword 1, in line with a role of context on learning. The two-way interaction between test session and keyword was not reliable ($\hat{\beta} = 0.070, SE = 0.069, z = 1.011, p = 0.312$), indicating that maintenance was equivalent across keywords. The two-way interaction between test session and training condition was not reliable between the talker identification to transcription training group ($\hat{\beta} = 0.002, SE = 0.115, z = 0.018, p = 0.986$), nor was the no feedback to transcription training group interaction ($\hat{\beta} = -0.107, SE = 0.114, z = -0.943, p = 0.346$), indicating that the learning was maintained equivalently across training groups. All other two-way and three-way interactions were not reliable. Thus, it appears that training promoted similar long-term retention across keywords and training groups, in line with results reported for the low-predictability items. Collectively, the maintenance analysis converges with the learning analysis: overall, listeners in all training groups showed relatively stable performance immediately after training as
compared to performance one-week later and this effect was sustained equally for words at the sentence-onset as compared to words at the sentence-offset.

Table 2. Mean accuracy (proportion correct), standard deviation, and standard error of the mean for low-predictability and high-predictability items for the maintenance analysis (n = 71).

<table>
<thead>
<tr>
<th>Test session</th>
<th>Sentence type</th>
<th>Mean accuracy</th>
<th>Standard deviation</th>
<th>Standard error of the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-01</td>
<td>Low-predictability</td>
<td>0.568</td>
<td>0.154</td>
<td>0.013</td>
</tr>
<tr>
<td>Test-02</td>
<td>Low-predictability</td>
<td>0.523</td>
<td>0.138</td>
<td>0.013</td>
</tr>
<tr>
<td>Test-01</td>
<td>High-predictability</td>
<td>0.735</td>
<td>0.138</td>
<td>0.012</td>
</tr>
<tr>
<td>Test-02</td>
<td>High-predictability</td>
<td>0.698</td>
<td>0.158</td>
<td>0.013</td>
</tr>
</tbody>
</table>

4. Discussion

In the current study, we examined the individual contributions of lexical influence on perceptual learning and maintenance of noise-vocoded sentences. All participants heard the same set of anomalous sentences during training and either completed a transcription task with lexical feedback, a transcription task without lexical feedback, or a talker identification task. At test, all listeners completed a transcription task without feedback. Learning was assessed prior to training (Test-00), immediately after training (Test-01), and one-week post-training (Test-02). The goal of the current design was to (1) parse out the influence of lexical training tasks and lexical feedback on perceptual learning, (2) examine perceptual learning in the immediate and

---

6 See addendum for an additional set of analyses using proportion correct keywords on trial-level data.
maintenance of gains one-week later, and (3) describe performance at test using an the understudied metric of word-initial learning. Collectively, we observed that all training task conditions, regardless of lexical feedback or task, demonstrated equivalent perceptual learning immediately after training and those gains were maintained to the same degree one-week later. For low-predictability test items, accuracy of sentence-initial test items and sentence-final test items did not differ at any test sessions – the magnitude of improvement was equivalent immediately after training and one-week later across all test conditions. Critically, for high-predictability test items, accuracy for sentence-final test items was higher before training relative to sentence-initial test items, but after training equivalent degree of learning and maintenance of gains was observed for words at both the beginning and end of the sentence. In the below sections we discuss each finding in detail.

4.1. Lexical influences on learning outcomes

We observed robust perceptual learning for participants in the transcription training group. Our finding replicates that of various other studies demonstrating a short training period can be leveraged to promote robust changes in perception for noise-vocoded speech signals (e.g., Davis et al., 2005; Hervais-Adelman et al., 2008; Loebach et al., 2008, Huyck et al., 2017). Because our test stimuli differed from the sentences presented at training, the gains observed during test reflected generalization to novel sentences, which replicates previous findings demonstrating generalization to noise-vocoded sentences (e.g., Loebach et al., 2008). To parse out the influence of lexically based training tasks and lexical feedback on perceptual learning, we compared degree of learning between the transcription with feedback group to the transcription no feedback training group. We observed equivalent changes in performance from Test-00 to Test-01 for both the no feedback and the transcription group. As the no feedback group
completed an identical task to the transcription group and only differed on the dimension of lexical feedback, our finding suggests that robust learning can occur irrespective of whether lexical feedback is provided or not. Our finding extends results reported by Davis et al. (2005) as our design allows for a direct conclusion about training task and feedback – in the absence of supporting sentence context and lexical feedback, learning can still be achieved using a lexically focused task.

We also isolated the role of task-based attention on perceptual learning. As reviewed, previous research has demonstrated that perceptual learning of noise-vocoded sentences can occur using the non-linguistic training task of talker identification (Loebach et al., 2008). It has been hypothesized that this observation of a linguistic benefit using a non-linguistic task reflects the attentional requirements of the task that allow for the listener to engage in a deep level of analysis with the sentences. However, a limitation to the finding reported by Loebach et al. (2008) was that training occurred using high-predictability sentences, leaving open the question that sentence context, as opposed to training task, could have promoted perceptual learning. Recall that in the current study, the training stimuli consisted of anomalous sentences (composed of real words, but semantically meaningless) thus removing a possible role of sentence context on learning. Strikingly, we observed that perceptual learning for the talker identification training group was equivalent to the perceptual learning observed for the transcription training group, suggesting that the learning effect was truly driven by the attentional requirements of the training task. Our finding replicates results reported by Loebach et al. (2008) demonstrating a means for non-linguistic training to be effective in promoting comprehension benefits for noise-vocoded sentences. That we observed no difference between the transcription and talker identification training group is in line with predictions made under a depth of processing framework (e.g.,
Namely, it appears the talker identification manipulation was sufficiently challenging to promote deep engagement towards the acoustic details of the signal, resulting in a linguistic benefit.

4.2. Maintenance of training gains

Following training, we found no difference in perceptual learning between the various training groups. Thus, by this measure, listeners across training groups gain the same benefit despite differences in how the task was implemented, supporting predictions made under a depth of processing framework (e.g., Craik & Lockhart, 1972) as opposed to a task-specific framework (e.g., Morris et al., 1977; Roediger et al., 1989). However, it is unknown how benefits obtained across training groups persist over time and neither theoretical framework makes explicit predictions about how learning is maintained. As very few training studies implement the inclusion of a test session displaced in time from the initial training period, we sought to characterize the maintenance of learning by testing participants one-week following training. Relative to Test-01 (immediately after training), participants showed a modest decline in performance at Test-02 (one-week post-training) and this slight decrease in performance was equivalent across training groups – there was no difference in how learning was maintained between groups. This finding suggests that task manipulations that promote improvements in the short-term generalize to improvements observed at least one-week later. As no group difference were observed, we can conclude that task-specificity was not necessary for gains to be maintained. Instead, it appears that the initial performance following training predicted performance at a later time interval. The inclusion of the one-week interval extends findings by Loebach et al. (2008) who found no difference between their transcription and talker identification training groups in the immediate. Our results at the one-week follow-up are
supported by the depth of processing framework in which comprehension benefits can be observed using non-linguistic tasks that sufficiently promote engagement towards the acoustic details of the speech signal.

4.3. Role of training on learning for words across the sentence

We sought to characterize performance at test using an understudied performance metric—accuracy for sentence-initial words. Recall that previous research in the speech learning domain calculated sentence accuracy by examining performance for words in the word-final position only (e.g., Loebach et al., 2008). However, using this metric alone to determine sentence accuracy introduces context enhancement effects that are unique to word-final high-predictability items (e.g., Elliot, 1995) and recency effects (e.g., Postman & Philips, 1965). In our design, we tested participants using low- and high-predictability sentences, allowing for context to be uninformative for low-predictability items and informative high-predictability items. As such, we expected differences in word-final to word-initial accuracy to differ across sentence type. Namely, for low-predictability sentences in which context was uninformative, we expected changes in keyword accuracy to be comparable as neither word position was enhanced from context. For high-predictability sentences, in which sentence context was informative, we expected accuracy for word-final items to be higher relative to word-initial items due to context enhancement at baseline. However, after training, we expected performance for word-initial items to improve, and approach performance of word-final items. This prediction was partially observed; at both Test-00 and Test-01 keyword accuracy did not differ across positions for low-predictability items, but did differ for high-predictability items in which accuracy was significantly higher for word-final compared to word-initial items. Critically, we observed no test session by keyword interaction for either low- or high-predictability items, indicating that the
learning effect was equivalent regardless of context. At the one-week post-test, the same pattern was observed, suggesting that both the learning and maintenance gains for sentence-initial and sentence-final items was comparable. Thus, performance for word-initial items remained poorer than word-final items for high-predictability sentences even after training, but training significantly improved accuracy in the word-initial position. That we observed changes in the word-initial item suggests that training promoted a global change in sentence understanding, that was enhanced with sentence context for word-final high-predictability items.

4.4. Clinical implications and future directions

Use of training-based rehabilitation paradigms offers a potential opportunity to enhance speech perception performance among CI users. Our findings converge with other work demonstrating comprehension benefits following a non-linguistic training task (Loebach et al., 2008) and extend those findings to demonstrate that gains are relatively stable even one-week following training. Moreover, we found that sentence-level feedback is not necessary for learning to occur – we observed equivalent learning for those trained on a transcription task without feedback as those trained using a transcription task with feedback. A challenge in the rehabilitation literature with respect to designing training paradigms is that there are currently no agreed upon set of recommendations for how training should be structured. Our work suggests that a variety of training could be used to support speech comprehension, allowing the opportunity to customize training to the patient needs. Namely, for pediatric patients or for new adult users, a cognitively easier task like talker identification may be more appropriate and may offer the same comprehension benefits as a transcription task. Future research is aimed at examining the role of non-linguistic training tasks on comprehension benefits in the patient population.
Finally, a challenge in examining gains obtained from training in both the lab and the clinic is the outcome metrics used to assess learning. Namely, in widely utilized clinical metrics like the Speech Perception In Noise (SPIN) test, accuracy is generally only measured as accuracy of the final word in a sentence (Kalikow et al., 1977). This is motivated by the types of stimuli used in the SPIN; for high-predictability sentences, final keyword accuracy is expected to be higher as compared to low-predictability sentences if the individual capitalizes on contextual cues. However, we argue that identifying changes at the sentence onset (initial keyword) may be a more sensitive metric to observe changes in processing that occur as a consequence of training. As the sentence-initial position in high-predictability items offers no contextual cues to enhance accuracy, training-based changes observed in this keyword position cannot be due to context enhancement or recency effects. The current study observed robust changes at both sentence-onset and sentence offset-items suggesting that training promotes a comprehensive improvement across the sentence structure. As previous work has suggested that some CI users do not appear to benefit from sentence context to the same degree as NH listeners and instead appear to process sentences as strings of unrelated words (e.g., Conway et al., 2008), comparing performance on both sentence-initial and sentence-final items in a high-predictability context might offer a way to determine how training improves use of sentence context and coarticulatory cues in the patient population.
Supplemental Material

Chapter II.

Full list of sentences used for training and test stimuli. Keywords for each sentence are underlined.

1. The box was thrown beside the parked truck.
2. The boy was there when the sun rose.
3. A pot of tea helps to pass the evening.
4. Read verse out loud for pleasure.
5. Take the winding path to reach the lake.
6. Wipe the grease off his dirty face.
7. The young girl gave no clear response.
8. The ship was torn apart on the sharp reef.
9. The lazy cow lay in the cool grass.
10. The rope will bind the seven books at once.
11. The frosty air passed through the coat.
12. The show was a flop from the very start.
13. Place a rosebush near the porch steps.
14. This is a grand season for hikes on the road.
15. The dune rose from the edge of the water.
16. The two men met while playing on the sand.
17. The walled town was seized without a fight.
18. The horn of the car woke the sleeping cop.
19. Bail the boat to stop it from sinking.
20. The bill was paid every third week.
21. Add the sum to the product of these three.
22. The pennant waved when the wind blew.
23. We find joy in the simplest things.
24. Type out three lists of orders.
25. The boss ran the show with a watchful eye.
26. It caught its hind paw in a rusty trap.
27. Two plus seven is less than ten.
28. Bring your problems to the wise chief.
29. Pure bred poodles have curls.
30. The tree top waved in a graceful way.
31. Mud was splattered on the front of his white shirt.
32. The empty flask stood on the tin tray.
33. A speedy man can beat this track mark.
34. The sofa cushion is red and of light weight.
35. Drop the two when you add the figures.
36. An abrupt start does not win the prize.
37. Wood is the best for making toys and blocks.
38. Steam hissed from the broken valve.
39. The child almost hurt the small dog.
40. The sky that morning was clear and bright blue.
41. Torn scraps littered the stone floor.
42. Sunday is the best part of the week.
43. Add the column and put the sum here.
44. We admire and love a good cook.
45. She has a smart way of wearing clothes.
46. The paper box is full of thumb tacks.
47. The cement had dried when he moved it.
48. The loss of the second ship was hard to take.
49. The fly made its way along the wall.
50. The large house had hot water taps.
51. It is hard to erase blue or red ink.
52. A pencil with black lead writes best.
53. Jazz and swing fans like fast music.
54. Rake the rubbish up and then burn it.
55. They are pushed back each time they attack.
56. Some ads serve to cheat buyers.
57. A waxed floor makes us lose balance.
58. The play began as soon as we sat down.
59. Add salt before you fry the egg.
60. These pills do less good than the others.
61. The rude laugh filled the empty room.
62. High seats are best for football fans.
63. A dash of pepper spoils beef stew.
64. The junk yard had a moldy smell.
65. The gold ring fits only a pierced ear.
66. Watch the log float in the wide river.
67. The heap of fallen leaves was set on fire.
68. Slide the box into that empty space.
69. The beam dropped down on the workman's head.
70. Pink clouds floated with the breeze.
71. The fight will end in just six minutes.
72. The purple tie was ten years old.
73. The plush chair leaned against the wall.
74. Nine rows of soldiers stood in line.
75. The beach is dry and shallow at low tide.
76. Pages bound in cloth make a book.
77. Try to trace the fine lines of the painting.
78. The zones merge in the central part of town.
79. Code is used when secrets are sent.
80. Pour the stew from the pot into the plate.
81. It takes a good trap to capture a bear.
82. Feed the white mouse some flower seeds.
83. Plead to the council to free the poor thief.
84. He crawled with care along the ledge.
85. Mark the spot with a sign painted red.
86. He wrote down a long list of items.
87. Roads are paved with sticky tar.
88. The sun came up to light the eastern sky.
89. The desk was firm on the shaky floor.
90. It takes heat to bring out the odor.
91. Raise the sail and steer the ship northward.
92. A cone costs five cents on Mondays.
93. Jerk the dart from the cork target.
94. The sense of smell is better than that of touch.
95. No hardship seemed to keep him sad.
96. The marsh will freeze when cold enough.
97. They slice the sausage thin with a knife.
98. The bloom of the rose lasts a few days.
99. Drop the ashes on the worn old rug.
100. Throw out the used paper cup and plate.
101. The couch cover and hall drapes were blue.
102. The clothes dried on a thin wooden rack.
103. The music played on while they talked.
104. The kite flew wildly in the high wind.
105. The tin box held priceless stones.
106. He offered proof in the form of a large chart.
107. They told wild tales to frighten him.
108. A man in a blue sweater sat at the desk.
109. The dusty bench stood by the stone wall.
110. Smile when you say nasty words.
111. The room was crowded with a wild mob.
112. The beetle droned in the hot June sun.
113. The black trunk fell from the landing.
114. His wide grin earned many friends.
115. Those last words were a strong statement.
116. He wrote his name boldly at the top of the sheet.
117. If you mumble your speech will be lost.
118. A brown leather bag hung from its strap.
119. A toad and a frog are hard to tell apart.
120. A break in the dam almost caused a flood.
121. The child crawled into the dense grass.
122. A round hole was drilled through the thin board.
123. Drive the screw straight into the wood.
124. Keep the hatch tight and the watch constant.
125. Slide the tray across the glass top.
126. Get the trust fund to the bank early.
127. Choose between the high road and the low.
128. A six comes up more often than a ten.
129. The early phase of life moves fast.
130. She flaps her cape as she parades the street.
131. Crouch before you jump or miss the mark.
132. Pack the kits and don't forget the salt.
133. They sang the same tunes at each party.
134. Pile the coal high in the shed corner.
135. A gold vase is both rare and costly.
136. Hang tinsel from both branches.
137. Pick a card and slip it under the pack.
138. A round mat will cover the dull spot.
139. The mail comes in three batches per day.
140. You cannot brew tea in a cold pot.
141. Put the chart on the mantel and tack it down.
142. We don't like to admit our small faults.
143. Take the match and strike it against your shoe.
144. The baby puts his right foot in his mouth.
145. The streets are narrow and full of sharp turns.
146. The big red apple fell to the ground.
147. The curtain rose and the show was on.
148. He sent the boy on a short errand.
149. Small children came to see him.
150. The grass and bushes were wet with dew.
Chapter III.

Full list of anomalous sentences used during training. Keywords for each sentence are underlined.

1. Ruth’s problems are made from bark.
2. Face the cop through a notch.
3. Heroes called lots of seeds.
4. Miss Brown charged her wood of sheep.
5. The accident washes a short beam.
6. The burglar was parked by an ox.
7. Discuss the sailboat on the bend.
8. Miss Smith was worn by Adam’s blade.
9. The low woman was gladly in the calf.
10. The turn twisted the cards.
11. Toss the boy into shipwrecked chunks.
12. Jane ate in the glass for a clerk.
13. The landlords stood for a clue.
14. Nancy was poured by the cops.
15. Consider the local floor about crumbs.
16. Mr. White hit the debt.
17. My man wiped a shirt for a pet.
18. The first man heard a feast.
19. The sink served with an easy flame.
20. For a bloodhound he had spoiled pie.
21. The round lion held a flood.
22. The coat is talking about six frogs.
23. They milked a frightened entry of gin.
24. It was beaten around with glue.
25. Betty buttered a sharp gown.
26. The stories covered the glass hen.
27. The seats were called about the host.
28. The folding hands drowned the pill.
29. The old cloud broke his lungs.
30. Water the worker between the pole.
Full list of low-predictability sentences used at test. Keywords for each sentence are underlined.

1. Betty has considered the bark.
2. Mr. Smith knew about the bay.
3. Tom has been discussing the beads.
4. Mr. Smith spoke about the aid.
5. She’s discussing the beam.
6. I’m talking about the bench.
7. Bob considered the tent.
8. Mary hasn’t discussed the blade.
9. She hopes Jane called about the calf.
10. Mr. Black has discussed the cards.
11. I did not know about the chunks.
12. Bob was considering the clerk.
13. The man spoke about the clue.
14. Ruth hopes Bill called about the cop.
15. She might discuss the crumbs.
16. We will consider the debt.
17. Peter could consider the dove.
18. She’s glad Bill called about the beak.
19. Mary can’t consider the tide.
20. Paul should have discussed the flock.
21. The class should consider the flood.
22. The woman talked about the frogs.
23. The girl talked about the gin.
24. Tom has not considered the glue.
25. The girl should not discuss the gown.
26. Bill didn’t discuss the hen.
27. We’re speaking about the toll.
28. Harry had thought about the logs.
29. The old man talked about the lungs.
30. Peter has considered the mat.
31. The woman considered the notch.
32. He has a problem with the oath.
33. The man should discuss the ox.
34. They heard I called about the pet.
35. Miss Smith knows about the tub.
36. Tom had spoken about the pill.
37. Bob could consider the pole.
38. She might consider the pool.
39. The boy would discuss the scab.
40. You’ve considered the seeds.
41. They’ve considered the sheep.
42. Tom won’t consider the silk.
43. Nancy didn’t discuss the skirt.
44. Mr. Brown thinks about the vault.
45. Mary had considered the spray.
1. The **doctor** prescribed the **drug**.
2. A **spoiled** child is a **brat**.
3. Ann **works** in the bank as a **clerk**.
4. **Banks** keep their money in a **vault**.
5. The **cabin** was made of **logs**.
6. Break the dry bread into **crumbs**.
7. Cut the meat into small **chunks**.
8. **Eve** was made from Adam’s **rib**.
9. **Follow** this road around the **bend**.
10. For **dessert** he had apple **pie**.
11. Get the **bread** and cut me a **slice**.
12. **Greet** the heroes with loud **cheers**.
13. He **rode** off in a cloud of **dust**.
14. He was scared out of his **wits**.
15. Her **entry** should win first **prize**.
16. Her **hair** was tied with a blue **bow**.
17. He’s **employed** by a large **firm**.
18. I **ate** a piece of chocolate **fudge**.
19. Instead of a fence, plant a **hedge**.
20. It was **stuck** together with **glue**.
21. I’ve got a **cold** and a sore **throat**.
22. **Keep** your broken arm in a **sling**.
23. **Kill** the bugs with this **spray**.
24. **Maple** syrup is made from **sap**.
25. The **chicken** pecked corn with its **beak**.
26. The **cow** gave birth to a **calf**.
27. Old **metal** cans were made with **tin**.
28. Our **seats** were in the second **row**.
29. Paul **hit** the water with a **splash**.
30. Paul **took** a bath in the **tub**.
31. Paul was arrested by the **cops**.
32. The **dealer** shuffled the **cards**.
33. **Playing** checkers can be **fun**.
34. Please **wipe** your feet on the **mat**.
35. **Raise** the flag up the **pole**.
36. Ruth had a **necklace** of glass **beads**.
37. Ruth **poured** herself a cup of **tea**.
38. She **cooked** him a hearty **meal**.
39. She **shortened** the hem of her **skirt**.
40. **Spread** some butter on your **bread**.
41. That **job** was an easy **task**.
42. The **bird** of peace is the **dove**.
43. The **detectives** searched for a **clue**.
44. The **boat** sailed across the **bay**.
45. The **bride** wore a white **gown**.
References


Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of


Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity...


Fu, Q. J., & Galvin, J. J. (2008). Maximizing cochlear implant patients’ performance with


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


Addendum

Chapter Two

In the experiment presented in chapter two, we chose to analyze our training and test data using an accuracy metric criterion based off of number of keywords correct in a sentence. Recall, that sentences in experiment two contained five or six keywords per sentence. We coded accuracy as a binomial variable, with an accurate sentence coded as two or more keywords correct in a sentence. While our accuracy metric is in line with previous research in which sentence accuracy was coded based only on the accuracy of the final keyword (e.g., Loebach et al., 2008), an alternative option for calculating accuracy is to use proportion correct keywords for each sentence. The use of this alternative accuracy procedure could overcome potential accuracy overestimations that occurs as a consequence of requiring only two of the keywords correct in a sentence, and thus allowing three (in the case of sentences with five keywords) or four (in the case of sentence with six keywords) keywords to be inaccurate. Below, we present the results for the data presented in chapter two, with accuracy on trial-level data coded as mean proportion correct.

Training

Figure 1a shows the accuracy distributions during training for the two training groups, which were determined by calculating mean proportion correct for each of the keywords identified on a given trial. In line with the previous analysis presented, visual inspection of this figure suggests that accuracy during training was equivalent between the two training groups.
To examine this pattern statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM); the Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using lmerTest (Kuznetsova et al., 2017). The fixed effect of the model was group (sum coded; morning = -1, evening = 1). The random effects structure consisted of random intercepts by subject and random intercepts by sentence. The main effect of group was not significant ($\hat{\beta} = -0.017$, $SE = 0.014$, $z = -1.246$, $p = 0.217$), providing evidence that accuracy between the two groups did not differ during the training phase.

**Test**

Each test sentence was scored for accuracy following the procedure outlined for the training phase. Figure 2a shows the accuracy distributions at each test for the two training groups separately for trained and novel items, which was determined by calculating mean accuracy across the appropriate test items for each participant. Consider first performance for the trained items. Visual inspection suggests that both groups improved from pretest to the first posttest, indicating that learning occurred following training. In addition, inspection of the group difference between posttest1 and posttest2 (which took place approximately 12 hours after
posttest1) suggests the presence of a sleep-based consolidation effect. Specifically, the morning group shows a decline in performance between posttest1 and posttest2, whereas the evening group does not. Moreover, visual inspection suggests that learning is maintained over time, in that performance for posttest3 is equivalent or improved relative to performance at posttest2 for both groups.

Now consider performance for the novel items. Overall, the magnitude of learning appears less robust relative to the trained items. Similar to the trained items, learning is observed such that performance for posttest1 is improved relative to the pretest. A sleep-based consolidation effect is also suggested, in that performance at posttest2 remains stable relative to posttest1 for the evening group, but a modest decline is observed for the morning group. Performance between posttest3 and posttest2 again suggests maintenance of learning, such that the evening group appears to show stable performance, while the morning group appears to improve for trained items at a level similar to the evening trained group.

**Figure 2a.** Boxplots showing the distribution of mean proportion correct responses during the test phase for the morning and night training groups.
To examine these patterns statistically, three sets of analyses were performed: the
learning analysis examined performance between pretest and posttest1, the consolidation
analysis examined performance between posttest1 and posttest2, and the maintenance analysis
examined performance between posttest2 and posttest3. Each is addressed in turn.

**Learning.** The learning analysis examined performance between pretest and posttest1. Recall that pretest took place immediately before training and posttest1 took place immediately after training. As can be seen in Figure 3a, both groups improved from pretest to posttest1, though it appears that they did so to a larger degree for trained compared to novel items. To examine these patterns statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM) with the fixed effects of group (morning = -1, evening = 1), test (pretest = -1, posttest1 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject, random intercepts by sentence, and random slopes by subject for test and type.

**Figure 3a.** Boxplots showing the distribution of mean proportion correct responses in the learning analysis (pretest vs. posttest1) for the morning and night training groups separately for trained and novel items. Lines show performance for the individual participants in each group.
The model revealed a main effect of test ($\hat{\beta} = 0.192$, $SE = 0.011$, $z = 16.861$, $p < 0.001$), with higher accuracy at posttest1 compared to pretest, and a main effect of type ($\hat{\beta} = -0.137$, $SE = 0.017$, $z = -7.830$, $p < 0.001$), reflecting higher accuracy for trained compared to novel items. There was also an interaction between test and type ($\hat{\beta} = -0.126$, $SE = 0.010$, $z = -12.398$, $p < 0.001$). Post-hoc paired comparisons (here and throughout) were performed to explicate the nature of the interaction. The paired comparisons showed no difference between trained and novel items at pretest ($\hat{\beta} = 0.022$, $SE = 0.040$, $z = 0.540$, $p = 0.949$), but significantly higher accuracy for trained compared to novel items at posttest1 ($\hat{\beta} = 0.524$, $SE = 0.040$, $z = 13.000$, $p < 0.001$). Moreover, accuracy improved from pretest to posttest1 for both trained ($\hat{\beta} = -0.634$, $SE = 0.013$, $z = -50.145$, $p < 0.001$) and novel ($\hat{\beta} = -0.132$, $SE = 0.041$, $z = -3.207$, $p = 0.007$) items. Thus, the interaction in the omnibus model reflects increased improvement between pretest and posttest1 for trained compared to novel items, and not a failure to improve for novel items.

The omnibus model showed no main effect of group ($\hat{\beta} = -0.003$, $SE = 0.008$, $z = -0.305$, $p = 0.762$), but there was an interaction between group and test ($\hat{\beta} = -0.019$, $SE = 0.006$, $z = -2.071$, $p = 0.043$). Paired comparisons showed an improvement from pretest to posttest1 for both the morning group ($\hat{\beta} = -0.407$, $SE = 0.026$, $z = -15.973$, $p < 0.001$) and the evening group ($\hat{\beta} = -0.359$, $SE = 0.026$, $z = -14.098$, $p < 0.001$). The beta estimate is larger in the former compared to the latter, which may suggest that the interaction between group and test in the omnibus model reflects increased improvement for the morning compared to the evening group. However, no reliable difference was observed between the two groups at either pretest ($\hat{\beta} = -0.019$, $SE = 0.014$, $z = -1.347$, $p = 0.533$) or posttest1 ($\hat{\beta} = 0.029$, $SE = 0.025$, $z = 1.160$, $p = 0.652$).

The omnibus model showed a marginal interaction between group and type ($\hat{\beta} = 0.007$, $SE = 0.004$, $z = 1.846$, $p = 0.070$). Paired comparisons showed that accuracy for trained items
was higher than accuracy for novel items for both the morning group ($\hat{\beta} = 0.288, SE = 0.036, z = 8.045, p < 0.001$) and the evening group ($\hat{\beta} = 0.258, SE = 0.036, z = 7.215, p < 0.001$). Paired comparisons showed no difference between the two groups for either trained ($\hat{\beta} = 0.020, SE = 0.020, z = 1.001, p = 0.749$) or novel items ($\hat{\beta} = -0.010, SE = 0.017, z = -0.579, p = 0.938$). The three-way interaction between group, test, and type was significant ($\hat{\beta} = 0.008, SE = 0.003, z = 2.904, p = 0.004$). To explicate the interaction, separate LMMs were constructed for trained and novel items with the same fixed and random effects structure of the omnibus model except for the removal of type as a fixed effect. The group by test interaction was significant for trained items ($\hat{\beta} = -0.019, SE = 0.009, z = -2.227, p = 0.0296$), but not for novel items ($\hat{\beta} = -0.004, SE = 0.005, z = 0.929, p = 0.357$). Collectively, the learning analysis demonstrates that the morning and night groups both showed improved perception of noise-vocoded signals following training. Training generalized to novel items, though the learning effect was stronger for trained compared to novel items.

**Consolidation.** The consolidation analysis examined performance between posttest1 and posttest2. Recall that posttest2 took place approximately 12 hours after posttest1, a period of time that included sleep for the evening group but not for the morning group. As can be seen in Figure 4a, the evening group appears to show increased accuracy at posttest2 compared to the morning group for both trained and novel items.
Figure 4a. Boxplots showing the distribution of mean proportion correct responses in the consolidation analysis (posttest1 vs. posttest2) for the morning and night training groups separately for trained and novel items. Lines show performance for the individual participants in each group.

To examine these patterns statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM) with the fixed effects of group (morning = -1, evening = 1), test (posttest1 = -1, posttest2 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject, random intercepts by sentence, and random slopes by subject for test and type. The results of the model showed a main effect of test ($\hat{\beta} = -0.041, SE = 0.020, z = -3.419, p < 0.001$), indicating higher accuracy at posttest1 compared to posttest2, and a main effect of type ($\hat{\beta} = -0.2486, SE = 0.0215, z = -11.541, p < 0.001$), indicating higher accuracy for trained compared to novel items. There was no main effect of group ($\hat{\beta} = 0.014, SE = 0.012, z = 1.123, p = 0.266$); however, a robust group by test interaction was observed ($\hat{\beta} = 0.028, SE = 0.003, z = 9.332, p < 0.001$). Post-hoc paired comparisons showed that accuracy for the morning group declined between posttest1 and posttest2 ($\hat{\beta} = 0.138, SE = 0.247, z = 5.598, p < 0.001$), whereas there was no difference in accuracy between
the two test sessions for the evening group ($\hat{\beta} = 0.026$, $SE = 0.025$, $z = 1.033$, $p = 0.730$).

Moreover, accuracy was higher for the night compared to the morning group at posttest2 ($\hat{\beta} = -0.084$, $SE = 0.025$, $z = -3.299$, $p = 0.005$), but no such difference was observed between the two groups at posttest1 ($\hat{\beta} = 0.029$, $SE = 0.025$, $z = 1.162$, $p = 0.651$).

In the omnibus model, there was no interaction between group and type ($\hat{\beta} = 0.003$, $SE = 0.008$, $z = -0.325$, $p = 0.746$) or between test and type ($\hat{\beta} = 0.013$, $SE = 0.012$, $z = 1.121$, $p = 0.265$). However, the three-way interaction between group, test, and type was reliable ($\hat{\beta} = -0.018$, $SE = 0.003$, $z = -5.808$, $p = 0.001$). To explicate the interaction, separate LMMs were constructed for trained and novel items with the same fixed and random effects structure of the omnibus model except for the removal of type as a fixed effect. The group by test interaction was significant for both trained ($\hat{\beta} = 0.046$, $SE = 0.005$, $z = 8.764$, $p < 0.001$) and novel items ($\hat{\beta} = 0.011$, $SE = 0.004$, $z = 2.909$, $p = 0.004$). The beta estimates for the interaction terms suggest a weaker consolidation effect for novel compared to trained items. Collectively, the results of the consolidation analysis are consistent with a facilitative effect of sleep-based consolidation on adaptation to noise-vocoded speech. Compared to listeners who were trained in the morning, listeners who were trained prior to sleep showed improved comprehension of noise-vocoded signals approximately 12 hours after training, which was the consequence of learning retention in the night group and a relative loss of learning in the morning group.

**Maintenance.** The maintenance analysis examined performance between posttest2 and posttest3. Recall that posttest3 took place one-week after posttest2. As can be seen in Figure 5a, the evening group appears to show stable performance at posttest3, while the morning group shows an improvement for trained and novel items.
To examine these patterns statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM) with the fixed effects of group (morning = -1, evening = 1), test (posttest2 = -1, posttest3 = 1), and type (trained = -1, novel = 1). The model also included random intercepts by subject, random intercepts by sentence, and random slopes by subject for test and type. The results of the model showed a main effect of group ($\hat{\beta} = 0.025$, $SE = 0.013$, $z = 2.001$, $p = 0.050$), indicating higher accuracy for the night compared to the morning group, and a main effect of type ($\hat{\beta} = -0.240$, $SE = 0.028$, $z = -10.524$, $p < 0.001$), indicating higher accuracy for trained compared to novel items. There was no main effect of test ($\hat{\beta} = 0.017$, $SE = 0.013$, $z = 1.320$, $p = 0.190$), no interaction between group and type ($\hat{\beta} = -0.010$, $SE = 0.008$, $z = -1.192$, $p = 0.238$), and no interaction between test and type ($\hat{\beta} = -0.004$, $SE = 0.026$, $z = -0.340$, $p = 0.735$).
The omnibus model did reveal a significant interaction between group and test ($\hat{\beta} = -0.017, SE = 0.003, z = -5.484, p = 0.001$), which was further mediated by type as indicated by a significant three-way interaction between group, test, and type ($\hat{\beta} = 0.010, SE = 0.003, z = 3.268, p < 0.001$). To explicate the interaction, separate LMMs were constructed for trained and novel items following the same fixed and random effects structure as the omnibus model except for removing the fixed effect of type. For the trained items, a significant group by test interaction was observed ($\hat{\beta} = -0.027, SE = 0.005, z = -5.235, p < 0.001$). Post-hoc paired comparisons showed no difference for trained items between the two groups at posttest3 ($\hat{\beta} = -0.017, SE = 0.040, z = -0.424, p = 0.974$), though the night grouped showed higher accuracy than the morning group at posttest2 ($\hat{\beta} = -0.124, SE = 0.040, z = -3.074, p = 0.011$). These between-subjects comparisons reflect improvement on trained items between posttest2 and posttest3 for listeners in the morning group ($\hat{\beta} = -0.095, SE = 0.014, z = -6.591, p < 0.001$), and equivalent performance on trained items between posttest2 and posttest3 for listeners in the night group ($\hat{\beta} = -0.012, SE = 0.014, z = 0.813, p = 0.848$). For the novel items, the effect of training group did not meet the threshold for statistical significance ($\hat{\beta} = 0.015, SE = 0.009, z = 1.681, p = 0.098$), suggesting similar performance between the morning and evening group for novel items. A significant group by test interaction was observed ($\hat{\beta} = -0.007, SE = 0.003, z = -2.041, p = 0.041$). Post-hoc paired comparisons showed no difference for trained items between the two groups at posttest3 ($\hat{\beta} = -0.016, SE = 0.019, z = -0.847, p = 0.832$), and the night grouped showed no difference in accuracy than the morning group at posttest2 ($\hat{\beta} = -0.044, SE = 0.019, z = -2.332, p = 0.091$). This reflects equivalent group performance on novel items between posttest2 and posttest3 for listener in the morning ($\hat{\beta} = -0.038, SE = 0.053, z = -0.726, p = 0.887$) and evening group ($\hat{\beta} = -0.011, SE = 0.053, z = -0.210, p = 0.997$). Collectively, the maintenance
analysis provides evidence that both groups of listeners maintained learning one-week after training, and that the morning group did not differ from the evening group for performance on either trained or novel items.

**Discussion**

Using a modified analysis based on proportion keyword correct (as opposed to a two or more keyword correct standard), our current analysis converges with the findings presented in chapter two. Namely, we observed that participants trained in the morning and evening group demonstrated robust and equivalent perceptual learning immediately after training for both trained and novel items. The magnitude of learning was significantly higher for trained relative to novel items. Upon returning approximately 12-hours later for posttest2, we observed a robust sleep consolidation effect that manifested as stable performance for listeners trained in the evening and a decline in performance for listeners trained in the morning. The decline in performance for the morning group and the stable performance for the evening group at posttest2 was observed for both trained and novel items. Critically, at posttest3, the time of day at which training initially occurred did not significantly affect performance one-week later. Namely, participants in the morning and evening groups did not differ in performance one-week later for trained items. With respect to the maintenance of novel items, we found that the effect of group did not meet the threshold of statistical significance, as it did in the results presented in chapter two, however the results patterned in the same direction as those reported in chapter two.

**Chapter Three**

In the experiment presented in chapter three, we chose to analyze our test data using a binomial accuracy metric and included word-initial and word-final keyword position as a predictor in the GLMM. In this follow-up analysis, we sought to examine the group differences
in sentence accuracy using proportion correct keyword. Proportion correct keyword was based off of trial-level proportion correct across three keywords (word-initial, word-medial, and word-final items). To isolate the effects of training task on proportion correct keyword for learning and maintenance, we removed the predictors of keyword and predictability from our model.

Test

**Learning.** The learning analysis examined performance between Test-00 and Test-01 (n = 108). As outlined previously, Test-00 occurred before training and Test-01 occurred immediately after training. In Figure 6a, we depict the distribution of mean proportion correct responses at Test-00 and Test-01. Visual inspection of Figure 6a reveals that performance at Test-01 was higher than performance at Test-00 for all training groups. Performance appears equivalent across training groups at each time point.

To examine these patterns statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM) with the fixed effect of training condition (treatment coded; transcription as the reference) and test session (Test-00 = -1, Test-01 = 1). The model also included random intercepts by subject, sentence, and talker, as well as random slopes by subject for test session.
The model revealed a main effect of test session ($\hat{\beta} = 0.075$, $SE = 0.008$, $z = 9.114$, $p < 0.001$), indicating that performance at test-01 was higher in accuracy relative to performance at test-00. With respect to training task, there was no difference between the talker identification and the transcription group ($\hat{\beta} = -0.024$, $SE = 0.023$, $z = -1.031$, $p = 0.305$), nor was there a reliable difference between the no feedback and transcription group ($\hat{\beta} = 0.016$, $SE = 0.023$, $z = 0.701$, $p = 0.485$). The two-way interaction between test session and training conditions was not reliable between the no feedback and transcription training group ($\hat{\beta} = 0.006$, $SE = 0.012$, $z = 0.475$, $p = 0.635$), or between the talker identification and transcription training group ($\hat{\beta} = -0.008$, $SE = 0.012$, $z = -0.726$, $p = 0.469$).

**Figure 6a.** Boxplots showing the distribution of mean proportion correct responses in the learning analysis (Test-00 vs. Test-01) for each training group.
**Maintenance.** The learning analysis examined performance between Test-01 and Test-02 (n = 71). As outlined previously, Test-01 occurred immediately after training and Test-02 occurred one-week following the initial training period. In Figure 7a, we depict the distribution of mean proportion correct responses at Test-01 and Test-02 for the subset of participants who returned for the one-week follow-up test. Visual inspection of Figure 7a reveals that performance at Test-01 appears slightly higher than performance at Test-02 for all training groups. Performance appears equivalent across training groups at each time point.

To examine these patterns statistically, trial-level responses on proportion correct keywords were submitted to a linear mixed effects model (LMM) with the fixed effect of training condition (treatment coded; transcription as the reference) and test session (Test-01 = -1, Test-02 = 1). The model also included random intercepts by subject, sentence, and talker, as well as random slopes by subject for test session.
The model revealed a main effect of test session ($\hat{\beta} = -0.025$, $SE = 0.008$, $z = -3.060$, $p = 0.003$), indicating that performance at test-01 was higher in accuracy relative to performance at test-02. With respect to training task, there was no difference between the talker identification and the transcription group ($\hat{\beta} = 0.008$, $SE = 0.028$, $z = 0.281$, $p = 0.780$), nor was there a reliable difference between the no feedback and transcription group ($\hat{\beta} = 0.024$, $SE = 0.028$, $z = 0.861$, $p = 0.392$). The two-way interaction between test session and training conditions was not reliable between the no feedback and transcription training group ($\hat{\beta} = 0.012$, $SE = 0.012$, $z = 1.022$, $p = 0.311$), or between the talker identification and transcription training group ($\hat{\beta} = 0.011$, $SE = 0.012$, $z = 0.880$, $p = 0.382$).
Discussion

Using a modified analysis based on proportion keyword correct, and removing keyword and item type as predictors, the current analysis converges with the findings presented in chapter three. Namely, we observed that across training groups, all participants showed a significant improvement from Test-00 to Test-01 that was equivalent across training groups. Our maintenance analysis revealed no difference in how the training groups maintained gains over time, which declined modestly at Test-02 relative to Test-01. We reported the same pattern of results in chapter three, with robust learning across groups, a slight decline in gains one-week following training, and no group differences in either the learning or maintenance analysis.