Perceptual Learning of Sine-Wave Speech:

Exploring, Generalization, Efficiency, and Individual Differences

A Thesis

Presented to

The Division of Philosophy, Religion, Psychology, and Linguistics

Reed College

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Arts

Kalin Mattern

May 2025

Approved for the Division

(Psychology)

Michael Pitts

Acknowledgments

This thesis wouldn't exist without the help and support from so many incredible people.

Thank you to all the professors at Reed who have had a positive impact on my time at Reed. Thank you to Victoria Fortuna, Kate Duffly, Bora Yoon, and Shohei Kobayashi for giving me space and inspiration to explore my creative identity that has so greatly influenced how I approach challenges. Thank you to Kevin Holmes for teaching me how to communicate difficult concepts and making me a better science writer in the process. Thank you to Jennifer Henderlong-Corpus for helping me realize my interest in psychology and for being such a kind and understanding professor. Most importantly, thank you to my advisor, Michael Pitts, for your careful line-editing, your ability to be calm even at points where I wasn't sure how to continue this project, and for pushing me to think deeper through every step of the process.

Thank you to the staff at Pappacino's for fueling my work. Thank you to everyone in the mLab for your help in getting my experiment up and running. Thank you to Yunkai Zhu and Andy Dykstra at the University of Central Florida for your collaboration with sorting out the nitty gritty details of my experiment and help creating my stimuli. Thank you to everyone who participated in my experiment, and to Karina and Helvetica for transcribing thousands of nonsense recordings into IPA, I would have no meaningful data if not for your help.

Thank you to my beautiful friends! Thank you, Alexis, Meagan, Walsh, and Rowan. You've all taught me resilience, how to find humor in the face of difficult feelings, and sisterhood. Camille, thank you for being my first and forever best friend, I can't wait to continue the tradition of Guillon-Mattern trips to Brigantine when we are real grown-ups and have our own families. To my long-time Reed friends, I feel so lucky to have met all of you and spent the past 4 years together, you are family to me. Gus and Parker, thank you for being your adventurous selves, our road trips together have been such a highlight of my time at Reed. Naomi, thank you for always sharing your laughter, and for always

being down to escape the horrors with some reality TV. I can't wait to keep listening to your music, even if it won't be with the exclusive access I've had from across the hall. Maggie, your commitment to playfulness in life has always been such an inspiration to me. I'm forever grateful that we sat next to each other on the first day of Hum, and I'm so excited to be living on the same coast as you after graduation. Ben, making music with you, even when it was perhaps grating (I'm thinking of our clarinet/bass recorder duet), has brought me so much joy. Someday soon we will make it to Dollywood together. Blix, thank you for being ridiculous always, but more importantly for always listening to my delusions/freak-outs and meeting them with sympathy and just the right amount of tough love. Thank you to every other one of my friends at Reed not mentioned here, I could fill an entire thesis with things that I appreciate about you.

Thank you to my family, for the endless support. Mom, thank you for listening to me complain for hours on the phone and always having some insider academia-knowledge trick up your sleeve. Dad, thank you for always telling me to "double-down" when things get hard, and for reminding me to take my vitamins (you know the one). Kai, thank you for diligently fulfilling your duties as a younger brother of always keeping me humble. Grandma and Grandpa, thank you for all your love, especially when expressed through happy birthday sung over voicemail.

Finally, thank you to Cedar, for being my favorite person. Thank you for loving and supporting me throughout all the doom and gloom that came with this project, as well as sharing in the excitement of progress and breakthrough moments. Through watching your dedication to your own thesis, you have amazed me, and helped me remember to always keep passion and curiosity by my side.

List of Abbreviations

EEG	Electroencephalogram
ERP	Event-related potential
fMRI	Functional magnetic resonance imaging
IPA	International Phonetic Alphabet
NVS	Noise-vocoded speech
OS	Original speech
RHT	Reverse Hierarchy Theory
STG	Superior temporal gyrus
SWS	Sine-wave speech

Table of Contents

List of Tables

List of Figures

Abstract

Past research in the field of speech perception has implemented a degraded-speech tool called sine-wave speech (SWS) to quickly modify perception using the same stimuli, from only hearing noise to perceiving speech. In these studies, stark differences in people's ability to perceive SWS as speech have been observed. Some people spontaneously perceive speech even without training, while others are never able to perceive speech at all. This study sought to examine learning, generalization, and potential sources of individual differences that have been previously demonstrated in the perception of SWS. In this study, participants were trained to perceive 3 different types of SWS: that derived from real English words (standard SWS), that derived from English language sounds but have no real meaning (pseudowords), and that which were so heavily degraded such to never be perceivable (flipped-frequency). Participants were assessed on their perceptual abilities before, throughout, and after training, including on untrained words to examine generalization to novel stimuli. Results raised some questions regarding traditional SWS experimental paradigms, but largely demonstrated learning of trained stimuli for both pseudowords and standard SWS tokens, while only learning of standard SWS tokens generalized to novel stimuli. In context of sensory perception theories, such as Reverse Hierarchy Theory, this experiment suggests that top-down associations are important for generalization of learning, but not for stimuli-specific learning, which can rely more heavily on bottom-up acoustic features.

Introduction

Background

After realizing that you have a little bit of extra time in your morning routine, you walk into a coffee shop to order your favorite beverage on the way to work. As you enter, your ears fill with the sounds of light jazz music playing over the speakers and other patrons chatting with each other. Ordering your coffee requires you to raise your voice, as the sound of the blenders and espresso machine behind the counter drown out your words. While you wait for your order, the person behind you clicks away at their laptop. Finally, as the bell attached to the door jingles when you exit, the barista yells out to "have a good day!" but you don't process what they said quite in time to respond. In nearly every environment that we interact with, just like this coffee shop, we are met with incredibly complex auditory scenes. Individual sounds are complex, layered, and varied in many auditory qualities such as volume, pitch, and timbre (the quality of a sound that makes a violin sound like a violin and a flute sound like a flute). Combine many sounds into a soundscape, and ambiguity and confusion arise as our brains must sort through the rich scene, identifying and assigning sonic sources to make sense of the world around us.

The world is filled with rich but ambiguous information that our brains are tasked with intaking, making sense of, and making use of, despite that ambiguity. Thus, a desire to understand the way that humans receive, process, and utilize auditory and other sensory information has become a key facet of the field of cognitive neuroscience (Aman et al., 2020; Calcus, 2024; Christison-Lagay et al., 2015; Snyder et al., 2012). To build this understanding, recent research has focused on discerning differences between when sensory information is either consciously perceived or not perceived, so that we might compare the neural and behavioral changes that characterize consciousness, or one's subjective experience. Though much of this research has been largely concentrated on visual perception, the interest in studying this comparison ranges across sensory modalities, with research in auditory perception steadily gaining more interest (Dykstra et al., 2017; Snyder et al., 2015;). Beyond broad auditory perception and scene analysis, the perception of speech-specific stimuli compared to non-speech is of particular interest (Calcus, 2024; Dykstra et al., 2017; Gohari et al., 2022; Viswanathan et al., 2012). In addition to speech's broad importance for facilitation of communication and expression of language, speech has been identified as a unique form of sensory input with a robust and intricate network of related psychological and neural processes, rendering speech a rich subject in studying auditory perception (Davis & Johnsrude, 2007; Norman-Haignere et al., 2015; Patterson & Johnsrude, 2008; Rosen et al., 2011; Uppenkamp et al., 2006).

Reverse Hierarchy Theory and Sensory Processing

In any perceptual processing, sensory input as a physical feature of one's environment is accompanied by an internal system that works to understand it. Though the physical and mental aspects of sensory processing are distinct, they work together. Returning to the coffee shop example, perhaps the light jazz playing over the speakers reminds you of the first concert you ever attended. Now, you are analyzing the auditory scene with that memory in mind, potentially influencing your perceptual experience. This interplay between sensory information and cognition and the role it plays in listeners' perception of auditory stimuli has thus led researchers to question how bottom-up and top-down processing function together in speech perception.

Bottom-up processing of stimuli is a type of processing that stems from the intake of low-level, objective and physical qualities of stimuli, relying on the most basic interaction with sensory organs and receptors. Top-down processing, however, is a cognitively based form of processing, forming predictions that are grounded in subjective experience and prior knowledge (also a result of top-down processing) that ultimately contextualize and make sense of bottom-up sensory feedback. Together, the dynamic relationship between bottom-up and top-down processing make up our conscious perceptual experiences, according to Reverse Hierarchy Theory (RHT). RHT has predominantly been used to describe visual processing but has grounds in other sensory modalities

(Hochstein & Ahissar, 2002; Shamma, 2008). Within the framework of RHT, bottom-up input informs top-down conceptions through feed-forward processing, and then top-down representations, in return, characterize bottom-up processing through a reverse feedback loop (Ahissar et al., 2009; Nahum et al., 2008; Shamma, 2008). The sensory information that we receive at the most basic level informs the way we think about and understand the world around us– a subjective perception of the world that then shapes how we interpret further information that we receive at sensory levels.

This model of the relationship between sensation and perception proposed by RHT has been observed in multiple experiments that have demonstrated specific top-down processes and their impacts on speech perception. For example, the way that individuals apply their attention to speech has been shown to influence the predictability of speech content, where insufficient attention to stimuli led to weaker predictions about the contents of degraded speech (Bhandari et al., 2022). Semantic relation, the similarity of words based on how often they appear in similar contexts, has also been shown to influence prediction and encoding of speech, where greater semantic relation between a priming sentence and a later target word led to greater prediction and encoding of acoustic features (Broderick et al., 2018). Both memory and categorization of speech have been shown to have relationships with extracting and understanding speech in noisy or otherwise difficult aural conditions (Hannemann et al., 2007; Viswanathan et al., 2012). Considering these findings, this study aims to further understand the role of top-down processes in conscious speech perception.

Sine-Wave Speech

One of the main challenges that research related to sensory processing faces is inconsistency of stimuli. To make causal claims, the best type of stimulus should be able to retain the same objective qualities before and after researcher manipulation. This way, any observed changes in behavioral or neural outcomes can truly be attributed to a difference in perception rather than a difference in a stimulus itself. In studying speech perception, a unique tool exists that allows researchers to manipulate participants' perception of stimuli without changing any of its bottom-up acoustic features– sine-wave speech (SWS). SWS is a form of artificially degraded speech that strips natural speech of many of its acoustic and harmonic qualities, only leaving behind a few areas of concentrated spectral energy "sine waves" that phonetically represent the original stimulus's formants, which are key frequencies in indicating qualities of vowels (Figure 1) (Remez et al., 1981). Upon first listening to SWS, listeners may hear a series of seemingly-random whistles and blips– overall unintelligible "noises." However, after a brief training, which simply involves listening to the original, undegraded, original speech (OS) sample, listeners are easily able to comprehend SWS as speech and understand its contents in a process of what has been characterized as "one-shot" learning (Remez et al., 1981; Whitehead et al., 2022). Given that SWS stimuli remain acoustically identical, yet individuals, depending on contextual information, can perceive the stimulus completely differently, renders it as a powerful tool for studying speech perception.

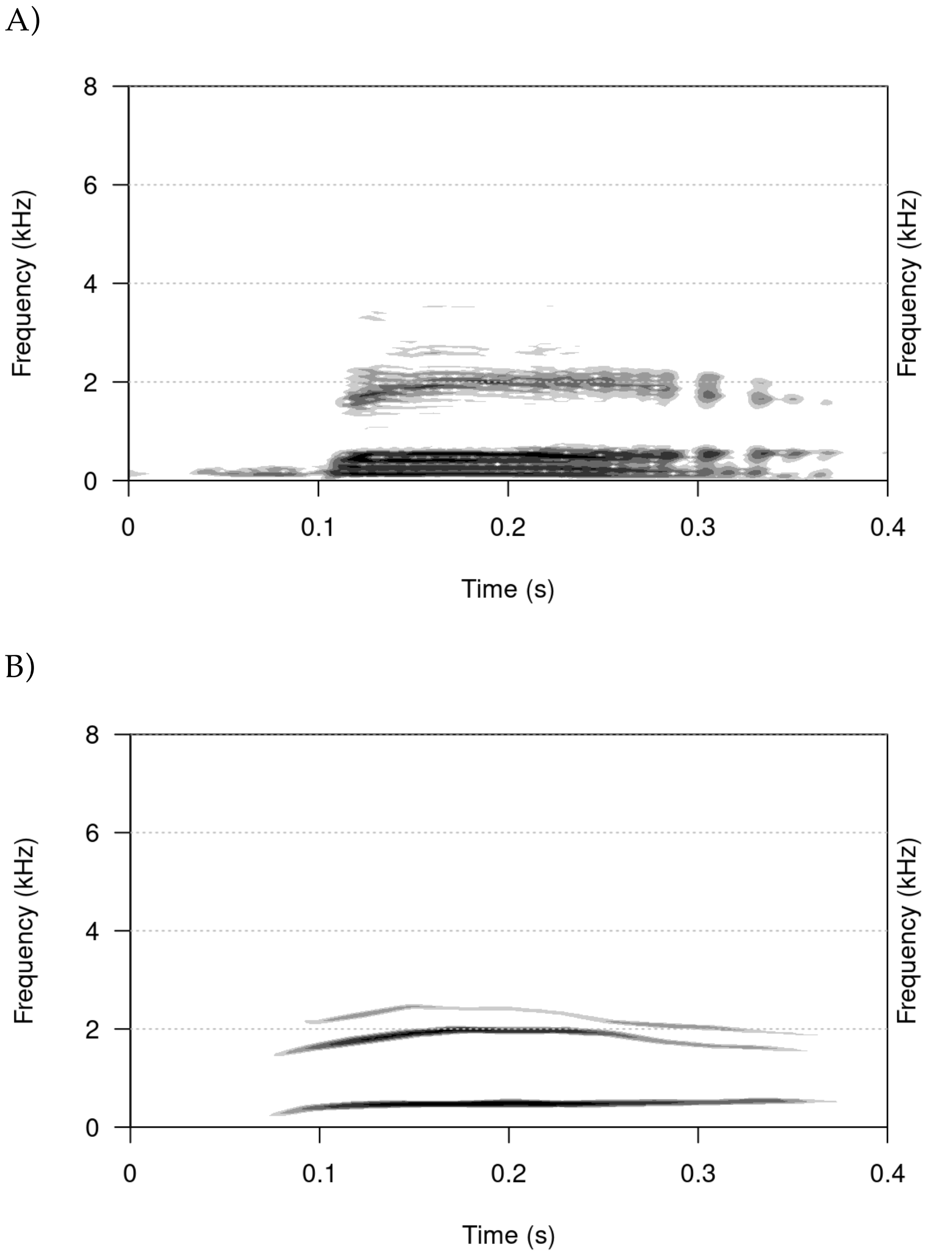


Figure 1: Spectrograms of the word "bear" before and after being converted to sine-wave speech.

Spectrograms are visualizations of all sonic frequencies (in kilohertz) present in any given audio sample, plotted over time (typically in seconds). *a)* This spectrogram depicts a speaker saying the word "bear." *b)* This spectrogram depicts the same clip of a speaker saying the word "bear," after being digitally degraded into sine-wave speech format. Notice that the "sine-waves" present in Figure 1b can be found in in Figure 1a, but they are surrounded by other frequencies.

Many studies have employed SWS in efforts to capture the neural correlates of conscious speech perception by using brain imaging techniques while participants undergo the perceptual shift, finding distinct neural changes associated with the onset of perception of SWS as speech. Studies using functional magnetic resonance imaging (fMRI) have identified increased activity the superior temporal gyrus and sulcus (STG) when SWS stimuli were perceived as speech rather than noise (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006). Using data from electrodes implanted in epilepsy patients' brains during surgery, researchers have corroborated the increased activity of the STG after comprehending SWS as speech, both in the perception of speech itself as well as the encoding of meaning. This study also cited the ventral sensorimotor cortex and inferior frontal gyri as active brain regions during the pre-comprehension phase, indicating the importance of these areas to bottom-up processing of low-level acoustic features (Khoshkhoo et al., 2018). A study using electroencephalogram (EEG), which measures electrical activity in the brain, even observed a perceptual awareness negativity—a specific event-related potential (ERP), or measured response to specific brain activity, that was associated with perception of speech rather than noise. This observation was also made through an experiment that employed a no-report paradigm, strengthening the association of that ERP to internal perception of a stimulus rather than any judgment or task-related neural activity (Zhu et al., 2024).

Aside from their contributions to uncovering different brain regions, networks, and types of activity that are implicated in speech perception, every single one of the studies mentioned above had noted distinct differences in participants' experience with perceiving SWS. In each study, researchers reported instances of participants spontaneously perceiving SWS as speech, even without any sort of training or priming to hear it as such (Deheane-Lamberts et al., 2005; Khoshkhoo et al., 2018; Möttönen et al., 2006; Zhu et al., 2024). In some studies, some participants did not ever perceive speech within the bounds of the experiment (Dehaene-Lambertz et al., 2005; Möttönen et al., 2006; Zhu et al., 2024). Even out of the participants who did undergo the perceptual switch within the bounds of the experiment, many experienced this shift at different efficiencies (Dehaene-Lambertz et al., 2005). Given that different people gained the perceptual ability of recognizing speech from noise at different rates indicates some sort of differences in perceptual learning across individuals.

Perceptual Learning and Generalizability

Perceptual learning is the way that a person's perception of stimuli may change over time based on their interaction with that stimulus, largely characterized by improvements in discrimination and detection abilities (Fahle & Poggio, 2002). Tightly tied to systems of memory, the process of perceptual learning forms implicit skills and associations over time, potentially due to plasticity in sensory cortices even in fully developed adult brains (Li & Gilbert, 2015; Weinberger, 1995). For a long time, it was generally agreed upon that perceptual learning is often highly specific to training conditions and hyper-specific qualities of stimuli, seldom generalizing to new contexts (Li & Gilbert, 2015; Sagi, 2010). However, some more recent studies have proposed that perceptual learning can indeed generalize, but that many sets of stimuli that participants are trained on are insufficient in the diversity of their bottom-up qualities to allow for generalization, such that top-down attention is too specified and thus hinders learning (Xiong et al., 2016). For example, stimulus uncertainty that arises from variations in target stimuli has been shown to limit learning and generalizability when the variations across stimuli exist and excite overlapping but not identical neural populations (Tartaglia et al., 2009). Conversely, when variations in stimuli are stronger and more distinct, where greater attentional shifts occur between perceptually engaging with stimuli, learning is facilitated (Zhang et al., 2008). Though both of those findings were considering visual perceptual learning, speech, with its temporal cohesion as a form of communication, yet distinct categorical representations between units, appears to be a good candidate for perceptual learning *and* generalization (Holt & Lotto, 2010).

The perceptual learning and generalization of speech stimuli has been demonstrated in natural speech environments, specifically surrounding non-native accents. When native American English speakers listened to native Chinese speakers (whose scores varied from low to high on baseline sentence intelligibility with regards to their accent when speaking English) read sentences in English, native English listeners' understanding of foreign-accented speech increased with exposure, regardless of baseline intelligibility. Though, more exposure was required to understand speakers with low baseline intelligibility, while less exposure was necessary for equal understanding of speakers with high baseline intelligibility (Bradlow & Bent, 2008). In natural speech contexts, listeners can perceptually adapt to differences in bottom-up acoustic features of speech while still implementing top-down associations to extract meaning.

Using noise-vocoded speech (NVS), another form of degraded speech, many studies have demonstrated such an effect in spectrally degraded speech as well. In one study, participants understood words and reported them correctly at significantly higher rates over the course of exposure to 30 NVS sentences (compared to the beginning of the experiment), indicating perceptual learning of degraded speech (Davis et al., 2005). Beyond establishing the presence of perceptual learning in degraded speech contexts, later studies demonstrated generalization across untrained words as well (Hervais-Adelman et al., 2008; Hervais Adelman et al., 2011). Instead of using full sentences, one study utilized single word tokens to control for structural or contextual clues. Participants were presented with a single word at a time and then asked to repeat it back as quickly and accurately as possible, receiving feedback on their accuracy with each response. Results displayed a distinct increase in perceptual ability over the course of the entire experiment, even though a completely new set of words was implemented in each of the multiple experimental blocks, indicating that perceptual learning of NVS generalized (Hervais-Adelman et al., 2008).

Given that perceptual learning to understand noise-vocoded speech is robust and has been shown to generalize, it is possible that similar patterns could be replicated with sine-wave speech. However, although NVS and SWS are both degraded forms of speech, they differ in which acoustic properties are degraded. As previously mentioned, SWS preserves spectral energy that represents formants, while removing broad-band frequencies that denote pitch and harmonic information. NVS, however, removes the details of spectral energy while retaining broad-band frequencies, so an essentially opposite manipulation of original speech. Whereas SWS is almost always unintelligible before top-down guidance through original speech due to its acoustic qualities, NVS is often used to simulate hearing with a cochlear implant, so is overall more intelligible (Hervais-Adelman et al., 2011). Although previous research has demonstrated robust perceptual learning and generalization with NVS, given the distinct differences in the structure and sonic output between NVS and SWS, further inquiry is required to understand whether such learning and generalization applies to SWS as well.

Individual Differences in Perceptual Learning Ability

With the spontaneous comprehension of SWS that has been seen in many studies employing the stimulus, the question of understanding how individual differences in perceptual learning and speech comprehension ability come about, and what factors might contribute to these differences, remains. When searching for patterns in variability across multiple unique perceptual learning tasks, researchers demonstrated that learning rates were highly varied across participants, but that variability was accounted for by a multivariate model that controlled for task, subject, and initial performance, indicating legitimate individual differences (as opposed to unrelated noise) in perceptual learning ability. Additionally, researchers collected data on and drew significant results from scores on personality traits contributing to individual differences (Yang et al., 2020). In a follow-up study by another group of interested researchers that sought to additionally assess generalization of learning, participants' learning rate across two visual perceptual learning tasks were related per individual, but not by task or other factors, indicating individual differences in perceptual learning abilities. This study also found that the degree to which this learning generalized varied by individuals, and that both perceptual learning and generalization outcomes were correlated with various cognitive, personality, and dispositional measures (Dale et al., 2021). Together, these studies exhibit evidence that different people experience perceptual learning in different ways based on unique learning abilities.

These studies, through their assessments of personality and dispositional traits, suggest that top-down, cognitive factors are implicated in producing learning differences. While these studies were focused on visual perception, the auditory perceptual modality offers a completely different set of top-down factors that could contribute to perceptual abilities. For example, in the large body of research dedicated to understanding the benefits of musical training and experience has demonstrated robust impacts on the neural encoding of speech. Engagement with music in many forms has been shown to improve overall auditory skills that are not exclusively beneficial to linguistic contexts, such as stronger neural encoding and cognitive understandings of pitch, timbre, rhythm, and timing (Kraus et al., 2009, Miendlarzewska & Trost, 2014). Beyond these basic aspects of listening, musicians have demonstrated better pitch distinction, understanding of melodic contours, and filtering of background noise, among many other factors in comparison to nonmusicians (Fujioka et al., 2004; Kraus & Chandrasekaran, 2010; Magne et al., 2006; Parbery-Clark et al., 2009; Patel, 2011, Yoo & Bidelman, 2019).

The benefits of musical experience on listening skills appears to transfer to linguistic realms as well. On the metric of pitch distinction, musician children performed better on detecting violations of pitch in both music *and* speech contexts than nonmusician children (Magne et al, 2006). In addition to pitch distinction, many more auditory skills that are relevant in experience with music are directly applicable to linguistic processing and production, such as pattern detection and vocal emotion recognition, revealed through enhancements in frequency encoding in auditory areas of the brainstem (Musacchia et al., 2007; Strait et al., 2009; Wong et al., 2007). Children who engaged with musical training for 6 months compared to visual art training showed improved reading and pitch discrimination in speech measured through ERPs and behavioral data (Moreno et al., 2008). Some studies have demonstrated overall enhanced linguistic perception of secondary languages in musicians compared to non-musicians, especially with regards to understanding statistical regularities in speech, which could potentially be transferable in learning to perceive other forms of speech that are initially unintelligible, such as SWS (Sadakata & Sekiyama, 2011; François & Schön, 2014).

Foreign language experience and bilingualism, like music, could have potential impacts on perceptual abilities related to speech as well. Different languages can include vastly different phonetic features, which exposes those who engage with those languages to a wider set of acoustic and phonetic knowledge. From a young age, children enrolled in dual language programs showed greater abilities in discriminating frequency modulation and had greater overall phonological awareness (the ability to both recognize and successfully manipulate speech) than their counterparts who did not partake in immersion programs (Jones et al., 2021). Greater phonological awareness and frequency modulation could perhaps positively impact overall speech perception, and therefore be applicable to the learning of degraded speech. However, in multiple instances, adult bilinguals showed greater difficulty in accurately perceiving speech from noisy conditions than monolinguals, which could instead hinder their perceptual learning (Bsharat-Maalouf & Karawani, 2022; Shokuhifar et al., 2024). These results could be due to bilinguals' later acquisition of language or splitting perceptual resources across multiple languages, so these greater acoustic vocabularies could potentially be detrimental to speech perception. Or perhaps, it depends on what language(s) one has experience with. For example, when comparing students at music conservatories in the United States, where non-tonal language is dominant, against those in China, where tonal language is dominant, more students at the Chines conservatories demonstrated absolute pitch, suggesting that speakers of tonal languages may attend to pitch variations with greater precision than non-tonal languages (Deutsch et al., 2006).

Given the somewhat conflicting evidence surrounding the auditory perceptual abilities of foreign language speakers, it is unclear whether experience with foreign language could be useful in learning to perceive SWS or instead, a detriment. Foreign language and musical experience are just two of many possible factors that could ultimately have an effect on the efficiency of one's perceptual learning of SWS but given each of their distinct impacts on auditory perception, it is reasonable to conduct further inquiry into their role.

The Current Study

Given the importance of auditory processing to our everyday lives, especially speech, a crucial aspect of language and communication, the current study aimed to investigate perceptual learning of sine-wave speech to gain further understanding of the processes and possibilities of perceptual ability, and generalization to new contexts. Because sine-wave speech can quickly be used to robustly manipulate one's conscious perception of a stimulus as either noise or speech, this study utilized SWS to assess perceptual learning over multiple series of one-shot trials, and how that learning generalized or failed to generalize to new contexts.

The experiment took place over 5 different blocks. The first block established a baseline perceptual ability of participants as they were exposed to degraded tokens, then immediately asked to speak back the word that they heard and complete a perceptual clarity rating assessing how clearly they could understand the speech they were presented with. While most of the tokens were real words in the English language, some tokens were spectrally flipped, meaning that they were modified such that they could never reasonably be perceived as speech, acting as the control. The final group of tokens words that were made up of combinations of sounds that follow the same rules and phonetic standards of English but hold no actual meaning—pseudowords.

The second, third, and fourth blocks then disambiguated the SWS by introducing the original speech samples to participants. Participants would hear the degraded token, complete their assessments, hear the original token, and then the same degraded token again followed by a final assessment of the stimuli (Trial, block, and full experimental flow outlined in Figure 3). In blocks 3 and 4, new sets of tokens were integrated with previously presented words to assess differences in the impact of prior exposure in the learning process. Key to blocks 2-4 was that the inclusion of flipped-frequency tokens was meant to act as a control, where speech could not be perceived neither pre-OS nor post-OS. Participants were instructed to only report back what they heard from the second degraded token and not the OS token, from which they could not possibly derive any helpful or accurate clues as to the contents of the degraded token. Therefore, they should not have been able to provide accurate responses for those trials.

In the fifth and final block, participants once again were only presented with the degraded token and immediately completed assessments, without hearing the original samples. In this final block, a group of words that participants had never heard before were presented. These words were key to assessing generalization, as participants never had these words disambiguated by original speech samples so had to rely solely on the perceptual abilities that they developed over the course of the experiment.

I first hypothesized that given the previous success of the SWS → OS → SWS paradigm in manipulating perception of speech versus noise, that if participants' perception of the SWS token is indeed altered by the presentation of the OS token, then in blocks 2-4, participants' verbal transcription accuracy and perceptual clarity ratings should be higher post-OS than pre-OS. Specifically, that this effect would be seen in standard SWS trials, but not in flipped-frequency trials, where speech cannot possibly be extracted. In the pseudoword condition, no exact hypothesis was made given the hitherto unknown roles of top-down vs bottom-up processing in the perceptual switch that occurs after hearing the original speech. Though, within the framework of RHT, top-down processes such as lexical representations (which standard SWS tokens contained, but pseudowords did not) inform and influence our bottom-up perception. To assess whether the process of one-shot learning of degraded speech fits this framework, I reasoned that if learning to comprehend SWS relies on top-down processes such as lexical association, then in the pseudoword condition there would be smaller differences between pre- and post- OS. Conversely, if top-down associations are not as important to learning to perceive SWS, then there might be similar pre-OS versus post-OS differences for both standard SWS and pseudowords.

Second, I assessed the degree to which perceptual learning of SWS occurred across the course of the experiment. Given that perceptual learning is the change of perception of a stimulus through interaction with that stimulus over time, exposure to and practice with the degraded tokens through this experimental paradigm should result in some perceptual learning. Thus, I hypothesized that participants would display greater pre-OS accuracy in later blocks compared to earlier blocks, with the starkest difference expected between the final block compared to the block 1 baseline, but a noticeable increase between blocks 2 and 3, after tokens were first disambiguated. Following the logic of the perceptual switch with regards to speech type, I hypothesized that learning could be most clearly seen for SWS tokens, and should not occur for flipped tokens. In the pseudoword condition, I hypothesized that even if the comparison between pre- and post-OS outcomes demonstrates equivalent results between SWS and pseudowords, that learning of pseudowords may occur at a less efficient rate. This might be the case even if the results of the pseudoword condition yield a result that indicates sufficiency of bottom-up cues in the perception of SWS, because the framework of RHT could explain a slower learning curve on tokens without top-down meaning. In other words, auditory processing of degraded speech may only *rely* on bottom-up acoustic features but could be aided and made more efficient by top-down associations.

To then examine whether any perceptual learning that occurred over the course of the experiment generalized to new stimuli, I compared transcription accuracy and perceptual clarity ratings across the course of the experiment in only novel tokens. I hypothesized that if perceptual learning does generalize, participants would still display higher learning outcomes toward the end of the experiment, even in unheard words. Though, it is possible that learning does not generalize to new words, in which case these outcomes would not show any sort of improvement across blocks.

Finally, to try and discern a picture of some possible influences on individual differences in perceptual learning abilities for auditory stimuli, I compared participants' learning outcomes with their prior experience with music or foreign language. Given that musical experience has been shown to improve auditory skills in ways that benefit speech stimuli, I hypothesized that musical experience would be positively correlated with perceptual outcomes. Regarding foreign language, the literature was more mixed. Studies have demonstrated some increased auditory skills associated with exposure to and experience with foreign language, though it is unclear whether these auditory skills necessarily support speech-specific stimuli. Thus, I did not have any specific hypothesis regarding the impact of foreign language experience on perceptual outcomes but reasoned that if the type of attention to auditory cues associated with foreign language experience is helpful to comprehending speech stimuli, that experience could be positively correlated with perceptual outcomes, whereas if it is detrimental, foreign language experience could potentially be negatively correlated with perceptual outcomes.

Methods

Participants

17 native English speakers with self-reported normal hearing and auditory processing ages 19-23 (*M* = 21.61) participated in this study. Participants received three entries into a lottery for $50 in return for their participation. All procedures were approved by the Reed College Institutional Review Board.

Stimuli

Auditory stimuli consisted of 196 total tokens across three conditions: standard SWS tokens, flipped-frequency tokens, and pseudoword tokens. In the standard condition, 81 monosyllabic English words representing either animals or inanimate objects such as "bear" or "spoon" were generated into spoken recordings to be used as OS tokens, using Google's WaveNet Text-to Speech AI, which allowed for control of duration and pitch across all tokens (Appendix 1). These original speech tokens were then converted to SWS, saved to be used as standard SWS tokens, and then additionally degraded to flipped-frequency tokens in MATLAB (Figure 2). Pseudoword tokens (33 total) were generated using UniPseudo, consisting of fake monosyllabic words that retain the same phonemes, structures, and patterns as English words, but are not real words (New et al., 2024) (Appendix 2). These tokens have the same bottom-up cues as English words and could reasonably be perceived as speech but are lacking top-down associations and thus remain meaningless. Individual participants were not presented with all 196 tokens, but a randomized subset that drew a specific number of tokens from each stimulus type (discussed further under "Procedure").

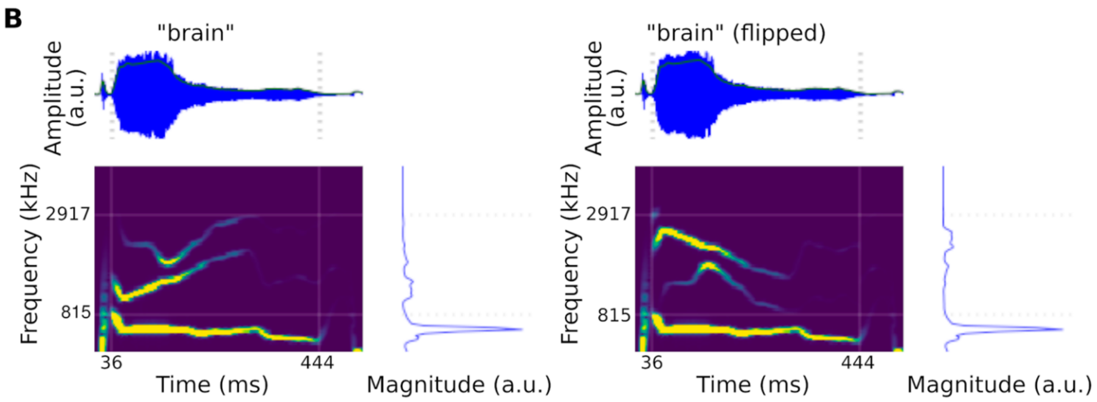


Figure 2: Conversion from standard SWS token to flipped-frequency token. *a)* Spectrogram, waveform, and magnitude of the SWS token, "brain." Waveforms and magnitude indicate intensity or loudness of sounds. Typically measured in decibels, a.u. stands for arbitrary unit and indicates the amplitude was normalized *b)* Spectrogram and waveform of SWS token after flipped-frequency modulation. Main differences can be seen in the top two formants (F2 and F3), which were flipped on a center axis. Magnitude and amplitude, as well as the first formant are similar to the non-flipped version (Zhu et al. 2024, Figure 1B).

Equipment and Software

The experiment was coded and run in PsychoPy on lab computer monitors. Stimuli were presented through Philips Audio Over-Ear Stereo Headphones at a sampling rate of 44.1 kHz, and verbal transcriptions were recorded using Yeti Blue Microphone, sampling rate capacity of 48 kHz. Recordings were subsequently transcribed into the International Phonetic Alphabet (IPA) by hired linguistics students and scored by the number of phonetic symbols that were correctly reported by participants in comparison to the target token. Data analyses were conducted in RStudio.

Procedure

Participants completed the experiment over the course of 1-1.5 hours in a quiet room in the Reed College psychology department building, wearing headphones to isolate experimental stimuli. Prior to participation, individuals signed consent forms.

Block 1

To assess participants' ability to learn to perceive SWS, both in one-shot trials and over-time settings, the experiment took place over the course of 5 different blocks. In the first block, which served as a baseline assessment of perceptual ability, participants completed 42 trials in random order where they were presented with a SWS token from one of the three conditions (30 standard SWS, 6 flipped-frequency, 6 pseudoword). Before starting the experiment, participants were notified that they would be listening to short auditory clips containing word tokens, but that not every single clip would contain real English words. After listening to the token, participants were asked to verbally repeat back the word they heard into a microphone. Responses were transcribed into the IPA and given an accuracy percentage indicating the number of symbols accurately reported compared to what was present in the target token. For example, if the participant was presented with the word "boat" (transcribed as boʊt), and responded with "coat" (transcribed as koʊt), they would receive an accuracy score of 75% as they accurately reported 3 out of 4 symbols present in the target.

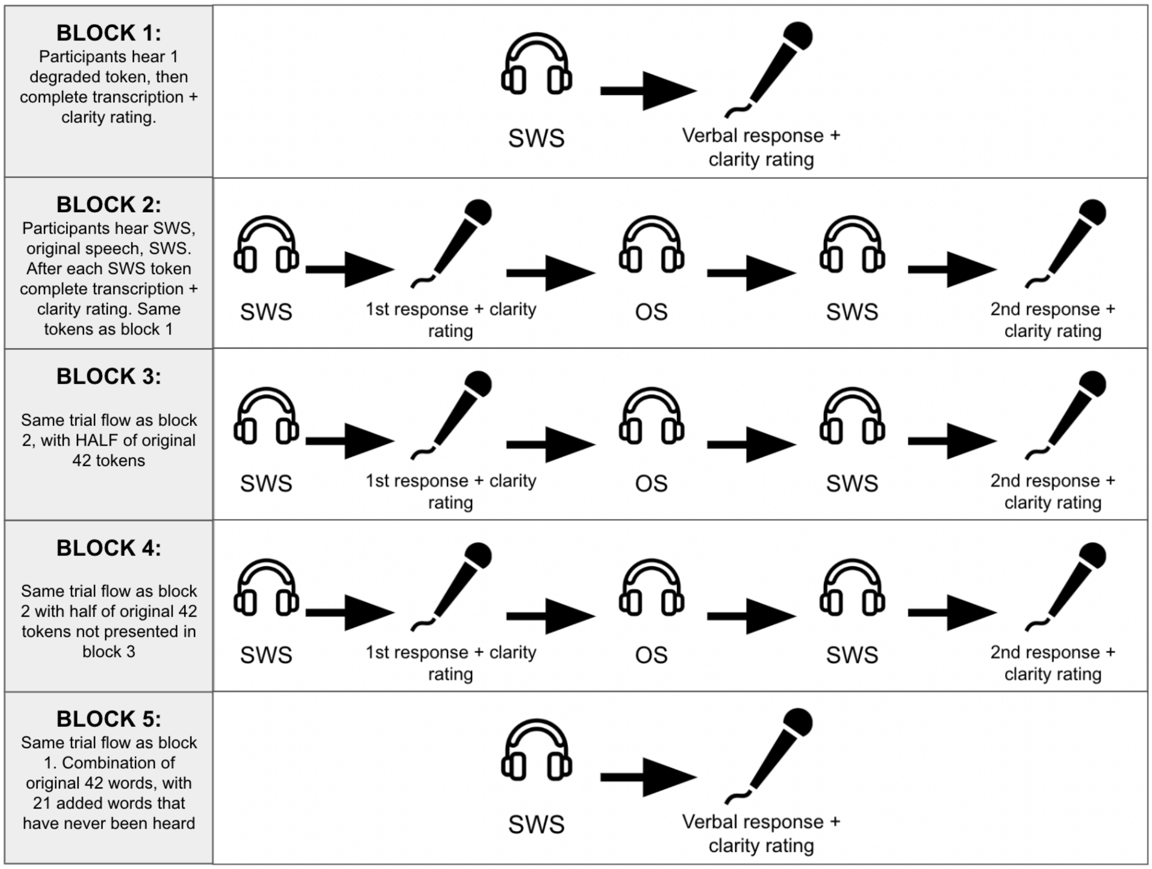
If the participant truly could not discern any word or parts of words from the stimuli, they were given the option to opt-out of the transcription for that trial by pressing a button indicating that they could "only hear the stimulus as noise." Using this button resulted in an automatic verbal response accuracy score of zero. The verbal transcription task was followed by a 4-point perceptual clarity rating indicating how clearly speech was perceived by participants (1 = extremely unclear, 4 = extremely clear). In this first block, participants did *not* hear the original speech samples associated with each token, meaning that their verbal responses and perceptual clarity ratings only reflect their perceptual abilities *prior* to any sort of training to hear SWS as speech.

Block 2

In the second experimental block, the same 42 tokens from the first block were re-randomized and repeated, this time disambiguated for the listener. After hearing the SWS token, followed by the verbal transcription and perceptual clarity rating, participants were presented with the original speech (OS) sample. Following the OS, participants once again heard the SWS token and completed another transcription and rating.

Blocks 3 and 4

Blocks 3 and 4 followed the same trial flow as block 2, with each token being disambiguated before being presented as SWS again. In these blocks, only half of the words present in block 1 (half in block 3, the other half in block 4) were repeated, and the other half replaced by a set of 15 new standard SWS tokens, 3 pseudoword, and 3 flipped.



BLOCK 1: Participants hear 1 degraded token, then complete transcription + clarity rating.

BLOCK 2: Participants hear SWS, original speech, SWS. After each SWS token complete transcription + clarity rating. Same tokens as block 1

BLOCK 3: Same trial flow as block 2, with HALF of original 42 tokens

BLOCK 4: Same trial flow as block 2 with half of original 42 tokens not presented in block 3

BLOCK 5: Same trial flow as block 1. Combination of original 42 words, with 21 added words that have never been heard

SWS
Verbal response + clarity rating
SWS
1st response + clarity rating
OS
SWS
2nd response + clarity rating
SWS
1st response + clarity rating
OS
SWS
2nd response + clarity rating
SWS
1st response + clarity rating
OS
SWS
2nd response + clarity rating
SWS
Verbal response + clarity rating

Figure 3: Experimental flow chart.
Broken down by trial flow and block structure, this figure depicts the overall experimental procedure.

Block 5

Block 5 returned to the same trial flow utilized in block 1, where participants heard only the SWS token, followed by the verbal transcription and perceptual clarity rating. Stimuli included all 42 tokens from block 1, as well as 15 completely new standard SWS tokens, 3 pseudoword, and 3 flipped. Participants were not presented with the original speech sample at any point during these trials. While most words had been heard by participants in previous blocks, the new words added in block 5 were never associated with an original speech sample.

Questionnaire

After completing the experimental blocks, participants completed a final questionnaire in Qualtrics to assess their musical and foreign language expertise. To assess musical experience, participants completed the General Musical Sophistication subscale of the Goldsmiths Musical Sophistication Index (Gold-MSI), which assesses on many factors including active engagement with music, perceptual abilities, musical training, singing abilities, and emotional responses to music, (Müllensiefen et al., 2014). For scoring information and items, see Appendix C. To assess foreign language experience, participants also completed the Bilingual Language Profile (BLP) which both accounts for experience with and exposure to language, as well identifying the specific languages that participants have interacted with (Birdsong et al., 2012). Scoring and items for the BLP are discussed in Appendix D.

Results

Confirming the Perceptual Switch

The first analysis was aimed at confirming the perceptual switch effect seen in previous studies, where SWS is perceived after disambiguation through listening to an OS sample. Given the replication of this effect in many studies since the seminal study on SWS by Remez et al., 1981, I hypothesized that there would be greater verbal response accuracy and higher clarity ratings post-OS compared to pre-OS, indicating stronger speech perception as a direct result of the disambiguation. To test this, one-way repeated measures ANOVAs compared clarity ratings and verbal response accuracies pre-OS versus post-OS for standard SWS tokens. All words from block 2, as well as those words from blocks 3 and 4 that had not been previously presented were included in this analysis, i.e., only words that had never been disambiguated before. As expected, post-OS clarity ratings (M = 2.632, SD = 0.848) were significantly higher than pre-OS ratings (M = 2.038, SD = 0.629), $F(1,16) = 45.627$, $p < .001$, and post-OS accuracy (M = 0.693, SD = 0.253) was significantly higher than pre-OS accuracy (M = 0.347, SD = 0.216), $F(1,16) = 68.823$, $p < .001$ (Figure 4), confirming the perceptual switch brought about by listening to the original speech sample.

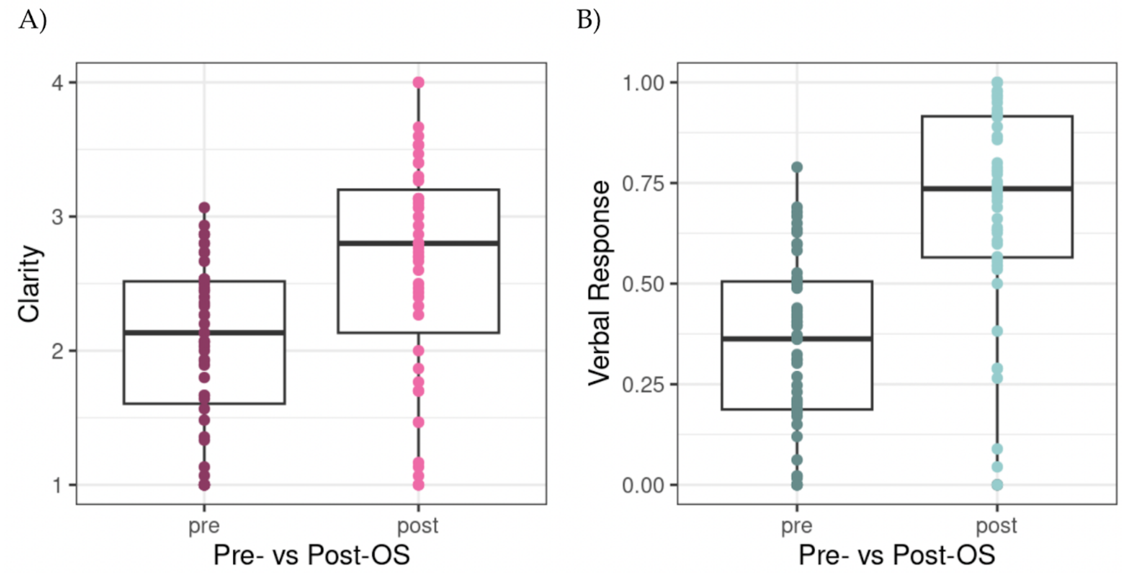


Figure 4: Pre-OS vs. Post-OS perceptual clarity ratings and verbal response accuracy
a) Perceptual clarity ratings (1-4) pre- and post- presentation of original speech (OS) stimulus in blocks 2-4, for new words. *b)* Verbal response accuracy (0-1) pre- and post- presentation of OS stimulus in blocks 2-4, for new words.

Along with confirming the perceptual switch, I aimed to assess the robustness of the one-shot learning effect, that is how much the one-shot learning effect truly happens in "one shot." To accomplish this, I assessed whether there were changes in perceptual outcomes across blocks 2-4 for post-OS outcomes on "old" (already heard) SWS tokens. This analysis enabled an assessment of additional perceptual learning over multiple instances of stimulus interaction and disambiguation, beyond the initial one-shot effect. A robust and exclusive one-shot learning effect would be indicated by unchanging post-OS scores across blocks, as this would represent a learning cap where perceptual abilities do not improve with further stimulus interaction after the one-shot trial. An additional effect would be characterized by increasing post-OS scores that represent a more gradual learning process brought about by multiple instances of disambiguation that build on each other, rather than one strong instance. A one-way repeated measures ANOVA assessing *only* old, *post*-OS SWS outcomes by blocks 2-4 revealed significant differences for both perceptual clarity, $F(2,32) = 12.976$, $p <$

.001, and verbal response accuracy, $F(2,32) = 19.243$, $p < .001$, across blocks, indicating support for more gradual learning rather than an exclusive one-shot learning effect (Figure 5). Post-hoc pairwise comparisons between blocks revealed that the ANOVA results were driven by significant differences between block 2 and block 4 (clarity: $p < .05$, verbal response: $p < .05$), with scores in block 3 falling numerically in between.

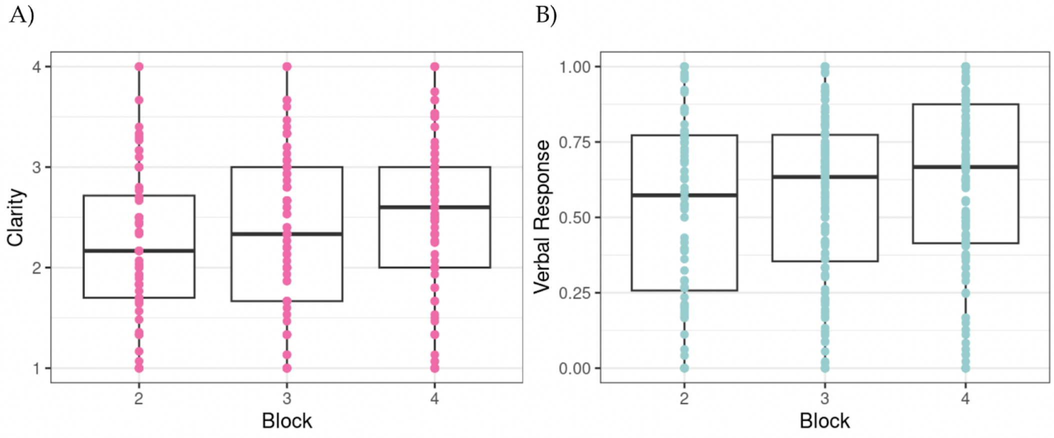


Figure 5*:* Perceptual clarity ratings and verbal response accuracy after hearing original speech.
a) Perceptual clarity ratings (1-4) for *only* old, post-OS SWS stimuli presented in blocks 2-4, split by block. *b)* Verbal response accuracy (0-1) for *only* old, post-OS SWS stimuli presented in blocks 2-4, split by block.

Next, to examine the differences in the one-shot learning effect associated with speech type, I further assessed pre- vs post-OS outcomes broken down by the stimulus type (standard SWS words, pseudowords, and flipped control stimuli). This assessment was aimed at answering the question of top-down versus bottom-up processing as a factor in the one-shot learning process. If learning primarily occurs at the acoustic level, there would be significantly higher post-OS outcomes in both pseudowords and standard SWS but not flipped tokens, because both of the former contain adequate acoustic features (they both reasonably sound like English speech), whereas flipped tokens do not. However, if lexically-driven top-down processing is either required to bring about or enhance the one-shot learning effect, then there would be that significant difference for standard SWS tokens, as they are the only stimulus type to retain any lexical associations.

A 2x3 repeated measures ANOVA was conducted with the factors: stimulus timing (pre-OS, post-OS), and stimulus type (standard, pseudo, and flipped) for each of the two perceptual outcome measures (clarity and accuracy) using scores collapsed across blocks 2-4. These analyses revealed main effects of both stimulus type (clarity ratings: $F(2,32) = 13.045$, $p < .001$; verbal response accuracy: $F(2,32) = 13.438$, $p < .001$) and stimulus timing (clarity ratings: $F(1,16) = 28.337$, $p < .001$; verbal response accuracy: $F(1,16) = 78.201$, $p < .001$), with no significant interaction between the two, meaning that the difference between pre-OS and post-OS outcomes did not seem to be affected by stimulus type (Figure 6). I hypothesized that there would be no significant differences between pre- and post-OS tokens in the flipped-frequency condition (in contrast to clear differences between pre- and post- in the standard SWS condition), as flipped tokens were designed to be unintelligible both before and after disambiguation. Interestingly, post-hoc Holm comparisons did show significantly higher outcomes post-OS (clarity: $M = 2.33$, $SD = 0.88$; verbal response: $M = 0.54$, $SD = 0.31$) than pre-OS in flipped tokens (clarity: $M = 1.89$, $SD = 0.68$; verbal response: $M = 0.25$, $SD = 0.19$), clarity: $p < .001$; verbal response: $p < .001$. This result raises the likelihood that participants could have been referencing the original speech token when responding to the post-OS stimulus, even if this was unintentional.

That being said, flipped token clarity and response accuracy in the pre- and post- conditions were significantly lower than their standard SWS counterparts, which drove much of the main effect of stimulus type. In other words, flipped-pre-OS tokens (clarity: $M = 1.89$, $SD = 0.68$; verbal response: $M = 0.25$, $SD = 0.19$) were significantly lower than SWS-pre-OS tokens (clarity: $M = 2.14$, $SD = 0.66$; verbal response: $M = 0.39$, $SD = 0.23$), clarity: $p < .001$; verbal response: $p < .001$, and flipped-post-OS tokens (clarity: $M = 2.33$, $SD = 0.88$; verbal response: $M = 0.54$, $SD = 0.31$) were significantly lower than SWS-post-OS tokens (clarity: $M = 2.70$, $SD = 0.86$; verbal response: $M = 0.71$, $SD = 0.24$), clarity: $p < .001$; verbal response: $p < .001$, indicating that SWS still had overall significantly better perceptual outcomes than flipped tokens. I revisit this finding about flipped tokens in the next section of results, "Assessing Learning and

Generalization." Regardless, no significant differences were found when making this same comparison between pseudowords and SWS tokens. Pseudowords still displayed the perceptual switch, just like flipped and SWS tokens, but their clarity and response accuracy in the pre- and post- conditions were not significantly different from their SWS counterparts.

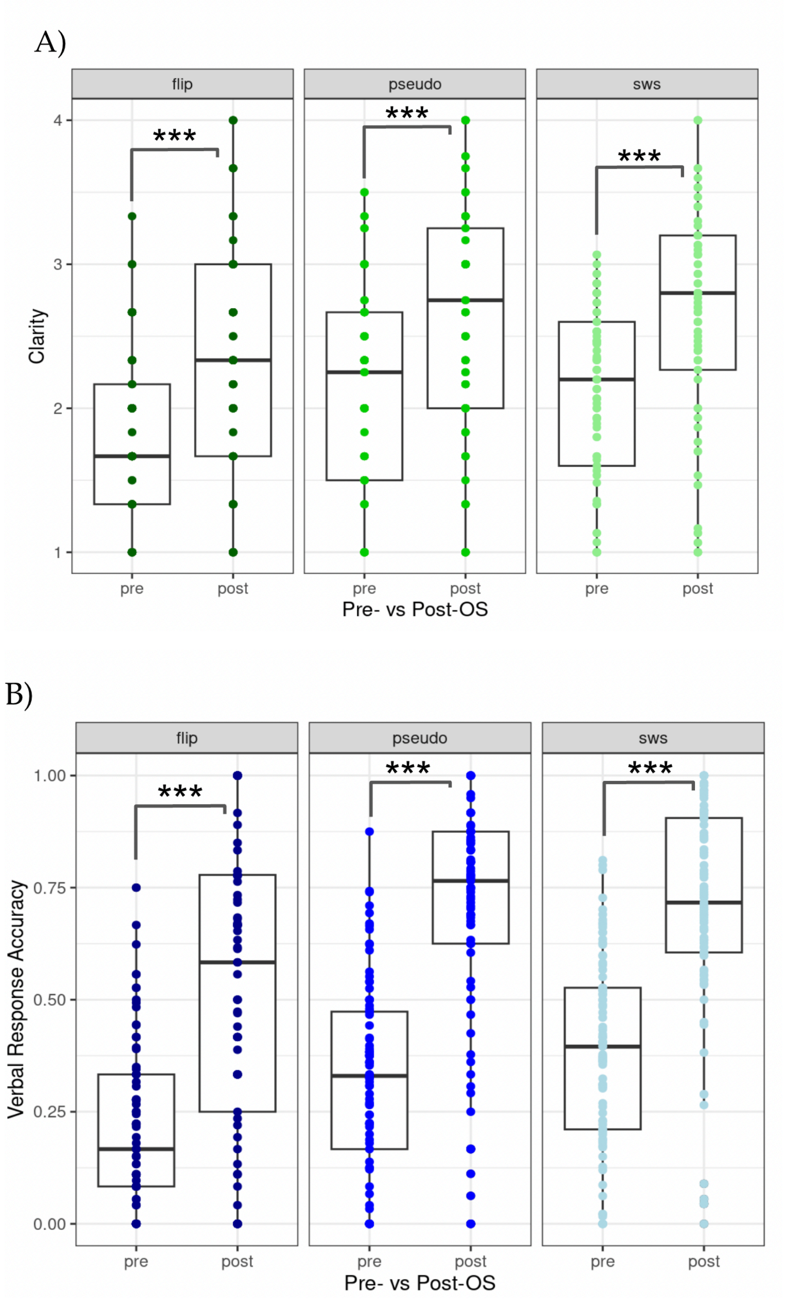


Figure 6*:* Pre-OS versus Post-OS tokens of all 3 speech types, with only tokens that have not been disambiguated before.
a) Perceptual clarity ratings (1-4) pre- and post- presentation of original speech stimulus in blocks 2-4, split by speech type. *b)* Verbal response accuracy (0-1) pre- and post- presentation of original speech stimulus in blocks 2-4, split by speech type. **p* < .05, ** *p* < .01, *** *p* < .001, **** *p* <.0001.

Assessing Learning and Generalization

Moving beyond the initial perceptual switch, the next set of analyses was aimed to assess whether learning occurred over the course of the whole experiment. I hypothesized that perceptual outcomes would increase from the beginning to the end of the experiment for standard SWS tokens but not for flipped tokens, with learning of pseudoword tokens potentially falling somewhere in the middle. To examine this, I first analyzed clarity ratings and verbal response accuracy in block 1 compared with those in block 5 (including both old and new tokens), split by speech type. Given that in block 1 participants had no training yet (no exposure to OS stimuli), performance in this block indicated participants' baseline ability to perceive SWS. Performance in block 5 reflects participants' ability *after* undergoing the training process in the intervening blocks. I thus hypothesized that for standard SWS tokens, participants would perform significantly better in block 5 compared to block 1. This test also allowed for a more careful assessment of the flipped frequency control tokens. While the previous test assessing pre-OS versus post-OS outcomes was potentially confounded by participants (knowingly or unknowingly) relying on the crutch of the original speech stimulus to inform their post-OS responses, both blocks 1 and 5 did not include any OS stimuli, therefore, this analysis captured a purer view of perceptual abilities.

A 2x3 repeated measures ANOVA was conducted with the factors: block (1 or 5), and stimulus type (standard, pseudo, and flipped) for each of the two perceptual outcome measures (clarity and accuracy). I expected that there would be an interaction of block and stimulus type. For clarity ratings, main effects were found for block, $F(1,16) = 57.606$, $p < .001$, and stimulus type, $F(2,32) = 57.606$, $p < .001$, separately, but no interaction was evident (Figure 7a). The main effect of block confirms that participants experienced greater perceptual clarity in block 5 ($M = 2.17$, $SD = 0.72$) than block 1 ($M = 1.47$, $SD = 0.49$) but the lack of an interaction indicates that stimulus type had no effect on these differences. That is, flipped, pseudoword, and standard SWS tokens all still resulted in increased perceptual clarity after training, once again challenging the logic that participants should not be able to perceive flipped frequency tokens.

However, clarity ratings are a distinctly subjective measure of perceptual ability. For verbal response accuracy, an objective measure of perceptual abilities, there was indeed an interaction between block and speech type, $F(2,32) = 6.760$, $p < 01$, such that significant increases from block 1 to block 5 were only found for pseudoword tokens (block 1: $M = 0.17$, $SD = 0.20$; block 5: $M = 0.39$, $SD = 0.21$) and standard SWS tokens (block 1: $M = 0.16$, $SD = 0.16$; block 5: $M = 0.48$, $SD = 0.22$), but not for flipped frequency tokens (Figure 7b). This outcome indicates that when perceptual abilities are represented by an objective measure of verbal response accuracy, the "learning" demonstrated by clarity ratings for flipped tokens goes away. That is not to discount the differences demonstrated by perceptual clarity ratings– it simply means that participants, on average, felt that they perceived tokens more clearly after training, including the flipped frequency ones, despite showing objective improvements only for the stimuli with genuine speech content.

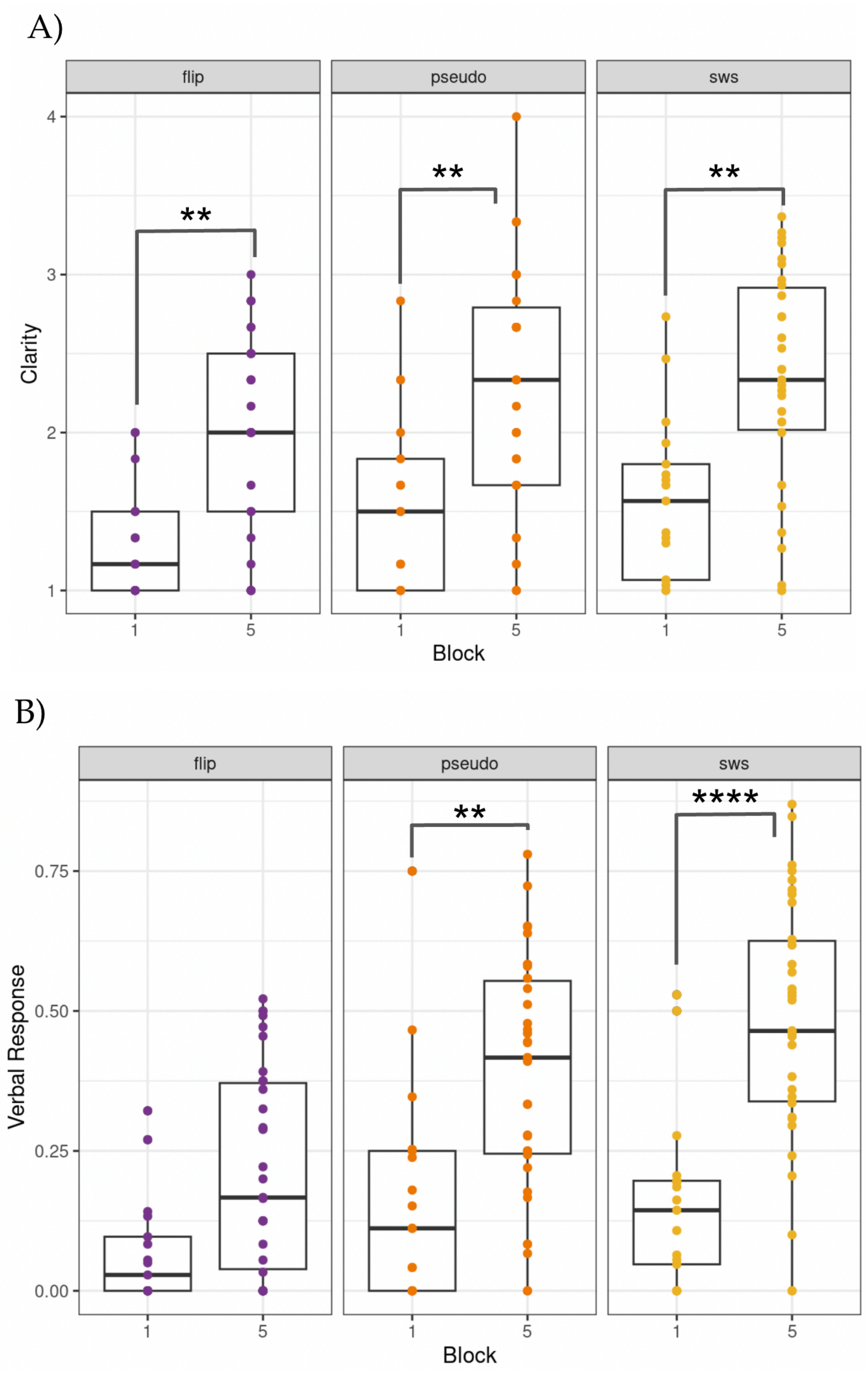


Figure 7: Clarity ratings and verbal response accuracy in block 1 versus block 5 to assess learning over the course of the experiment, for each stimulus type. *a)* Clarity ratings in block 1 versus block 5 for flipped tokens, pseudoword tokens, and standard SWS tokens. block 5 tokens included both old and new tokens. *b)* Verbal response accuracy in block 1 versus block 5 for flipped tokens, pseudoword tokens, and standard SWS tokens. Again, block 5 tokens included both old and new tokens. *$p < .05$, ** $p < .01$, *** $p < .001$, **** $p <.0001$.

After confirming that learning did occur from block 1 to 5 for standard SWS and pseudoword tokens, I next investigated the timing of this learning more precisely by analyzing perceptual outcomes across all five blocks for these stimulus types. I included only the pre-OS data from blocks 2-4, and only old words in blocks 3-5 to track learning of the repeated tokens across time. Because the largest increase in clarity and accuracy was expected after the first time participants heard the OS (with potential further improvements thereafter), I expected to see scores increase most from blocks 2 to 3, with possible further increases from blocks 3 to 4, and 4 to 5. I assessed this separately for regular SWS tokens and pseudoword tokens to see if there were learning differences between the two, once again to understand the role of top-down versus bottom-up processing in not just immediate perception, but learning over time. Furthermore, I only ran these analyses on verbal response accuracy given the previous results that demonstrated it as a more reliable and objective measure of learning than perceptual clarity.

Repeated measures ANOVAs indicated a significant effect of block on standard SWS verbal response accuracy, $F(4,64) = 43.224$, $p < .001$ (Figure 8a). Post-hoc tests revealed significant differences between specific block pairings, outlined in Table 1. Similar findings were evident for pseudoword tokens, where there was a significant effect of block on verbal response accuracy, $F(4,64) = 10.087$, $p < .001$ (Figure 8b). Though these pairwise comparisons did not display the exact pattern of differences that aligned with my block-by-block hypotheses, they still revealed useful information about the difference between standard SWS and pseudoword learning. While significant differences between block 1 and subsequent blocks did not begin until block 4 for pseudowords, standard SWS tokens began to show significant differences slightly earlier, in block 3, suggesting more efficient learning of standard SWS words.

A)

Blocks	*p*	Significance
1 vs. 2	.465	ns
1 vs. 3	.002	**
1 vs. 4	< .001	***
1 vs. 5	< .001	***
2 vs. 3	.125	ns
2 vs. 4	.078	ns
2 vs. 5	.004	**
3 vs. 4	.080	ns
3 vs. 5	.628	ns
4 vs. 5	.629	ns

B)

Blocks	*p*	Significance
1 vs. 2	.961	ns
1 vs. 3	.506	ns
1 vs. 4	.045	*
1 vs. 5	.006	**
2 vs. 3	.955	ns
2 vs. 4	.162	ns
2 vs. 5	.030	ns
3 vs. 4	.955	ns
3 vs. 5	.370	ns
4 vs. 5	.961	ns

Table 1: Post-hoc comparisons for repeated measures ANOVAs assessing verbal response accuracy by block, for both standard SWS tokens and pseudoword tokens.
a) Post-hoc pairwise comparisons for a repeated measures ANOVA on verbal response accuracy in pre-OS, already heard standard SWS token trials, by block.
b) Post-hoc pairwise comparisons for a repeated measures ANOVA on verbal response accuracy in pre-OS, already heard pseudoword token trials, by block. *p < .05, ** p < .01, *** p < .001, **** p <.0001.

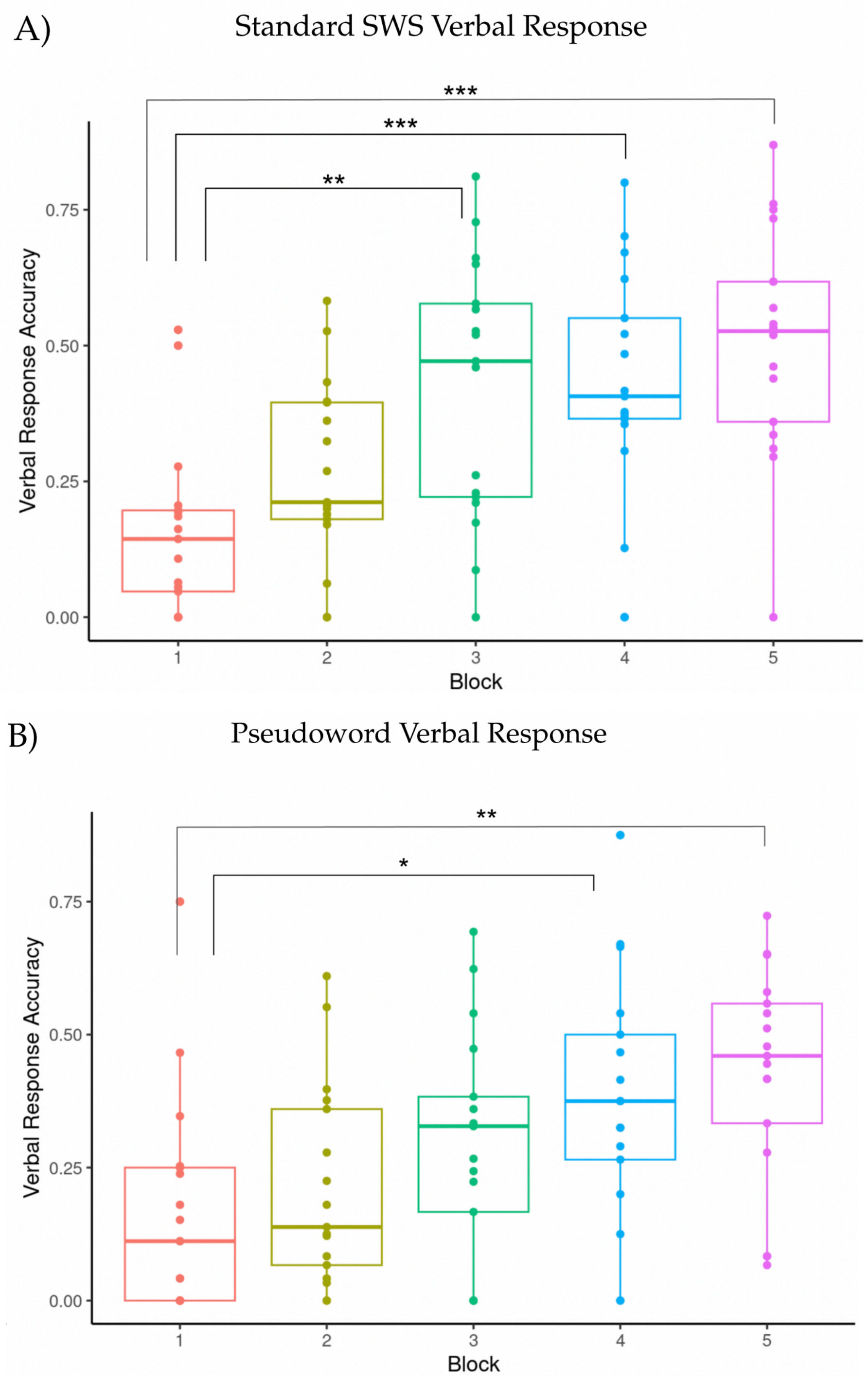


Figure 8: Clarity and verbal response accuracy for both standard SWS and pseudoword tokens across each block of the experiment.

a) Verbal response accuracy for standard SWS tokens across all blocks, filtered when applicable (pre-OS for blocks 2-4, old words for blocks 3-5). *b)* Verbal response accuracy for pseudoword tokens across all blocks, filtered when applicable (pre-OS for blocks 2-4, old words for blocks 3-5). **p* < .05, ** *p* < .01, *** *p* < .001, **** *p* <.0001.

Given the learning exhibited by participants across the experiment, the next question to assess was whether this learning generalized to novel stimuli. In other words, could participants learn to perceive SWS "in general" such that their perceptual abilities for new stimuli that they had never heard before improved across the course of the experiment? To test for generalization of learning, I compared block 1 versus block 5, split by speech type, though this time limited to only new words. Once again, I limited this analysis to only verbal response accuracy for an objective measure of generalization. I ran a 2x3 repeated measures ANOVA with the factors of block (1 or 5), and stimulus type (standard, pseudo, and flipped) and hypothesized that if learning generalizes, it would be represented by significantly higher verbal response accuracy in block 5 than block 1. I also hypothesized that flipped frequency tokens would not demonstrate this generalization while standard SWS tokens would, which should be picked up by this analysis as a significant interaction. I had no specific predictions about pseudowords. The ANOVA revealed a significant interaction between block and stimulus type, *F*(2,32) = 4.399, *p* < .05 (Figure 9), which post-hoc comparisons revealed to be driven by significant differences between standard SWS tokens in block 1 (*M* = 0.16, *SD* = 0.16) versus 5 (*M* = 0.45, *SD* = 0.23), *p* < .001, and the absence of these differences for flipped and pseudoword tokens. These results suggest that perceptual learning does generalize, but only for standard SWS tokens.

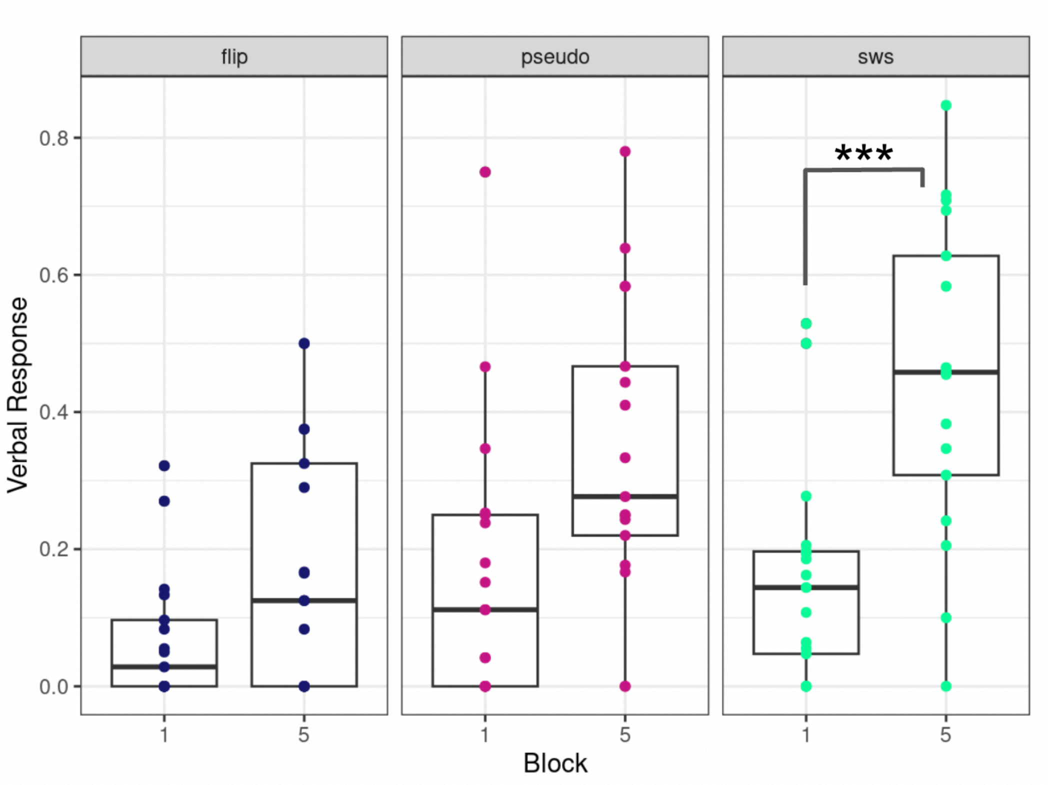


Figure 9: Verbal response accuracy in block 1 versus block 5 to assess generalization of learning to new tokens, for each stimulus type.
Verbal response accuracy in block 1 versus block 5 for flipped tokens, pseudoword tokens, and standard SWS tokens. Block 5 only included new tokens. *p < .05, ** p < .01, *** p < .001, **** p <.0001.

Following my strategy to assess overall learning, I further broke down the generalization of learning effect to understand its efficiency. For this follow-up analysis, I only used standard SWS tokens because pseudowords did not demonstrate a generalization of learning effect. A repeated measures ANOVA with block (1,3,4,5) and verbal response accuracy (only on new tokens, pre-OS for blocks 3 and 4) revealed a main effect of block for standard SWS tokens, $F(3,48) = 23.561$, $p < .001$. Only a few pairwise comparisons were significant, and these are indicated by the brackets in Figure 10. Verbal response accuracy was significantly higher than in baseline from the very first presentation of novel tokens (in block 3), indicating an efficient generalization process. There does seem to be a cap on the generalized learning though, just as with the overall learning, as outcomes did not improve block-by-block after that.

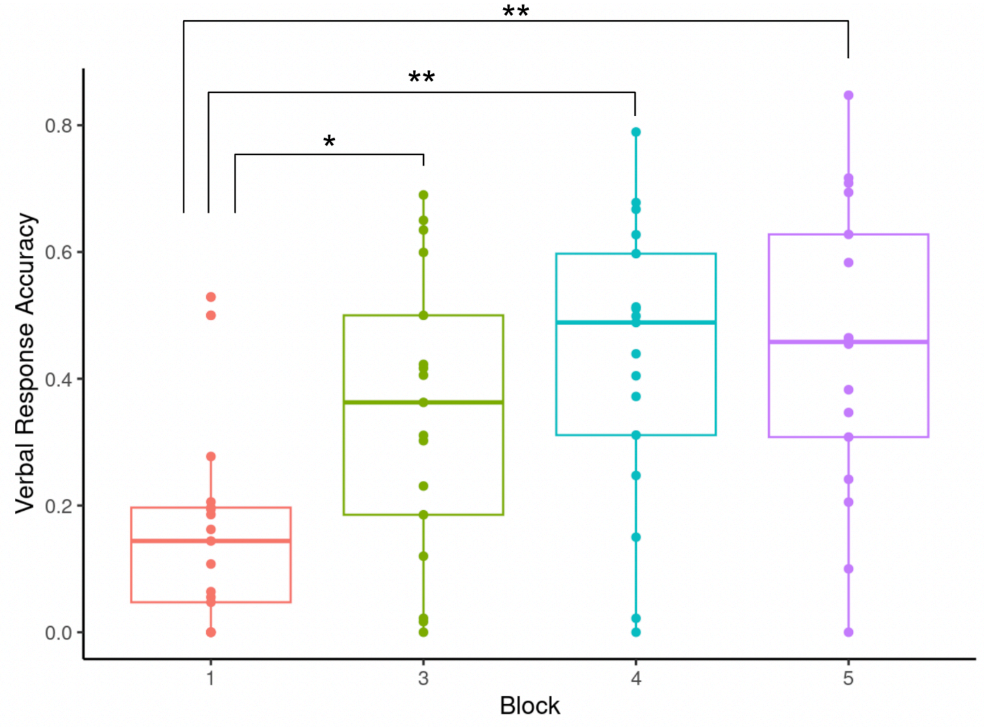


Figure 10: Breaking down generalized learning of standard SWS tokens
Verbal response accuracy in standard SWS tokens over blocks 3-5 where new tokens were introduced, compared to block 1 baseline. **p* < .05, ** *p* < .01, *** *p* < .001, **** *p* <.0001.

Considering Potential Sources of Individual Differences

The final question this experiment sought to address was whether some potential sources of individual differences (musical experience and/or foreign language proficiency) might impact baseline perceptual abilities or learning. This could potentially begin to illuminate what has driven past instances of spontaneous SWS perception, or the absence of perception even after training. I first compared participants' block 1 perceptual outcomes for standard SWS tokens, with their scores on the Gold-MSI (musical experience measure) and BLP (foreign language experience measure), using Pearson's correlation, to examine participants' immediate abilities to perceive SWS as a result of their prior musical and/or foreign language experience. Scores on the Gold-MSI (*M* = 71.71, *SD* =

15.50) were positively correlated with both clarity ratings, *r* = 0.22, *p* < .001, and verbal response accuracy, *r* = 0.27, *p* < .001 (Figure 11a-b). Scores on the BLP (*M* = 249.24, *SD* = 34.65) were also positively correlated with clarity ratings, *r* = 0.28, *p* < .001, and verbal response accuracy, *r* = 0.17, *p* < .01 (Figure 11c-d). These results suggest that musical experience and foreign language experience are both associated with stronger initial perceptual abilities for perceiving SWS.

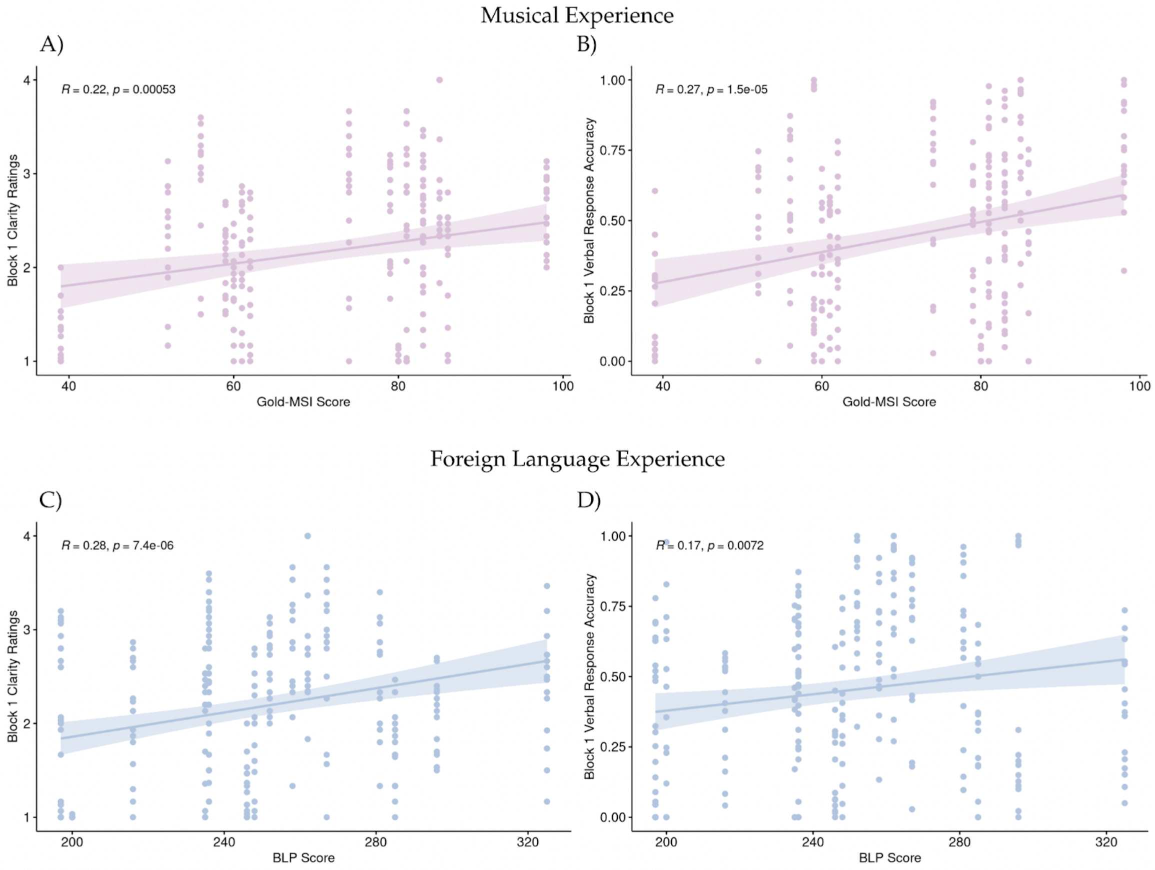


Figure 11: Pearson's correlations to assess musical and foreign language experience on baseline perceptual abilities.
a) Gold-MSI scores plotted against block 1 perceptual clarity ratings, where each vertical line of dots represents a single participants' block 1 trials. *b)* Gold-MSI scores plotted against block 1 verbal response accuracy, where each vertical line of dots represents a single participants' block 1 trials. *c)* BLP scores plotted against block 1 perceptual clarity ratings, where each vertical line of dots represents a single participants' block 1 trials. *d)* BLP scores plotted against block 1 verbal response accuracy, where each vertical line of dots represents a single participants' block 1 trials.

These correlations provided useful information regarding participants' initial perceptual abilities, though it did not assess anything about learning. To test for any differences in learning efficiency, participants' perceptual outcomes in block 1 were subtracted from those for old tokens in block 5 to create a learning difference-score. The same calculation was done with new tokens in block 5 to create a generalization learning-difference score. These scores were then separately evaluated through Pearson's correlations with Gold-MSI and BLP scores. No significant correlations were found, meaning that neither musical experience nor foreign language experience, despite both being associated with higher initial perceptual abilities, had any association with learning or generalization.

Discussion

This study revealed several interesting findings regarding the nature of perceptual learning of SWS. First, I was able to confirm the perceptual jump brought about by hearing the OS sample of a degraded token, which has been seen in many studies before, and is the basis of SWS perceptual learning being thought of as "one-shot." Though, participants still increased their perceptual abilities throughout multiple interactions with the same words, which does not align with previously held assumptions that one-shot learning reflects an instant achievement of maximum learning capacity. Additionally, when participants were given an OS token in the training blocks, they appeared to "learn" flipped-frequency tokens through disambiguation, which should not have been possible given the way they are degraded. This effect went away in block 5 when the OS was no longer presented, more accurately reflecting pure perceptual abilities. In this more pure context, learning was still demonstrated by standard SWS and pseudoword tokens, indicating that participants had sufficiently trained with those words so that they could perceive them even when they were no longer disambiguated. However, only SWS standard tokens demonstrated any generalization to new tokens, indicating some difference in top-down versus bottom-up cues as a part of the generalization process. Finally, no associations were found between either musical experience or foreign language experience when it came to perceptual learning or generalization abilities, though there was a positive relationship between both of those qualities and initial perceptual ability.

Revisiting the one-shot learning effect and traditional SWS paradigm

Sine-wave speech has been long hailed as an exemplar of a one-shot learning effect, where one instance of training produces a robust perceptual learning effect. The specific paradigm used to employ the one-shot learning (presenting the degraded SWS token followed by its original speech counterpart, followed once again by the degraded version) is a popular approach used in studies where manipulating perception without changing the physical stimulus is essential to control for low-level sensory differences, i.e. EEG and fMRI studies. However, the results of this experiment challenge the view that SWS learning is purely "one-shot," and raise questions regarding the validity of the SWS → OS → SWS paradigm.

In this experiment, the perceptual switch from hearing noise to perceiving speech that is brought about by participants hearing an undegraded speech sample was replicated. This perceptual switch, however, might not be as one-shot as previously thought. I found that participants continue to improve their perception of SWS with more accuracy and clarity after multiple instances of disambiguation on the same words. This finding suggests a more gradual learning process, where after one instance of disambiguation there is a steep increase in perception, but this jump does not reach maximum learning capabilities. Given that there is further learning occurring *after* disambiguation, this might contextualize previous findings where individuals who were reported to have experienced one-shot learning experienced this perceptual switch at varying time points (DeHaene-Lambertz et al., 2005; Khoshkhoo et al., 2018; Möttönen et al., 2006; Zhu et al., 2024). This, in turn, might point to a pre-OS learning process and/or to innate perceptual abilities (further discussed in the section, "Musical and/or foreign language experience as potential source of individual differences in initial perception").

Another important result that is relevant for existing views on the traditional SWS paradigm is the finding that participants demonstrated an apparent perceptual switch even for flipped-frequency tokens (albeit to a lesser extent than pseudoword or standard SWS tokens). At a baseline acoustic level, speech cannot possibly be perceived for these flipped frequency control stimuli, let alone learned. Of course, this supposed perceptual switch disappeared for flipped tokens (but not for pseudoword or standard SWS tokens) when participants were no longer given the OS token (i.e., when comparing blocks 1 to

5). Thus, we might consider that the traditional SWS → OS → SWS paradigm allows participants to “cheat” to some extent, by using the OS token they just heard to respond to the second SWS stimulus, even if they did not perceive anything intelligible. This is especially important for studies that rely on stimuli such as the flipped-frequency tokens to provide a reliable control where speech cannot be perceived.

Although the OS confound is undesirable, it ultimately did not get in the way of assessing learning that occurred from the beginning to the end of the experiment. In assessing overall learning, I found that when facilitated by multiple instances of disambiguation over time (i.e. the same words across blocks), participants were able to perceive the same tokens without disambiguation, but only for standard SWS and pseudoword tokens. So despite the fact that post-OS outcomes in blocks 2-4 may not be so trustworthy, when measured without the possibility for “cheating,” perceptual outcomes tell us that flipped-frequency tokens are indeed an adequate control, but that their use may need to be reconceived outside of the typical SWS paradigm.

With all of these findings in mind, to continue using SWS as a tool to drastically alter perception without using different stimuli, future studies might consider developing alternative paradigms. For example, a paradigm that accounts for post-OS learning would be beneficial. If we continue to operate with the assumption that one-shot learning reflects maximum learning capacity, then we may miss out on interesting findings about the full spectrum of perceptual learning of SWS. Such paradigms may also be able to solve the confound associated with flipped-frequency tokens, and find other ways of disambiguating stimuli that do not offer participants the ability to rely on the OS token, leading to more accurate learning information and more reliable control stimuli. This would be particularly beneficial to future studies that hope to employ concurrent brain measures, where being able to use the same stimulus for both non-perception and perception is key to making causal claims about the switch in perception itself, and not any of the confounding factors that the OS may bring about.

Generalization, and its implications for top-down versus bottom-up processing in speech perception

Given questions in the literature about specificity versus generalizability in perceptual learning, a major motivation for this study was to understand whether learning achieved through training could generalize to novel stimuli. The finding that specific tokens can be learned and perceived even when disambiguation is no longer available, is in line with previous research on perceptual learning, where repeated interaction with and exposure to specific stimuli over time produces changes in one's perception of that stimulus. Though, traditionally it has been posited that perceptual learning can only only be applied to trained stimuli. In other words, if participants were only trained to hear a degraded version of "bear" with the presence of the OS, they would not be able to apply that learning to perceive other words. This view of perceptual learning (that will henceforth be referred to as domain-specific learning), thus does not account for or expect any generalization of learning. Yet, the results of the current study demonstrated that, at least for standard SWS words, learning generalized to new words.

Not only did learning generalize, but it generalized seemingly rather efficiently, with perception being higher than baseline at the very first onset of novel stimuli. Though, this was only for standard SWS tokens, while pseudowords only seemed to produce domain-specific learning. Given the small sample size of this study, we may consider potential future replications in interpreting the current results. If generalization for *only* standard SWS tokens holds up, this may point to generalization of learning being completely reliant on stored representations of words gained through lexical knowledge, whereas domain-specific learning could be more related to short-term memory. After intaking low-level acoustic features such as pitch or intensity, top-down processes work to match that information with pre-existing representations of words. In this model, pseudowords do not display generalization because pseudowords do not exist within the participants' personal lexicon, and therefore no matches to stored representations can be made when low-level acoustic features are processed. Though, trained pseudowords can still exhibit domain-specific learning potentially because instead of referencing stored lexical knowledge, short-term memories of their specific makeup of acoustic features are instead referenced. To test if this is the case, future studies could take a longitudinal approach and include some sort of pre-assessment stimulus training of pseudowords. If participants are familiarized with undegraded versions of pseudowords over time, such that those pseudowords then become part of participants' vocabularies, then perhaps learning would generalize to those tokens during an assessment period at the end of the overall study where they function as "new" tokens did in the current study.

Alternatively, if a future replication of this study had more participants and included more pseudoword trials, maybe they would have eventually been able to generalize their learning to novel pseudowords. After all, in this study, there were very few pseudoword trials compared to standard SWS trials, meaning that participants had much less practice with pseudowords. And, when looking at block-by-block comparisons for overall learning, standard SWS tokens showed a significant perceptual increase from baseline earlier in the training process than pseudowords. So, if participants were given further practice with pseudowords, maybe generalization would eventually be demonstrated, just at a slightly lower efficiency than standard SWS words. This could point to low-level acoustic features being sufficient information for both domain-specific and generalized learning to occur, but that access to pre-existing top-down lexical representations facilitates more efficient learning and generalization. Instead of a model where pseudowords can never generalize because they do not have lexical associations, this model would account for the fact that pseudowords do still have phonemic associations, but not associations for those precise combinations of sounds. These phonemic associations could potentially be pieced together through more exposure and lead to generalizable learning, assuming the capacity to piece together matched representations of phonemes into a cohesive prediction exists.

If this model proves to be true in future research, then there is even further inquiry to be done, examining when and how preferences for different stored knowledge are formed, and whether some information contributes to more efficient learning than others. I hypothesize that if lexical knowledge leads to the most efficient learning/generalization but both can occur less efficiently with phonemic knowledge, then there may be some sort of hierarchical preference for larger or broader units of speech when referencing stored knowledge to inform predictions, only relying on the smallest units when stored representations of all greater units are unavailable. This thought is especially driven by the fact that when working with monosyllabic words in isolated contexts, lexical knowledge has the capacity to provide the most immediate "match" for the full set of acoustic features. Moving down the line, syllables provide less information than full words (lexical knowledge) are able to, but more information than phonemes, which are the smallest perceptually distinct units of speech sounds that can provide distinguishable knowledge, and presumably the last reliable resource that would be implemented. Of course, the role of syllabic information would need to be considered in a future study, as all stimuli used for this study contained only one syllable, and therefore syllabic and lexical knowledge operated at the same level.

Whether the findings of this study accurately represent the full learning/generalization processes or future studies reveal eventual generalization of learning for pseudowords, RHT proves to be an excellent framework for assessing auditory processing of speech, despite its most frequent applications to the visual modality. Perhaps domain-specific learning can be accomplished through either accessing stored representations of lexical knowledge or, when those representations are unavailable, turn to short-term memory to supply adequate information to make accurate predictions. When memory does not suffice, predictions cannot be made accurately, and learning cannot be generalized to novel stimuli. Or, perhaps short-term memory is not needed, because referencing stored representations to inform predictions can also happen at a phonemic or syllabic level, with lexical knowledge simply being the most efficient route. In either case, as is posited by RHT, bottom-up sensory processing that happens at the most basic, physical level both informs and is informed by top-down cognitive processes.

Musical and/or foreign language experience as potential source of individual differences in initial perception

Finally, this study attempted to begin explaining some of the potential sources of individual differences that have been seen in perceptual learning and efficiency of that learning. Given previous research that suggested both musical experience and foreign language experience can have tangible effects on one’s auditory skills, I reasoned that these skills might translate to perception of degraded speech in some way. When examining the extent to which participants learned to perceive standard SWS from the beginning to end of the experiment (and generalization), neither musical experience nor foreign language showed any significant associations in either direction. However, when looking at baseline perceptual abilities, participants with higher musical experience and/or foreign language experience displayed stronger perception of SWS from the onset of the experiment. This suggests that although neither musicianship nor multilingualism seem to influence perceptual learning efficiency or ability, the specific auditory skills that many musicians/multilinguals gain through their experience, such as pitch discrimination or pattern distinction, may give them a slight advantage in initial perception of degraded speech.

Of course, there are many other potential sources of differences, both in immediate perception and learning. For example, follow-ups on past research where there was a musician benefit in speech-from-noise detection have demonstrated that musicians also had higher IQ, working memory, and attention (Yoo & Bidelman, 2019). The direction of these associations is unknown, meaning that musicianship could either produce those qualities, or be promoted by already possessing those qualities. Outside the realms of musicianship or multilingualism, some studies have suggested general perceptual learning abilities across sensory modalities, or have found positive associations between certain basic cognitive abilities and personality dispositions in perceptual learning (Dale et al., 2021; Yang et al., 2020). More research is needed to explore potential sources of individual differences in perceptual learning, as musical experience and foreign language experience are just two of many possible factors that could either on their own or in convergence, influence perceptual abilities.

Limitations & future directions

This study was strong in reevaluating preexisting paradigms, assessing domain-specific and generalized learning with connections to broader sensory and perceptual research, and exploring potential sources of individual differences, all in the perceptual learning of SWS. Nevertheless, there were a few limitations to this research.

First, this is an incredibly rich data set with many possibilities for analysis, but because this study required developing a completely new experimental design, time was limited for statistical analyses. The real limitation here is that when assessing learning and generalization at a block-by-block level for the entire experiment, the statistical analyses used may not have adequately captured the most nuanced model of learning possible. More traditional statistical models of learning, such as linear regression or time series analyses, could potentially reveal more detailed information about learning efficiency. For example, maybe "blocks," while useful for the experimental paradigm itself, in their broad and amalgamated state do not accurately reflect when perceptual learning took place. Perhaps considering perceptual abilities trial-by-trial or even splitting blocks into beginning, middle, and end, could be more informative. An additional consideration is that despite all being monosyllabic, different tokens had different numbers of phonemes (represented as IPA symbols to evaluate verbal response accuracy), and so taking into account that words with more phonemes would be more difficult to report accurately would be an interesting avenue for analysis.

Additionally, the use of isolated words presented without context and in randomized order may not provide the best approximation of naturalistic perception of ambiguous speech. In everyday contexts when speech is employed, words rarely exist in isolation and instead are embedded in greater semantic contexts, like sentences or paragraphs at the structural level, or narratives and arguments at the conceptual and logical level. These contexts draw from a whole new host of top-down processes, and create expectations that could potentially assist in disambiguation. Contextual expectations can simultaneously rule out certain possibilities and provide specific direction for predictions (Ex. Many people would probably have a strong guess as to what is missing from the question, “Why did the chicken cross the __?”).

Despite these limitations, this study raises some exciting questions to be addressed in future research. One aspect of perceptual learning that has not been thoroughly investigated yet is the role of memory. While discussed with relation to aiding in perception of trained stimuli in the absence of lexical knowledge for pseudowords, memory could be important in other ways, and is interesting with regards to the degree to which perceptual learning holds up over time. Many studies have only employed short-term experiments, meaning that only short-term learning can be evaluated. Future researchers might consider a more longitudinal approach to see if perceptual skills that are learned in experimental settings can be maintained over longer periods of time, and what factors might lead to retention or attrition of those skills. Language is another potentially fruitful avenue. While the Bilingual Language Profile was ideal for this study and gaining a surface-level view of foreign language experience, does not specify breadth versus depth. In other words, a high score on the BLP could mean complete fluency in one non-native language, or surface-level knowledge of many. It is possible that there are differences between knowing two languages really well and knowing many languages very little when it comes to perception of ambiguous speech. Or, perhaps speakers of one language are better at perceiving SWS than speakers of another. Information like this could have important implications for understanding language acquisition and its influence on overall speech perception. Finally, to continue picking apart the relationship between top-down and bottom processing, perhaps a new type of stimulus could be designed with the goal of maintaining shared semantic or other top-down qualities of speech, but without any shared acoustic properties– essentially opposite to pseudowords. A stimulus such as this could reveal information about how top-down processes in greater states of isolation color novel sensory materials (specifically that of low-level acoustic qualities), as well as what information at the acoustic level is minimally sufficient for perception of speech.

With more perspective on the relationship between top-down and bottom-up processing and its role in specified and generalized learning, we now have more information that can be applied to understanding how people derive meaning from ambiguous environmental stimuli. Next time you can't quite make out the exact string of words someone just spoke to you, take note of how you try to fill in the gaps. First, you may try to rely on your memory– maybe you have heard this exact person say something in that cadence before and it triggers a memory of what those words were. Or maybe, you are in a completely new setting talking to a person you've never met before, and the context clues of your conversation allow you to make an educated guess that they are praising their favorite author. In either case, your brain will take in the ambiguous sensory material and turn it into something that you can understand, by applying different, pre-existing cognitive lenses to inform your interpretation. And maybe in the process, you learn something new from the conversation, which will be stored away and used in the future to fill another gap. This sensory-cognitive feedback loop is the makeup of our experience of the world around us, which is why this research and hopefully many studies in the future will continue to try and expand our knowledge of its intricacies.

Appendix A: List of English Tokens

This list includes the 81 real English words included in the experiment. These same words were first generated as original speech tokens, then degraded to standard SWS tokens, and finally turned into flipped-frequency tokens.

1. Ant
2. Badge
3. Bat
4. Bear
5. Bed
6. Bee
7. Bird
8. Blimp
9. Boat
10. Book
11. Bowl
12. Box
13. Brush
14. Cake
15. Can
16. Car
17. Card
18. Cat
19. Chair
20. Clip
21. Coin
22. Conch
23. Cow
24. Crab
25. Croc
26. Cup
27. Deer
28. Desk
29. Dice
30. Dog
31. Door
32. Dove
33. Duck
34. Fan
35. Fish
36. Flag
37. Flea
38. Floor
39. Fly
40. Fork
41. Fox
42. Frog
43. Goat
44. Goose
45. Harp
46. Hat
47. Hen
48. Horse
49. Key
50. Knife
51. Lamp
52. Leech
53. Mask
54. Mole
55. Moose
56. Pen
57. Phone
58. Pig
59. Plate
60. Rat
61. Roach
62. Shark
63. Sheep
64. Shelf
65. Shirt
66. Shoe
67. Shrimp
68. Sink
69. Skunk
70. Snake
71. Soap
72. Sock
73. Squid
74. Swan
75. Sword
76. Toad
77. Wall
78. Wasp
79. Watch
80. Wolf
81. Worm

Appendix B: List of Pseudoword Tokens

This list includes the 33 pseudowords. These same words were first generated as original speech tokens, then degraded to SWS tokens.

1. Bape
2. Brips
3. Cuch
4. Femps
5. Flirs
6. Foof
7. Frip
8. Glat
9. Greds
10. Gurnt
11. Gurt
12. Lall
13. Leck
14. Leff
15. Losh
16. Lusk
17. Mups
18. Onks
19. Plame
20. Rast
21. Shis
22. Sonk
23. Spuch
24. Spunt
25. Stams
26. Stoid
27. Stowl
28. Theal
29. Trid
30. Wearn
31. Wilp
32. Woins
33. Woog

Appendix C: Goldsmiths Musical Sophistication Index

To assess musical experience, the General Musical Sophistication subscale of the Goldsmiths Musical Sophistication Index was used. Participants responded to each item on a scale of 1-7, where 1 is completely disagree, and 7 is completely agree. Some items were reverse-scored.

1. I spend a lot of my free time doing music-related activities.
2. I enjoy writing about music, for example on blogs and forums.
3. If somebody starts singing a song I don't know, I can usually join in.
4. I can sing or play music from memory.
5. I am able to hit the right notes when I sing along with a recording.
6. I can compare and discuss differences between two performances or versions of the same piece of music.
7. I have never been complemented for my talents as a musical performer. (reverse-scored)
8. I often read or search the internet for things related to music.
9. I am not able to sing in harmony when somebody is singing a familiar tune. (reverse-scored)
10. I am able to identify what is special about a given musical piece.
11. When I sing, I have no idea whether I'm in tune or not. (reverse-scored)
12. Music is kind of an addiction for me—I couldn't live without it.
13. I don't like singing in public because I'm afraid that I would sing the wrong notes. (reverse-scored)
14. I would not consider myself a musician. (reverse-scored)
15. After hearing a new song two or three times, I can usually sing it by myself.

Appendix D: Bilingual Language Profile

To assess foreign language experience, the Bilingual Language Profile was used. The full questionnaire has 4 sub-sections, which are outlined below including sample items. The highest score one can earn when only reporting one language is 218 points. Participants were allowed to report up to 5 languages, leading to a highest-possible total of 1,090 points, which would indicate full fluency from a very young age, so is an extremely unlikely score. In the "Limitations & Future Directions" section of the discussion, I discuss some limitations of using this scale.

Language History: 6 questions, each worth between 0 and 20 (120 possible) for each language

Ex. "At what age did you start learning X language?" or "How many years have you spent in a family where X language was spoken?"

Language Use: 5 questions, each worth between 0 and 10 (50 possible) for each language

Ex. "In an average week, what percentage of the time do you use X language with friends?" or "When you count, how often do you count in X language?"

Language Proficiency: 4 questions, each worth between 0 and 6 (24 possible) for each language

Ex. "How well do you speak X language?" or "How well do you understand X language?"

Language Attitudes: 4 questions, each work between 0 and 6 (24 possible) for each language

Ex. "I feel like myself when I speak X language."

Bibliography

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1515), 285–299. https://doi.org/10.1098/rstb.2008.0253

Aman, L., Picken, S., Andreou, L.-V., & Chait, M. (2021). Sensitivity to temporal structure facilitates perceptual analysis of complex auditory scenes. *Hearing Research*, *400*, 108111. https://doi.org/10.1016/j.heares.2020.108111

Bhandari, P., Demberg, V., & Kray, J. (2022). Predictability effects in degraded speech comprehension are reduced as a function of attention. *Language and Cognition*, *14*(4), 534–551. https://doi.org/10.1017/langcog.2022.16

Birdsong, D., Gertken, L.M., & Amengual, M. (2012). Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism. *COERLL, University of Texas at Austin.* https://sites.la.utexas.edu/bilingual/

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

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *The Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Bsharat-Maalouf, D., & Karawani, H. (2022). Bilinguals' speech perception in noise: Perceptual and neural associations. *PLOS ONE*, *17*(2), e0264282. https://doi.org/10.1371/journal.pone.0264282

Calcus, A. (2024). Development of auditory scene analysis: A mini-review. *Frontiers in Human Neuroscience*, *18*, 1352247. https://doi.org/10.3389/fnhum.2024.1352247

Christison-Lagay, K. L., Gifford, A. M., & Cohen, Y. E. (2015). Neural correlates of auditory scene analysis and perception. *International Journal of Psychophysiology*, *95*(2), 238–245. https://doi.org/10.1016/j.ijpsycho.2014.03.004

Dale, G., Cochrane, A., & Green, C. S. (2021). Individual difference predictors of learning and generalization in perceptual learning. *Attention, Perception, & Psychophysics*, *83*(5), 2241–2255. https://doi.org/10.3758/s13414-021-02268-3

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21–33. https://doi.org/10.1016/j.neuroimage.2004.09.039

Deutsch, D., Henthorn, T., Marvin, E., & Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *The Journal of the Acoustical Society of America*, *119*(2), 719–722. https://doi.org/10.1121/1.2151799

Dykstra, A. R., Cariani, P. A., & Gutschalk, A. (2017). A roadmap for the study of conscious audition and its neural basis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160103. https://doi.org/10.1098/rstb.2016.0103

Fahle, M., & Poggio, T. A. (Eds.). (2002). *Perceptual Learning*. The MIT Press. https://doi.org/10.7551/mitpress/5295.001.0001

François, C., & Schön, D. (2014). Neural sensitivity to statistical regularities as a fundamental biological process that underlies auditory learning: The role of musical practice. *Hearing Research*, *308*, 122–128. https://doi.org/10.1016/j.heares.2013.08.018

Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., & Pantev, C. (2004). Musical Training Enhances Automatic Encoding of Melodic Contour and Interval Structure. *Journal of Cognitive Neuroscience*, *16*(6), 1010–1021. https://doi.org/10.1162/0898929041502706

Gohari, N., Hosseini Dastgerdi, Z., Bernstein, L. J., & Alain, C. (2022). Neural correlates of concurrent sound perception: A review and guidelines for future research. *Brain and Cognition*, *163*, 105914. https://doi.org/10.1016/j.bandc.2022.105914

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, *1153*, 134–143. https://doi.org/10.1016/j.brainres.2007.03.069

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283–295. https://doi.org/10.1037/a0020772

Hochstein, S., & Ahissar, M. (2002). View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron,* 36(5), 791-804. https://doi.org/10.1016/S0896-6273(02)01091-7

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, *72*(5), 1218–1227. https://doi.org/10.3758/APP.72.5.1218

Jones, C., Collin, E., Kepinska, O., Hancock, R., Caballero, J., Zekelman, L., Vandermosten, M., & Hoeft, F. (2021). Auditory Processing of Non-speech Stimuli by Children in Dual-Language Immersion Programs. *Frontiers in Psychology*, *12*, 687651. https://doi.org/10.3389/fpsyg.2021.687651

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*, *187*, 83–91. https://doi.org/10.1016/j.bandl.2018.01.007

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605. https://doi.org/10.1038/nrn2882

Kraus, N., Skoe, E., Parbery-Clark, A., & Ashley, R. (2009). Experience-induced Malleability in Neural Encoding of *Pitch* , *Timbre* , and *Timing*: Implications for Language and Music. *Annals of the New York Academy of Sciences*, *1169*(1), 543–557. https://doi.org/10.1111/j.1749-6632.2009.04549.x

Li, W., & Gilbert, C. D. (n.d.). *Perceptual Learning: Neural Mechanisms*.

Magne, C., Schon, D., & Besson, M. (n.d.). *Musician Children Detect Pitch Violations in Both Music and Language Better than Nonmusician Children: Behavioral and Electrophysiological Approaches*. *18*(2).

Miendlarzewska, E. A., & Trost, W. J. (2014). How musical training affects cognitive development: Rhythm, reward and other modulating variables. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00279

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2009). Musical Training Influences Linguistic Abilities in 8-Year-Old Children: More Evidence for Brain Plasticity. *Cerebral Cortex*, *19*(3), 712–723. https://doi.org/10.1093/cercor/bhn120

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage*, *30*(2), 563–569. https://doi.org/10.1016/j.neuroimage.2005.10.002

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences*, *104*(40), 15894–15898. https://doi.org/10.1073/pnas.0701498104

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-Level Information and High-Level Perception: The Case of Speech in Noise. *PLoS Biology*, *6*(5), e126. https://doi.org/10.1371/journal.pbio.0060126

New, B., Bourgin, J., Barra, J., & Pallier, C. (2024). UniPseudo: A universal pseudoword generator. *Quarterly Journal of Experimental Psychology (2006)*, *77*(2), 278–286. https://doi.org/10.1177/17470218231164373

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical Experience Limits the Degradative Effects of Background Noise on the Neural Processing of Sound. *The Journal of Neuroscience*, *29*(45), 14100–14107. https://doi.org/10.1523/JNEUROSCI.3256-09.2009

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00142

Patterson, R. D., & Johnsrude, I. S. (2008). Functional imaging of the auditory processing applied to speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1023–1035. https://doi.org/10.1098/rstb.2007.2157

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception Without Traditional Speech Cues. *Science*, *212*(4497), 947–950. https://doi.org/10.1126/science.7233191

Rosen, S., Wise, R. J. S., Chadha, S., Conway, E.-J., & Scott, S. K. (2011). Hemispheric Asymmetries in Speech Perception: Sense, Nonsense and Modulations. *PLoS ONE*, *6*(9), e24672. https://doi.org/10.1371/journal.pone.0024672

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*(1), 1–10. https://doi.org/10.1016/j.actpsy.2011.03.007

Sagi, D. (2011). Perceptual learning in Vision Research. *Vision Research*, *51*(13), 1552–1566. https://doi.org/10.1016/j.visres.2010.10.019

Shamma, S. (2008). On the Emergence and Awareness of Auditory Objects. *PLoS Biology*, *6*(6), e155. https://doi.org/10.1371/journal.pbio.0060155

Shokuhifar, G., Javanbakht, M., Vahedi, M., Mehrkian, S., & Aghadoost, A. (2024). The relationship between speech in noise perception and auditory working memory capacity in monolingual and bilingual adults. *International Journal of Audiology*, 1–8. https://doi.org/10.1080/14992027.2024.2328556

Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, Awareness, and the Perception of Auditory Scenes. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00015

Snyder, J. S., Yerkes, B. D., & Pitts, M. A. (2015). Testing domain-general theories of perceptual awareness with auditory brain responses. *Trends in Cognitive Sciences*, *19*(6), 295–297. https://doi.org/10.1016/j.tics.2015.04.002

Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience and neural efficiency – effects of training on subcortical processing of vocal expressions of emotion. *European Journal of Neuroscience*, *29*(3), 661–668. https://doi.org/10.1111/j.1460-9568.2009.06617.x

Tartaglia, E. M., Aberg, K. C., & Herzog, M. H. (2009). Perceptual learning and roving: Stimulus types and overlapping neural populations. *Vision Research*, *49*(11), 1420–1427. https://doi.org/10.1016/j.visres.2009.02.013

Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W., & Patterson, R. D. (2006). Locating the initial stages of speech–sound processing in human temporal cortex. *NeuroImage*, *31*(3), 1284–1296. https://doi.org/10.1016/j.neuroimage.2006.01.004

Viswanathan, V., Shinn-Cunningham, B. G., & Heinz, M. G. (2022). Speech Categorization Reveals the Role of Early-Stage Temporal-Coherence Processing in Auditory Scene Analysis. *The Journal of Neuroscience*, *42*(2), 240–254. https://doi.org/10.1523/JNEUROSCI.1610-21.2021

Weinberger, N. M. (1995). Dynamic Regulation of Receptive Fields and Maps in the Adult Sensory Cortex. *Annual Review of Neuroscience*, *18*(1), 129–158. https://doi.org/10.1146/annurev.ne.18.030195.001021

Whitehead, P. S., Pfeuffer, C. U., & Egner, T. (2020). Memories of control: One-shot episodic learning of item-specific stimulus-control associations. *Cognition*, *199*, 104220. https://doi.org/10.1016/j.cognition.2020.104220

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420–422. https://doi.org/10.1038/nn1872

Xiong, Y.-Z., Zhang, J.-Y., & Yu, C. (2016). Bottom-up and top-down influences at untrained conditions determine perceptual learning specificity and transfer. *eLife*, *5*, e14614. https://doi.org/10.7554/eLife.14614

Yang, J., Yan, F.-F., Chen, L., Xi, J., Fan, S., Zhang, P., Lu, Z.-L., & Huang, C.-B. (2020). General learning ability in perceptual learning. *Proceedings of the*

National Academy of Sciences, *117*(32), 19092–19100. https://doi.org/10.1073/pnas.2002903117

Yoo, J., & Bidelman, G. M. (2019). Linguistic, perceptual, and cognitive factors underlying musicians' benefits in noise-degraded speech perception. *Hearing Research*, *377*, 189–195. https://doi.org/10.1016/j.heares.2019.03.021

Zhang, J.-Y., Kuai, S.-G., Xiao, L.-Q., Klein, S. A., Levi, D. M., & Yu, C. (2008). Stimulus Coding Rules for Perceptual Learning. *PLoS Biology*, *6*(8), e197. https://doi.org/10.1371/journal.pbio.0060197

Zhu, Y., Li, C., Hendry, C., Glass, J., Canseco-Gonzalez, E., Pitts, M. A., & Dykstra, A. R. (2024). Isolating neural signatures of conscious speech perception with a no-report sine-wave speech paradigm. *The Journal of Neuroscience*, e0145232023. https://doi.org/10.1523/JNEUROSCI.0145-23.2023