PHONETIC CONVERGENCE AND AUDITORY IMAGERY IN READING

By

Josue E. Rodriguez-Zamora

A Thesis Presented to

The Faculty of Humboldt State University

In Partial Fulfillment of the Requirements for the Degree

Master of Arts in Psychology: Academic Research

Committee Membership

Dr. Kauyumari Sanchez, Committee Chair

Dr. Christopher Aberson, Committee Member

Dr. Amber Gaffney, Committee Member

Dr. Christopher Aberson, Graduate Coordinator

May 2019

Abstract

PHONETIC CONVERGENCE AND AUDITORY IMAGERY IN READING

Josue E. Rodriguez-Zamora

This study aimed to address whether phonetic convergence (speech imitation) and auditory imagery in reading are fundamentally governed by the same process — episodic encoding (c.f., Goldinger, 1998). A set of participants (*talkers*; N = 12) were recorded speaking sentences at a baseline level. Talkers were then exposed model speaker with either a *fast* or *slow* speech rate and then engaged in a reading phase where they read sentences thought to be written by that speaker. If episodic encoding predicts effects of phonetic convergence and auditory imagery in reading style, then talkers should be influenced by a speaker on three dimensions: pronunciation of words, duration of words, and duration of sentences. A different set of participants (*raters*; N = 68) engaged in an AXB perceptual similarity ratings task. *Raters* were presented with three sets of recordings of individual target words in a row — A (baseline), X (model), and B (reading) — and made perceptual similarity ratings, indicating whether A or B is more similar in pronunciation to X. If episodic encoding predicts effects of phonetic convergence then *talkers* should be rated as being perceptually similar to the speaker. The results of the study suggest that episodic may not play a role in either phonetic convergence or auditory imagery and speech.

Acknowledgements

To my father and mother: thank you for sacrificing yourselves so that I could be here today. I will never be able to adequately thank you.

To my sister: thank you for keeping me grounded.

A kind thank you to Dr. Mari Sanchez, who guided me through the painful process of learning how to be an academic and who set me up in a position to succeed.

A special gratitude to Dr. Chris Aberson, who introduced me to R and sparked my passion for statistics. Things I can no longer live without.

A special gratitude to Dr. Amber Gaffney, who showed me how to be headstrong in academia, a skill I highly value and admire.

A special mention to Ben Chu and Olivia Kuljian, I may have never gotten through my first year if not for you.

Table of Contents

List of Tables

Phonetic Convergence and Auditory Imagery in Reading

Imagine that you are conversing with someone familiar, such as your mother or your best friend. You may find that your speech might start to be affected. Specifically, you may find that during the conversation, and potentially shortly thereafter, that you start to say certain words more similarly to how your mother or best friend would say them. This spontaneous change in speech production indicates that we are influenced by the unique way a particular talker speaks. This phenomenon is referred to as phonetic convergence — when the sounds (phonemes) of your speech start to move toward (converge) the pronunciation style of another.

Now, imagine you receive a written message (e.g. text) from the same familiar individual that reads “I found parking. I will see you soon!” Whose voice do you “hear” as you read it? Chances are that you “hear” the voice of the author of the message as you read it. This phenomenon is called auditory imagery, which is the ability to mentally simulate sound. With the ability to experience auditory imagery of voices and our tendency to imitate the way others speak, one has to wonder whether these two phenomena overlap. More specifically, do we phonetically converge to voices when we evoke auditory imagery, such as when simply reading aloud text written by familiar people? The aim of this thesis is to identify the relationship between phonetic convergence and auditory imagery and whether they are governed by the same processes.

Phonetic Convergence

Our ability to perceive and imitate spoken language may stem from the fact that humans are extraordinarily good at identifying and understanding spoken language. Listeners of speech can distinguish various aspects of speakers such as dialect, status, and health (Labov, 1972), emotional state (Frick, 1985; Murray & Arnott, 1993), and talker identity (Van Lancker, Kreiman & Emmory, 1985; Van Lancker, Kreiman, & Wickens, 1985). These aspects of speech are unique to each person and are referred to as talker-specific characteristics (TSCs). For example, you may notice a person speaks with a certain accent, speaks at a certain rate, or has a certain pitch. These aspects of the person's speech constitute components of that person's talker-specific characteristics. Our tendency to unconsciously identify and imitate a person's TSCs is referred to as phonetic convergence.

Phonetic convergence is explained by two main theories: the direct realist theory (see Fowler, 1986; Sancier & Fowler, 1997; Fowler, Brown, Sabadini, & Weihing, 2003; Galantucci, Fowler, & Turvey, 2006) and the episodic theory (see Goldinger, 1998; Goldinger & Azuma, 2004; Namy, Nygaard, & Sauerteig, 2002). Whereas the direct realist approach relies on the perception-production link, the episodic approach relies on stored memories of talker-specific characteristics.

Direct realist theory. The direct realist approach is a gestural theory. It contends that speech perception does not occur through the auditory signals of speech, but rather through perception of articulatory gestures (or the kinematics involved in creating the sound) which "causally and distinctively structure the acoustic speech signal" (Fowler et al., 2003). In turn, the gestures of the speaker provide information about the acoustic speech signals, such as how the word was physically produced (e.g. kinematics of the gestures) and similarly, how to produce it oneself. Consider a listener who hears a particular phoneme, which is the smallest unit of sound in speech. Phonemes can be used to distinguishes one word from another (e.g. the /i/ in "beet" and the /a/ in "bat"). Now if a person hears the vowel /i/ in the word "beet", the listener would automatically have access to the gestures that are involved in producing the /i/, such as the high-forward positioning of the tongue (e.g. the vowel /i/ in "beet"), the specific speech rate, and the tone in which it was produced. The gestures responsible for producing the speech sound would carry information about how the word was produced (i.e. the position of the articulators, such as the lips and tongue, etc.). Under direct realist theory, phonetic convergence occurs as a result of perceiving an interlocutor's articulatory gestures which then inform how the listener produces her articulatory gestures.

There is evidence to suggest that gestural information is present not only in auditory speech, but that the same gestural information is also present in visual speech. As such, research has found that people will converge to the unique style of a speaker when presented with only the speaker's articulating face (Miller, Sanchez, & Rosenblum, 2010; Sanchez, Miller, & Rosenblum, 2010). In these studies, participants are asked to

lip-read the speech from a talker and to say the words uttered by the silent talker. When participants say the silent talker's words out-loud, the participants' own utterances shift in line with the silent talker's talker-specific characteristics. This means that the articulating face contains the same gestural information on how to produce speech. As a consequence, a perceiver is influenced in their own speech productions by the style of the silent talker, resulting in phonetic convergence.

It has additionally been shown that listeners quickly access and extract information from an interlocutor's articulatory gestures (Fowler et al., 2003). Gestural information extracted from speech can be used to identify a talker, which means that individuals can match a voice to a speaking face when auditory and gestural stimuli are presented separately, but cannot match a voice to a static face (Lachs & Pisoni, 2004a, 2004b). In effect, an individual can extract information from a speaker's articulatory gestures and use that information to articulate sounds in a similar fashion (e.g. converge), when the individual has visual access to that speaker. An individual cannot extract gestural information from a person to which they do not have access to kinematic information, be it in an auditory or visual form. In this way, perception of the gestures of speech lead to phonetic convergence within the direct realist theory.

The perception-production link. A mechanism that is thought to be responsible for speech convergence in the direct realist approach is the perception-production link. It has been suggested that perception and production share the same mechanisms, at least partially (e.g. Chartrand & Bargh, 1999; Chen, Chartrand, Lee Chai, & Bargh, 1998; Prinz, 1990). This idea suggests that incoming signals to the brain as a result of perceived stimuli do not need to be "translated" into outgoing signals to produce a behavior in response to the stimuli. Rather, perception and behavior share a "common code" in which perception automatically influences action. The perception-production link sets up a framework for imitation in which the perception of stimuli directly and automatically influences how behavior is produced. Thus, if one is exposed to the gestures of speech, one is likely to be influenced in their own gestural realizations, in essence, phonetic convergence.

The perception-production link: Neurological basis. There is neurophysiological evidence for the perception-production link, which was first discovered in primates (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). Di Pellegrino et al. found that neurons located in the F5 region of the premotor cortex — which are responsible for goal-directed hand movements — were similarly activated when: (a) conducting goal directed movements, (b) observing another monkey conduct the same movements, and (c) observing a human experimenter conduct the same movements. Specifically, the same regions of the monkeys' brains activated when they grasped food or observed another monkey or human grasp food. The activation of neurons in response to observed action suggest that perception of a behavior automatically activates neurons needed to produce that behavior. These neurons are known as mirror neurons because they activate in response to the observation of behavior (Rizzolatti, Fogassi, & Gallese, 2001) and their discovery provide neural confirmation for the existence of the perception-production link (Lotto, Hickok, & Holt, 2009). Importantly, mirror neurons and the perception-production link exist in humans (Fadiga, Fogassi, & Rizzolati, 1995).

The existence of a perception-production link in humans holds implications for speech imitation. Specifically through a class of mirror neurons known as echo neurons, which respond both to the execution of an action and its resulting sound (e.g. the action and sound of opening a can; Kohler et al., 2002). For example, Fadiga, Craighero, Buccino, & Rizzolatti (2002) examined how the perception-production link affects the excitability of tongue muscles. They monitored activity of tongue muscles while participants passively listened to human speech. The results suggested that passive

listening of speech was enough to excite muscles in the tongue of a human listener. That is, the perception of auditory signals activated the muscles in the tongue needed to produce the heard sound productions.

Gentilucci and Bernardis (2005) examined whether speech imitation was observable at a phonemic level when speaking with an interlocutor. In this experiment, participants were measured on their lip movements and their pitch. Participants were asked to repeat strings of phonemes (i.e. /aba/) spoken by model speakers, that were presented visually and auditorily. The results indicated that both the participants' lip movements and their pitch shifted toward the model speaker's realizations. This suggests that a person shifts the way they produce sound, both in frequency and physical articulation of the lips, when repeating sounds. These studies lend neurophysiological support for phonetic convergence within a direct realist perspective.

Most studies that use a direct realist perspective are agnostic to the role of memory involved in speech convergence. Direct realist studies often employ a shadowing task, where participants quickly repeat utterances spoken by a model talker. In this task the perception-production link is viable, as participants are directly exposed to a model talker's gestures and are required to immediately repeat the models' speech. However, the current study shall not employ a shadowing task, but instead rely on stored memory traces to inform the speech productions of participants. As such, a theory that utilizes a memory component is warranted for explanation: the episodic theory.

Episodic theory. Under the episodic theory of speech perception and production, our ability to store memories of others' TSCs is what allows us to engage in phonetic convergence. The episodic theory of speech convergence posits that each word we perceive leaves a trace in memory which contains information regarding the episode (including TSCs of the speaker) and therefore influences the mental representation of that word (Goldinger, 1998). Each time we produce that word, we activate its associated traces, and access the information stored in those traces. Consequently, the activation of these traces influence the way a word is produced.

When we activate stored memory traces, the number of traces we have affect what we will subsequently speak. To illustrate this phenomenon, consider two examples at the phrase level from the Star Wars movies: "May the Force be with you" and "Help me Obi-Wan Kenobi. You're my only hope". Chances are you do not attribute the former phrase to any single character, as many Star Wars characters have uttered that phrase. However, you are likely to attribute the latter phrase only to Princess Leia, as she is the only character to have uttered that phrase. Under the episodic theory of speech, each time you hear Princess Leia say her classic phrase you store an episodic trace which is associated with both her TSCs and that particular phrase. This means that if you were to say "Help me Obi-Wan Kenobi. You're my only hope." yourself, the largest portion of episodic traces that activate contain Princess Leia's TSCs. You are then more likely to be influenced in the way you produce this phrase by the way Princess Leia produced it (e.g. you are more likely to converge to Princess Leia's voice). On the other hand, many characters have said "May the Force be with you". Episodic traces are stored every time

you hear this phrase and each trace is associated with the TSCs of the character who said it. Therefore, no single character has a dominant amount of episodic traces associated with their TSCs when saying this phrase. Consequently, when you say “May the Force be with you”, you are not likely to be influenced by the TSCs of any given character and are less likely to converge to any one character. This phenomenon may extend even further such that we may converge to the overall speaking style of any one person (e.g. Princess Leia) in an everyday setting given enough exposure (see Sancier & Fowler, 2003; Sanchez, Hay, & Nilson, 2015).

Neither the episodic theory nor the direct realist theory provides a holistic explanation of phonetic convergence. However, the episodic theory better accounts for known effects such as word frequency effects, repetition effects, and persistent convergence (Fowler, 1986; Gambi & Pickering, 2013; Goldinger, 1998). For the purposes of this paper, an episodic approach will be emphasized at the word level.

Word frequency effects refer to how the ubiquity of a word in everyday language affects phonetic convergence. For example, words which are not commonly used in English (e.g. “portal”) are considered low-frequency words and words that are common in English (e.g. “hello”) are considered high-frequency words. Low-frequency words have less traces associated with them due to not being encountered frequently. This makes low-frequency words more susceptible to be influenced by traces associated with a specific talker and their TSCs. Alternatively, high-frequency words have many traces associated with them and are less likely to be influenced by the traces of any single speaker. The episodic theory predicts higher rates of convergence when a person is

exposed to a speaker using low-frequency words than when the speaker uses high-frequency words.

Repetition effects refer to the fact that repeatedly hearing a word affects convergence within a speech experiment. If you repeatedly hear a specific person say "portal" many times, and a trace is stored for each repetition, then the amount of traces associated with this person's TSCs increases. Thus, when you produce the word "portal" you are more likely to be influenced by that specific talker in the way you say it. Conversely, if "portal" is not repeatedly spoken by a particular person, then stored traces of that word are not associated with any particular individual's TSCs and you are less likely to be influenced in your production of that word. The episodic theory predicts higher rates of convergence as a person hears a word repeated more often from a speaker. Repetition effects differ from frequency effects in that they refer to how the amount of times a word is heard affects phonetic convergence, while frequency effects have to do with how the prevalence of a word in everyday language affects convergence. Repetition effects can interact with frequency effects so that convergence has the highest chance of occurring when the word being uttered is both a low-frequency word and has been heard many times by a speaker, and less so when a word is high-frequency and has been heard few times.

The episodic theory also provides an explanation for why persistent convergence occurs. Persistent convergence is when an individual continues to converge to a speaker despite the passage of time (Pardo, 2006), and has been observed up to six days after exposure to a model speaker (Goldinger & Azuma, 2004). Under the episodic theory, this

occurs because stored traces associated with a talker's TSCs might persist in their influence on phonetic convergence over time. Episodic theory cannot wholly account for persistent convergence, but can account for it better than the direct realist theory (see Fowler, 1986; Gambi & Pickering, 2013). This is because the direct realist theory requires direct interaction with a speaker in order for phonetic convergence to occur.

Word frequency and repetition are variables that are often considered in speech convergence studies, but not many investigate the persistence of convergence. However, Goldinger and Azuma's (2004) study combined frequency and repetition effects as they persist over time. In the study, Goldinger and Azuma investigated different types of word frequency categories: high (HF), medium high (MHF), medium low (MLF), and low (LF). Words from these different frequency categories were presented with different amounts of repetitions (e.g. 0, 2, 6, 12). Under the episodic theory, both word frequency and repetitions are important, as the episodic theory predicts that words with lower frequencies that have many repetitions will lead to higher rates of convergence. In the test of persistence over time, it is these factors that are also predicted to demonstrate the most convergence.

The Goldinger and Azuma (2004) experiment unfolded over three phases. Participants first engaged in a baseline phase, where they were recorded reading all stimulus words. The baseline recordings serve to provide a natural example of the speaking style of each participant. Most studies, including the current study, use baselines because they provide a reference point to compare participant's different utterances and test whether phonetic convergence has occurred.

In the second phase of the Goldinger and Azuma (2004) study, participants engaged in a listening task. In the listening task, participants were exposed to the auditory speech of a model speaker. These words were the same ones the participant read aloud during the baseline phase and were presented at different repetition rates, where the words were presented 0, 2, 6, or 12 times within the experiment. Again, repetition is important because the episodic theory predicts that rates of convergence increase as a word is increasingly repeated. At no time were participants told to try to remember the words or how they were said by the model speaker. In the listening phase, participants were asked to identify each word they heard by clicking on the text version of the word on their computer, which was displayed on the computer screen in a grid that contained all words they would hear.

The third phase of the study occurred one week after engaging in the listening task of phase two. This one-week delay aimed to test if stored traces continue to influence phonetic convergence despite the passage of time (i.e. persistence). The third phase was identical to the baseline phase: participants were recorded reading all words again. However, given the exposure to the speech of the model in phase two, these participants were expected to utter the words in this last phase in a manner similar to the model, or in other words to converge. According to the episodic theory, participants should have stored traces associated with the words and TSCs of the model speakers heard in the second phase. The words that were LF and repeated the most were expected to show the most evidence of convergence.

To identify whether the participants converged to the model speakers they listened to one week prior (in phase two), a perceptual rating task was employed (see next section "Measuring phonetic converge" for specifics). In the rating task, a separate group of participants listened to and rated the perceptual similarity of sets of recordings. Specifically, participants were asked to rate the similarity of the recorded subjects' baseline utterances and utterances post-exposure to the model's utterances. The results of the perceptual rating task suggest that word frequency and repetition affect speech convergence even after a delay of one week, suggesting the persistence of speech convergence. Convergence was measured to be highest for LF words, followed by MLF, MHF, and HF words (in decreasing order of convergence). This result is predicted by episodic theory because LF words are more likely to be influenced by stored traces of a word and therefore have higher rates of convergence. Repetition affected speech convergence. Convergence was highest for words repeated 12 times, followed by 6, 2, and 0 times (in order of decreasing convergence). This result is predicted by episodic theory because convergence occurs at a higher rate when a word is repeatedly heard. These results lend support to the notion that LF words are more likely to be influenced by traces of a speaker's TSCs. They also support the notion of repetition effects on phonetic convergence. Individuals are more likely to converge to a speaker as word frequency lowers and as word repetitions increase (see also Goldinger, 1998). The current study will make use of LF words as target words to measure phonetic convergence and will expose each word three times to each participant as two repetitions have been shown to be sufficient to induce phonetic convergence.

Measuring phonetic convergence. The AXB perceptual rating task is a valid measure for phonetic convergence (see Goldinger, 1998; Goldinger & Azuma, 2004). It measures convergence in studies operating under the direct realist theory as well as studies operating under the episodic theory.

Generally, convergence studies have multiple phases: baseline and exposure phases. At baseline, participants who will be referred to henceforth as *talkers,* are presented with words in text format and are asked to say the words out-loud. These baseline recordings aim to capture the *talker's* normal, uninfluenced manner of speaking the target words. During the exposure phase, *talkers* are exposed to utterances from a model speaker and are either asked to shadow the utterances (e.g. say the word out loud immediately after the utterance is heard) from the model speaker or say the words at some time point after having been exposed to the model's speech while being recorded.

The *talker's* baseline and post-exposure recordings are then compared to the model's recordings via a perceptual rating AXB task. The perceptual rating task is performed by a different set of participants who will henceforth be referred to as *raters*. *Raters* are presented with three auditory utterances successively — A, X, and B. Raters are asked to decide, which utterance, A or B, is most like X in pronunciation. The X stimulus is typically an utterance (e.g. a word like "portal") produced by the model talker. The A and B stimuli are the same word (e.g. "portal") produced by the *talker* at different phases of the experiment: the baseline utterance and the post-exposure utterance. For example, a *rater* may hear the word "portal" uttered by a *talker* at baseline (e.g. A), followed by the same word uttered by the model speaker (e.g. X), and lastly hear "portal"

uttered by the *talker* after the exposure phase (e.g. B). *Raters* would then judge whether the stimulus in position A or the stimulus in position B sounded most like the stimulus in position X. Using the AXB task, evidence for phonetic convergence is found when raters judge utterances from the post-exposure phase to be more similar to the model's utterances than the baseline utterances. The current study will employ an AXB task similar in structure to what was described above.

Phonetic convergence has also been measured using acoustic features such as duration — which is an acoustic dimension that varies reliably in convergence studies (Pardo, Gibbons, Suppes & Krauss, 2011; Pardo et al., 2013; Pardo et al., 2017). Duration is measured using differences-in-distance (DID) estimates. DID estimates are calculated by comparing differences in the durations of speech between *talkers'* baseline speech and models' speech (baseline - model), and then comparing differences in durations of speech between *talkers'* shadowed speech and model's speech (shadowed - model). The shadowed differences are then subtracted from the baseline differences. If these differences yield a positive value, they can be interpreted as convergence occurring. The current study will use differences in DID estimates to measure duration of speech.

Degree of Convergence. The episodic theory predicts differences in phonetic convergence due to frequency and repetition effects. However, other factors which have been found to influence differences in phonetic convergence include the sex of model talkers (Pardo, 2017), conversational role (Pardo, 2006), and context (Sanchez, Hay, & Nilson, 2015).

Research has suggested that the sex of the model speaker plays a role in the degree of convergence, but the results are mixed. Some research suggests that convergence has a higher chance of occurring when the speaker is a female (Namy, Nygaard, & Sauerteig, 2002; Dias & Rosenblum, 2011). There is also evidence to suggest that females converge more readily to talkers of their own sex (Namy et al., 2002; Pardo, 2006; Miller, Sanchez, & Rosenblum, 2010). Others have found that convergence is higher when the speaker is a male (Miller et al., 2010). More recently, some studies suggest that sex has no effect on the degree to which an individual converges (Pardo, Jordan, Mallari, Scalon, & Lewandowski, 2013; Pardo, Urmanche, Wilman, & Wiener, 2016; Pardo, 2017). The current study will use only female participants as well as a female model speaker based on evidence that females have higher rates of convergence and converge more readily to same-sex talkers.

An individual's role in a conversation may affect the degree to which an individual converges. Pardo (2006, also Pardo, Jay, & Krauss, 2010) enrolled pairs of participants in a task where participants were required to converse. Each participant was assigned the role of either a giver of directions or receiver of directions. The giver described a route on a map labeled with landmarks to the receiver. The receiver had to

converse with the giver in order to draw the described route on a separate unlabeled map. Participants were recorded uttering the names of the landmarks before, during, and after the task. These recordings were then used to assess convergence in an AXB task. Both studies found that givers converged more towards receivers than vice-versa. This suggests that an individual's role within a conversation may influence the degree to which they converge. The current study will control for this by not having the participants engage in conversation, but instead they will read sentences aloud, without engaging in conversation.

The context in which speech takes place may also affect phonetic convergence. Sanchez, Hay, and Nilson (2015) conducted an analysis on a corpus of New Zealand English. Specifically, they identified sections of the corpus where speakers had speech referencing Australia (i.e. Australian context for speech) as well as speech without Australian context. They analyzed vowels (/ɪ/, /æ/, and /ɛ/) that distinguish New Zealand English and Australian English and compared how individuals produced each vowel in the Australian context versus a neutral context. They found that in a setting where an Australian context naturally occurred in a conversation, that New Zealand speakers would spontaneously adopt more Australian-like speech realizations, suggesting that context matters when shifting one's speech.

Sanchez, Hay, and Nilson (2015) also investigated the role of context in speech convergence in an experiment that did not use conversational speech, but instead had participants read words from a computer screen. In an experiment, the researchers recruited New Zealand English speakers and recorded them uttering a sequence of words

in two phases — baseline and experimental. Each sequence had three words. The first word was a word meant to prime the context of speech and the next two words were words which contained the vowels of interest. In the baseline phase, participants were recorded uttering words with a neutral context (e.g. “marmalade”) followed by words containing the vowels of interest (e.g. “skit”, “peck”). In the experimental phase, participants first uttered words with an Australian context (e.g. “koala”) before uttering the words with the vowels of interest. An analysis comparing the recordings from the baseline and experimental phases revealed that participants shifted the way they produced the vowels of interest. Participants in the experimental phase that activated an Australian context resulted in more Australian-like speech realizations. These results suggest that the context in which we are speaking may influence the way speech is produced, even if one is simply reading words from a computer screen. The current study will expose participants to the auditory speech of a model and then place them in a context where participants will read sentences attributed to the model which they previously heard, thus creating a model context.

Auditory Imagery

Auditory imagery is a phenomenon which refers to what we “hear” when we imagine sounds (e.g. what you “hear” when you imagine your favorite song being played). There is evidence to suggest that it has similar characteristics to physically perceiving acoustic signals. Specifically, the auditory cortex is activated both when experiencing auditory imagery for an event and during the direct perception of those

events (Halpern & Zatorre, 1999; Halpern, Zatorre, Bouffard, & Johnson, 2004). Auditory imagery for features of a talker's voice, such as speech rate has been shown to be experienced while reading text attributed to that talker and may influence how that text is read (Alexander & Nygaard, 2008; Kosslyn & Matt, 1977; Zhou & Christianson, 2016a, 2016b).

In the auditory imagery literature, speech rate is often used as a measure of the influence of one's auditory imagery affecting silent reading and reading aloud. For example, Alexander and Nygaard (2008) exposed participants to the speech of two model speakers — one fast speaker and one slow speaker — by presenting a recording of the speakers conversing with each other. After this exposure phase, participants were then asked to read two passages, one attributed to the fast speaker and one attributed to the slow speaker (counterbalanced). The results from the experiment showed no differences in reading speed when reading silently, but found significant differences when participants read aloud. In the reading aloud condition, reading rates were in line with the respective attributed author, fast or slow. This suggests that readers experience auditory imagery of a speaker's voice that can lead to changes in the spoken speech rate of participants who read aloud a text attributed to a given speaker. This study suggests that readers experience auditory imagery of a model's TSCs while reading a text aloud and that the model's TSCs influences how one reads, insofar as speech rate is concerned. However, no studies to date have investigated whether the words in the sentences attributed to a model will also influence one's speech at a more fine-grained level, such as the level of how words are pronounced (i.e. phoneme level).

The episodic theory offers an explanation for auditory imagery speech rate effects (Alexander & Nygaard, 2008; Kurby et al., 2009). This explanation is identical to the explanation offered by research in the speech convergence literature (Goldinger, 1998). TSCs experienced in auditory imagery reflect the speaking styles of specific talkers, which include speech rate. A key difference is that auditory imagery research tends to focus on a more macroscopic level of TSCs (e.g. speech rate) while phonetic convergence literature focuses on a microscopic level (e.g. phonemes). However, it is unknown whether these two bodies of literature are fundamentally addressing the same phenomena under different levels.

Whereas the episodic theory has been offered as an explanation for why reading rates may be affected after exposure to a speaker, there may be alternative explanations. For example, it could be that readers adopt a general response strategy where they simply mimic the speaker (e.g. "the talker was speaking fast, therefore I should read fast"; Alexander & Nygaard, 2008). Another explanation is offered by the theory of embodied cognition. Under this theory, after exposure to a speaker (fast or slow), a reader would gain a physical sensation of quickness or slowness. In turn, this physical sensation of quickness or slowness affects the reading rates of the participants, in line with their physical sensation. However, the episodic theory would better account for this effect if auditory imagery influenced the production speech on other levels (e.g. phonemes).

Current studies on auditory imagery in reading have focused on reading speeds as a measure of auditory imagery. However, the current literature does not address whether this occurs because readers are simply influenced by the temporal characteristics of a

speaker's speech, or whether they are experiencing auditory imagery of a specific talker's characteristics. For example, suppose an individual is instructed to read the following sentence "Students apply using an online portal" and that this sentence is attributed to a known fast talker. Will the individual read this sentence quickly as a result of simply copying the talker on only the dimension of speech rate (e.g. "the talker was speaking fast, therefore I should read fast"), experiencing embodied cognition, or will they read the sentence quickly as a result of experiencing auditory imagery of TSCs along several dimensions (i.e. speech rate and the unique way phonemes are uttered in a word), which would suggest that the participant may experience "hearing" the talker's voice while they read, by activating stored memory traces of the talker along with specific words uttered by the talker. An investigation that examines changes at the sentence level and word level would provide some insight into this issue.

The Current Study

The current literature on phonetic convergence and auditory imagery fails to address the relationship between the two phenomena. Within the episodic theory, if an individual gains sufficient experience with a speaker's voice, they should accumulate episodic traces which are associated with the speaker's TSCs, such as the speech rate of the speaker or how the speaker pronounces certain phonemes. If this individual reads aloud text thought to be written by a fast talker (e.g., "Students apply using an online portal"), then they should have a faster speech rate and pronounce certain words (i.e. low-frequency words) similar to the speaker. However, it is possible that when reading

content attributed to a model speaker, as is done in auditory imagery studies, that speech rate takes precedence in the TSCs, where participants will shift in line to the macro level of speech rate, but not the micro level of phoneme. This issue has yet to be investigated in the literature.

The current study examines the link between phonetic convergence and auditory imagery by testing whether auditory imagery of a speaker extends beyond speech rate to other characteristics of that speaker (e.g., TSCs), such as how they pronounce individual words. In the current study, two groups of participants will be used — *talkers* and *raters*. *Talkers* will engage in three phases: baseline, exposure, and reading. During the baseline phase, *talkers* will be recorded reading sentences aloud from a computer screen to assess their natural speech rate and to obtain an example of how they naturally say (target) words in general. Each sentence will contain a low-frequency target word which will later be used to test for phonetic convergence and auditory imagery. In the exposure phase, *talkers* will be exposed to a model speaker's auditory speech. The model they hear will either have a fast or slow speech rate. *Talkers* will listen to the model utter sentences (different sentences from the baseline) containing half of the target words from the baseline task in her sentences. In the reading phase, *talkers* will engage in a task similar to the baseline phase where they will be recorded reading the same sentences, but in this case, each sentence will be attributed to the model speaker from the exposure phase. The recordings will be used to identify changes in the *talker's* speech as compared to the model's, as measures of phonetic convergence and auditory imagery.

Hypotheses

The variables in this study are used to establish an understanding of the relationship between phonetic convergence and auditory imagery. The two independent variables used in this study are word exposure and model speech rate. Because it is the premise of this thesis that phonetic convergence and auditory imagery are fundamentally governed by the same processes, but are simply indicating differences in the level of observation (e.g. micro vs. macro), it is not expected for the variables word exposure and speech rate to interact for either phonetic convergence or auditory imagery. However, the main effects stemming from these independent variables will illuminate the differences between the levels of interest when conducting phonetic convergence research (e.g. micro level; at level of phoneme) compared to auditory imagery research (macro level; at level of duration of a sentence or passage), and where the two areas, phonetic convergence and auditory imagery, overlap (e.g. mid-level; at level of duration of a word). The hypotheses of interest are the following:

Hypothesis 1. A main effect of word exposure is expected to demonstrate evidence of phonetic convergence for exposed words over unexposed words when measuring perceptual ratings via an AXB task.

Rationale. If low-frequency words which are repeatedly heard influence *talkers* in their speech production, then *talkers* will converge in their pronunciation of those words toward the model speaker when reading a sentence attributed to that speaker. Specifically, they will converge to exposed words (words previously spoken by the model talker) as compared to unexposed words not previously uttered by the model talker. This prediction is in line with findings that suggest that exposure to a talker will leave episodic traces of that talker's TSCs and lead to phonetic convergence (Goldinger, 1998; Goldinger & Azuma, 2004; Pardo, 2006).

Hypothesis 2. A main effect of word exposure is expected to demonstrate evidence for both phonetic convergence and auditory imagery when measuring duration at the level of the word.

Rationale. Exposed words uttered by the subjects should be most similar in duration to the model's duration of the words compared to unexposed words. If *talkers* are influenced in the way they say individual words by experiencing auditory imagery of a speaker, and the auditory imagery preserves details of a speakers' TSCs, then *talkers* will have a shorter duration for words to which they were previously exposed at a rate that is more in line with the model speaker they experienced. Exposed words will have longer or shorter durations (in line with the speaker's duration of the target words) than words to which they were not exposed, relative to baseline. This prediction is in line with the notion that auditory imagery may be explained by the episodic theory. (Alexander & Nygaard, 2008; Goldinger, 1998; Kurby et al., 2009). This hypothesis would provide a link between phonetic convergence (exposed vs. unexposed words) and auditory imagery (word duration). However, it could just be that the participants read words quickly as a result of simply copying the model talker's overall speed or experiencing embodied cognition instead of actually being influenced by that talker's TSCs.

Hypothesis 3. When measuring duration at the level of the word, a main effect of speech rate is expected to demonstrate evidence for auditory imagery.

Rationale. Those exposed to the model with a fast speech rate will say target words quicker than those exposed to the model with a slow speech rate. When *talkers* are exposed to different rates of speech (e.g. fast or slow) by specific model speakers, if people are then presented with a target word which is attributed to that model speaker, then auditory imagery whilst reading that word aloud will lead to reading it either fast or slow, in line with the rate of the specific model. This prediction is in line with evidence that suggests listeners shift in reading speed to align with the speech rate of a model speaker to whom the text is attributed (Alexander & Nygaard, 2008; Zhou & Christianson, 2016a, b). However, individual words have not been tested individually as the literature tends to use passages of text.

Hypothesis 4. When measuring duration at the sentence level, a main effect of speech rate is expected to demonstrate evidence for auditory imagery.

Rationale. Those exposed to the model with a fast speech rate will say the sentences quicker than those exposed to the model with a slow speech rate. When people are exposed to different rates of speech (e.g. fast or slow) by specific model speakers, if people are then presented with a sentence that is attributed to that model speaker, then auditory imagery whilst reading the sentence will lead to reading rates that are either fast or slow, in line with the rate of the specific model. Thus, it is predicted that those exposed to the fast talker condition will have faster reading times than those exposed to the slow talker condition. This prediction is in line with evidence that suggests listeners shift in reading aloud speed to align with the speaking rate of a model speaker when a text is attributed to that model speaker (Alexander & Nygaard, 2008; Zhou & Christianson, 2016a, b).

Method

Participants

There were two participant groups for this experiment, *talkers* (N = 12) and *raters* (N = 68). The sample size for *talkers* was determined based upon previous studies in relevant literature. The sample size for *raters* was determined through a power analysis which yielded power of .807 (Judd, Westfall, & Kenny, 2016; Westfall, 2018). Participants who were *talkers* were disqualified from being *raters*. Only female participants were used as *talkers*, but people of all sexes and genders will were allowed to participate as *raters*. All participants were students from Humboldt State University and

participated via the SONA systems recruitment pool for class credit or extra credit. All participants will were at least eighteen years of age, native English speakers, had normal to corrected-to-normal vision, and had no reported hearing impairments. This project was approved by the Institutional Review Board (IRB) at Humboldt State University (IRB number: 17-209).

Materials

The materials in this experiment consist of a word list, 60 sentences containing words from the word list, and recordings of model talkers saying all sentences out loud.

Word list. The word list for this experiment consisted of 40 low frequency words that are bisyllabic (from Goldinger (1998)). Low-frequency bisyllabic words were used in this experiment because this class of stimuli have shown stronger rates of phonetic convergence compared to other sets of stimuli (Goldinger, 1998; Goldinger & Azuma, 2004).

Sentences. Sixty sentences — model sentences and *talker* sentences — were constructed by the author. All sentences will contained one target word. The target words were placed in a clause- or sentence-final position (Pardo, 2006). This was done to match the recorded utterances of the *talkers* and model in the same context (e.g. sentence-final positions). Twenty unique target words from the word list were reserved for the 20 unique model sentences. All of the 40 target words were used to create 40 *talker* sentences. All *talker* sentences were distinct from the model sentences.

Model recordings. One female model speaker was recorded reading all words from the word list, twenty model sentences, and all *talker* sentences. The model speaker recorded each sentence twice – once at a fast pace and another time at a slow pace (see Alexander & Nygaard, 2008). Additionally, the model speaker was recorded saying all *talker* sentences in order to compare them to the *talkers'* recordings. However, only the 20 model sentences were presented to the *talkers*. The recordings were done in a sound attenuated booth with a Beyerdynamic TG H55c microphone and saved onto a computer. All recordings were amplitude adjusted using the software Audacity.

Talker recordings. All *talkers* were recorded reading all *talker* sentences twice - once during the baseline phase and once during the reading phase. The recordings were done in a sound attenuated booth with a Beyerdynamic TG H55c microphone and saved onto a computer. All recordings were amplitude adjusted using the software Audacity.

Design

This study was an experimental 2 (word exposure) x 2 (model speech rate) experimental design. Word exposure was a within-subjects factor with two levels: exposed words and unexposed words. Exposed words were target words embedded in the model sentences produced by the model speaker during the exposure (listening) phase. Unexposed words were target words not embedded in the model sentences during the exposure phase. Model speech rate was a between subjects factor with two levels: fast and slow. Half of the talkers were exposed to a model speaker with a fast speech rate. The remaining talkers were exposed to a model speaker with a slow speech rate.

The dependent variables in this experiment were perceptual ratings of phonetic convergence, word duration, and duration of sentences. Perceptual ratings of phonetic convergence is the percent of naïve listeners who judge a talker's post-exposure utterance as more similar to the utterance of a model speaker relative to the *talker*'s baseline. Talker speech rate is measured in two ways: target word duration and sentence duration. Duration measurements for word and sentence length were made by using the computer program Praat (Boersma et al., 2018). Duration of target words were measured as the duration of individual words from the onset to offset of vocalizations. Duration of

sentence length were measured as the duration starting with the initial onset vocalization of the first word in a sentence to the offset of vocalization of the last word in the sentence. DID estimates between the baseline utterance, post-exposure utterance, and model utterance will provide a duration measure of convergence

Procedure

There were two parts to this experiment with different sets of participants – *talkers* and *raters*. All stimuli was presented with the experimental program E Prime (Psychology Software Tools, Pittsburgh, PA).

Talkers. In the first part of this experiment, participants, referred to as *talkers*, engaged in the experiment in a sound attenuated booth in the psychology department. The utterances of *talkers* in the various tasks were recorded with a Beyerdynamic TG H55c microphone and saved onto a computer. *Talkers* engaged in three phases. The first phase was the baseline phase, where participants were asked to read the *talker* sentences aloud whilst being recorded. The sentences were presented one at a time on a computer screen at two second intervals. *Talkers* were asked to read the sentences clearly. The sentences of the baseline phase were intended to reflect the normal way the *talkers* speak and will be used as baselines.

In the second phase (exposure) of the experiment, *talkers* were asked to listen and pay attention to the model speaker uttering the 20 model sentences through Beyerdynamic DT 770 Pro headphones. The model speaker was given a name in order to

refer to her in the third phase. Each model sentence was repeated two times during the course of the phase. All 20 sentences were presented in blocks, where the specific sentence within each block were presented in a random order.

In the reading phase of the experiment, *talkers* were asked to read sentences out-loud, off of a computer screen. The sentences were identical to those in the baseline phase and were attributed to the model speaker from the exposure phase. Each sentence was prefaced with a notice attributing the sentence to the model speaker. The target words from the *talker* sentences, and each sentence were saved into individual files. All recordings were amplitude adjusted using the software Audacity(R) (Audacity Team, 2017).

Raters. Raters engaged in the experiment in a room in a lab suite at Humboldt State University. Raters were asked to make perceptual similarity ratings in an AXB task. They were presented with three sets of recordings of individual target words in a row — A, X, and B. Participants were tasked with making perceptual similarity ratings, indicating whether A or B was more similar in pronunciation to X.

The X stimulus was a target word from the model sentences uttered by the model speaker. The A and B stimuli were the same utterance as the model speaker, but uttered by the *talker* during the baseline and reading phase. For example, a *rater* may have hear the word “portal” uttered by a *talker* at baseline, followed by the same word uttered by the model talker, and lastly heard “portal” uttered by the *talker* in the reading phase. *Raters* then judged whether the first or third versions of “portal” sounded most like the

middle version. The A and B stimuli were counterbalanced such that half the recordings at baseline were placed in the A stimulus position and half were placed in the B stimulus position.

Results

Data Cleaning

The data were collected via E-Prime (Psychology Software Tools, Inc., 2016) and analyzed with R (R Core Team, 2016). Audio samples from *talkers* which had excessively poor quality (i.e., contained incomplete or indecipherable speech) were excluded from subsequent analyses (n = 41 words). Talking durations were averaged across sentences and also averaged across words. Talking durations which fell three standard deviations above or below the mean were excluded from subsequent analyses (n = 11 words). Table 1 shows summary statistics for word durations as measured by differences-in-distance (DID) estimates. Table 2 and Table 3 show average word and sentence durations as measured by the raw difference in average duration from the reading to baseline phases for both word and sentences.

Model Speaker

There was a difference in the length of words (in seconds) spoken by the model speaker. Such that words spoken at a fast pace (M = 0.47, SD = 0.12) did indeed have shorter durations than those spoken at a slower pace (M = 0.61, SD = 0.18), t(65.98) = -

4.0, $p < .001$, $d = 0.9$. There was also a difference in the length of sentences (in seconds) spoken by the model speaker. Such that sentences spoken at a fast pace ($M = 3.44$, $SD = 0.64$) did indeed have shorter durations than those spoken at a slower pace ($M = 4.26$, $SD = 0.74$), $t(38) = -3.75$, $p < .001$, $d = 1.19$.

Analyses

All data were analyzed within the R statistical computing environment using the R packages lme4 (Bates et al., 2015), lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017), and the tidyverse (Wickham, 2017). Each hypothesis was tested at an alpha level of .05 using a mixed effect model with random effects specified for individual subjects. Each analysis employed a χ^2 test comparing the specified model to a null model specified with random effects to measure whether the specified model provides an improvement over a null model.

Hypothesis 1: Effect of word exposure on talker convergence via AXB. It was expected that exposed words uttered by the *talkers* would be rated by *raters* as having converged more often than words which were unexposed would be rated as having converged. The dependent variable was a binary value (zero or one) indicating whether a *rater* judged a *talker*'s utterance as having converged to the model speaker. A value of zero indicates that the *rater* judged the *talker* as not having converged and a value of one indicates that the *rater* judged the *talker* as having converged.

The results from this analysis suggest that there was no effect of word exposure on convergence ratings, $B = -0.002$, $z = -0.674$, $p = .5$, marginal $R^2 = .0001$. This suggests that *talkers* did not converge to the model speaker at a higher rate when uttering an exposed target word ($M = 0.49$, $SD = 0.50$) compared to an unexposed target word ($M = 0.50$, $SD = 0.50$). Overall, the model was not found to add much information over and above a null model, $\chi^2(1) = 0.45$, $p = .5$ (see Table 4).

Hypothesis 2: Main effect of word exposure on talker speech via differences-in-distance Estimates. It was expected that exposed words uttered by the subjects should be most similar in duration to the model's duration of the words compared to unexposed words. If a subject was exposed to target words being spoken slowly, then the subject would slow down in their speech. If a subject was exposed to a target word being spoken quickly, then they would speed up in their speech. The dependent variable was *talker* word duration as measured by differences-in-distance (DID) estimates. A positive value indicates the participants shifting their speech in line with model speaker. The fixed effects for this model was word exposure (exposed or unexposed).

The results from this analysis suggest that there was no effect of word exposure on DID estimates, $B = -0.004$, $t(432.32) = -0.78$, $p = .44$, marginal $R^2 = 0.001$. This suggests that *talkers* did not shift their speech at the word level when uttering an exposed target word ($M = -0.001$, $SD = 0.057$) compared to an unexposed target word ($M = -0.006$, $SD = 0.068$). Overall, the model was not found to add much information over and above a null model, $\chi^2(1) = 0.61$, $p = .43$ (see Table 5).

Hypothesis 3: Effect of model speaker speech rate on talker speech via raw differences in word duration. It was expected that model speech rate would have an effect on *talkers* word duration utterances. Participants that listened to the model speaking quickly were expected to say target words quicker than those who listened to the model with a slow speech rate. The dependent variable for this model was the raw difference in duration of target words (regardless of whether they were exposed or unexposed words) between the reading and baseline phase. For the raw difference in duration, a negative value indicates that the participant sped up in their speech rate and a positive value indicates that the participant slowed down in their speech rate. The fixed effect for this model was model speech rate (fast or slow). This analysis was performed on a subset of the data which included only included target words to which the participant was exposed.

The analysis indicated that model speech rate had no effect on the duration of target words, $B = 0.01$, $t(9.63) = 0.96$, $p = .36$, marginal $R^2 = .01$ (see Table 6). This result suggests that *talkers* who listened the model speaker with fast speech rate ($M = -0.013$, $SD = 0.074$) and *talkers* who listened to the model speaker with a slow speech rate ($M = -0.001$, $SD = 0.075$) did not shifted their overall speaking style in line with the model. This model did not provide additional information over and above a null model, $\chi^2(1) = 1.04$, $p = .31$ (See Table 6).

Hypothesis 4. Main effect of model speaker speech rate on talker speech rate via raw differences in sentence duration. It was expected that model speech rate would have an effect on talker utterances of sentences in overall duration. Those exposed to the model speaking quickly were expected to read sentences quicker than those exposed to the model who spoke slowly. The dependent variable for this model was the raw difference in duration of sentences between the reading and baseline phase. For the raw difference in duration, a negative value indicates that the participant sped up in their speech rate and a positive value indicates that the participant slowed down in their speech rate. The fixed effect for this model was model speech rate (fast or slow). The analysis suggests that model speech rate had no effect on the duration of sentences of *talkers*, B = 0.02, $t(10.01) = 0.171$, $p = .87$, marginal $R^2 = 0.0001$ (see Table 7). This result suggests that *talkers* who listened to the model speaker with fast speech rate (M = -0.14, SD = 0.81) and *talkers* who listened to the model speaker with a slow speech rate (M = -0.16, SD = 0.67) did not shift their utterances in line with the model. This model did not provide additional information over and above a null model, $\chi^2(1) = 0.03$, $p = .86$.

Discussion

The current study aimed at investigating whether phonetic convergence and auditory imagery are governed by the same processes. This question was addressed by combining the designs and hypotheses from extant literature on phonetic convergence (Goldinger, 1998; Goldinger & Azuma, 2004) and auditory imagery (Alexander & Nygaard, 2008). The results of this study suggest that there may be different mechanisms

that underlie each process, however further investigation is needed in order to verify this conclusion.

Hypothesis 1: Main Effect of Word Exposure via AXB

Goldinger and Azuma (2004) found that if a person is repeatedly exposed to words which are low-frequency (i.e., do not often appear in everyday language) then a person will converge to that speaker (i.e., imitate their TSCs when uttering that word). The current study expected that participants repeatedly exposed to low-frequency words uttered by a model speaker would converge to those words. This hypothesis was not supported by data. This result suggests that episodic encoding may not play a role in phonetic convergence. Contrary to previous findings on episodic encoding's influence on speech (Goldinger, 1998; Goldinger & Azuma, 2004; Pardo, 2006), these results do not provide evidence for *talker*s storing traces in memory associated with the TSCs of the model speaker influencing their speech.

Hypothesis 2: Main Effect of Word Exposure on Talker Speech via Differences-in-Distance Estimates

Alexander and Nygaard (2008) found that participants exposed to a model speaker reading sentences quickly or slowly would shift their speech rate to match the model speaker when asked to read a passage purported to be written by that model speaker. Specifically, this occured when the passages were read aloud, but not when read silently. Based on this finding, in combination with Goldinger and Azuma's (2004)

finding, it was expected that *talkers* would converge to repeatedly exposed low-frequency words with respect to duration. It was expected that *talkers* would utter these words quickly or slowly, in line with the model speaker experienced, relative to their baseline speech rate. However, this hypothesis was not supported by the data. This result suggests that episodic encoding is not a shared mechanism between auditory imagery in reading and phonetic convergence. Readers experiencing auditory imagery may indeed experience "hearing" the voice of a familiar author while reading due to stored memories of that author's talker-specific characteristics (TSCs) (Alexander & Nygaard, 2008; Kurby, Magliano, & Rapp, 2009; Zhou & Christianson, 2015, 2016), but these memories may not actually influence the reader's speaking style. An explanation for why episodic encoding alone does not explain how auditory imagery affects reading is that familiarity with a voice extends beyond simple word/voice pairings or brief interactions with a voice, as is commonly seen in research. It is suggested that one must also take into account the goals, relationship, and context with respect to a voice in order to understand how the relationship between auditory imagery and reading arises naturalistically (Kurby, Magliano, & Rapp, 2009; Sanchez, Hay, & Nilson, 2015). While the current study created a context related to the speaker, it may not have been a strong enough context for auditory imagery to occur. Alexander and Nygaard (2008) exposed participants to passages of speech where a model context was created where each passage had a specific theme (one concerned a family vacation and the other concerned plans for a new business). It could be that Alexander and Nygaard's manipulation provided a more salient

context for which participants could experience auditory imagery compared to the current study.

Hypotheses 3 and 4: Main Effect of Speech Rate via Word and Sentence Duration

The current study expected to find that *talkers* would shift their speech in line with a model speaker (speaking quickly or slowly) at both the word and sentence level. At the word level, this was hypothesized to occur regardless of whether the target word was exposed to the participant. However, *talkers* did not demonstrate any shift in their speech at the word level after being exposed to a model speaker. Similarly, *talkers* also did not shift their speech rate in line with the model's speech rate at the sentence level. These findings are contrary to the findings in Alexander and Nygaard (2008) who, at least for the sentence level, did find shifts in the speech rate of talkers after presented with a model's speech rate. The current study suggests that exposure to a fast or slow speaker's speech does not influence reading style when reading text thought to be written by that speaker. Currently, the data provides support for the abstractionist view of speech perception, where perceived speech is thought to be stripped away from its nonlinguistic properties (Pisoni, 1997; Tenpenny, 1995). This view contends that we perceive speech as context-free, independent of the identification, recognition, and storage of nonlinguistic properties of speech (i.e., a talker's voice). However, there were some limitations to this study that may need to be reconciled before a stance is made.

Limitations

This study faced limitations in several facets of the experiment. For example, at the word level, *talkers* only received two repetitions for each sentence (and each exposed target word) spoken by the model speaker. This may not have been enough exposure to the model speaker for convergence to occur. At the word level, the effects of phonetic convergence has been observed at a minimum of two repetitions per word, but the effect is more pronounced as repetitions increase (Goldinger, 1998; Goldinger & Azuma, 2004). However, the work on phonetic convergence has found its evidence primarily via the AXB method, though there are some studies that have looked at some speech dimensions, like duration , which may serve as a proxy for speech rate, (Pardo, Jay, & Krauss, 2010; Pardo et al., 2009). In these cases, duration is measured using differences-in-distance estimates (see Measuring phonetic convergence) as well as articulation rate (e.g., words per second). At the sentence level, the effects of auditory imagery on *talker* speech has been observed with about four minutes of exposure to a model speaker (Alexander & Nygaard, 2008). However, in the current study, each *talker* was exposed to about 2.5 minutes of speech from a model speaker, which may not have been enough exposure to a voice for auditory imagery to occur.

Another limitation of this study is that the sentences spoken by the model speaker were presented to participants one at a time. This stands in contrast to previous research on auditory imagery where participants were exposed to complete passages of speech in the form of a conversation between two speakers (Alexander & Nygaard, 2008; Kurby,

Magliano, & Rapp, 2009). Speech is typically not spoken one sentence at a time with large pauses in between. It is possible that auditory imagery may be more likely to occur for speech spoken naturalistically (i.e., conversational speech) as opposed to speech presented as unrelated isolated sentences.

Furthermore, the current study only used one model speaker with one sex. This limits the generalizability of the study with respect to the effects of sex. This is because it has been found that speech convergence is modulated by sex of the speaker such that people may more readily converge to female speakers (Namy, Nygaard, & Sauerteig, 2002). However, there is also evidence that people more readily converge to male speakers (Miller et al., 2010). Because of this opposing literature, it is important to use model speakers of all sexes.

Lastly, the current study instantiated a context related to the model speaker by having *talkers* listen to the model speaker's speech and by having them read sentences aloud purported to be written by that speaker. However, it may be that the context in the current study was not rich enough to induce auditory imagery. For example, speech presented to participants in previous studies have had specific themes throughout passages, made multiple references to the model speaker's name, and emphasized the model's speech rate by contrasting them to another speaker (e.g., having the "fast" model speaker speak to someone speaking slowly; Alexander & Nygaard, 2008; Kurby, Magliano, & Rapp, 2009; Zhou & Christianson, 2015, 2016). Because the current study did not include these aspects, it may be that the context was not rich enough for auditory imagery to occur.

Future Directions

The results of the current study oppose several studies regarding phonetic convergence and auditory imagery. Thus, further research should be conducted in order to provide evidence that auditory imagery with respect to reading and phonetic convergence are indeed not governed by episodic encoding. Further work should aim to understand these two phenomena by integrating various aspects of studies done to date. These include presenting the model's speech in a more naturalistic manner (i.e., conversational speech), testing target words at different levels of repetitions and frequencies, and having a more salient context for the model speech, and using different model speakers. If it is found that participants' speech does not shift after interacting with the voice of a model speaker and reading text thought to be written by the model speaker with respect to repetition of target words, different levels of word frequency, and context, then it would provide strong evidence that auditory imagery and phonetic convergence do not have a shared mechanism through episodic encoding. However, if participants' speech does shift, it would provide evidence that there episodic encoding plays a key role in both auditory imagery and phonetic convergence.

Studies on auditory imagery typically present the model speakers' speech in the form of passages or scripts (Alexander & Nygaard, 2008; Kurby, Magliano, & Rapp, 2009; Zhou & Christianson, 2015, 2016). Additionally, it has been found that phonetic convergence occurs during the course of conversational speech (Pardo, 2006; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo et al., 2013). Thus future work should integrate

these aspects of both literatures by having participants engage with a model speaker in conversational speech. Ideally, this conversational speech integrates aspects of both auditory imagery and phonetic convergence. For example, by having model speakers vary in their speech rate (fast vs. slow), including target words of varying frequency (low-, medium-, and high-frequency), and having these target words have varying repetitions (between zero and twelve). If there is some interaction whereby participants show differing degrees of convergence to a model speaker according to word repetition and word frequency when reading a passage thought to be written by a model speaker, then it would provide strong evidence for episodic encoding as a shared mechanism between auditory imagery and phonetic convergence.

Moreover, speech presented to participants in future studies should have enriched contexts. This can be done several ways. For instance, by having consistent themes throughout the passages or scripts which are used or by making repeated references to the model speaker's name. It is important to have a salient context at it has been suggested that context influences convergence (Sanchez, Hay, & Nilson, 2015) and because speech is contextualized when it is encountered naturally (Goldinger, 1998; Pisoni, 1993, 1997). Thus, having a salient context would increase reliability by simulating an aspect of natural speech.

Lastly, future work should make use of multiple male and female model speakers. Exposing participants to different voices means exposing them to varying pitches, tones, and rates of speech. This would increase the generalizability of the findings beyond a single model speaker's voice.

Conclusions

The current study sought to find a link between auditory imagery and phonetic convergence by combining methods from both bodies of literature. Exposure to a model speaker's speech did not result in any *talkers* shifting their speech to more closely align with the model speaker with respect to speech rate at the word or sentence level. While the current study provides no support for any link between these two phenomena, there is a host of literature suggesting that they may be linked. Thus, further work should be conducted before concluding that there is no relationship between phonetic convergence and auditory imagery.

To the author's knowledge, the present results are the first to examine the link between auditory imagery and phonetic convergence and their effect on speech. Studying the link between these two phenomena may allow insight into how memory, specifically episodic memory, play into the perception and production of speech. Whether there is (or is not) a relationship between auditory imagery and phonetic convergence must be further investigated for a stronger understanding of the interplay of memory and language.

Tables

Table 1

Summary statistics for talkers' word durations via DID estimates

	M	*SD*
Condition		
Fast	0.011	0.061
Slow	-0.003	0.064
Word Exposure		
Exposed	0.006	0.068
Unexposed	0.001	0.057

Table 2

Summary statistics for talkers' word durations via raw estimates

	M	*SD*
Condition		
Fast	-0.013	0.074
Slow	0.001	0.075
Word Exposure		
Exposed	-0.001	0.065
Unexposed	-0.013	0.082

Table 3

Summary statistics for talkers' sentence durations via raw estimates

	M	*SD*
Condition		
Fast	-0.14	0.81
Slow	-0.16	0.67
Word Exposure		
Exposed	-0.11	0.67
Unexposed	-0.19	0.81

Table 4

Hypothesis 1: Effect of word exposure on talker convergence via AXB

Fixed Effects	Estimate	SE B	*z*	*p*(*z*)	χ^2(*df*)	*p*(χ^2)
(Intercept)	-0.002	0.05	-0.05	.96		
Exposed: Yes	-0.036	0.05	-0.67	.5		
					0.45(*1*)	.5

Note. Converged ~ Word_Exposure + (1|Word) + (1|Talker)

Table 5

Hypothesis 2: Effect of word exposure on talker speech via word duration: Differences-in-distance model

Fixed Effects	Estimate	SE B	*t*	*p*(*t*)	χ^2(*df*)	*p*(χ^2)
(Intercept)	0.01	0.01	1.08	.30		
Exposed: Yes	-0.004	0.006	-0.78	.44		
					0.61(*1*)	.43

Note. DID ~ Exposed + (1|Word) + (1|Talker)

Table 6

Hypothesis 3: Effect of model speaker speech rate on talker speech via word duration: raw differences model

Fixed Effects	Estimate	SE B	*t*	*p*(*t*)	χ^2(*df*)	*p*(χ^2)
(Intercept)	-0.01	0.01	-1.39	.18		
Speech Rate: Slow	0.01	0.01	0.96	.36		
					1.04(*1*)	.31

Note. Raw_diff ~ Speech_Rate + (1|Word) + (1|Talker)

Table 7

Hypothesis 4: Effect of model speaker speech rate on talker speech via raw differences in sentence duration

Fixed Effects	Estimate	SE B	*t*	*p*(*t*)	χ^2(*df*)	*p*(χ^2)
(Intercept)	-0.14	0.1	-2.05	.07		
Speech Rate: Slow	-0.02	0.1	0.17	.87		
					0.03(*1*)	.86

Note. Raw_diff ~ Speech_Rate + (1|Talker)

References

Alexander, J. D., & Nygaard, L. C. (2008). Reading voices and hearing text: Talker-specific auditory imagery in reading. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 446–459. doi:10.1037/0096-1523.34.2.446

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660. Retrieved from https://www-cambridge-org./core/journals/behavioral-and-brain-sciences

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01

Blank, A. & Koch, P. (1999). *Historical semantics and cognition*. Berlin, Germany: Mouton de Gruyter

Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.38, retrieved 29 March 2018 from http://www.praat.org/

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893–910. doi: 10.1037/0022-3514.76.6.893

Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical*

Neurophysiology, *107*, 287–295. doi: 10.1016/S0013-4694(98)00071-6

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*, 176–180. doi: 10.1007/BF00230027

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception*, *40*, 1457–1466. doi:10.1068/p7071

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics*, *78*, 317–333. doi:10.3758/s13414-015-0982-6

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, *73*, 2608–2611. doi: 10.1152/jn.1995.73.6.2608

Fowler, C. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics*, *14*, 3–28. Retrieved from https://www.journals.elsevier.com/journal-of-phonetics

Fowler, C., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks.

Journal of Memory and Language, *49*, 396–413. doi:10.1016/S0749-596X(03)00072-X

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97, 412. doi: 10.1037/0033-2909.97.3.412

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*, 361–377. doi: 10.3758/BF03193857

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493–501. doi:10.1016/S1364-6613(98)01262-5

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00340

Gentilucci, M., & Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia*, *45*, 608–615. doi:10.1016/j.neuropsychologia.2006.04.004

Giles, H., & Ogay, T. (2007). *Explaining Communication: Contemporary Theories and Exemplars*. Mahwah, NJ: Lawrence Erlbaum.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *48*, 76–95. doi:10.1037/0033-295x.105.2.251

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*, 716–722. doi: 10.3758/BF03196625

Halpern, A. R., & Zatorre, R. J. (1999). When that tune runs through your head: A PET investigation of auditory imagery for familiar melodies. *Cerebral Cortex*,

9, 697–704. doi:10.1093/cercor/9.7.697

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*, 1281–1292. doi:10.1016/j.neuropsychologia.2003.12.017

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*, 846–848. doi: 10.1126/science.1070311

Kosslyn, S. M., & Matt, A. M. (1977). If you speak slowly, do people read your prose slowly? Person-particular speech recoding during reading. *Bulletin of the Psychonomic Society*, *9*, 250–252. doi:10.3758/bf03336990

Kurby, C. A., Magliano, J. P., & Rapp, D. N. (2009). Those voices in your head: Activation of auditory images during reading. *Cognition*, *112*, 457–461. doi:10.1016/j.cognition.2009.05.007

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*. *82*, 1-26. doi: 10.18637/jss.v082.i13

Judd, C. M., Westfall, J., & Kenny, D.A. (2016). Experiments with more than one random factor: Designs, analytic models, and statistical Power. *Annual Review of Psychology*, *68*. doi:10.1146/annurev-psych-122414-033702

Labov, W. (1972). Sociolinguistic patterns (No. 4). University of Pennsylvania Press.

Lachs, L., & Pisoni, D. B. (2004a). Crossmodal Source Identification in Speech Perception. *Ecological Psychology*, *16*, 159–187.

doi:10.1207/s15326969eco1603_1

Lachs, L., & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *The Journal of the Acoustical Society of America*, *116*, 507–518. doi:10.1121/1.1757454

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461. doi: 10.1037/h0020279

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences, 13*, 110–114. doi:10.1016/j.tics.2008.11.008

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics*, *72*, 1614–1625. doi:10.3758/APP.72.6.1614

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*, 1097-1108. doi: 10.1121/1.405558

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation:: The role of perception. *Journal of Language and Social Psychology*, *21*, 422–432. doi:10.1177/026192702237958

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception & Psychophysics, 72*, 2254–2264. doi:10.3758/APP.72.8.2254

Pardo, Jennifer S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*, 2382–2393. doi:10.1121/1.2178720

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40, 190–197. doi:10.1037/e520562012-541

Pardo, Jennifer S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). Influence of Role-Switching on Phonetic Convergence in Conversation. *Discourse Processes, 50*, 276–300. doi:10.1080/0163853X.2013.778168

Pardo, Jennifer S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*, 183–195. doi:10.1016/j.jml.2013.06.002

Pardo, Jennifer S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*, 637–659. doi:10.3758/s13414-016-1226-0

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In Johhnson K. & Mullenix, J.W. (Eds.), *Talker variability in speech processing* (pp. 9-32). Morgan Kaufmann San Francisco.

Prinz, W. (1990). A common coding approach to perception and action. In Neumann, O. & Prinz, W. (Eds.), *Relationships between perception and action* (pp. 167–201). Springer Berlin Heidelberg.

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://www.pstnet.com

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*, 661. doi: 10.1038/35090060

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics*, *75*, 1359–1365. doi:10.3758/s13414-013-0534-x

Sanchez, K., Hay, J., & Nilson, E. (2015). Contextual activation of Australia can affect New Zealanders' vowel productions. *Journal of Phonetics*, *48*, 76–95. doi:10.1016/j.wocn.2014.10.004

Sanchez, K., Miller, R. M., & Rosenblum, L. D. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research, 53*, 262–272. doi:10.1044/1092-4388(2009/08-0247)

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian

Portuguese and English. *Journal of Phonetics, 25,* 421–436. doi: 10.1006/jpho.1997.0051

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66,* 422–429. doi: 10.3758/BF03194890

Westfall, J. (2018) Power analysis with random targets and participants [Shiny application]. Retrieved from http://jakewestfall.org/two_factor_power/

Wickham, H. (2017). Tidyverse: Easily install and load the 'Tidyverse'. *R package version 1.2.1*. https://CRAN.R-project.org/package=tidyverse

Van Lancker, D., Kreiman, J., & Wickens, T. (1985). Familiar voice recognition: Parameters and patterns. Part II. Recognition of rate-altered voices. J*ournal of Phonetics, 13*, 39-52.

Zhou, P., & Christianson, K. (2015). Auditory perceptual simulation: Simulating speech rates or accents? *Acta Psychologica*, *168*, 85–90. doi:10.1016/j.actpsy.2016.04.005

Zhou, P., & Christianson, K. (2016). I “hear” what you’re “saying”: Auditory perceptual simulation, reading speed, and reading comprehension. *Quarterly Journal of Experimental Psychology*, *69*, 972–995. doi:10.1080/17470218.2015.1018282