What Turns Speech into Song? Investigations of the Speech-to-Song Illusion

By

Nichol Castro

Submitted to the graduate degree program in Psychology
and the Graduate Faculty of the University of Kansas in partial fulfillment
of the requirements for the degree of Master of Arts.

Chairperson: Dr. Michael S. Vitevitch

Dr. Susan Kemper

Dr. Lila Chrysikou

Date Defended April 21, 2014

The Thesis Committee for Nichol Castro certifies

that this is the approved version of the following thesis:

What Turns Speech into Song? Investigations of the Speech-to-Song Illusion

Chairperson: Dr. Michael S. Vitevitch

Date approved: April 21, 2014

Abstract

In the Speech-to-Song Illusion a spoken phrase is presented repeatedly and begins to sound more like it was being sung. The present study used several different types of stimulus phrases to elicit the Speech-to-Song Illusion to determine the underlying cognitive mechanism responsible for this illusion. Previous stimuli used to elicit the Speech-to-Song Illusion have been phrases extracted from real sentences. We examined whether having a meaningful phrase (in Experiment 1), or even real words (in Experiment 2), were necessary for eliciting this illusion. We also examined how the length of the stimulus impacts the occurrence of the Speech-to-Song Illusion by manipulating the number of words in the repeated phrases (in Experiment 3) and by manipulating the number of syllables in each word of the phrases (in Experiment 4). The results of these 4 experiments support an underlying mechanism as described in Node Structure Theory: satiation of lexical nodes allows for the syllables to "pop out" as music-like beats.

Acknowledgement

First, I would like to thank my advisor, Dr. Michael Vitevitch, for his support and encouragement throughout this entire process. Mike has been a great influence in my work, and his insight has been invaluable. I would also like to thank my committee members, Drs. Susan Kemper and Lila Chrysikou, for their helpful comments, as well as my undergraduate advisor, Dr. Lori James, who nudged me into my current path. And lastly, Bryan “Kip” Haaheim for pointing us to “It’s gonna rain” by Steve Reich.

Also, I would like to thank Kenny and my entire family for their unending support and confidence. Without them none of this would be possible.

Table of Contents

List of Figures

List of Tables

List of Appendices

What Turns Speech into Song? An Investigation of the Speech-to-Song Illusion

Perceptual illusions occur when there's a mismatch between what is actually in the environment and what is perceived. Essentially, there is an error in interpreting correctly the sensory information from the environment (Gregory, 1968; Warren, 1976). Illusions can occur in any sense, with the most common illusions in vision dating all the way back to Aristotle (Gregory, 1968; Coren & Girgus, 1978). However, the focus of this paper will be on an auditory illusion in which what is presented to the ears perceptually transforms into another percept. Examples of this type of auditory illusion include the Verbal Transformation Effect (Warren & Gregory, 1958), phonemic restoration (Warren, 1970), and verbal satiation (Bassett & Warne, 1919). Although seemingly problematic, these illusions provide a window into "normal" cognitive processing, and some argue that the lack of experiencing a well-known perceptual illusion sometimes warrants claims of "abnormal" processing (Gregory & Wallace, 1963; Coren & Girgus, 1978). Perceptual illusions provide an experimental tool for understanding *how* and *why* things go awry in our perception of the environment.

Illusions are fascinating in that they arise not from the stimulus, but rather from the mind. Our minds are wired to be highly efficient and to do so limit information intake to only key features, become predictive when capable, and continue to function in times of limited or no input (Gregory, 1968). The downside to this efficiency, however, is that sometimes our brains overcompensate. John Locke noted that "the mind cannot fix long on one invariable idea," but rather shifts to new or modified ideas (Locke, 1894, p. 244). Given Locke's thoughts about the mind continually shifting to new ideas, it is perhaps not surprising that many perceptual illusions occur from prolonged stimulation or over-stimulation of a constant stimulus (Gregory, 1968). The unchanging stimuli will perceptually change or even fade and disappear (Warren, 1976).

Our brains attempt to be efficient in an unchanging environment by providing alternatives (e.g., hearing multiple words when only one word is played to a listener) to what we are sensing.

Speech-to-Song Illusion

The Speech-to-Song Illusion is an auditory illusion in which spoken language (i.e., words) comes to be perceived as rhythmical or musical in quality. The illusion can be elicited by continuously repeating a spoken phrase, without changing the stimulus in any other way. Although this illusion was a well-known technique used by musicians who made loops of magnetic recording tape (e.g., "It's gonna rain" by Steve Reich in the 1960's), the earliest report of this illusion in the scientific literature appeared in Deutsch, Henthorn, & Lapidis (2011). Deutsch "discovered" the Speech-to-Song Illusion while making instructional recordings describing other musical illusions (Deutsch, 1995; 2003). When the spoken phrase "sometimes behave so strangely" was played over and over, listeners indicated hearing a change in the stimulus from sounding like speech to sounding something more song-like (Deutsch, Henthorn, & Lapidis, 2011). Through repetition, the words in a phrase are perceptually transformed into music-like beats, even though the stimulus never actually changes.

As is the case for all perceptual illusions, it is only a listener's percept that changes during the Speech-to-Song Illusion; the actual speech signal is not altered in any way. Indeed, Deutsch et al. (2011) failed to elicit a song-like transformation in listeners when altering the speech stimulus. Listeners heard ten presentations of the phrase "sometimes behave so strangely" as either untransformed (i.e., no change made to the stimulus), jumbled (i.e., the syllables were arranged in random order from one presentation to the next), or transposed (i.e., the semitone of the stimulus was raised or lowered from one presentation to the next). In the two altered conditions, listeners heard an untransformed phrase on the initial and final presentations with altered presentations intervening. After each presentation, participants rated the stimuli on a scale from 1 (sounds like speech) to 5 (sounds like song), but only the initial and final presentation ratings were analyzed. The untransformed condition produced higher song-like ratings than the transposed and jumbled conditions on the final presentations, with no significance difference between ratings on the initial presentations of each condition (Deutsch et al., 2011). Physically altering the stimulus results in a bottom-up change of processing, whether by re-arranging the syllables or adjusting the tone. However, because the Speech-to-Song Illusion occurs due to top-down processing, these altered stimuli do not result in as great of a song-like transformation as hearing an untransformed phrase.

The Speech-to-Song Illusion has been examined in German (Falk & Rathcke, 2010) and in Mandarin Chinese (Zhang, 2011) demonstrating that the illusion is not specific to the single English phrase used in Deutsch et al. (2011). Also, in English, the use of other phrases than “sometimes behave so strangely” was examined in Tierny, Dick, Deutsch, & Sereno (2013). The phrases used as stimuli in all of the experiments vary considerably, for example, in the number of words, the number of syllables, and the meaning of the words. Although this variability in the stimuli demonstrates the robustness of the illusion, this variability makes it difficult to determine which characteristics of the stimuli successfully elicit the Speech-to-Song Illusion.

The transition in perception from speech to song in the Speech-to-Song Illusion suggests that overlapping neural regions may be involved in the processing of language and music. Neuroimaging data suggests that several regions of the brain associated with speech processing are also activated while listening to musical stimuli. Tierney et al. (2013) conducted an fMRI study in which participants listened to phrases that they reported as either sounding more like speech or sounding more like song after repetition, even though all phrases were spoken. Six regions were more activated for the song-like stimuli than the speech-like stimuli: anterior superior temporal gyrus bilaterally, right midposterior superior temporal gyrus, right lateral precentral gyrus, middle temporal gyrus bilaterally, left supramarginal gyrus, and left inferior frontal gyrus. These regions are associated with pitch processing, vocalization, and auditory-motor integration, suggesting that spoken phrases perceived as song activate regions of the brain tied to music processing (Tierney, Dick, Deutsch, & Sereno, 2013).

The previous research on the Speech-to-Song Illusion in multiple languages and the neurophysiological evidence establish the generalizability of and a neurological foundation for the Speech-to-Song illusion. However, few studies have examined the parameters of the stimulus that are crucial for eliciting the Speech-to-Song Illusion. Furthermore, there are no known accounts of the cognitive mechanisms that underlie the Speech-to-Song Illusion. The present study will explore some of the stimulus parameters that contribute to the Speech-to-Song Illusion, and attempt to account for the illusion in terms of a cognitive mechanism found in Node Structure Theory (NST; MacKay, 1987). The cognitive mechanism provided by NST underlies several other aspects of language, so an account in NST would be parsimonious because the same cognitive mechanism would also account for the Speech-to-Song Illusion.

Intersection of Music and Speech

The similarities between music and speech continue to draw attention in the psychological, linguistic, music, and even evolutionary literature despite differences in the components of language and music (e.g., words and notes, respectively; Jackendoff, 2009). Recent evidence suggests that for humans the learning of music impacts speech processing, and vice versa, (Asaridou & McQueen, 2013; Christiner & Reiterer, 2013), which can be beneficial when learning foreign languages (Ludke, Ferreira, & Overy, 2014). However, for tone languages, Mok & Zuo (2012) suggest that the benefit of music training is limited to non-tone language speakers only, with no benefit for native tone speakers. The vast array of research suggests that some processing is shared between music and speech and some processing is distinct, but determining the exact boundary of overlap between music and speech continues to be contested.

Language consists of sentences with syntax, phrases, and words, whereas music consists of melodies, sequences of beats, and notes (Jackendoff, 2009). Although the component pieces are not identical, similar cognitive processes occur in both language and music. For example, rhythm in music emerges from the repetition of groupings or segments of the musical stream, while rhythm in speech comes from intonational phrasing (Jackendoff, 2009). Poetry, for instance, brings together aspects of music in a speech context by mapping rhyme and intonational phrasing of language onto the recursive, rhythmical structure of music (Jackendoff, 2009; Christiner & Reiterer, 2013). The use of rhythm and recursion allow for music-like qualities to apply to language. Repetition of spoken words, and more specifically the syllables of those words, may produce a rhythmic, music-like quality in listeners.

Rhythm and beat processing is not unique to humans. In an attempt to determine the evolutionary origins of music, comparative studies have indicated that other nonhuman animals can entrain to musical beats in varying degrees. Beat perception and synchronization has been examined in a sulfur-crested cockatoo (Patel, Iversen, Bregman, & Schulz, 2009), a California sea lion (Cook, Rouse, Wilson, & Reichmuth, 2013), and chimpanzees (Hattori, Tomonaga, & Matsuzawa, 2013), among other nonhuman animals. Of particular interest, a bonobo was also able to synchronize to the drum beats of a human at a preferred beat tempo of 280 beats per minute (Large, 2014). Large (2014) suggests that this preferred tempo of the bonobo falls within the range of syllable production rates found in human languages. These comparative analyses provide additional evidence of an evolutionary history of beat perception, and even a possible link to linguistic rhythm in humans.

With the use of neuroimaging techniques, researchers can examine the amount of overlap in speech and music processing; that is, which brain locations are domain-specific (i.e., only apply to music or speech) and which are domain-general (i.e., apply to both music and speech). For instance, some research suggests hemispheric differences between music and speech (Zatorre, Belin, & Penhune, 2002), but other research shows that certain processes, and therefore brain regions, are common to both music and speech (Koelsch et al., 2002). For instance, particular structures in the brain often associated with language-specific tasks (e.g., Broca's and Wernicke's areas) can be activated during music processing (Koelsch et al., 2002; Maess, Koelsch, Gunter, & Friederici, 2001). Speech and music processing may not be as different and as hemispherically distinct as once speculated. Particularly, the Speech-to-Song Illusion suggests that at least some brain locations are shared by speech and music (Tierney et al., 2013), suggesting some domain-general processing for the perception of speech and music.

The Speech-to-Song Illusion is a unique illusion in which speech transforms into sounding like song. Importantly, the transformation emerges from the repetition of a phrase over and over, quite similar to another auditory illusion, the Verbal Transformation Effect (VTE). VTE has been extensively studied with many attempts at identifying its underlying cognitive mechanism (Warren & Gregory, 1958; Swales & Evans, 1967; Zuck, 1992; Snyder, Calef, Choban, & Guller, 1993; Kaminska, Pool, & Mayer, 2000; Pitt & Shoaf, 2002). In particular, the proposal by MacKay, Wulf, Yin, & Abrams (1993) is the most comprehensive, and appeals to mechanisms that are also employed in other cognitive processes, therefore the NST approach will be examined as a possible explanation of the Speech-to-Song illusion. At present, an underlying cognitive mechanism of the Speech-to-Song Illusion has not been proposed. In the studies reported here we will examine the mechanism proposed by NST for the VTE to see if the same mechanism might also account for the Speech-to-Song illusion.

Verbal Transformation Effect

The Verbal Transformation Effect (VTE) was first described by Warren & Gregory (1958). This auditory illusion occurs when a single word is repeated over and over and over. As a target word plays repeatedly, a person may indicate hearing words other than the actual target word that was presented. For example, the word BASE repeats continuously. Perceptually, however, a person begins to hear other similar sounding words, like *case* and *face*, and even nonwords, like *dace* (MacKay, Wulf, Yin, & Abrams, 1993). Again, the stimulus never changes; a single word repeats over and over. However, listeners perceive a transformation from one word to another with continued repetition. Thus, VTE relies upon stimulus repetition to produce word changes, or transformations (Warren & Gregory, 1958; Warren, 1961; MacKay et al., 1993).

The rapid repetition of a word elicits the VTE in listeners. What appear to be slight modifications in the presentation of the stimulus (e.g., slowing presentation rate of the word) actually can diminish the occurrence of the VTE. Longer stimuli (e.g., a short sentence) did not produce as many word transformations as shorter stimuli (e.g., a single word; Warren, 1961). Also, as the time between repetitions increased from 0.2 s to 0.9 s to 4 s, the number of word transformations decreased (Warren, 1961). Both manipulations decreased the number of perceived transformations. Time, as manipulated by adding more phonetic elements or longer pauses between repetitions, impacts the elicitation of the VTE.

Kaminska & Mayer (2002) also found that other types of stimuli transform with repetition, like music (e.g., well-known classical music) and everyday sounds (e.g., a coin dropping). Listeners indicated any perceptual transformation during four minutes of repetition for three categories of stimuli: words, musical motifs, and sounds. Participants' responses were classified as either identity (i.e., a distinctly new stimulus is perceived) or non-identity (i.e., changes in the surface characteristics of the original stimulus) transformations. All three categories of stimuli elicited more identity transformations than non-identity transformations, with no significant difference between stimulus categories in the total number of transformations (i.e., identity and non-identity transformations combined). These findings suggest that transformations can occur for stimuli that do not contain linguistic information. Particularly, the identity transformations of words in Kaminska & Mayer (2002) would be classified as verbal transformations in a typical VTE experiment. Interestingly, the number of word identity transformations did not significantly differ from the number of music or sound identity transformations. Thus, the rate of transformations in non-linguistic stimuli occurs at the same rate as that for words. The repetition of a constant stimulus produces the perceptual illusion in any repetitive auditory stream.

Theories of the Verbal Transformation Effect

Two competing theories of the underlying mechanism of the VTE exist: habituation and connectionist models. Both theories account for how a continuously repeated stimulus disrupts perceptual processing. Habituation, in particular, results in a decline in responding to a stimulus with continued exposure (Snyder, Calef, Choban, & Guller, 1993). More specific to the VTE, habituation theory suggests that as a word repeats continuously, the listener habituates, or responds less to that word, allowing other words to be perceived instead (Snyder et al., 1993).

Snyder, Calef, Choban, & Guller (1993) suggest that words that require minimal processing, or "non-arousing" words, are more susceptible to habituation effects and elicit more transformations. With an increase in arousal (either psychologically or physiologically), the number of transformations should decrease. For instance, Calef, Calef, Kesecker, & Burwell (1974) found that taboo words had fewer verbal transformations than neutral words when given in a typical VTE task. The increase in arousal elicited by taboo words warded off the effects of habituation by allocating more attention to that word, and thereby reducing the number of perceived verbal transformations. Snyder et al. (1993) also predicted that as a word continues to repeat for an extended period of time that word becomes monotonous and less arousing. When given a repeating word for 6 min, participants indicated experiencing more verbal transformations during the second 3 min than the first 3 min. As time passed, the stimuli became less arousing and increased the number of verbal transformations. In sum, habituation theories propose that "arousal" is the underlying mechanism of the VTE.

Node Structure Theory (NST), on the other hand, is a connectionist model similar to other spreading activation theories that is used to describe processes of perception (e.g., speech perception) and action (e.g., language production; MacKay, 1987). NST has been used to account for accurate memory and language processes (e.g., word retrieval and production; MacKay, 1987), dysfunctional processes (e.g., tip-of-the-tongue states; Burke, MacKay, Worthley, Wade, 1991), and differences in processes due to aging (e.g., MacKay & Burke, 1990) and cognitive deficits (e.g., amnesic-patient H.M.; MacKay, Stewart, Burke, 1998), among other cognitive processes. In NST, nodes and links are used to organize the mental lexicon (i.e., that part of memory that stores all the words a person knows). A node represents a piece of information (e.g., a phoneme or a word), and links connect related nodes together. The network of nodes is organized into different levels, like the sentential system and the phonological system, and the nodes are linked within systems and across systems (MacKay, 1987). For example, the node for the word *frisbee* would be connected to the phonological node /f/ and the semantic node “frisbees are thrown”. The interplay of nodes within a level and across different levels results in processing throughout the mental lexicon.

General top-down processing through the network of nodes allows for language production, while general bottom-up processing allows for speech perception. The movement of processing through the network occurs by priming and activating nodes. To become activated, a node must receive enough priming from across its connections to surpass an activation threshold (MacKay, 1987). Once activated, the node sends priming across all of its connections, and the process continues (MacKay, 1987). The continued priming and activation of nodes allows a listener to retrieve spoken words from an auditory stimulus in speech perception. The sound signal activates phonological nodes, which then activate a connected lexical node. Activation of a lexical node gives the listener conscious awareness of that particular word.

A third process, node satiation, is a critical component to the VTE. When a node satiates, the activity level of the node falls below its normal resting level (MacKay, 1987). This decrease in activity level makes it very difficult for that node to become sufficiently primed and activated. Instead, other nodes will activate through the most-primed-wins principle; within a given pool of possible nodes to become activated, the node with the most priming will win and activate (MacKay, 1987; MacKay et al., 1993). The lowered activity level of a satiated node (e.g., BASE) makes it difficult for that node to become the most primed, allowing other nodes to activate instead (e.g., *case, face)*. Words like *case* and *face* are within the pool of possible nodes to activate as they share some primed phonological nodes. Through satiation and the most-primed-wins principle, verbal transformations emerge.

In contrast to NST, spreading activation theories propose the underlying mechanism of the VTE lies in the spread of activation through a network of nodes. A target node, representing the repeated word, has connections with semantically and phonologically related nodes (Kaminska, Pool, & Mayer, 2000). Once the target node is activated, activation continues to spread to all other connected nodes. However, as the repetition continues, there is a reduction in activation threshold of all the repeated word's connected nodes (Kaminska et al., 2000). This reduction allows the connected nodes to more easily activate than the repeated word.

As activation spreads through the network, a node with many connections will have more opportunities for transformations than a node with fewer connections (Kaminska et al., 2000). Although Kaminska, Pool, & Mayer (2000) suggest that stimuli with more connections will elicit more verbal transformations, this prediction was not supported by the data from their study. Participants listened to words that varied in visual imagery (high vs low) and the number of syllables (1 vs 3), such that words high on visual imagery will have more semantic connections than words low on visual imagery, and words with more syllables will also have more phonological connections than words with fewer syllables. No significant effect was found for either imagery or syllable length on the total number of verbal transformations. Semantic connectedness may not be crucial to the VTE, and syllable length may have a different impact than what Kaminska et al. (2000) predict. For instance, manipulating the number of semantic connections (i.e., visual imagery) may not impact verbal transformations since a satiated lexical node cannot pass priming onto that higher-level information. Also, increasing the number of syllables, although located in the lower-level phonological system, may actually decrease verbal transformations. The lexical node appears to be the locus for eliciting verbal transformations. Therefore, increasing lexical connectivity may be a better predictor of increasing verbal transformations than increasing semantic or phonological nodes. Indeed, Yin & MacKay (1992) found that increasing phonological neighborhood density (or the number of lexical nodes of a target word that differ by one phoneme) results in an increase in the number of perceived verbal transformations.

In NST repetition of a stimulus (in this case, a word) results in satiation of a node allowing other nodes to activate instead, providing an explanation of the VTE. Given the important role that the repetition of the stimulus plays in the Speech-to-Song Illusion, perhaps the same underlying mechanism found in NST (i.e., satiation of a lexical node, and activation of other connected nodes) could be used to account for the Speech-to-Song Illusion. However, the Speech-to-Song Illusion transforms an entire phrase rather than a single word. Repetition may satiate the lexical node leaving the syllabic structure of the word accessible and more prominent. The repeated activation of the syllable nodes may maintain the metrical pattern found in the phrase, and produce the song-like quality experienced by listeners.

In the following studies I will examine if the satiation of nodes in NST used to explain the VTE can also account for the Speech-to-Song Illusion. If repetition of words in a phrase leads to satiation of the lexical nodes, but maintains activation of the syllable nodes (producing a metrical, music-like percept), then repetition of a meaningful phrase should not be required to elicit the Speech-to-Song Illusion. To test this hypothesis I will construct lists composed of four words, thereby eliminating syntactic information, to elicit the Speech-to-Song Illusion in Experiment 1. In Experiment 2, I will use nonwords to eliminate semantic information (in addition to grammatical information) from the repeated phrase. In Experiment 3 and 4, I will examine how lexical nodes can recover from satiation with time and diminish the experience of the perceptual illusion. In Experiment 3, I will manipulate the time between repetitions by increasing the number of words per stimulus, with word-lists ranging from one to ten words. In Experiment 4, I will manipulate both the number of words and the number of syllables in each word-list. Examining these different parameters of the Speech-to-Song Illusion will provide support for the underlying cognitive mechanism of this perceptual illusion.

Experiment 1

Previous research only used grammatical phrases taken from actual sentences (e.g., "sometimes behave so strangely"; Deutsch et al., 2011) to elicit the Speech-to-Song Illusion. However, if NST provides an account of Speech-to-Song Illusion, then the repetition of any four words should elicit the song-like transformation. During the repetition of four words the lexical nodes become satiated, eliminating the need for higher-level information (e.g., semantic and syntactic node activation). What becomes important is the repetition of beat-like syllables, not the semantic or grammatical meaning of the phrase. As the syllable nodes repeat over and over, a beat-like quality emerges. Any combination of words in a phrase should result in the song-like percepts experienced by listeners, thus making a meaningful, grammatical phrase unnecessary in eliciting the Speech-to-Song Illusion.

NST predicts that as a single word is repeated continuously, the lexical node corresponding to that word satiates and can no longer become activated. Other similar nodes, which share priming, become activated instead allowing for the perception of verbal transformations. Similarly, NST would also predict that the lexical nodes corresponding to each word in the phrase would satiate with repetition of the entire phrase. Higher-level nodes in the sentential system (e.g., syntactic nodes) would not impact the Speech-to-Song Illusion, since priming ceases to pass from the lexical node without activation. However, the lower-level nodes in the phonological system, particularly the syllable nodes, would continue to activate with repetition. As the syllable nodes repeat over and over, a beat-like percept emerges. Given the length of a typical four-word phrase, enough beats are prominent to elicit the Speech-to-Song Illusion.

Rather than use 4 randomly selected words in each word-list, I will also examine the prediction of spreading activation theories that having more connections increases the number of verbal transformations (Yin & MacKay, 1992; Kaminska et al., 2000). Different from Kaminska et al. (2000), another method of measuring the number of connections will be used, namely phonological neighborhood density, which is a lexical measure of connectivity (in contrast to connected semantic or phonological nodes). Phonological neighborhood density is the number of words, or neighbors, that differ by one phoneme through addition, subtraction, or substitution from a target word (e.g., BATTLE has neighbors *cattle*, *bat*, *beetle*; Vitevitch, Stamer, & Sereno, 2008). In the present experiment, word-lists will contain either dense words (i.e., words that have many neighbors) or sparse words (i.e., words that have few neighbors). In regards to the VTE, Yin and Mackay (1992) found that dense words elicited more verbal transformations than sparse words during a typical VTE task. I also predict that dense words will elicit the Speech-to-Song Illusion more so than sparse words. Dense words, with more connections, will satiate more quickly than sparse words due to less priming going to each connecting node (e.g., with two connections priming is divided in half, with five connections priming is divided in fifths). Thus, I predict that repeated word-lists will elicit stronger Speech-to-Song ratings than non-repeating word-lists, particularly with the dense word-lists.

Method

Participants

Ninety-five undergraduate Psychology students enrolled at the University of Kansas received partial course credit for their participation. None of the participants reported any speech or hearing disorders, or participated in the other experiments reported in this paper. All of the participants were native English speakers. Five of the participants failed to follow the instructions of the task; their data were excluded from analyses.

Materials

The 56 bisyllabic words used in the present experiment were the same stimuli used in Vitevitch, Stamer, & Sereno (2008). These items were recorded by a female, native English speaker at a normal speaking rate. Words were recorded in an IAC sound-attenuated booth using a high-quality microphone onto a digital recorder at a sampling rate of 44.1 kHz. The words were edited into individual sound files using Sound Edit 16 (Macromedia, Inc.).

The words were divided equally into two conditions: words with dense and words with sparse phonological neighborhoods. Phonological neighborhood density was determined by counting the number of phonological neighbors (i.e., words that differed by the addition, substitution, or deletion of one phoneme). As reported in Vitevitch et al. (2008), dense words had a mean of 11.71 (SD = 1.58) phonologically similar words and sparse words had a mean of 4.43 (SD = 1.99) phonologically similar words. The difference between the two conditions was statistically significant, F (1, 54) = 229.88, p < .0001. Although the words differed in neighborhood density, they were controlled for word frequency, neighborhood frequency, phoneme length, and uniqueness points (all ps > .40).

The twenty-eight dense words were grouped into 7 lists, such that each list consisted of four words and each word was used only once. The twenty-eight sparse words were also grouped into 7 lists in a similar manner. Word-lists were matched between dense and sparse conditions in phoneme onset of each word (see Appendix A for word-lists). Audacity 2.0.2 digital audio editor was used to concatenate the four separate sound files for each of the words into a single sound file. No additional time was included at the beginning or end of each sound file in order to mimic a natural speaking rate. Participants listened to both the dense and sparse word lists.

As a between-subjects factor, there were three repetition conditions: no repetition, repetition with pausing, and repetition without pausing. In the no repetition condition, participants heard each word-list played once as in Deutsch et al. (2011). Also, in Deutsch et al. (2011) stimuli were repeated ten times with a 2300 ms pause between repetitions. This extended pause was given to allow enough time for participant ratings after each presentation, but only the ratings after the first and last presentation were analyzed. In the present study, a shorter pause was used as participants only rated stimuli after all ten repetitions. In the repetition with pausing condition, participants heard the word-lists repeated ten times with a 750 ms pause between each repetition. A final condition, not examined by Deutsch et al. (2011), was the repetition without pausing condition. Participants heard each list repeated 10 times consecutively with no pause between repetitions. For the repetition conditions, Audacity 2.0.2 digital audio editor was used to produce a single sound file containing all the repetitions with or without pauses. There was no significant difference in total stimulus duration between the dense and sparse word-lists in each of the three repetition conditions (all $ps > .25$). Participants only participated in one of the three repetition conditions.

Participants were tested in groups of up to three and seated individually at an iMac computer running PsyScope 1.2.2 (Cohen, MacWhinney, Flatt, & Provost, 1993). Participants wore a set of Beyerdynamic DT 100 headphones and used a computer keyboard to indicate their ratings. PsyScope controlled stimulus presentation, played recordings to participants, and collected their responses.

Procedure

Participants were asked to listen to and rate each stimulus as sounding like speech or sounding like song on a scale from 1 (sounding like speech) to 5 (sounding like song); this was the same rating scale used in Deutsch et al. (2011). Higher ratings on the scale indicate experiencing a song-like percept, whereas lower ratings on the scale indicate perceiving the stimulus as sounding more like normal speech.

Participants were prompted with the word READY for 500 ms on the screen before the stimulus was presented over the headphones. After the stimulus was done playing, participants were prompted with “>” and typed in their rating using the computer keyboard. Participants then pressed the “return” key, which initiated the next trial. Participants heard each of the 14 stimuli (7 lists of dense words and 7 lists of sparse words) in a randomized order.

Results

Participant ratings were analyzed in a 2 (density: dense and sparse) x 3 (repetition: no repetition, repetition with pauses, and repetition with no pauses) mixed ANOVA. Phonological neighborhood density was a within-subjects condition (i.e., participants heard both dense and sparse word-lists). Repetition was a between-subjects condition (i.e., participants were only in one of the three repetition conditions). A significant interaction between word density and repetition condition was found, F (2, 87) = 3.47, $p < .05$ (see Figure 1). The effect size for the interaction (η^2) was .06, which is considered a small effect size according to Fay & Boyd (2010) and a medium effect size according to Cohen (1988). Because main effects are not interpretable with a significant interaction, simple effects were instead calculated to further explore how word density and repetition influenced the Speech-to-Song Illusion.

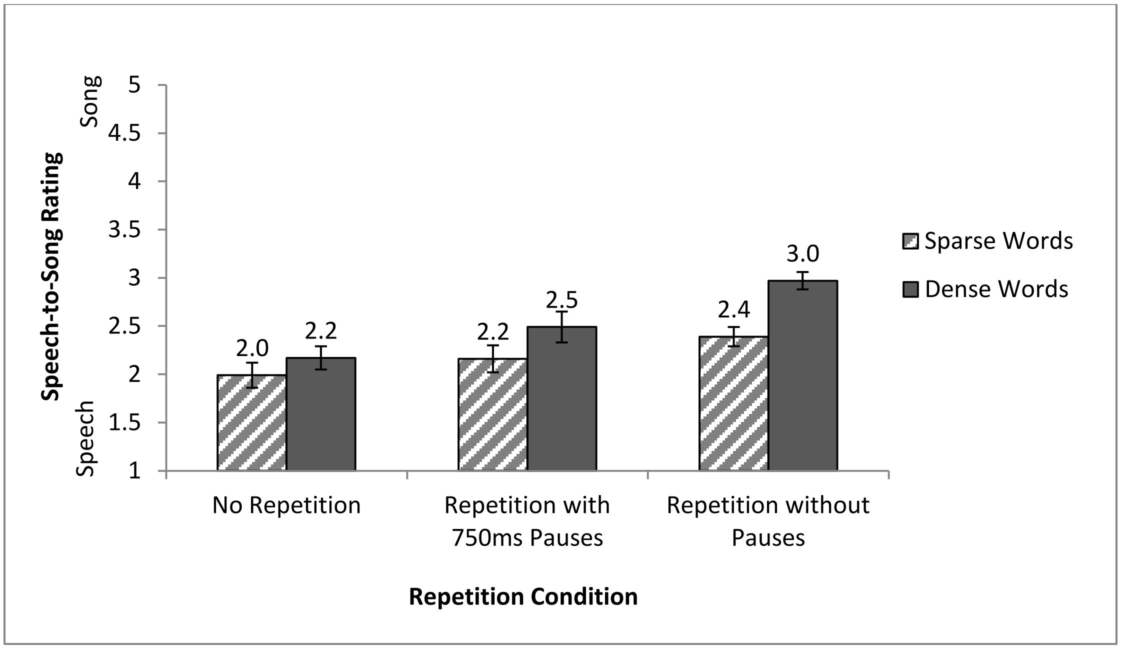

Figure 1. Ratings of dense and sparse word-lists on a scale from 1 (sounding like speech) to 5 (sounding like song) in three repetition conditions (no repetition, repetition with pauses, and repetition without pauses). Mean ratings for each condition are listed above each column, with standard error of the mean bars.

The simple effect of word density was examined at each level of repetition condition. These pairwise comparisons have been adjusted for multiple tests through a Bonferroni correction. For the no repetition condition, the mean difference between the average rating for dense word-lists (M = 2.17, SD = 0.68) and the average rating for sparse word-lists (M =1.99, SD = 0.71) was not significant, p = .09. For the repetition with pauses condition, the mean difference between the average rating for dense word-lists (M =2.50, SD = 0.86) and the average rating for sparse word-lists (M =2.16, SD = 0.74) was significant, $p < .01$. Also, for the repetition without pausing condition, the mean difference between the average rating for dense word-lists (*M* =2.97, *SD* = 0.51) and the average rating for sparse word-lists (*M* =2.39, *SD* = 0.55) was significant, $p < .001$. In the two repetition conditions, dense word-lists were rated as sounding more like song than sparse word-lists. Without repetition, phonological neighborhood density did not impact Speech-to-Song ratings.

The simple effect of the repetition condition was examined at each level of word density. These pairwise comparisons have been adjusted for multiple testing through a Bonferroni correction. For dense words, the no repetition condition is not significantly different from the repetition with pauses condition, $p = .23$, but is significantly different from the repetition without pauses condition, $p < .001$. Also, the repetition with pauses and the repetition without pauses conditions are significantly different, $p < .05$. Repetition without pauses elicits the strongest song-like percepts with dense word-lists. For sparse words, the no repetition condition is not significantly different from both the repetition with pauses condition, $p = 1.00$, and the repetition without pauses condition, $p = .07$. Also, the repetition with pauses and the repetition without pauses conditions are not significantly different, $p = .55$. There is no difference in Speech-to-Song ratings amongst the repetition conditions for sparse words. Higher word density elicits greater song-like perception in listeners, especially after rapid repetition.

Discussion

The results of the present experiment showed several important things. First, a word list composed of four unrelated words produces the Speech-to-Song Illusion in listeners, suggesting that a grammatical phrase, like “sometimes behave so strangely” (Deutsch et al., 2011), is not necessary to elicit the illusion. The phrase used by Deutsch et al. (2011) was taken from an actual, syntactically meaningful sentence. However, the results of the present experiment suggest that elicitation of the Speech-to-Song Illusion does not require the stimulus to convey meaning or to be syntactically well-formed.

Second, repetition plays a key role in eliciting the Speech-to-Song Illusion, especially repetition without a pause. Participants that heard the stimuli repeated 10 times without a pause rated the stimuli as sounding more song-like than participants who heard the stimuli presented only once. However, the repetition condition with a 750 ms delay diminished the perception of the Speech-to-Song Illusion. Having no pause between repetitions produced the strongest song-like percepts. With enough time, satiated nodes can recover and be successfully retrieved in NST. Perhaps this 750 ms delay provides enough time to elapse to allow the nodes to recover somewhat from satiation, thereby diminishing the Speech-to-Song Illusion in the delay condition.

Lastly, the dense words (i.e., with more connections in NST) increased the song-like percepts with repeated stimuli more than sparse words (i.e., with fewer connections in NST). As the number of connections increases, the amount of priming transmitted at a given time is reduced due to a division of priming across all connections. The dense words satiate more quickly than sparse words allowing for the beat-like rhythm of the syllable nodes to emerge. Interestingly, the ratings of dense and sparse word-lists did not differ in the "no repetition" condition. Without repetition, the lexical nodes for each word do not satiate, allowing for accurate word retrieval.

With rapid repetition, a word-list can elicit the Speech-to-Song Illusion. Also, having more phonologically-related lexical connections increases the perception of song in listeners. As lexical nodes satiate, priming fails to pass onto higher-level nodes, thus sentential nodes (e.g., syntactic nodes) do not seem to impact the Speech-to-Song Illusion. Rather, the repeated activation of syllable nodes seems to be the source of song-like percepts. To further explore the idea that syntactic and sentential nodes are not critical to the Speech-to-Song Illusion, the use of nonwords in the repeated words-lists will be examined. Nonwords do not have syntactic or semantic node connections, allowing for an examination of stimuli with only phonological nodes.

Experiment 2

In Experiment 1 a word-list composed of four words elicited the Speech-to-Song Illusion in listeners, demonstrating that a meaningful phrase was not required to produce the illusion. This finding suggests that higher-level nodes in NST, such as sentential nodes, do not impact the elicitation of the Speech-to-Song Illusion. To further explore the necessity of sentential nodes I used nonwords as stimuli in the present experiment. The use of nonwords not only eliminates grammatical information from the phrase, but also semantic information. Although NST would suggest that priming cannot be transmitted to sentential nodes of a satiated lexical node (i.e., only with activation does priming spread), choosing stimuli without syntactic and semantic nodes ensures that the aforementioned nodes cannot impact the Speech-to-Song Illusion. This yields a strong test of the hypothesis that the Speech-to-Song Illusion emerges from repetition leading to the prominence of lower-level phonological nodes, particularly the syllable nodes.

Since the phonological nodes are of interest in the present experiment, one method of examining them is through the manipulation of phonotactic probability (PP). Phonotactic probability is the frequency of phonological segments and sequences of segments that occur in words in English (Vitevitch & Luce, 2005). Nonwords with high-PP contain sequences of phonological nodes whose connections are more likely to already be connected and be stronger than nonwords with low-PP. With a stronger pathway of connections, priming transmits faster and in greater amounts from node to node (MacKay, 1987) allowing high-PP nonwords to be retrieved more quickly and easily than low-PP nonwords (e.g.,Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Vitevitch & Luce, 1998, Vitevitch & Luce, 1999; Vitevitch & Luce, 2005). Potentially, the low-PP nonwords would elicit stronger transformation effects than the high-PP nonwords since they are more difficult to access and retrieve. Thus, I predict that repeated nonword-lists will also elicit song-like percepts in listeners, particularly with low-PP nonword-lists.

Method

Participants

Sixty undergraduate Psychology students enrolled at the University of Kansas received partial course credit for their participation. None of the participants reported any speech or hearing disorders, or participated in any of the other experiments reported in this paper. All of the participants were native English speakers.

Materials

Fifty-six nonwords (28 high-PP and 28 low-PP) from Vitevitch and Luce (2005) were used in this experiment. All of the nonwords were recorded by a male, native English speaker at a normal speaking rate. Stimuli were recorded in an IAC sound-attenuated booth using a high-quality microphone and recorded onto a digital audiotape at a sampling rate of 44.1 kHz. The nonwords were edited into individual sound files using Sound Edit 16 (Macromedia, Inc.).

The nonwords were divided into two conditions: low phonotactic probability and high phonotactic probability. Two measures of phonotactic probability were computed according to Vitevitch and Luce (1998): the sum of the segments and the sum of the sequences of segments.

The sum of the segments for high-PP nonwords ($M = .16$, $SD = .03$) was significantly greater than the sum of the segments for low-PP nonwords ($M = .09$, $SD = .02$), $F(54) = 10.12$, $p < .0001$. The sum of the sequences of segments for high-PP nonwords ($M = .007$, $SD = .005$) was also significantly greater than the sum of the sequences of segments for low-PP nonwords ($M = .001$, $SD = .001$), $t(54) = 5.81$, $p < .0001$. All of the nonwords were 3 phonemes in length, with an equal number of onset phonemes in the high- and low-PP nonword-lists.

The twenty-eight high-PP nonwords were grouped into seven lists, such that each list consisted of four nonwords, and each nonword was only used once. The twenty-eight low-PP nonwords were also grouped into 7 lists in a similar manner. Nonword-lists were matched between high-PP and low-PP conditions in the phoneme onset of each nonword (see Appendix A for nonword-lists). Audacity 2.0.2 digital audio editor was used to concatenate the four separate sound files for each of the nonwords into a single sound file. No additional time was included at the beginning or end of each sound file in order to mimic a natural speaking rate. Participants listened to both the high-PP and low-PP nonword-lists.

There were only two repetition conditions in the present experiment: no repetition and repetition (with no pause). In the no repetition condition, participants heard each list once. In the repetition condition, participants heard each list repeated 10 times consecutively. Audacity 2.0.2 digital audio editor was used to combine the four nonwords into lists, and to produce a single sound file containing all the repetitions, if any. There was no significant difference in the total stimulus duration between high-PP and low-PP nonword-lists in each of the repetition conditions (all *p*s > .80). Participants only participated in one of the two repetition conditions.

Procedure

The same equipment and procedure used in Experiment 1 was used in the present experiment. Participants heard each of the 14 stimuli (7 high-PP and 7 low-PP nonword-lists) in a randomized order.

Results

Participant’s ratings were analyzed in a 2 (PP: low and high) x 2 (repetition condition: no repetition and 10 times repetition) mixed ANOVA. PP was a within-subjects factor and repetition condition was a between-subjects factor. Neither the interaction between PP and repetition condition (F (1, 58) = 0.13, $p > .05$), nor the main effect of PP (F (1, 58) = 0.10, $p > .05$) were significant. However, a main effect of repetition condition was significant, F (1, 58) = 13.36, $p < .001$ (see Figure 5). For the main effect of repetition condition, the effect size (η^2) was .23, which is considered a large effect size according to Fay & Boyd (2010) and Cohen (1988). Participants who heard the repeated stimuli rated the nonword-lists as more song-like than participants who heard the stimuli played only once.

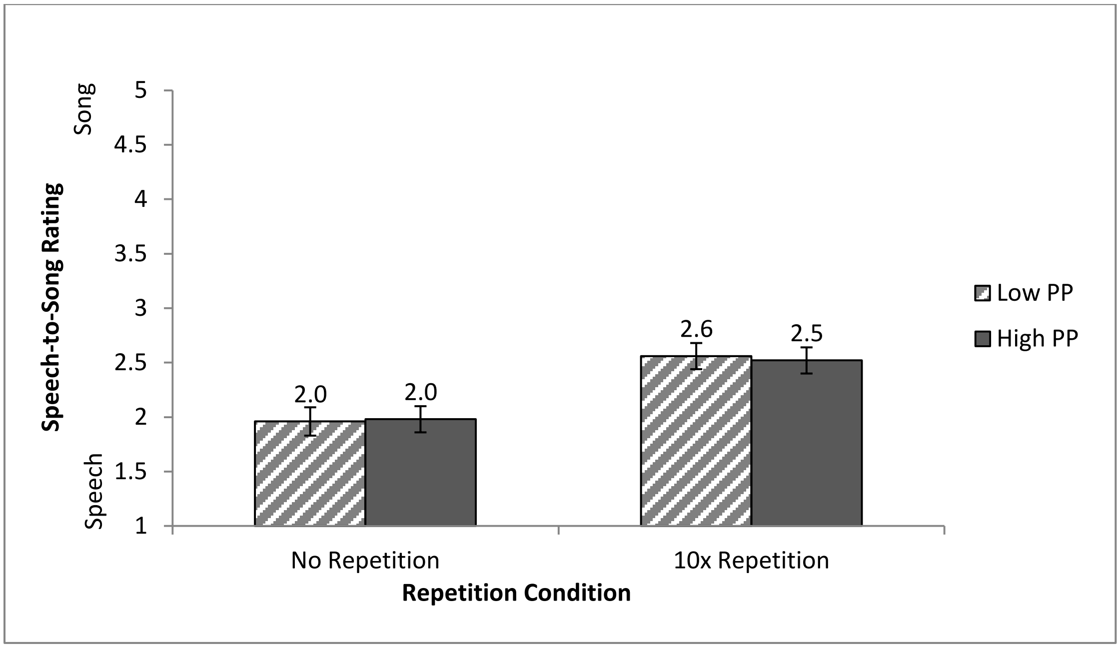

Figure 2. Speech-to-Song Rating of Low and High Phonotactic Probability (PP) word-lists with and without Repetition. Mean ratings for each condition are listed above each column, with standard error of the mean bars.

Discussion

The results of the present experiment further indicate that repetition is a key component in eliciting the Speech-to-Song Illusion. Somewhat surprisingly, phonotactic probability, which has been shown to influence a variety of language related processes (e.g., Vitevitch et al., 1997; Vitevitch, Armbruster, & Chu, 2004; Vitevitch & Luce, 2005), did not seem to influence the elicitation of the Speech-to-Song Illusion. The failure to observe an influence of PP on the Speech-to-Song Illusion may be due to a number of reasons. For example, nonwords likely do not have a lexical node representation in the sentential system. Unlike in Experiment 1 where a lexical measure of neighborhood density impacted perception of the Speech-to-Song Illusion, a sublexical measure of phonotactic probability was not beneficial or harmful. Even though low-

PP nonwords would have weaker connections in the phonological system than high-PP nonwords, the syllable nodes of all nonwords were still accessed. And, as the repetition of the syllable nodes continued, the percepts of listeners transformed from sounding like speech to sounding more song-like.

Future research could examine at what point listeners begin hearing the Speech-to-Song Illusion when listening to lists of nonwords. Potentially, if the difference between high-PP nonwords and low-PP nonwords is a matter of the strength of connections amongst phonological nodes, the high-PP nonword-lists might elicit the Speech-to-Song Illusion *sooner* than low-PP nonword-lists. The syllable node for high-PP nonwords would be activated sooner than the syllable node for low-PP nonwords, since priming would spread more quickly amongst the phonological nodes. As soon as the syllable nodes are activated, the rhythmical beat of the syllables would become more evident to listeners. Thus, measuring the time at which the illusion first appears, rather than the strength of the illusion (i.e., ratings of song-like perception), may be a better way of examining the effect of phonotactic probability on the Speech-to-Song Illusion.

Experiment 3

Experiment 1 found no difference between hearing the stimulus played only once and hearing the stimulus with a 750 ms pause between repetitions. Although it is difficult to interpret a null-result, this finding suggests that adding time between the activations of word-nodes decreases the perception of song. A reduction in song-like quality, or more speech-like perception, as time increases between repetitions resembles a similar finding with the Verbal Transformation Effect. Warren (1961) observed that increasing the time between repetitions from 0.2 s to 0.9 s and from 0.9 s to 4 s decreased the number of word transformations. NST would predict that nodes recover from satiation during that short pause between repetitions. In the present experiment, I created a "filled" pause between repetitions by manipulating the number of words in each list. Lists containing many words will allow more time to pass between subsequent activations of a particular word in that list than lists containing fewer words. The words in longer lists will recover from satiation more quickly than the words in shorter lists, thereby reducing the perception of song. To maximize our chance of observing the Speech-to-Song Illusion, I used only dense words and repeated all stimuli ten times (without a silent pause), as this was the condition from Experiment 1 that elicited the greatest Speech-to-Song ratings. I predict that as the number of words in the list increases, Speech-to-Song ratings will decrease.

Method

Participants

Twenty-nine undergraduate Psychology students enrolled at the University of Kansas received partial course credit for their participation. None of the participants reported any speech or hearing disorders, or participated in the other experiments reported in this paper. All of the participants were native English speakers. One participant failed to follow the instructions of the task; their data were excluded from analyses.

Materials

The 28 dense words used in Experiment 1 were also used for this experiment. The words were grouped into lists that varied in the number of words, from one word up to ten words per list (see Appendix C for word-lists). There were five lists at each length yielding a total of 50 stimuli. Due to the limited number of dense words, the words were repeated across lists. Each list of words was repeated 10 times with no pause. Audacity 2.0.2 digital audio editor was used to combine the individual words into lists and repeat the stimuli to create a single sound file for each word-list.

Procedure

The same equipment and procedure used in Experiment 1 was used in the present experiment. Participants heard each of the 50 stimuli in a randomized order.

Results

A simple linear regression was calculated to determine the relationship between the number of words per stimulus and participants' Speech-to-Song rating. Participants' ratings were aggregated across word-lists of the same length and used to predict Speech-to-Song ratings. The aggregated mean and standard deviation for each word-list length is listed in Table 2. A negative correlation between the number of words per list and participant ratings was found, $R = -.77$, $F(1, 9) = 11.76$, $p < .01$ (see Figure 2). Listeners perceived the stimuli as sounding more like song when there were fewer words per list and more like speech when there were more words per list. Increasing the length of the list reduced the perception of the Speech-to-Song Illusion.

Table 1

Mean and Standard Deviations of Speech-to-Song Ratings for Each Word-List Length

Number of Words per List	Speech-to-Song Rating	
	M	*SD*
1	2.50	1.22
2	2.84	1.04
3	2.78	0.86
4	2.41	0.86
5	2.34	0.79
6	2.26	0.81
7	2.20	0.78

8	2.45	0.82
9	2.13	0.80
10	2.18	0.77

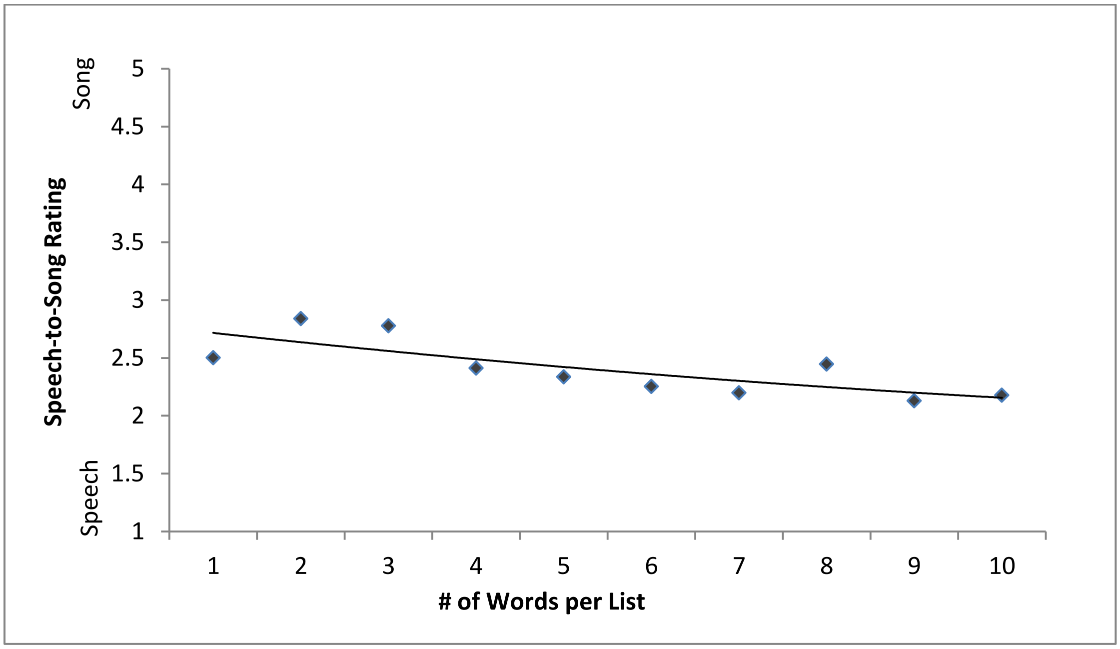

Figure 3. Scatterplot of Speech-to-Song ratings and the number of words per list.

Discussion

Lists with more words were perceived as sounding more speech-like, rather than song-like. Having more words in a list increased the amount of time between subsequent activations of each word in the list. For instance, as a stimulus repeats over and over, the time between subsequently activating the first word in a three-word list is much shorter than the time between subsequently activating the first word in a nine-word list. NST states that satiation diminishes with time, allowing for successful retrieval (MacKay, 1987). The longer word-lists allow enough time to pass for node recovery, thereby reducing the song-like illusion. Although this manipulation of time in the present experiment utilizes "filled pauses" in contrast to a silent pause, the shorter word-lists elicited stronger song-like percepts than longer word-lists.

By increasing the number of words per word-list, the Speech-to-Song Illusion decreased. But increasing the number of words also increases the number of syllables in each word-list. Potentially, increasing the number of syllables may have a different effect on eliciting the Speech-to-Song Illusion than increasing just the number of words, especially since the syllable nodes appear to play a crucial role in eliciting the illusion. To further examine the role of syllable nodes in the elicitation of the Speech-to-Song Illusion, I manipulated in Experiment 4 the number of words per list and the number of syllables in each word.

Experiment 4

As time lapses, nodes recover from satiation and their activity levels rise to normal resting levels allowing for normal processing. Without satiation, lexical nodes would activate successfully and the Speech-to-Song Illusion would disappear. By using "filled" pauses in Experiment 3, increasing the number of words in each word-list reduced listeners' Speech-to-Song ratings. However, in the Speech-to-Song Illusion, lexical nodes satiate while syllable nodes activate. Potentially, increasing the number of syllable nodes activated in each word-list will have a more direct impact in reducing the perception of the Speech-to-Song Illusion than increasing the number of words alone. The present experiment will further examine the impact of satiation on eliciting the Speech-to-Song Illusion by manipulating both the number of words and the number of syllables in each list.

Method

Participants

Thirty undergraduate Psychology students enrolled at the University of Kansas received partial course credit for their participation. None of the participants reported any speech or hearing disorders, or participated in any of the other experiments reported in this paper. All of the participants were native English speakers.

Materials

Eighty bi-syllabic words and thirty quad-syllabic words were used in this experiment. The words were combined into eight lists that varied in the number of words and the number of syllables (see Appendix D for word-lists). For the bi-syllabic words, the lists consisted of 1 word (2 syllables), 2 words (4 syllables), 3 words (6 syllables), 4 words (8 syllables), and 6 words (12 syllables). For the quad-syllabic words, the lists consisted of 1 word (4 syllables), 2 words (8 syllables), and 3 words (12 syllables). There were 5 lists for each list-length, for a total of 40 stimuli. Each word list was repeated 10 times consecutively. All of the words were recorded by the author (a female, native English speaker) at a normal speaking rate in an IAC sound-attenuated booth using a high-quality microphone onto a digital recorder at a sampling rate of 44.1 kHz. The words were edited into individual sound files using Sound Edit 16 (Macromedia, Inc.). Audacity 2.0.2 digital audio editor was used to concatenate the words into a single sound file containing all repetitions.

The bi-syllabic and quad-syllabic words were controlled for phonological density, word frequency, and word familiarity. All words had no phonological neighbors. The log frequency (Kucera & Francis, 1967) for bi-syllabic words (M = 1.20, SD = 0.42) and quad-syllabic words (M = 1.26, SD = 0.44) did not differ, t (108) = -0.57, p = .57. Words were also rated on familiarity using a 7-point scale, with higher ratings indicating greater familiarity (Nusbaum, Pisoni, & Davis, 1984). Bi-syllabic (M = 6.93, SD = 0.16) and quad-syllabic (M = 6.88, SD = 0.30) words did not differ in word familiarity, t (108) = 0.96, p = .34.

Procedure

The same equipment and procedure used in Experiment 1 was used in the present experiment. Participants heard each of the 40 stimuli in a randomized order.

Results

A cross-classified random effects mixed model was used to analyze the effect of the number of words and the number of syllables per stimulus on Speech-to-Song ratings. This particular type of multilevel modeling allows examination of both the variability between subjects, *j*, and the variability between words, *i*. To create the model, the number of words was centered at 1 and the number of syllables was centered at 2 for easier interpretation (the shortest stimulus consisted of 1 word with 2 syllables). The quadratic effect of words and syllables was also examined. Quadratic effects were chosen because the data in Experiment 2 showed a potential curve between one and four words. Including the linear and quadratic predictors will allow a complete model for predicting the Speech-to-Song ratings of participants.

A hierarchical model building procedure and deviance test was used to determine significance of each effect. The fixed, linear effects of the number of words, γ_{10}, and the number of syllables were added, γ_{20}, respectively. Both linear effects were significant, and the quadratic effects were then added. Only the quadratic effect of words, γ_{30}, was significant; the quadratic effect of syllables was excluded from the model. Once determining these fixed effects, the random effects of words, u_{1i}, syllables, u_{2i}, and words-squared, u_{3i}, were included in that order. The final model, $Y_{ij} = \gamma_{00} + \gamma_{10} + \gamma_{20} + \gamma_{30} + u_{0i} + u_{1i} + u_{2i} + u_{3i} + u_{0j} + e_{ij}$, was used to predict the Speech-to-Song rating of individual i on word list j.

All of the fixed effects were significant. The intercept, or the expected average Speech-to-Song rating across participants and items when words and syllables equals one, was 2.26 (*t* =

17.67, $p < .0001$). A significant positive linear trend of the number of words ($t = 6.64$, $p < .0001$) and a significant negative quadratic trend of the number of words was found ($t = -6.21$, $p < .0001$). The linear trend shows that as the number of words increased by one (controlling for the number of syllables), the speech-to-song rating also increased by 0.79. However, the quadratic trend of -0.12 shows that as the number of words increased, the speech-to-song rating became more negative, and eventually became a negative slope (see Figure 3). There was also a significant negative linear trend of the number of syllables ($t = -5.55$, $p < .0001$) such that as the number of syllables increased by one (controlling for the number of words), the speech-to-song ratings also decreased by -0.11 (see Figure 4).

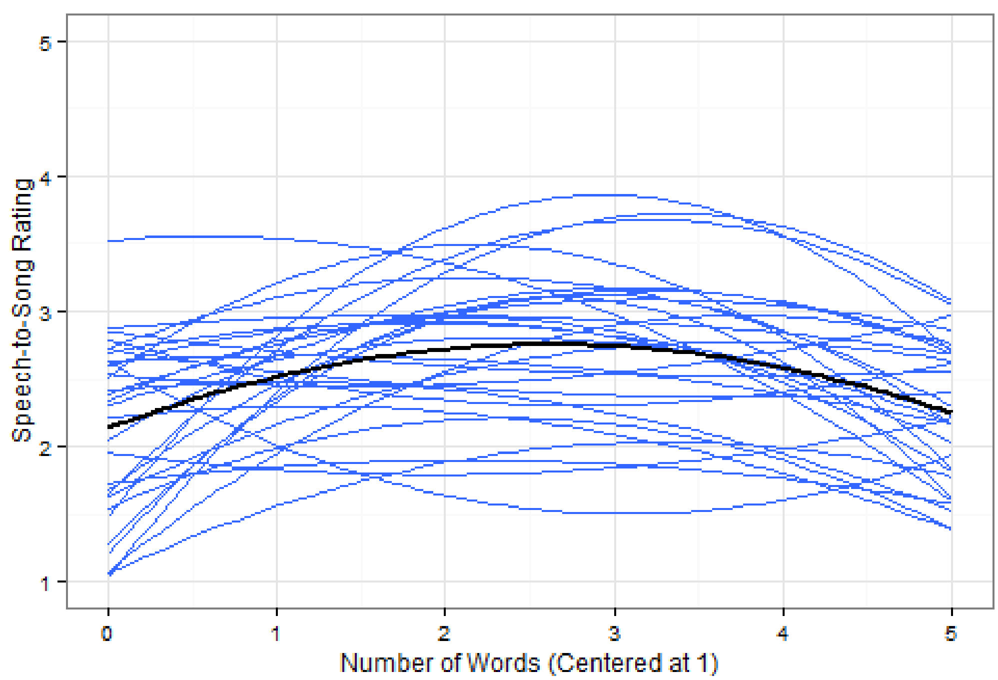

Figure 4. Spaghetti Plot of the Quadratic Effect of the Number of Words on Speech-to-Song Ratings for Each Participant. Blue lines represent each individual's trajectory with the black line representing the average of all individuals.

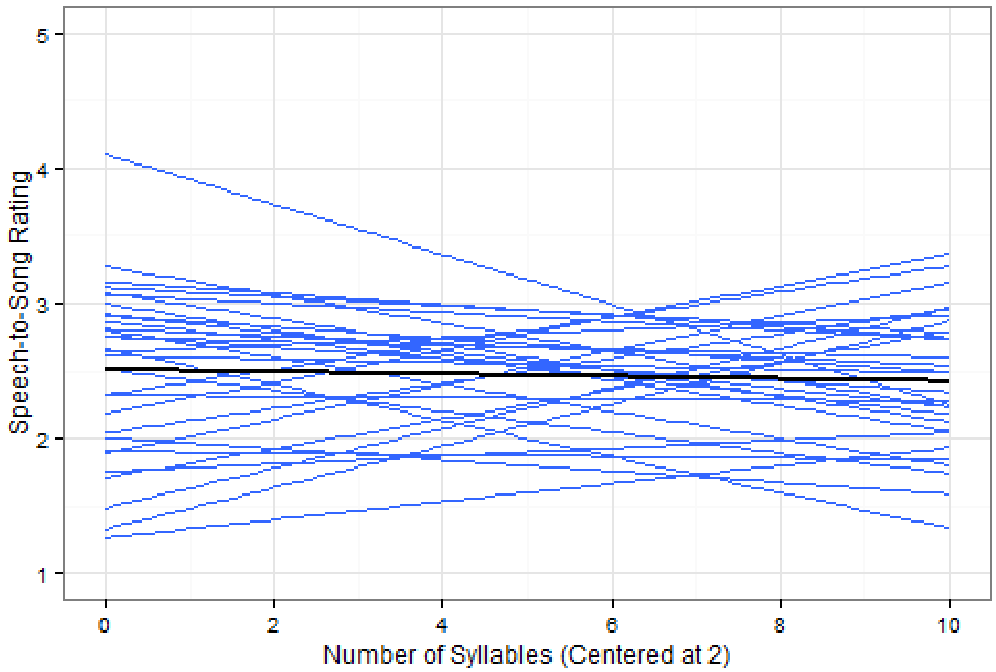

Figure 5. Spaghetti Plot of the Linear Effect of the Number of Syllables on Speech-to-Song Ratings for Each Participant. Blue lines represent each individual's trajectory with the black line representing the average of all individuals.

Discussion

A quadratic curve was found for the number of words per list, such that at approximately 3-4 words participants perceived the stimuli as sounding the most song-like. Shorter and longer word-lists produced more speech-like percepts. The linear trend also suggests that as the number of words continues to increase, the Speech-to-Song ratings will continue to decrease as seen in Experiment 3.

Only a negative linear trend was found for the number of syllables per list. Participants perceived the stimuli as less song-like (i.e., Speech-to-Song ratings decreased) as the number of syllables increased.

As NST suggests, with continual repetition words become satiated and more difficult to retrieve (MacKay, 1987). However, by adding more words or more syllables to the stimulus as a "filled pause," time between subsequent activations of each word increases allowing recovery from node satiation. As nodes recover from satiation, successful retrieval occurs and the perception of the Speech-to-Song Illusion diminishes.

Conclusion

In the Speech-to-Song Illusion repetition of a spoken phrase elicits in listeners a perceptual transformation resulting in spoken words sounding like they are being sung. The present study examined how changes in the stimulus impact the ability to elicit the Speech-to-Song Illusion. Two major findings were found. First, a grammatically or semantically meaningful phrase is not necessary to elicit the Speech-to-Song Illusion. Second, the amount of satiation, and thus the degree of song-like perception, changes as a result of repetition frequency and time between repetitions. Importantly, these findings can be explained in the context of Node Structure Theory, providing the first account of the cognitive mechanism that might underlie the Speech-to-Song Illusion.

Initial research on the Speech-to-Song Illusion used a grammatical phrase extracted from a sentence. However, having a syntactically correct phrase is not necessary to elicit the Speech-to-Song Illusion. Listeners experienced a song-like transformation with a list of four words in Experiment 1. Furthermore, Experiment 2 showed that using lists of nonwords with no semantic meaning also elicited the Speech-to-Song Illusion. Taken together, these findings indicate that higher-level information (i.e., syntactic and semantic information) does not impact the occurrence of the Speech-to-Song Illusion.

As a listener perceives each word, phonological nodes are primed and activated, which then send priming to the lexical node of each word. As the words continue to repeat, the lexical node becomes satiated, lowering the activity level of that node well below normal resting levels. The satiated lexical nodes, which cannot be activated, can no longer continue in the transmission of priming to the next higher-level, supporting the present finding that syntactic and semantic information is not necessary to elicit the Speech-to-Song Illusion. Since the satiated lexical nodes do not activate themselves, the next lowest-level, the syllable nodes, are the last nodes to activate in that series of connections. The syllable nodes continue to activate over and over as the phrase continues to repeat over and over, eliciting a beat-like perception in listeners.

One of the findings of Deutsch et al. (2011) must be explained as it seemingly contradicts the findings of this study. That is, it appears that the jumbled phrases (i.e., randomization of syllables in the phrase) did not result in a change in ratings from the initial to final presentation (although the statistical analysis to make this claim was not conducted, it appears that the means are not significantly different in Fig. 2). However, in the present explanation, the repetition of syllables is the locus of the Speech-to-Song Illusion and the present findings suggest that even a list of nonwords produces the illusion. The participants in Deutsch et al. (2011) were given the correct phrase (i.e., "sometimes behave so strangely"), then eight repetitions of that phrase with different combinations of syllable order (e.g., "strange have be times so –ly some"), then the correct phrase again. Potentially, being given the opportunity to hear the correct grammatical phrase and then jumbled versions of the same phrase did not allow for satiation to occur. Each of the syllables in the phrase, excluding "-ly" still contain lexical nodes that only satiate with repetition. Having a constantly changing stimulus from presentation to presentation possibly does not allow the nodes to satiate. Depending on the ordering of subsequent activations, there may be sufficient time to allow for node recovery. The present study removed this variability and instead repeated identical phrases over and over to listeners, thereby strengthening the impact of node satiation. By removing syntactic and semantic information with the use of words and nonwords, the present findings suggest that repetition of the syllable nodes does elicit the Speech-to-Song Illusion.

Through all of the studies on the Speech-to-Song Illusion, including the present study, repetition is essential for eliciting the illusion. Hearing the stimulus repeated ten times elicited higher song-like ratings than hearing the stimulus only once. With repetition, the lexical nodes of each phrase satiate, but the syllable nodes continue to activate. Even with nonwords which do not have a lexical node, the repetition of the syllables still elicited a song-like transformation. Importantly, Jackendoff (2009) suggests the music and language have similar rhythmic structures. In music the smallest metrical unit is a beat or note, while in language the smallest metrical unit is the syllable. As beats repeat in music, rhythm is evoked. Similarly, it is possible that as syllables repeat in speech, rhythm is also evoked leading to the Speech-to-Song Illusion.

NST suggests that the effects of satiation diminish with time as the node rebounds back to normal resting levels (MacKay, 1987). Once unsatiated, the node can activate under normal processes. One way to manipulate the time a node remains satiated is through the use of pauses, such that with longer amounts of time between subsequent activations, the effects of satiation will diminish. In the present study, pausing to reduce node satiation was manipulated in two ways: unfilled pauses and “filled” pauses.

Warren (1961) found that by increasing the amount of time between repetitions of a stimulus, verbal transformations decreased. Specifically, Warren increased pause durations from 0.2 s to 0.9 s and then from 0.9 s to 4s. In the present study, a 750 ms pause was added after each repetition in the repetition with pauses condition of Experiment 1. The Speech-to-Song ratings of participants who heard the word-lists only once was not significantly different than the ratings of participants who heard the word-lists played ten times with the 750 ms pause between repetitions. The 750 ms pause appears to be a sufficient amount of time to reduce the perception of song. In contrast to an unfilled pause placed at the end of each word-list, Experiments 3 and 4 used “filled” pauses. Even with the use of “filled” pauses, similar results were produced, such that Speech-to-Song ratings decreased as more time passed between subsequent activations of each individual word. Of important consideration is the fact that these “filled” pauses were created by adding additional lexical nodes. With a shorter phrase, activity remains localized, leading to satiated lexical nodes and more prominent beat-like syllables. However, with a longer phrase, lexical nodes are given the time to recover from satiation. With diminished satiation of the lexical nodes, the Speech-to-Song Illusion is less prominent.

Although the findings of the present study demonstrate specific stimulus characteristics that impact the elicitation of the Speech-to-Song Illusion, and that NST can provide an underlying mechanism for this illusion, there are many questions about this illusion that remain to be answered. For instance, syllables appear to be the locus of the Speech-to-Song Illusion. However, syllable structure varies across languages. English is a stress-timed language with lexical stress tending to fall on the second syllable, while in Spanish, a syllable-timed language, stress falls on the penultimate syllable. Even more distinct, Japanese is a mora-timed language. Future research could examine how these different timing structures impact the perception of the Speech-to-Song Illusion.

There are other factors that need to be addressed in eliciting the Speech-to-Song Illusion. Tierney et al. (2013) used a variety of phrases as stimuli which were judged during a pilot study to either sound more like speech or sound more like song after repetition. Although detailed information of the pilot study was not given, the question remains of why some phrases were perceived to be more speech-like while others more song-like. According to the account from NST, repetition of syllable nodes in any phrase should elicit a Speech-to-Song Illusion. Further examination of the phrases used in Tierney et al.'s pilot study may provide evidence for particular arrangements of words to elicit stronger song-like perceptions. For instance, the phrase "sometimes behave so strangely" consists of a sequence of words varying in syllables: 2 syllables (sometimes), 2 syllables (behave), 1 syllable (so), 2 syllables (strangely). Potentially, certain combinations of beats (e.g., 2, 2, 1, 2) may elicit a stronger song-like perception than some other combinations. Also, the speaking rate of the phrases may impact the perception of beats. Syllable production rates coincide with brain rhythms in the theta range of 4-8 Hz, including the preferred 280 beats per minute tempo of the bonobo (E. W. Large, personal communication, February 20, 2014). Manipulating the rate of syllable production in the phrases may elicit stronger or weaker perceptions of song.

A final question remains in determining the boundary between the Verbal Transformation Effect and the Speech-to-Song Illusion. Both perceptual illusions rely upon repetition of a linguistic stimulus, with the difference lying in the number of words. Often, one and sometimes two word stimuli are examined in the Verbal Transformation Effect (Warren, 1961). In the present study, Experiments 3 and 4 also examined stimuli of similarly short lengths. It is possible that the shorter stimuli in the present study might have elicited verbal transformations, and may even have done so more than a song-like transformation. It could also be possible that when conducting a typical VTE task, participants may also experience the Speech-to-Song Illusion. These suggestions seem plausible given that one of the non-identity changes described by participants in Kaminska & Mayer (2002) during a VTE task for word stimuli was rhythm.

Future research could examine if there is a transition from verbal transformations to the Speech-to-Song Illusion (or vice versa), or if listeners experience both simultaneously.

In sum, the Speech-to-Song Illusion is elicited as a list of words repeats over and over. The spoken words transform into sounding like song, even though the stimulus is unchanged. The lists of words necessary for eliciting the Speech-to-Song Illusion do not need to be grammatically meaningful phrases, or even lists of real words. Also, by increasing the number of words, the number of syllables, or the time between repetitions, the Speech-to-Song Illusion diminishes. Using NST, the underlying mechanism of the Speech-to-Song Illusion resides in the repetition of syllable nodes, which elicits a beat-like perception. As the syllable beats repeat over and over, the transformation from speech to song emerges in listeners' percepts.

References

Bassett, M. F., & Warne, C. J. (1919). On the lapse of verbal meaning with repetition. *The American Journal of Psychology, 30,* 415-418.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip-of-the-tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30,* 542-579.

Calef, R. S., Calef, R. A. B., Kesecker, M. P., & Burwell, R. (1974). Verbal transformations of "stabilized" taboo and neutral words. *Perceptual and Motor Skills, 38,* 177-178.

Christiner, M., & Reiterer, S. M. (2013). Song and speech: Examining the link between singing talent and speech imitation ability. *Frontiers in Psychology, 4,* 1-11.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25,* 257-271.

Cook, P., Rouse, A., Wilson, M., Reichmuth, C. (2013). A California sea lion (*Zalophus californianus*) can keep the beat: Motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *Journal of Comparative Psychology, 127*, 412-427.

Coren, S., & Girgus, J. S., (1978). *Seeing is deceiving: The psychology of visual illusions.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Deutsch, D. (1995). *Musical Illusions and Paradoxes* [CD]. La Jolla, CA: Philomel Records.

Deutsch, D. (2003). 'But they sometimes behave so strangely.' *Phantom Words and Other Curiosities* [CD]. La Jolla, CA: Philomel Records.

Deutsch, D., Henthorn, T., Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America, 129,* 2245-2252.

Falk, S., & Rathcke, T. (2010). On the Speech-to-Song Illusion: Evidence from German. *Speech Prosody 2010*, *100169*, 1-4.

Fay, K., & Boyd, M. J. (2010). Eta-squared. In. N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 422-425). Thousand Oaks, CA: SAGE Publications Inc.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 171*, 279-296.

Gregory, R. L., & Wallace, J. G., (1963). *Recovery from early blindness: A case study*. Cambridge, England: Heffner and Sons.

Hattori, Y., Tomonaga, M., & Matsuzawa, T. (2013). Spontaneous synchronized tapping to an auditory rhythm in a chimpanzee. *Scientific Reports, 3,* 1566.

Kaminska, Z., & Mayer, P. (2002). Changing words and changing sounds: A change of tune for verbal transformation theory? *European Journal of Cognitive Psychology, 14,* 315-333.

Kaminska, Z., Pool, M., & Mayer, P. (2000). Verbal transformation: Habituation or spreading activation? *Brain and Language*, *71*, 285-298.

Koelsch, S., Gunter, T. C., von Cramon, D. Y., Zysset, S., Lohmann, G., Friederici, A. D. (2002). Bach speaks: A cortical “language-network” serves the processing of music. *NeuroImage, 17,* 956-966.

Large, E. W. (2014, February). Spontaneous tempo matching and synchronization in a bonobo. In Gray, P. M. (Chair), *Rhythmic Entrainment in Non-Human Animals: An Evolutionary Trail of Time Perception.* Symposium conducted at the annual meeting of the American Association for the Advancement of Science, Chicago, IL.

Locke, J. (1894). An essay concerning human understanding (Vol. 1). Ed and annotated by A. C. Fraser. Oxford: Clarendon Press.

Ludke, K. M., Ferreira, F., & Overy, K. (2014). Singing can facilitate foreign language learning. *Memory and Cognition*, *42*, 41-52.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer-Verlag.

MacKay, D. G., & Burke, D. M. (1990). Cognition and aging: A theory of new learning and the use of old connections. In T. Hess (Ed.), *Aging and Cognition: Knowledge organization and utilization* (pp. 213-263). Amsterdam, Holland: Elsevier.

MacKay, D. G., Stewart, R., Burke, D. M. (1998). H. M.'s language production deficits: Implications for relations between memory, semantic binding, and the hippocampal system. *Journal of Memory and Language, 38,* 28-69.

MacKay, D. G., Wulf, G., Yin, C., & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language*, *32,* 624-646.

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D., (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature Neuroscience, 4,* 540-545.

Mok, P. K. P., & Zuo, D. (2012). The separation between music and speech: Evidence from the perception of Cantonese tones. *Journal of the Acoustical Society of America, 132,* 2711-2720.

Patel, A. D, Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology, 19,* 827-830.

Pitt, M. A., & Shoaf, L. Linking verbal transformations to their causes. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 150-162.

Roediger, H. L., & McDermott, K. B. (2000). Tricks of memory. *Current Directions in Psychological Science, 9,* 123-127.

Snyder, K. A., Calef, R. S., Choban, M. C., & Geller, E. S., (1993). Effects of word repetition and presentation rate on the frequency of verbal transformations: Support for habituation. *Bulletin of the Psychonomic Society, 31*, 91-93.

Tierney, A., Dick, F, Deutsch, D., & Sereno, M. (2012). Speech versus song: Multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, *23*, 249-254.

Vitevitch, M. S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effect of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Language, 30,* 514-529.

Vitevitch, M. S., & Luce, P. A. (1998).When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374-408.

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52,* 193-204.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40,* 47-62.

Vitevitch, M. S., Stamer, M. K., & Sereno, J. A. (2008). Word length and lexical competition: Longer is the same as shorter. *Language and Speech, 51,* 361-383.

Warren, R. M. (1961). Illusory changes of distinct speech upon repetition – The verbal transformation effect. *British Journal of Psychology, 52,* 249-258.

Warren, R. M. (1968). Verbal transformation effect and auditory perceptual mechanisms. *Psychological Bulletin, 70,* 261-270.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 392-393.

Warren, R. M. (1976). Auditory illusions and perceptual processes. In. N. J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 389-417). New York, NY: Academic Press, Inc.

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology, 71*, 612-613.

Yin, C., & MacKay, D. G. (1992, April). *Auditory illusions and aging: Transmission of priming in the verbal transformation paradigm.* Poster presented at the Cognitive Aging Conference, Atlanta, GA.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Science, 6,* 37-46.

Zhang, S. (2011, August). *Speech-to-song illusion in MC: Acoustic Parameter vs. perception*. Poster presented at the biennial meeting of the Society for Music Perception and Cognition, Rochester, NY.

Appendix A. Words-Lists Used In Experiment One

Dense Word-Lists				Sparse Word-Lists			
lever	battle	furry	candle	lumber	badger	formal	cancer
letter	muscle	berry	babble	lawyer	mother	button	barrel
polar	bubble	money	ladder	person	beggar	movie	lucky
cattle	banner	tackle	hurry	cashew	burden	tower	hero
leather	valley	puddle	candy	lady	vapor	powder	camel
dairy	meter	body	lighter	devil	mighty	bottom	lotion
paddle	shallow	mayor	worry	purple	shower	mitten	water

Appendix B. Nonword-Lists Used In Experiment Two

High PP				Low PP			
bæz	mɛk	faɪs	pɪm	bɔʃ	moɪk	fɑutʃ	pɝg
mos	tæs	dɛm	dʒoɹ	mɝz	teʃ	dɔb	dʒaɪg
gɛl	bɛs	nʌs	ɹɛn	gɝp	bɝʃ	nɑub	oɪk
ɹɛs	nɪd	gɪd	des	ɹɑub	nɝg	gɑub	deð
pɪz	dʒɪt	bɪθ	faɪd	pʊtʃ	dʒɝʃ	baɪdʒ	fɝp
tes	fɪm	mɑun	gaɪn	todʒ	foɪz	mɑub	gaɪθ
dʌp	paɪd	dʒæd	nɛs	dɝf	pudʒ	dʒeʃ	nɛʃ

Appendix C. Word-Lists Used In Experiment Three

1 Word per List			
cattle			
candle			
polar			
banner			
furry			
2 Words per List			
ladder	money		
worry	leather		
hurry	meter		
tackle	battle		
babble	dairy		
3 Words per List			
lever	shallow	paddle	
dairy	babble	worry	
bubble	lighter	money	
ladder	valley	battle	
cattle	banner	candy	
4 Words per List			
furry	mayor	leather	tackle
polar	letter	candy	battle

lever	cattle	tackle	hurry			
candle	banner	puddle	berry			
hurry	lighter	babble	ladder			
5 Words per List						
battle	muscle	polar	money	hurry		
shallow	furry	puddle	lever	body		
worry	mayor	babble	paddle	banner		
lighter	candy	muscle	tackle	letter		
berry	meter	leather	candle	bubble		
6 Words per List						
worry	babble	letter	cattle	dairy	candy	
paddle	berry	valley	mayor	letter	candle	
money	puddle	leather	banner	furry	babble	
bubble	lighter	body	lever	worry	puddle	
shallow	tackle	battle	polar	leather	hurry	
7 Words per List						
paddle	hurry	banner	ladder	valley	berry	mayor
babble	lighter	muscle	cattle	tackle	meter	hurry
cattle	battle	candy	lever	dairy	polar	candle
furry	leather	mayor	body	shallow	valley	bubble
letter	money	dairy	puddle	ladder	worry	paddle
8 Words per List						

leather	tackle	candy	polar	bubble	lever	meter	battle		
battle	banner	paddle	ladder	shallow	dairy	money	cattle		
letter	shallow	mayor	berry	muscle	valley	worry	lighter		
mayor	paddle	babble	shallow	hurry	dairy	polar	money		
candle	muscle	valley	furry	puddle	cattle	lever	meter		
9 Words per List									
dairy	puddle	lighter	berry	bubble	body	candle	furry	letter	
bubble	hurry	furry	valley	money	berry	leather	lever	banner	
body	berry	ladder	dairy	candy	tackle	paddle	furry	puddle	
meter	leather	money	babble	lever	candle	polar	muscle	cattle	
shallow	battle	worry	paddle	letter	mayor	lighter	valley	tackle	
10 Words per List									
candle	banner	valley	puddle	worry	muscle	meter	lighter	babble	shallow
ladder	muscle	meter	puddle	candy	lighter	bubble	tackle	polar	mayor
leather	valley	battle	tackle	shallow	ladder	meter	candle	bubble	muscle
hurry	candy	ladder	paddle	berry	banner	furry	meter	dairy	worry
letter	berry	mayor	bubble	cattle	money	candy	lever	muscle	polar

Appendix D. Word-Lists Used In Experiment Four

1 Word, 2 Syllables per List	
derby	
cocoon	
essence	
logic	
throttle	
1 Word, 4 Syllables per List	
anatomy	
immediate	
recreation	
biography	
ultimatum	
2 Words, 4 Syllables per List	
elbow	brother
react	angry
humor	achieve
courage	measure
puppet	fatigue
2 Words, 8 Syllables per List	
voluntary	diameter
economy	monopoly

generalize	obligation		
eventual	ordinary		
secondary	initiate		
3 Words, 6 Syllables per List			
damage	fluid	observe	
network	baton	devote	
convey	purchase	luggage	
wallet	empty	corrupt	
moisture	junior	acid	
3 Words, 12 Syllables per List			
supervision	academic	morality	
alleviate	technology	phenomenon	
colonial	publicity	dietary	
reality	limitation	tolerable	
mandatory	valuable	comedian	
4 Words, 8 Syllables per List			
signal	ethnic	kitchen	awkward
bazaar	donate	napkin	review
dainty	violet	govern	chamber
blossom	creature	drama	option
improve	sincere	jagged	buffet
6 Words, 12 Syllables per List			

bundle	message	assume	lagoon	exit	organ
neutral	remove	bishop	cradle	sugar	joyous
unite	chimney	wagon	magic	admire	rodent
orchard	lecture	giant	impose	unique	crucial
furnace	routine	hygiene	alarm	cigar	voyage