Rapid Learning and Long-term Memory in the Speech-to-song Illusion

Benjamin M. Kubit & Christine Deng
Princeton University

Adam Tierney
Birkbeck, University of London, London, United Kingdom

Elizabeth H. Margulis
Princeton University

The speech-to-song illusion is a perceptual transformation in which a spoken phrase initially heard as speech begins to sound like song across repetitions. In two experiments, we tested whether phrase-specific learning and memory processes engaged by repetition contribute to the illusion. In Experiment 1, participants heard 16 phrases across two conditions. In both conditions, participants heard eight repetitions of each phrase and rated their experience after each repetition using a 10-point scale from "sounds like speech" to "sounds like song." The conditions differed in whether the repetitions were heard consecutively or interleaved such that participants were exposed to other phrases between each repetition. The illusion was strongest when exposures to phrases happened consecutively, but phrases were still rated as more song-like after interleaved exposures. In Experiment 2, participants heard eight consecutive repetitions of each of eight phrases. Seven days later, participants were exposed to eight consecutive repetitions of the eight phrases heard previously as well as eight novel phrases. The illusion was preserved across a delay of one week: familiar phrases were rated as more song-like in session two than novel phrases. The results provide evidence for the role of rapid phrase-specific learning and long-term memory in the speech-to-song illusion.

Received: July 14, 2023, accepted January 4, 2024.

Key words: learning, memory, music perception, auditory perception, repetition

The speech-to-song illusion is a perceptual transformation in which spoken phrases can sound increasingly like song as they are repeated (Deutsch et al., 2011). The transformation is usually measured by participant reports that a phrase sounds more song-like on the final repetition compared to the first. While repetition is commonly deployed in music composition (Margulis, 2014), the cognitive mechanisms by which repetition contributes to the perceptual experience of music remain unclear.

The transformation from speech to song influences participants' abilities to recite a phrase and shapes their perceptual expectations upon hearing the phrase again. Deutsch et al. (2011) asked participants to repeat back phrases and found that phrases for which the illusion was experienced were sung rather than spoken as speech, with the pitches distorted to conform to the melodic expectations of Western tonal music. Once a phrase is perceived as song, participants are also less likely to notice changes in pitch when the phrase is heard again, so long as the changes conform to familiar musical scale structures (Vanden Bosch der Nederlanden et al., 2015). Similarly, participants are better at detecting temporal irregularities when transformed phrases are heard again compared to phrases that continue to be perceived as speech (Graber et al., 2017). Consistent with behavioral work, Tierney et al. (2013) found that auditory-motor brain regions previously associated with music perception increase blood-oxygen-level-dependent signal during repetitions of transformed speech, compared to untransformed speech. Overall, the findings provide converging evidence that the illusion entails updating perceptual expectations across repetitions based on prior knowledge for Western tonal and rhythmic structures.

Critically, when a phrase is transposed or the syllables are scrambled across repetitions, the illusion doesn't occur (Deutsch et al., 2011). These results suggest that the illusion and the underlying changes in perceptual expectations driven by repetition are tailored to a particular phrase, and do not merely reflect a general tendency to perceive music when varying short segments of

Music Perception, volume 41, issue 5, pp. 348–359, issn 0730-7829, electronic issn 1533-8312. © 2024 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, https://www.ucpress.edu/journals/reprints-permissions. DOI: https://doi.org/10.1525/mp.2024.41.5.348

speech are repeated out of context. In the present study, we were interested in determining whether phrase-specific knowledge (e.g., about the specific sounds and the transitions between them) learned across repetitions of spoken phrases contributes to the illusion.

Early work on the illusion hypothesized that increases in perceived musicality are the result of the interaction between cognitive processes differentially engaged by speech and music perception. Deutsch et al. (2011) proposed that repetition frees cognitive resources devoted to pitch processing that are inhibited during speech perception, resulting in the emergent salience of perceived pitches. More recently, Castro et al. (2018) suggested that the tension between pitch and speech processing can be explained by the dynamics between the lexical and syllable nodes in a connectionist model of language (Vitevitch et al., 2021). Neither of these hypotheses can explain why the illusion is found for non-speech stimuli like tones (Margulis & Simchy-Gross, 2016; Tierney et al., 2018a) and environmental sounds (Rowland et al., 2019; Simchy-Gross & Margulis, 2018), given that such stimuli don't engage lexical processing. Moreover, hypotheses that appeal to the competing dynamics between cognitive processes most often study the effects of repetition over short periods of time (seconds) and do not provide a clear framework for examining the role of learning and memory processes in the illusion, for example, by testing whether stimulus-specific knowledge is learned across repetitions and consolidated in long-term memory.

Several other hypotheses have been proposed, including two variations of a template matching process, according to which perceptual stimuli are continuously compared and matched to existing music templates of common tonal and rhythmic structures found in Western music (Rowland et al., 2019; Tierney et al., 2018a). Presumably, once a match is found, perceptual expectations are updated and a stimulus is perceived as more song-like. According to Tierney et al. (2018a), the cognitive mechanisms by which repetition contributes to the illusion include two components: short-term memory for the storage of a phrase's melodic structure, and a mechanism for comparing the structure held in short-term memory to music templates. We suggest that the latter mechanism engages working memory processes (Naveh-Benjamin & Cowan, 2023; van Ede & Nobre, 2023) to maintain and manipulate the melodic structure across repetitions until a match is made. Alternatively, Rowland et al. (2019) suggest that the interaction between existing music templates and attentional processes gives rise to the illusion by biasing perception towards musical structure. These hypotheses can explain why the illusion occurs for non-speech stimuli but would not be able to explain how the illusion could persist despite delays and interference from other stimuli between repetitions.

Studies have found the illusion can occur even when the final repetition comes at the end of the study following exposure to other stimuli and a short delay period (Graber et al., 2017; Margulis & Simchy-Gross, 2016). If too many other stimuli are heard between repetitions, or if enough time has passed since the last repetition, phrase-specific information cannot be maintained in short-term or working memory, because both processes are limited in the duration and quantity of information that can remain active (Brem et al., 2013; Cowan et al., 2012; Miller, 1956). Based on the hypothesis put forth by Tierney et al. (2018a), the illusion shouldn't be experienced when intervening stimuli are heard between repetitions of the same stimulus. Soehlke et al. (2022), however, demonstrated the opposite effect: interleaved repetitions of different spoken phrases produced the illusion. Even if some manner of template matching is biasing attention (Rowland et al., 2019), evidence that the illusion is experienced despite interleaved presentations suggest that the results of the template matching process are temporarily stored and redeployed after attention has been focused on intervening stimuli, implicating a role for learning processes. Although Soehlke et al. (2022) provided preliminary evidence that learning and long-term memory contribute to the illusion, further work is needed to test phrase-specific learning, whether interleaved and consecutive stimulus repetitions produce an illusion of similar strength, and how long the illusion lasts.

In the present study, we examined the role of phrase-specific learning and long-term memory in the illusion by designing two experiments that required phrase-specific knowledge to be encoded and retained across delays for the illusion to occur. As a result, changes in ratings of perceived musicality, measured within-participants at the level of individual phrases, served as a measure of phrase-specific memory in experiment conditions that precluded sustained contributions from attention and working memory processes. Previous work shows that repetition drives learning of the note and chord sequences particular to a piece of music (Hébert & Peretz, 1997; Janata & Grafton, 2003; Kubit & Janata, 2022a, 2022b) and that the resulting veridical representations in memory interact with schematic knowledge about Western music to shape listeners' expectations (Bharucha, 1987; Tillmann & Bigand, 2010; Vuust et al., 2022). Repetition is also likely to drive the learning of the structure particular to a spoken phrase. We hypothesize the phrase-specific knowledge plays an import role in perceiving musicality in speech. Such learning may be differentially engaged by music perception compared to speech perception because the higher frequency of repetition characteristic of music stimuli (Margulis, 2014) affords more opportunities to learn the time-varying structure. In Experiment 1, we tested the hypothesis that phrase-specific knowledge learned across repetitions influences perception such that a phrase sounds more musical.

Anecdotally, people report that the illusion can persist across long temporal delays, but this hasn't yet been tested in an experimental setting. Evidence that the illusion persists across delays greater than a day would further implicate learning and memory processes and provide evidence that undercuts previous hypotheses that only consider attention, short-term memory, and working memory: such processes cannot influence perception across time periods spanning multiple days unless learning takes place and the results are stored in long-term memory. In Experiment 2, we tested the hypothesis that phrase-specific knowledge learned across repetition is consolidated in long-term memory and increases perceived musicality upon hearing a phrase after a seven-day delay period.

Experiment 1

In Experiment 1, we extend previous work by Soehlke et al. (2022) by examining, within participants at the level of individual spoken phrases, the effects of interleaved compared to blocked (consecutive) repetitions on the illusion. Interleaved repetitions, during which different phrases are heard between repetitions of a phrase, prevent the same phrase-specific content from remaining the focus of attention and being maintained in working memory across repetitions (see hypotheses described in Rowland et al., 2019 and Tierney et al., 2018a). Thus, results showing the illusion is reliably experienced across interleaved repetitions would provide evidence for the role of phrase-specific learning processes that result in the rapid encoding and retrieval of knowledge despite delays and interference between exposures. We hypothesized that interleaved repetitions would still produce the illusion, but that the illusion would be stronger after blocked repetitions. Blocked repetition may lead to a stronger illusion by engaging the attention and working memory processes previously suggested to underlie the illusion (Rowland et al., 2019; Tierney et al., 2018a) that are limited in contribution during interleaved repetition. Alternatively, blocked repetition may lead to a stronger illusion because it is more conducive to learning, as interleaved and blocked presentation schedules are known to differentially influence learning rates (Brunmair & Richter, 2019; Schorn & Knowlton, 2021; Shea & Morgan, 1979). In the case of the speech-to-song illusion, if phrase-specific learning contributes to the transformation, then the difference in learning rates between blocked and interleaved presentations would produce illusions of different strength.

METHOD

Participants

All data collection procedures used in the study were approved by the Princeton University Institutional Review Board. We estimated the required number of participants for Experiment 1 based on the results of a previous experiment in which participants heard repetitions of spoken phrases and provided musicality ratings after every repetition (Experiment 1 in Tierney et al., 2021). Using the partial eta squared of the *F*-test for the repetition variable as the effect size, we determined that 22 participants would suffice for power of .80. We quadrupled the expected number of participants in Experiment 1 to counterbalance condition order and to help account for extra noise inherent to the online study environment.

Eighty-two Princeton University undergraduate students (47 females, 19–23 years; mean age = 21 years) participated in Experiment 1 after providing informed consent. Participants reported neither neurological nor hearing impairments and declared English to be their primary language. Fifty-two participants reported having more than 1 year of formal musical instrument training ("2" – "10 or more years"; mean training = "4 – 5 years"). Participants were compensated with research credits for completing each of the two days of the experiment.

Materials

Equipment. Participants were tested online using their own desktop or laptop computer in a location of their own choosing. Responses were made using a mouse and keyboard. Participants were instructed to wear headphones and to find a quiet and comfortable place to complete the study. The experiment was hosted by Pavlovia (https://pavlovia.org) and controlled by jsPsych (de Leeuw, 2015).

Speech Stimuli. Phrase stimuli were chosen from a stimulus set used in previous studies of the speech-to-song illusion (Graber et al., 2017; Tierney et al., 2013, 2018a, 2018b, 2021). We selected the 16 phrases from the set of 24 "illusion" stimuli that had, on average, the greatest increase in musicality ratings from the 1st to the

8th repetition in Experiment 1 from Tierney et al. (2021). The mean length of stimuli was 6.3 syllables (minimum = 4, maximum = 8) and 1.3 seconds (minimum = 0.84, maximum =1.80). Speakers featured in the phrase stimuli were three different males who were native speakers of American or British English.

Procedure

Participants heard 16 phrases across two conditions in a single session. In both conditions, participants heard eight presentations of each phrase and after each repetition clicked buttons labeled 1 through 10 to indicate the extent to which a phrase was perceived as music, where "1" indicated completely speech-like and "10" indicated completely song-like. A trial comprised a single phrase presentation and musicality rating, and eight trials formed a block. Participants pressed a button to start each block. Within a block, participants were given two seconds to respond on each trial, after which time the experiment automatically went on to the next trial. The two conditions differed in whether repetitions of a phrase were heard consecutively within a block (blocked condition) or interleaved (interleaved condition) such that participants heard other phrases between each repetition (Figure 1). Phrase order was randomized within each interleaved condition block, which was always comprised of a single presentation of eight different phrases. As a result, each phrase was heard eight times either within a single block (blocked condition) or interleaved across eight blocks (interleaved condition). The assignment of phrases to conditions and order of stimulus presentation were randomized across participants, and the starting condition was counterbalanced across participants. Participants were instructed to take a short break (no more than five minutes) after completing the first condition.

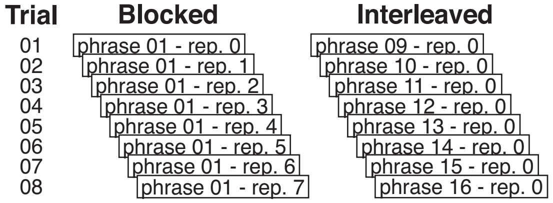
Trial
Blocked
Interleaved
01 phrase 01 - rep. 0 phrase 09 - rep. 0
02 phrase 01 - rep. 1 phrase 10 - rep. 0
03 phrase 01 - rep. 2 phrase 11 - rep. 0
04 phrase 01 - rep. 3 phrase 12 - rep. 0
05 phrase 01 - rep. 4 phrase 13 - rep. 0
06 phrase 01 - rep. 5 phrase 14 - rep. 0
07 phrase 01 - rep. 6 phrase 15 - rep. 0
08 phrase 01 - rep. 7 phrase 16 - rep. 0

FIGURE 1. First blocks (eight trials) in the blocked and interleaved conditions from Experiment 1. Within a block, participants heard a phrase and were given two seconds to respond on each of the eight trials. In the blocked condition, participants heard and responded to the same phrase on all eight trials and a different phrase was heard during each block. In the interleaved condition, participants heard and responded to a different phrase on each of the eight trials, and the same eight phrases were heard during each block. rep. = repetition number; rep. 0 represents the first presentation of a phrase.

After blocks 3, 6, 11, and 14, participants encountered catch trials during which they heard the sentence, "Don't rate this speech, instead choose response" followed by a spoken number ("two," "four," "five," or "eight"). Participants were expected to click the button on the musicality scale that corresponded to the spoken number. Half of the catch trials were spoken by a female voice and the rest by a male voice. The purpose of these trials was to identify participants who were listening attentively to the phrases. Fourteen participants incorrectly answered more than one catch trial and were excluded from analyses. Ten participants were also excluded for missing multiple responses in a single block more than once. On average, each participant missed 1.39 trials (*SD* = 1.53) and 1.46 trials (*SD* = 1.55) in the blocked and interleaved conditions, respectively.

At the end of the session, participants filled out the Goldsmiths Musical Sophistication Index (GMSI) (Müllensiefen et al., 2014) to measure individual differences in music aptitude (see Supplementary Materials accompanying this paper online at online.ucpress.edu/mp; Experiment 1: Individual differences in illusion strength and Table S1). Participants also indicated the extent to which they agreed with the statement, "I paid attention throughout the experiment" using a 5-point scale (e.g., "Strongly Disagree," "Disagree," "Neither Agree nor Disagree," "Agree," "Strongly Agree"). Two participants responded that they didn't agree with the statement and were excluded from analyses.

Analyses

Mixed models were estimated in R (https://www.R-project.org/) using the lmer(), glmer(), and bootMer() functions from the *lme4* package (Bates et al., 2014). Descriptions of all analyses are provided in Supplementary Materials Table S2. For each experiment, separate linear mixed models (LMMs) were used to estimate the effect of repetition number and condition on musicality ratings. LMMs were estimated using maximum likelihood based on Laplace approximation. We modeled participant as a random intercept for every model to account for variance resulting from repeated measurements.

For every LMM we performed 2,000 parametric bootstraps of the model using the bootMer() function in lme4. Both fixed and random effects were estimated for every bootstrapped sample. The fixed effect coefficients were extracted from a bootstrapped model to create a sampling distribution for each coefficient. The median

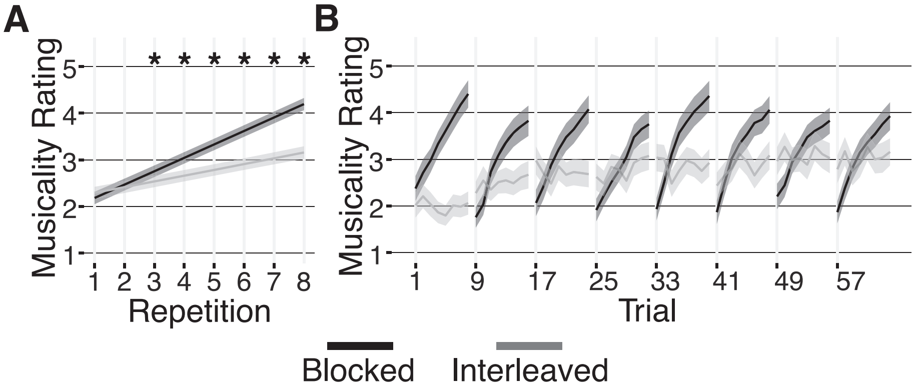

FIGURE 2. Effects of conditions across repetitions in Experiment 1. (A) Modeled change in musicality ratings across repetitions based on exposure condition. Asterisks denote significant contrasts between conditions. (B) Average change in musicality ratings across trials based on exposure condition. The median value of the musicality scale is 4.5, minimum = 0 ("sounds exactly like speech"), maximum = 9 ("sounds exactly like song"). Shaded regions represent the SEM of the bootstrapped marginal estimates in A and the SEM of the participant means in B.

value from the sampling distribution served as the estimated fixed effect coefficient and values from the 2.5 and 97.5 percentiles served as the lower and upper bounds for the 95% confidence interval (CI). For all models we considered a coefficient to be significant if the 95% CI did not include zero. Prior to model evaluation, the data were preprocessed so that model coefficients served as effect sizes that can be interpreted as the estimated change in the dependent variable for a standardized increase in the predictor variable.

To aid comprehension, for each bootstrap iteration we also extracted the estimated marginal means for a model across the full range of observed values by using the predict() function in the *multcomp* package (Hothorn et al., 2008). This allowed us to create a sampling distribution for each marginal mean. Error bars representing the standard error of the mean of the distribution are included in all figures to convey the variability in our samples, while the CIs reported in text provide information on whether an effect was statistically significant. When a hypothesis warranted further comparisons we used the estimated marginal means at each iteration to calculate the desired contrasts. Because the contrasts were based on the estimated marginal means, the resulting contrast coefficients served as effect sizes that describe the difference between levels of a predictor in the units of the dependent variable. The contrast coefficients were extracted from each bootstrap iteration and the median value from the sampling distribution served as the estimated contrast coefficient. Values from the 2.5 and 97.5 percentiles served as the lower and upper bounds for the 95% confidence interval (CI). For all models we considered a contrast to be significant if the 95% CI did not include zero.

RESULTS AND DISCUSSION

We predicted that repetition would increase the perceived musicality of phrases heard in both conditions. We used a linear mixed model to estimate ratings for a phrase as a function of condition and repetition (see Supplementary Materials Table S2 for descriptions of analyses). Results show participants perceived phrases as more musical on the final repetition compared to the first during both conditions (Figure 2A). For each additional repetition, the musicality rating for a phrase increased by 0.287 (95% CI [0.261, 0.315], $p < .001$) in the blocked condition and 0.123 (95% CI [0.096, 0.151], $p < .001$) in the interleaved condition. Phrases heard in the blocked condition demonstrated a greater rate of increase in musicality than phrases heard in the interleaved condition (difference = 0.165, 95% CI [0.127, 0.204], $p < .001$).

Post hoc contrasts suggest the effects of blocked and interleaved conditions resulted in similar musicality ratings during the first two repetitions. Across repetitions three through eight, musicality ratings were greater in the blocked, compared to the interleaved condition. By the last repetition, phrases heard in the blocked condition were, on average, perceived as 1.05 rating scale values more song-like than the interleaved condition phrases (Table 1, Figure 2A). The different rates of change across repetitions are also clearly visible in the unmodeled data using participants' average ratings (Figure 2B). Overall, the pattern of results demonstrate that phrase-specific learning contributes to the illusion across interleaved repetitions, but that the illusion is strongest after consecutive repetitions. The design of Experiment 2 directly tested phrase-specific learning during blocked presentations by probing long-term memory for phrases.

TABLE 1. *Post hoc Contrast Comparing Experiment 1 Ratings Between Conditions at Each Phrase Repetition*

Contrast	Coefficient	95% CI	*p*
rep. 1 BLK - INT	−0.109	[−0.271, 0.048]	.189
rep. 2 BLK - INT	0.055	[−0.078, 0.186]	.412
rep. 3 BLK - INT	0.219*	[0.116, 0.326]	< .001
rep. 4 BLK - INT	0.383*	[0.295, 0.475]	< .001
rep. 5 BLK - INT	0.549*	[0.457, 0.640]	< .001
rep. 6 BLK - INT	0.714*	[0.606, 0.819]	< .001
rep. 7 BLK - INT	0.878*	[0.747, 1.007]	< .001
rep. 8 BLK - INT	1.043*	[0.884, 1.201]	< .001

Note: Values are bootstrapped contrast coefficients and 95% CIs. Contrasts from all models reflect the difference between estimated marginal means. Asterisks denote significant effects. Blocked condition (BLK); Interleaved condition (INT); rep (repetition).

Experiment 2

In Experiment 2, we examine whether the illusion persists, within participants at the level of individual spoken phrases, across longer delays. Re-exposure to a phrase days after the illusion was experienced prevents the same phrase-specific content from remaining the focus of attention and being maintained in working memory across the delay (see hypotheses described in Rowland et al., 2019, and Tierney et al., 2018a). Thus, results showing that phrases are still perceived as more song-like after several days would provide strong evidence for the role of phrase-specific long-term memory in the illusion. Based on the results of Experiment 1, we hypothesized that phrase-specific knowledge initially learned across blocked repetitions would be consolidated in long-term memory, resulting in previously heard phrases sounding more song-like at the start of a second session seven days later, compared to novel phrases heard for the first time.

METHOD

Participants

Fifty-two Princeton University undergraduate students (32 females, 19–24 years; mean age = 21 years) participated in Experiment 2 after providing informed consent. Participants reported neither neurological nor hearing impairments and declared English to be their primary language. Thirty participants reported having more than 1 year of formal musical instrument training (“2” – “10 or more years”; mean training = “6 – 9 years”). Participants were compensated with research credits for completing each of the two days of the experiment.

Materials

The same apparatus was used as in Experiment 1. Phrase stimuli were the same as those used in Experiment 1.

Procedure

Participants heard 16 phrases across three conditions and two sessions. During the first session, all participants heard eight consecutive presentations of each of eight phrases (day 1 condition). Seven days later, participants again heard eight consecutive presentations of the eight familiar phrases first heard during session one (day 2 familiar condition) as well as eight novel phrases (day 2 novel condition). The structure of trials and blocks was the same structure used in Experiment 1. In all three conditions, repetitions of a phrase were heard consecutively within a block, as in the blocked condition from Experiment 1. The assignment of phrases to conditions and order of stimulus presentation were randomized across participants. Participants were instructed to take a short break (no more than five minutes) after completing the first eight blocks during the second session.

After blocks 3 and 6 in session one and blocks 3, 6, 11, and 14 in session two, participants encountered catch trials during which they heard the sentence, “Don’t rate this speech, instead choose response” followed by a spoken number (“two,” “four,” “five,” or “eight”). As in Experiment 1, participants were expected to click the button on the musicality scale that corresponded to the spoken number. Half of the catch trials on each day were spoken by a female voice and the rest by a male voice. Eight participants incorrectly answered more than one catch trial on a given day and were excluded from analyses. Three participants were also excluded for missing multiple responses in a single block more than once. On average, each participant missed 1.54 trials (*SD* = 1.53) on day 1, 0.88 trials (*SD* = 1.46) during the day 2 familiar condition, and 0.58 trials (*SD* = 0.98) during the day 2 novel condition.

At the end of the second session, participants filled out the GMSI (see Supplementary Materials, Experiment 2: Individual differences in illusion strength and Table S3). Participants also indicated the extent to which they agreed with the statement, “I paid attention throughout the experiment” using a 5-point scale (e.g., “Strongly Disagree,” “Disagree,” “Neither Agree nor Disagree,” “Agree,” “Strongly Agree”). A single participant responded that they didn’t agree with the statement and was excluded from analyses.

RESULTS AND DISCUSSION

We first examined whether repetition increased the perceived musicality of phrases heard in all three conditions. We used a linear mixed model to estimate ratings for a phrase as a function of condition and repetition (see Supplementary Material Table S2 for descriptions

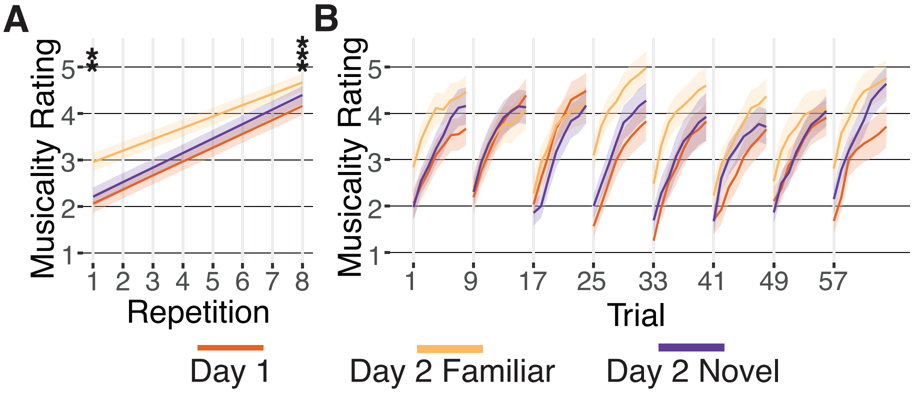

FIGURE 3. Effects of conditions across repetitions in Experiment 2. (A) Modeled change in musicality ratings across repetitions based on exposure condition. Asterisks denote significant contrast between conditions. (B) Average change in musicality ratings across trials based on exposure condition. The median value of the musicality scale is 4.5, minimum = 0 ("sounds exactly like speech"), maximum = 9 ("sounds exactly like song"). Shaded regions represent the SEM of the bootstrapped marginal estimates in A and the SEM of the participant means in B.

of analyses). Results show participants perceived phrases as more musical on the final repetition compared to the first in all conditions (Figure 3A). For each additional repetition, the musicality rating for a phrase increased by 0.301 (95% CI [0.268, 0.337], $p < .001$) in the day 1 condition, and by 0.244 (95% CI [0.209, 0.278], $p < .001$), and 0.312 (95% CI [0.279, 0.345], $p < .001$) in the day 2 familiar and novel conditions, respectively. Phrases heard in the day 2 novel and day 1 conditions demonstrated similar rates of increase in musicality (difference = 0.011, 95% CI [0.037, 0.058], $p = .647$), while day 2 familiar phrases increased at a slightly slower rate than day 2 novel phrases (difference = -0.067, 95% CI [-0.115, -0.023], $p = .004$) and day 1 phrases (difference = -0.057, 95% CI [-0.104, -0.011], $p = .018$). Post hoc analysis suggests the lower rate of change in perceived musicality in the day 2 familiar condition is the result of phrases starting off as more song-like on the first presentation.

TABLE 2. *Post hoc Contrast Comparing Experiment 2 Ratings Between Conditions at the First and Final Phrase Repetitions*

Contrasts	Coefficients	95% CI	*p*
rep. 1 D2 Familiar - D1	0.896*	[0.704, 1.097]	< .001
rep. 1 D2 Familiar - D2 Novel	0.745*	[0.543, 0.927]	< .001
rep. 1 D2 Novel - D1 Familiar	0.157	[−0.038, 0.349]	.124
rep. 8 D2 Familiar - D1	0.500*	[0.298, 0.690]	< .001
rep. 8 D2 Familiar - D2 Novel	0.274*	[0.067, 0.455]	.006
rep. 8 D2 Novel - D1 Familiar	0.226*	[0.038, 0.420]	.022

Note: Values are bootstrapped contrast coefficients and 95% CIs. Contrasts from all models reflect the difference between estimated marginal means. Asterisks denote significant effects. Day 1 condition (D1); Day 2 Familiar condition (D2 Familiar); Novel condition (D2 Novel); rep (repetition).

We predicted that phrase-specific representations encoded in long-term memory during the first session would increase the perceived musicality when the same phrase was heard again after the one-week delay period. Post hoc contrasts show that during the second session, perceived musicality was greater for the first repetitions of familiar phrases compared to the first presentations of novel phrases (difference = 0.745; Table 2 and Figure 3A). The first repetitions of phrases in the day 2 familiar condition were also perceived as more song-like compared to the first time the phrases were heard in the day 1 condition (difference = 0.896). Though we found evidence for an effect of long-term memory for phrases, perceived musicality still, on average, decreased by 1.214 rating scale values (95% CI [1.019, 1.408], $p < .001$) between the last repetition of a phrase in session one and the first repetition in session two. Day 2 familiar phrases heard for the second time in session two were also perceived as more song-like on the final repetition compared to day 2 novel phrases as well as in comparison to the final repetition of the same phrases at the end of session 1 (difference = 0.274 and difference = 0.500, respectively; Table 2). The different rates of change across repetitions are also clearly visible in the unmodeled data using participants' average ratings (Figure 3B). Overall, despite a slightly smaller rate of change across repetitions, day 2 familiar phrases started off and ended up being perceived as more song-like than phrases heard for the first time. The pattern of results demonstrates phrase-specific long-term memory contributes to the illusion.

Unexpectedly, even though we didn't find a difference in ratings between the first phrase presentations in the day 1 and day 2 novel conditions, day 2 novel phrases were rated as more song-like on the final repetition (difference = 0.226; Table 2). Though the difference was small, the finding suggests that task-related knowledge not specific to a stimulus but still conducive to the illusion can be learned and transferred to novel phrases, much like the transfer of skills in various learning paradigms (Kóbor et al., 2020; Mosha & Robertson, 2016; Schorn & Knowlton, 2021). Though we used phrases with different words spoken in distinct voices, the transfer effect may reflect pitch and or rhythmic similarities between phrases that weren't directly manipulated in the current study. Importantly, the current study was not designed to test for a transfer effect and the result could reflect changes in participant behavior across sessions unrelated to learning, e.g., an upward drift in participants' musicality ratings. Future work measuring the illusion across multiple days is required to establish the transfer effect.

General Discussion

The present study provides evidence for the role of rapid phrase-specific learning and long-term memory in the speech-to-song illusion. Experiments 1 and 2 required phrase-specific knowledge to be encoded and retained across delays for the illusion to occur. As a result, the magnitude of the change in perceived musicality across repetitions also served as a measure of phrase-specific memory. In Experiment 1, learning that took place across interleaved repetitions of different phrases led to increases in perceived musicality. The information encoded was sufficient to produce a more song-like perception even though the interleaved presentations prevented a phrase from being maintained by attentional and working memory processes. In Experiment 2, phrase-specific knowledge learned during the first session led to increases in perceived musicality at the start of the second session one week later. Even though participants only heard eight repetitions, the knowledge contributing to the illusion was consolidated into long-term memory and biased subsequent perception towards a more song-like experience. In both experiments, repetition provided the opportunity for phrase-specific learning to take place which was sufficient to produce the speech-to-song illusion.

Overall, participants in the present study experienced levels of musicality in the speech excerpts comparable to that experienced in previous studies using the same speech stimuli. Tierney et al. (2018b) reported average musicality rating values for the final repetitions in Experiments 1–3 of approximately 5 (using a 1–10 response scale) and a similar value was reported in Tierney et al. (2018a). The average musicality rating values for the final repetition in the present study correspond to a value slightly more song-like than the median scale value: 5.20 (Experiment 1 blocked condition), 4.16 (Experiment 1 interleaved condition), 5.16 (Experiment 2 day 1 condition), 5.66 (Experiment 2 day 2 familiar condition), 5.39 (Experiment 2 day 2 novel condition). Note that to facilitate comparison between previous studies that used a 1-10 response scale, the median values reported in this paragraph were shifted from the 0-9 response scale we used to analyze and report the results. Participants in the present study also experienced the illusion to an extent comparable to that experienced in previous studies using the same speech stimuli. The average subject-level difference between the last and first presentation in the present set of experiments was approximately 2 rating scale values: 2.01 (Exp. 1 blocked condition), 0.86 (Exp. 1 interleaved condition), 2.11 (Exp. 2 day 1 condition), 1.71 (Exp. 2 day 2 familiar condition), 2.18 (Exp. 2 day 2 novel condition). The same value, reported by Tierney et al. (2021) was approximately 1.9.

Once a phrase transforms, participants make clear predictions about the syllable sequence and are differentially sensitive to pitch and timing deviations when it recurs (Graber et al., 2017; Vanden Bosch der Nederlanden et al., 2015). We suggest that repetition transforms speech into song because the illusion requires learning how a phrase unfolds over time, much like how repetition drives learning of the tonal and temporal (rhythmic) sequence of a particular piece of music (Kubit & Janata, 2022a, 2022b). Previous work examined the illusion after removing temporal structure and found the illusion strength to be unaffected by the introduction of random jitter between syllables in a phrase (Falk et al., 2014; Graber et al., 2017; Tierney et al., 2018b) and between repetitions (Margulis et al., 2015). While beat structure and meter are undoubtedly important features in music, humans also learn to represent higher-order structure in temporal sequences (Dehaene et al., 2015; Janata & Grafton, 2003). For example, ordinal knowledge about a syllable sequence can be learned even if meter is lacking and more complex structures like melodic contour are independent of timing (Dowling, 1978; Hébert & Peretz, 1997). We hypothesize that repetition drives the learning of higher-order sequence structure in phrases and that the cognitive mechanisms that support sequence learning in the illusion are likely not unique to music listening (Janata & Grafton, 2003; Zatorre et al., 2007) or speech perception (Christiansen & Chater, 2008; Conway & Pisoni, 2008).

In Experiment 1, phrases heard in the interleaved condition were perceived as more song-like by the final repetition than they had been on the initial presentation; however, the final repetition of the blocked condition phrases was perceived as more song-like than the final repetition of the interleaved condition phrases (Table 1, Figure 2A). One explanation of these results is that phrase learning is facilitated by blocked repetition. Although the blocked condition precludes a direct measure of phrase-specific learning within a single experiment session free from the influence of attention and working memory processes, we found in Experiment 2 that familiar phrases that had been presented in session one were perceived as more song-like on the first repetition within session two, seven days later. The endurance of the perceptual transformation from speech to song suggests that phrase-specific learning influenced the illusion during blocked conditions in both experiments. Research on visuo-motor sequence learning provides evidence that blocked presentations lead to superior short-term retention but poor long-term memory of the sequences, compared to interleaved presentations (Schorn & Knowlton, 2021). During interleaved presentations, interference arising from other stimuli inhibits performance during learning, but the same interference eventually helps produce a more stable memory trace (Robertson et al., 2004; Shea & Morgan, 1979). The pattern of short-term learning observed in such studies resembles the pattern of musicality ratings between conditions in Experiment 1 and suggests that the perceptual illusion experienced for interleaved phrases would be better preserved over time. An important question for future work is whether additional interleaved repetitions can further strengthen the illusion such that the magnitude of the effect is comparable to that experienced after blocked repetitions. Finding that additional interleaved repetitions produce an illusion comparable to blocked repetitions would provide evidence that differences in learning rates lead to differences in illusion strength between conditions. Additionally, while the results of Experiment 1 demonstrated a reliable increase in perceived musicality in the interleaved condition, future work is needed to clarify whether some threshold exists at which an increase in perceived musicality is experienced as the illusion.

Participants tend to experience greater changes in perceived musicality across repetitions when phrases conform to melodic features typically found in Western music (Tierney et al., 2018b). Previous work has interpreted such results as support for the hypothesis that the illusion results from warping phrases to fit internalized templates of common musical patterns (Rowland et al., 2019; Tierney et al., 2018b). However, given the present results showing that the illusion entails the learning of phrase-specific knowledge, findings on stimulus-related differences can be explained in the context of their influence on learning. We hypothesize that differences in stimuli such as within- and between-syllable pitch slopes (Tierney et al., 2018b), rhythmic stability (Falk et al., 2014), and stimulus length (Rowland et al., 2019) influence the extent to which the structure of a temporal sequence can be learned. For example, shorter phrases and phrases composed of syllables that have flat rather than steep pitch contours produce a stronger illusion and may be easier to learn when heard repeatedly. Indeed, the mnemonic benefit of structure has been well documented in learning paradigms using visuo-motor sequences (Howard et al., 2004; Kóbor et al., 2020; Nissen & Bullemer, 1987) and music (Bharucha & Krumhansl, 1983; Dowling, 1991; Lévêque et al., 2022). One way of distinguishing between sequence learning and mechanisms that entail music-specific template matching (Rowland et al., 2019; Tierney et al., 2018a) is to test whether structures that benefit learning but are atypical music patterns still strengthen the speech-to-song illusion.

Music listening is an active process, during which predictions about what's next are continuously updated according to prior schematic knowledge of typical music structure as well as representations of veridical sequences in memory (Bharucha, 1987; Tillmann & Bigand, 2010; Vuust et al., 2022). The present study provides evidence that stimulus-specific learning shapes listeners' expectations across repetitions of spoken phrases and suggests that the influence of such expectations on perception underlies the transformation from speech to song. While the illusion manifests as an explicit awareness of a change in perceived musicality, the underlying changes in perceptual expectations may not require awareness of the learned knowledge or any overt effort by the listener. Sequence learning has been shown to improve task performance even when participants aren't explicitly aware of the sequences (Robertson, 2007) much like how implicit memory for music reflects prior exposure even when participants don't explicitly recognize the music (Halpern & Müllensiefen, 2008; Thorpe et al., 2019). Thus, perceiving a stimulus as music may not require active engagement, but simply the opportunity to learn sequence structures found in music—made possible by the repetition inherent to much of it (Margulis, 2014).

Author Note

The authors would like to thank Madeline Kushan for her help with setting up the online experiment, and Mara Breen and Aniruddh Patel for helpful early conversations about this work.

All audio files, data, and analysis code are available through the Open Science Framework at https://osf.io/7dxwu/ (https://doi.org/10.17605/OSF.IO/7DXWU).

Preliminary findings from this study were presented at the Society for Music Perception and Cognition 2022 Meeting in Portland, Oregon. All audio files, data, and analysis code are available through the Open Science Framework (https://osf.io/7dxwu/).

Correspondence concerning this article should be addressed to Benjamin Kubit, Department of Music, Princeton University, Princeton, NJ 08544. E-mail: bkubit@princeton.edu

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, *5*(1), 1–30. https://doi.org/10.2307/40285384

Bharucha, J. J., & Krumhansl, C. L. (1983). The representation of harmonic structure in music: Hierarchies of stability as a function of context. *Cognition*, *13*, 63–102. https://doi.org/10.1016/0010-0277(83)90003-3

Brem, A. K., Ran, K., & Pascual-Leone, A. (2013). Learning and memory. *Handbook of Clinical Neurology*, *116*, 696–737. https://doi.org/10.1016/B978-0-444-53497-2.00055-3

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Castro, N., Mendoza, J. M., Tampke, E. C., & Vitevitch, M. S. (2018). An account of the speech-to-song illusion using node structure theory. *PLOS ONE*, *13*(6), 1–32. https://doi.org/10.1371/journal.pone.0198656

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–558. https://doi.org/10.1017/S0140525X08004998

Conway, C. M., & Pisoni, D. B. (2008). Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences*, *1145*, 113–131. https://doi.org/10.1196/annals.1416.009

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review*, *119*(3), 480–499. https://doi.org/10.1037/a0027791

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*, 2–19. https://doi.org/10.1016/j.neuron.2015.09.019

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America*, *129*(4), 2245–2252. https://doi.org/10.1121/1.3562174

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354. https://doi.org/10.1037/0033-295X.85.4.341

Dowling, W. J. (1991). Tonal strength and melody recognition after long and short delays. *Perception and Psychophysics*, *50*(4), 305–313. https://doi.org/10.3758/BF03212222

Falk, S., Rathcke, T., & Bella, S. D. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1491–1506. https://doi.org/10.1037/a0036858

Graber, E., Simchy-Gross, R., & Margulis, E. H. (2017). Musical and linguistic listening modes in the speech-to-song illusion bias timing perception and absolute pitch memory. *Journal of the Acoustical Society of America*, *142*(6), 3593–3602. https://doi.org/10.1121/1.5016806

Halpern, A. R., & Müllensiefen, D. (2008). Effects of timbre and tempo change on memory for music. *Quarterly Journal of Experimental Psychology*, *61*(9), 1371–1384. https://doi.org/10.1080/17470210701508038

Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory and Cognition*, *25*(4), 518–533. https://doi.org/10.3758/BF03201127

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Howard, D. V., Howard Jr., J. H., Japikse, K., DiYanni, C., Thompson, A., & Somberg, R. (2004). Implicit sequence learning: Effects of level of structure, adult age, and extended practice. *Psychology and Aging*, *19*(1), 79–92. https://doi.org/10.1037/0882-7974.19.1.79

Janata, P., & Grafton, S. T. (2003). Swinging in the brain: Shared neural substrates for behaviors related to sequencing and music. *Nature Neuroscience*, *6*(7), 682–687. https://doi.org/10.1038/nn1081

Kóbor, A., Horváth, K., Kardos, Z., Nemeth, D., & Janacsek, K. (2020). Perceiving structure in unstructured stimuli: Implicitly acquired prior knowledge impacts the processing of unpredictable transitional probabilities. *Cognition*, *205*, 104413. https://doi.org/10.1016/j.cognition.2020.104413

Kubit, B. M., & Janata, P. (2022a). Spontaneous mental replay of music improves memory for incidentally associated event knowledge. *Journal of Experimental Psychology: General*, *151*(1), 1–24. https://doi.org/10.1037/xge0001050

Kubit, B. M., & Janata, P. (2022b). Spontaneous mental replay of music improves memory for musical sequence knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. https://doi.org/10.1037/xlm0001203

Lévêque, Y., Lalitte, P., Fornoni, L., Pralus, A., Albouy, P., Bouchet, P., et al. (2022). Tonal structures benefit short-term memory for real music: Evidence from non-musicians and individuals with congenital amusia. *Brain and Cognition*, *161*(May). https://doi.org/10.1016/j.bandc.2022.105881

Margulis, E. H. (2014). *On repeat: How music plays the mind*. Oxford University Press.

Margulis, E. H., & Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music Perception*, *33*(4), 509–514. https://doi.org/10.1525/mp.2016.33.4.509

Margulis, E. H., Simchy-Gross, R., & Black, J. L. (2015). Pronunciation difficulty, temporal regularity, and the speech-to-song illusion. *Frontiers in Psychology*, *6*, 1–7. https://doi.org/10.3389/fpsyg.2015.00048

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Mosha, N., & Robertson, E. M. (2016). Unstable memories create a high-level representation that enables learning transfer. *Current Biology*, *26*(1), 100–105. https://doi.org/10.1016/j.cub.2015.11.035

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS ONE*, *9*(2), 1–23. https://doi.org/10.1371/journal.pone.0089642

Naveh-Benjamin, M., & Cowan, N. (2023). The roles of attention, executive function and knowledge in cognitive ageing of working memory. *Nature Reviews Psychology*, *2*, 151–165. https://doi.org/10.1038/s44159-023-00149-0

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cogntive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-0285(87)90002-8

Robertson, E. M. (2007). The serial reaction time task: Implicit motor skill learning? *Journal of Neuroscience*, *27*(38), 10073–10075. https://doi.org/10.1523/JNEUROSCI.2747-07.2007

Robertson, E. M., Pascual-Leone, A., & Miall, R. C. C. (2004). Current concepts in procedural consolidation. *Nature Reviews Neuroscience*, *5*(7), 576–582. https://doi.org/10.1038/nrn1426

Rowland, J., Kasdan, A., & Poeppel, D. (2019). There is music in repetition: Looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychonomic Bulletin and Review*, *26*(2), 583–590. https://doi.org/10.3758/s13423-018-1527-5

Schorn, J. M., & Knowlton, B. J. (2021). Interleaved practice benefits implicit sequence learning and transfer. *Memory and Cognition*, *49*(7), 1436–1452. https://doi.org/10.3758/s13421-021-01168-z

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187. https://doi.org/10.1037//0278-7393.5.2.179

Soehlke, L. E., Kamat, A., Castro, N., & Vitevitch, M. S. (2022). The influence of memory on the speech-to-song illusion. *Memory and Cognition*, 1–12. https://doi.org/10.3758/s13421-021-01269-9

Thorpe, L., Cousins, M., & Bramwell, R. (2019). Implicit knowledge and memory for musical stimuli in musicians and non-musicians. *Psychology of Music*, *48*(6), 836–845. https://doi.org/10.1177/0305735619833456

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus song: Multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, *23*(2), 249–254. https://doi.org/10.1093/cercor/bhs003

Tierney, A., Patel, A. D., & Breen, M. (2018a). Repetition enhances the musicality of speech and tone stimuli to similar degrees. *Music Perception*, *35*(5), 573–578. https://doi.org/https://doi.org/10.1525/MP.2018.35.5.573

Tierney, A., Patel, A. D., & Breen, M. (2018b). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, *147*(6), 888–904. https://doi.org/10.1037/xge0000455

Tierney, A., Patel, A. D., Jasmin, K., & Breen, M. (2021). Individual differences in perception of the speech-to-song illusion are linked to musical aptitude but not musical training. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(12), 1681–1697.

Tillmann, B., & Bigand, E. (2010). Musical structure processing after repeated listening: Schematic expectations resist veridical expectations. *Musicae Scientiae*, *14*(2_suppl), 33-47. https://doi.org/doi:10.1177/10298649100140S204

van Ede, F., & Nobre, A. C. (2023). Turning attention inside out: How working memory serves behavior. *Annual Review of Psychology*, *74*, 137–165. https://doi.org/10.1146/annurev-psych-021422-041757

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Finding the music of speech: Musical knowledge influences pitch processing in speech. *Cognition*, *143*, 135–140. https://doi.org/10.1016/j.cognition.2015.06.015

Vitevitch, M. S., Ng, J. W., Fatley, E., & Castro, N. (2021). Phonological but not semantic influences on the speech-to-song illusion. *Quarterly Journal of Experimental Psychology*, *74*(4), 585–597. DOI: https://doi.org/10.1177/1747021820969144

Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*, *23*(5), 287–305. https://doi.org/10.1038/s41583-022-00578-5

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547–558. https://doi.org/10.1038/nrn2152