sempre:
Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Article

Music & Science
Volume 1: 1–6

DOI: 10.1177/2059204317731992
journals.sagepub.com/home/mns
SAGE

The sound-to-music illusion: Repetition can musicalize nonspeech sounds

Rhimmon Simchy-Gross[1] and Elizabeth Hellmuth Margulis[1]

Abstract

The speech-to-song illusion tracks a perceptual transformation across repetitions where a stimulus that originally sounded like speech comes to sound like song. This article examines whether the illusion also generalizes to other kinds of nonspeech sounds. Participants heard each of 20 environmental sound clips repeated in either original or jumbled form. They rated the musicality of the clips on a 5-point scale where 1 represented *sounds exactly like environmental sound* and 5 *sounds exactly like music*. Average ratings increased significantly across repetitions, suggesting that the speech-to-song illusion is one form of a more general sound-to-music illusion produced by repetition. This illusion occurred regardless of whether the clips were repeated in original or jumbled form, marking a difference compared to speech, for which the illusion only occurred if the repetitions were exact.

Keywords

Environmental sound, music and language, repetition, sound-to-music illusion, speech-to-song illusion

Submission date: 28 June 2017; Acceptance date: 28 August 2017

The speech-to-song illusion occurs when a brief segment of a spoken utterance is digitally excised and repeated. In some cases, after this series of reported exposures, listeners report that the excised segment, which had initially sounded like speech, now sounds like song (Deutsch, Henthorn, & Lapidis, 2011). The illusion occurs in people with and without musical training (Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015a). Tierney et al. (2013) used neuroimaging to demonstrate that brain areas subserving pitch extraction and song production were more active in response to utterances subject to this illusion than utterances that were not. Perceptions of musicality, thus, depend not merely on the characteristics of the acoustic signal, but also on the contextual frame within which it is presented. In the case of the speech-to-song illusion, the relevant contextual frame consists of a series of repetitions of the same segment.

Prior to the discovery of this illusion by science, composers had been exploiting it for years. American composer Steve Reich used excerpts of recorded speech in compositions from the mid-1960s such as "It's Gonna Rain" and "Come Out," looping them repeatedly in a way that foregrounded the prosodic and rhythmic elements of the speech—framing them as music. The same kind of looping was also applied to other nonspeech recorded sounds, to much the same effect.

Compositional practice thus tends to suggest that perceptions of musicality arise not merely out of the acoustic characteristics of sounds themselves, but also out of the contextual usage of these sounds. It is particularly common for musical cultures throughout the world to create musical contexts rich with repetition (Fitch, 2006). Margulis (2014) theorizes that repetition plays a special role in music—in particular, that it encourages a musical orientation to sound by drawing attention to its sonic characteristics over its everyday meanings and by drawing listeners into a more participatory attitude. People exposed to random sequences of tones either once or on a 6-time loop later rated the ones they had heard on loop as more musical (Margulis & Simchy-Gross, 2016).

The speech-to-song illusion demonstrates the capacity of repetition to transform speech into music, but speech is already comprised of human voices communicating intentionally, much like conventional song. This article asks:

[1] University of Arkansas, USA

Corresponding author:
Elizabeth Hellmuth Margulis, Department of Music, University of Arkansas, 201 Music Building, Fayetteville, AR 72701, USA.
Email: ehm@uark.edu

Are environmental sounds, produced by actions, objects, and animals, also subject to transformation into music across repetitions? If no, it would suggest that speech and music are uniquely intertwined. Under this scenario, given the proper context (e.g., a string of repetitions), speech can relatively easily sound like music, but other sorts of sounds lack this latent potential for musicalization, remaining tied instead to their everyday or environmental function. If yes, on the other hand, it would suggest that speech's relationship to music is less unique, with repetition capable of musicalizing even sounds arbitrarily generated by a shovel being dragged across rocks or water dripping from a faucet. A finding that nonspeech songs could transform to music would strengthen the case for repetition's general power to musicalize.

Studies on the speech-to-song illusion suggest that when linguistic processing areas fully capture a sound stimulus, it can be harder for repetition to transform it into song. Speech transforms more readily to song when listeners don't speak the language of the utterance (Jaisin, Suphanchaimat, Candia, & Warren, 2016), and when the utterance is spoken in a language more difficult for them to pronounce (Margulis, Simchy-Gross, & Black, 2015)—both situations in which speech circuitry might possess a less tight grip on the acoustic signal from the beginning. If this hypothesis were correct, then environmental sounds, which presumably activate speech circuitry even less, should transform to music more easily than spoken utterances. Neuroimaging studies show that speech and environmental sounds activate different regions of the temporal lobe (Binder et al., 2000; Leaver & Rauschecker, 2010; Specht, Osnes, & Hugdahl, 2009). A recent paper by Norman-Haignere, Kanwisher, and McDermott (2015) identifies distinct cortical pathways for processing speech, music, and environmental sounds.

In the experiment presented here, we used the methodology from the original speech-to-song illusion paper (Deutsch et al., 2011) to investigate whether repetition can musicalize environmental sounds the same way it can musicalize speech.

Method

Participants

A total of 58 undergraduate students (34 females; ages ranging from 18 to 39 years; $M_{age} = 19.5$; $SD = 2.8$) who were enrolled at the University of Arkansas volunteered to participate in this experiment in exchange for course credit. All of the participants were recruited from an introductory class in general psychology. Forty of them reported having some level of musical training ($M_{years} = 2.9$; $SD = 3.1$) and three were music majors, but the musical training they reported consisted mostly of participation in school groups such as band and choir. All of the participants gave informed consent before participating in this experiment. The experiment was approved by the University of Arkansas Institutional Review Board.

Table 1. Clip total duration and fundamental frequency.

Clip	Duration (s)	Median F0 (Hz)
Bee buzz	2.59	127.14
Bubbles	3.05	482.50
Chicken cackle	3.94	481.19
Door noise	3.52	457.68
Ice cracks	3.26	432.37
Frogs and sheep	3.08	525.58
Jungle animals I	2.37	281.62
Jungle animals II	4.26	443.71
Machine noise	3.03	532.45
Seagulls	3.63	331.42
Nightingale song	3.19	489.18
Owl and birds	2.63	537.73
Rally car	2.96	152.00
Shovel drag	3.08	239.59
Water drops I	2.98	551.42
Water drops II	3.24	N/A
Pebbles into water	3.26	421.99
Whale song	3.99	170.85
Wind I	2.98	304.81
Wind II	3.41	103.51

Note. The median F0s were calculated in Praat. F0 information for water drops was "Undefined."

Materials

Twenty 10 s clips of environmental sound were taken from free sound effects websites. Listed in Table 1, they featured sounds such as water dripping, a shovel being dragged across a rock, a rally car driving by, or a whale vocalizing. None of the clips included human vocal sounds or traditional musical instruments.

Following Deutsch et al. (2011), we digitally excised a segment from the second half of each clip using Audacity 2.0.6. The mean segment length was 3.2 s ($SD = 0.5$). The segment was presented either in untransformed form or in jumbled form. We created eight different jumbled versions by splicing each segment into seven pieces, and then jumbling the order of these seven pieces as follows: 6, 4, 3, 2, 5, 7, 1; 7, 5, 4, 1, 3, 2, 6; 1, 3, 5, 7, 6, 2, 4; 3, 6, 2, 5, 7, 1, 4; 2, 6, 1, 7, 4, 3, 5; 4, 7, 1, 3, 5, 2, 6; 6, 1, 5, 3, 2, 4, 7; and 2, 5, 4, 3, 7, 1, 6. We aimed to make our segmentation method as analogous as possible to that of Deutsch et al. (2011). They segmented the clips based on the spoken syllables. Our clips did not have spoken syllables, so we segmented the clips by marking the six most salient points of auditory separation (e.g., by identifying amplitude peaks, sound onsets, and perceptual groupings). These points did not always take place at silent intervals, and we did not use ramping to guard against transients.

Procedure

Participants were tested individually in a $4' \times 4'$ WhisperRoom Sound Isolation Enclosure (MDL 4848E/ENV). They wore Sennheiser HD 600 headphones facing a $22''$

Dell P2212H monitor and made responses using a computer keyboard and mouse. The auditory stimuli were presented binaurally at a comfortable listening level. The experiment was presented using Medialab (Version 2016.1.104; Jarvis, 2016) on a Dell OptiPlex 7010 desktop computer running Windows 7.

The design of this experiment was modeled on the first experiment in Deutsch et al. (2011). As in that study, for each trial we first presented the full 10 s clip. After a 3 s pause, we presented 10 clip segments, each of which was also followed by a 3 s pause. During each pause that followed a clip segment, participants rated the clip on a Likert-type scale from 1 to 5, where 1 indicated *sounded exactly like environmental sound* and 5 indicated *sounded exactly like music*.

We randomly assigned participants to one of two conditions—untransformed and jumbled. In both conditions, the initial and final (first and tenth) segments were presented in their original, untransformed form. In the untransformed condition, the eight intervening segments were identical—they consisted of the same untransformed segment. In the jumbled condition, however, the eight intervening segments consisted of the eight different jumbled segment versions (presented in the order listed above, following Deutsch et al., 2011). Participants completed a short demographic questionnaire to conclude the experiment. The experiment lasted about 25 min.

Results

To account for random effects within participants, we ran a linear mixed model (Baayen, Davidson, & Bates, 2008). The within-subjects fixed effect was Presentation (initial or final). The between-subjects fixed effect was Condition (untransformed or jumbled). The random effects were Subject ($n = 58$) and Item (the 20 environmental sound clips). The data consisted of 2320 ratings, 27 of which were identified as error ratings (responses not registered within the 3 s time limit) and excluded from the analysis. We first ran the maximal model with all of the factors, interactions, and random slopes with the subject and item grouping variables in order to account for random slope variance and obtain model convergence (see Barr, Levy, Scheepers, & Tily, 2013). The final converged model included the full Presentation by Condition interaction, and the random slope of Presentation within the Subject and Item grouping variables.

Figure 1 shows the main effect of Presentation. The mean rating of the final presentations ($M = 2.45$, $SD = 1.31$) was significantly higher than mean rating of the initial presentations ($M = 1.84$, $SD = 1.17$), $F_{(1, 52)} = 40.53$, $p < .0001$. These results suggest that the excerpts sounded more like music (and less like environmental sound) on the final presentation than on the initial presentation.

We did not find a main effect of Condition, $F_{(1, 56)} = 2.02$, $p = .160$, or an interaction between Presentation and Condition ($F < 1$; see Figure 2). These results suggest that the transformation from sound to music across repetitions was unaffected by whether the segments were presented identically or in jumbled form. These findings contrast with those in Deutsch et al. (2011), which show that speech transformed to song only when it was presented in a series of identical (rather than jumbled) repetitions.

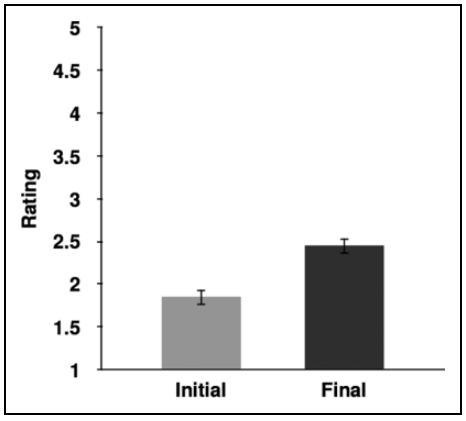

Figure 1. Mean musicality ratings (±1 standard error of the mean) on the initial and final presentations.

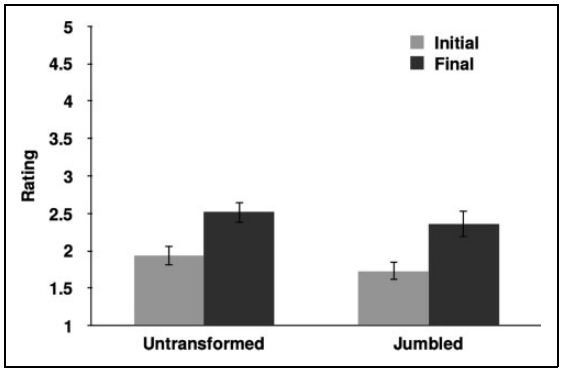

Figure 2. Mean musicality ratings (±1 standard error of the mean) on the initial and final presentations as a function of Condition.

Following Deutsch et al. (2011), our principal interest was the transformation effected between the first and final repetition; however, we also examined the mean rating changes during the intervening repetitions. Figure 3 traces these changes across all 10 of the repetitions. On average, the segment was perceived as increasingly musical across each of these repetitions in the untransformed condition. In the jumbled condition, however, ratings increased across the first 9 repetitions, but dipped slightly on the 10th, when the original, untransformed version of the utterance recurred.

Discussion

This study shows that the speech-to-song illusion can be viewed as one example of a more generalized sound-to-music illusion.

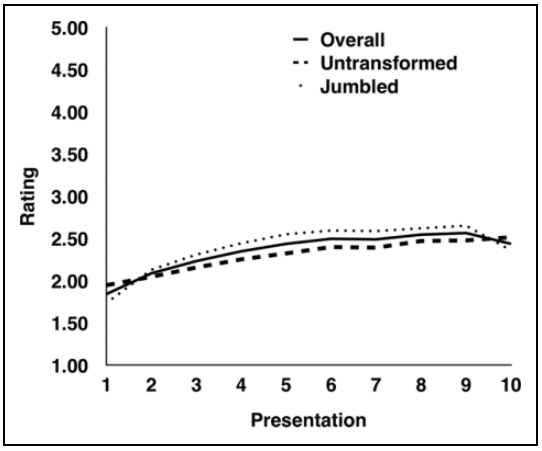

Figure 3. Mean ratings for each presentation, collapsed across clips, as a function of conditions (untransformed and jumbled) and collapsed across conditions (overall).

Across repetitions, environmental sounds can come to seem more music-like than they did on initial presentation. This sound-to-music illusion differs from the speech-to-music illusion in that it can occur regardless of whether the stimuli are repeated exactly or in jumbled form. Jumbling speech excerpts by scrambling the order of the syllables disrupts their semantic meaning—the level at which speech is typically apprehended; but jumbling excerpts of environmental sounds does not disturb any such semantic dimension. A succession of drops of water that has been rearranged is still just a succession of drops of water.

Environmental sounds are typically apprehended at the level of source identification (Gaver, 1993)—listeners hear that a certain string of sounds comes from a dripping faucet and another from a shovel being dragged across rocks. Rearranging individual components of the sound does not tend to alter this source identification. Moreover, listeners possess keen categorical representations for phonemes within speech (Liberman, Harris, Hoffman, & Griffith, 1957; Schouten & van Hessen, 1992), but lack such stable categories for the individual components of environmental sounds. Given these differences, the jumbled repetitions likely continued to sound like iterations of the same fundamental stimulus—the sound of water dripping or a shovel being dragged, for example—but the jumbled repetitions of speech in Deutsch et al. (2011) fundamentally altered what was being heard. Words that have had their individual phonemes rearranged no longer sound like repetitions of the same word and no longer possess the same prosodic contour. The differential effect of jumbling on speech and sound excerpts underscores that differences in what constitutes a perceived repetition for various types of sounds might impact their susceptibility to musicalization.

Although musicality ratings increased, on average, across each of the 10 repetitions in both conditions, they dipped slightly on the last one in the jumbled condition. Although repetitions 2 through 9 were scrambled, repetition 10 returned to the untransformed version, in which the segments followed one another in the actual order of the original sound file. Musicality ratings for the last statement were higher than musicality ratings for the first one, but they were lower than ratings for some of the intervening versions. This pattern would result if two different factors were influencing ratings: (1) a tendency to perceive increased musicality with each repetition and (2) a tendency to perceive increased musicality in jumbled compared to untransformed versions. This tendency to hear more musicality in jumbled versions could arise from the sense that they have already been digitally manipulated—a potential mark of human artistic intent—in comparison to the unaltered recordings of naturally occurring environmental sounds. Future work could disentangle these effects by varying the amount of manipulation in the original sound file.

The speech-to-song illusion might depend on semantic satiation (Severance & Washburn, 1907) to suppress semantic associations before musical listening can emerge. The sound-to-music illusion, by contrast, might depend on the suppression of the saliency of the source identification to allow musical attending to emerge. In the speech-to-song illusion (Deutsch et al., 2011), only the untransformed condition allowed illusory transformation to music to occur, likely because semantic satiation (or the perceptual deterioration of semantic meaning following repeated exposure) can only occur across repetitions of words with distinct semantic meanings. But in the sound-to-music illusion, both the untransformed and the jumbled condition allowed illusory transformation to music to occur, likely because the source identification can be preserved even in the case of segment jumbling.

Studies on the speech-to-song illusion suggest that musicalization occurs more easily when speech circuitry captures a sound sequence less firmly. Extending this interpretation to environmental sound might imply that because environmental sound captures speech circuitry even less than speech, musicalization should occur more easily. Yet we found that strings of environmental sounds did not transform to music more effectively than the speech stimuli in Deutsch et al. (2011).

Mean ratings in the untransformed condition in Deutsch et al. (2011) were below 1.5 after the initial presentation and above 3.5 after the final presentation. Mean ratings in the untransformed condition in our study, on the other hand, were 1.84 after the initial presentation and 2.45 after the final presentation. Although mean ratings of environmental sound after the final presentation increased significantly compared to the initial presentation, they did not extend over the rating threshold—3.0—the value midway between the scalar endpoints (1 for *sounds exactly like environmental sound* and 5 for *sounds exactly like music*), beyond which a rating would indicate that the segment sounded more like music than like environmental sound.

Cursory examination would seem to suggest that the illusory effect was therefore weaker for environmental sounds than for speech. Yet the single speech segment used in the first Deutsch et al. (2011) study was likely one that transformed particularly successfully.

Five stimuli in our study underwent transformations that exceeded the 3.0 threshold—two examples of water dripping, the excerpt featuring whale song, the excerpt featuring ice cracking, and the one featuring a dragged shovel. To examine the size of more typical speech-to-song transformations, it can help to look at Margulis et al. (2015), which—unlike a number of other speech-to-song papers—reported initial and final means. The speech stimuli in the category that transformed most easily started with a mean initial rating of 1.56 and transformed to a mean final rating of 2.4. Transformations of speech in that study, like the environmental sound stimuli reported here, often did not surpass the threshold of 3.0 and were comparable in effect size to the transformations reported in the present study. The speech-to-song and sound-to-music illusions, however, are ideally compared within subjects, and it would help to study them within the same pool of participants to draw firm conclusions about comparative effect size.

Future work could use a service like Amazon's Mechanical Turk to collect large amounts of data on transformations from speech to song and from sound to music to determine whether either of these categories transforms more readily. Falk, Rathcke, and Dalla Bella (2014) found that utterances with stable tonal targets and recurring durational contrasts transformed to song more readily than utterances without these features. Merrill and Larrouy-Maestri (2017) reinforced the central role of pitch and suggested that timbre and register play an important role as well—qualities that may be especially important to the border between nonspeech sounds and music. A study with sufficient data could also make it possible to analyze in more detail the acoustic features that lead to the transformation from speech to song and ascertain whether they are the same acoustic features that lead to the transformation from sound to music.

It would also be useful to obtain cross-cultural responses on these tasks. The pervasiveness of particular environmental sounds differs across cultures. By comparing how easily sound sequences transform to music in cultures within which the sequences are common or uncommon, research can identify whether speech's tendency to transform more easily when spoken in an unfamiliar language extends to a tendency for environmental sounds to transform more easily when they are unfamiliar. Vanden Bosch der Nederlanden, Hannon, and Snyder (2015b) showed that listeners were more sensitive to pitch changes that violated rather than conformed to familiar musical structures in utterances that transformed to music—suggesting a role for enculturated notions of typical pitch sequences.

Finally, the sound-to-music illusion makes possible a number of investigations of individual differences. It has previously been established that not everyone experiences the transformation from speech to song in the classic illusion (Deutsch et al., 2011). Are the people who are not susceptible to the speech-to-song illusion the same people who are not susceptible to the sound-to-music illusion? Does experience with 20th-century music that uses environmental sound as materials influence illusion susceptibility?

This article documents the existence of a sound-to-music illusion that can be thought of as a generalization of or interesting comparison case to the speech-to-song illusion. It provides behavioral evidence for the effects of a common 20th-century compositional tool—repetition—allowing for future scientific investigations that expose the domain specificity or generality of aspects of auditory processing. In general, psychologists have thought about the speech-to-song illusion as a phenomenon at the nexus of language and music; however, this study suggests that repetition musicalizes nonspeech sounds with similar ease. The speech-to-song illusion may function less as evidence of some special overlap between language and music and more as an index of the power of repetition to encourage a different and more musical orientation to the sound. Both the speech-to-song and sound-to-music illusion raise questions about what kinds of acoustic characteristics allow this transformation to occur—Are they the same regardless of whether the string of sounds started out being heard as speech or as sound? What are the limits of what can be musicalized in this way, and how do they vary culturally? One source for insight into these questions might be compositional practice. Here, as in many other cases, artistic innovation tends to precede scientific understanding.

Contributorship

EHM conceived of the study. RS and EHM designed the study. RS analyzed the data. RS and EHM co-wrote the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Peer review

Adam Tierney, Birkbeck College.
Daniela Sammler, Max Planck Institute for Human Cognitive and Brain Sciences, Otto Hahn Group Neural Bases of Intonation in Speech and Music.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*, 512–528.

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

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1491–1506.

Fitch, W.T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*, 173–215.

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, *5*, 1–29.

Jaisin, K., Suphanchaimat, R., Candia, M. A. F., & Warren, J. D. (2016). The speech-to-song illusion is reduced in speakers of tonal (vs. non-tonal) languages. *Frontiers in Psychology*, *7*, 662.

Jarvis, B. (2016). *MediaLab*. [Computer Software]. New York: Empirisoft Corporation.

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience*, *30*, 7604–7612.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

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

Margulis, E. H., & Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music Perception*, *33*, 509–514.

Margulis, E. H., Simchy-Gross, R., & Black, J. L. (2015). Pronunciation difficulty, temporal regularity, and the speech-to-song illusion. *Frontiers in Psychology*, *6*, 48.

Merrill, J., & Larrouy-Maestri, P. (2017) Vocal features of song and speech: Insights from Schoenberg's Pierrot lunaire. *Frontiers in Psychology*, *8*, 1108.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*, 1281–1296.

Schouten, M. E. H., & van Hessen, A. J. (1992). Modeling phoneme perception. I: Categorical perception. *The Journal of the Acoustical Society of America*, *92*, 1841–1855.

Severance, E., & Washburn, M. F. (1907). The loss of associative power in words after long fixation. *The American Journal of Psychology*, *18*, 182–186.

Specht, K., Osnes, B., & Hugdahl, K. (2009). Detection of differential speech-specific processes in the temporal lobe using fMRI and a dynamic "sound morphing" technique. *Human Brain Mapping*, *30*, 3436–3444.

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*, 249–254.

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015a). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology: General*, *144*, e43–e49.

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015b). Finding the music of speech: Musical knowledge influences pitch processing in speech. *Cognition*, *143*, 135–140.