A Rational Analysis of the Speech-to-Song Illusion

Raja Marjieh[1], Pol van Rijn[2], Ilia Sucholutsky[3], Harin Lee[2,4], Thomas L. Griffiths[1,3,*], Nori Jacoby[2,*]
[1]Department of Psychology, Princeton University
[2]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics
[3]Department of Computer Science, Princeton University
[4]Max Planck Institute for Human Cognitive and Brain Sciences
[*]Equal Contribution
{raja.marjieh, is2961, tomg}@princeton.edu; {pol.van-rijn, harin.lee, nori.jacoby}@ae.mpg.de

Abstract

The speech-to-song illusion is a robust psychological phenomenon whereby a spoken sentence sounds increasingly more musical as it is repeated. Despite decades of research, a complete formal account of this transformation is still lacking, and some of its nuanced characteristics, namely, that certain phrases appear to transform while others do not, is not well understood. Here we provide a formal account of this phenomenon, by recasting it as a statistical inference whereby a rational agent attempts to decide whether a sequence of utterances is more likely to have been produced in a song or speech. Using this approach and analyzing song and speech corpora, we further introduce a novel prose-to-lyrics illusion that is purely text-based. In this illusion, simply duplicating written sentences makes them appear more like song lyrics. We provide robust evidence for this new illusion in both human participants and large language models.

Keywords: speech-to-song, music cognition, rational analysis, Bayesian modeling

Introduction

First published as an audio demonstration just over two decades ago (Deutsch, 2003, 2019), the speech-to-song illusion concerns a curious phenomenon whereby spoken phrases can be made to sound more song-like simply by repetition. This observation has been replicated and elaborated on in numerous studies (Deutsch, Henthorn, & Lapidis, 2011; Castro, Mendoza, Tampke, & Vitevitch, 2018; Tierney, Patel, & Breen, 2018a, 2018b; Margulis, Simchy-Gross, & Black, 2015; Simchy-Gross & Margulis, 2018; Falk, Rathcke, & Dalla Bella, 2014; Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015; Rowland, Kasdan, & Poeppel, 2019), and it taps into the deep relationship between speech and music which itself has been the center of considerable inquiry (Zatorre, Belin, & Penhune, 2002; Albouy, Benjamin, Morillon, & Zatorre, 2020; Ozaki et al., 2022; Albouy, Mehr, Hoyer, Ginzburg, & Zatorre, 2023; Ding et al., 2017) tracing as far back as the works of eighteenth century philosophers such as Rousseau and Herder (Rousseau, 1782; Herder, 2002).

Despite this rich research tradition, there is still no unifying theoretical account for the full phenomenology of the illusion, including, (i) what causes repetition on average to transform spoken phrases to song? (Deutsch et al., 2011) (ii) why does the strength of the effect increase with the number of repetitions? (Tierney et al., 2018a) (iii) why is it that certain phrases transform whereas others do not? (Tierney, Dick, Deutsch, & Sereno, 2013) and (iv) how does this connect to the statistics of speech and music learned throughout life and culture?

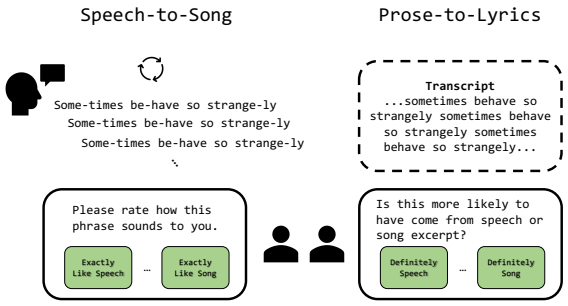

Figure 1: The speech-to-song paradigm and the newly proposed prose-to-lyrics paradigm.

(Margulis et al., 2015; Jaisin, Suphanchaimat, Figueroa Candia, & Warren, 2016)

In the present work, we propose one such framework by analyzing the abstract computational problem underlying the speech-to-song illusion and recasting it as rational statistical inference (Anderson, 1990; Griffiths & Tenenbaum, 2007). Specifically, we argue that the phenomenon can be understood in terms of the extent to which a given stimulus (or its repetitions) is more likely to originate from a generative model of speech versus song. Using Bayesian inference, the problem of determining whether a stimulus is more speech-like or song-like reduces to analyzing the probability of that stimulus under suitable speech and song corpora. We show that textual transcription datasets of speech and songs, both naturalistic and synthetically generated by large language models (LLMs), are indeed sufficient to reproduce some of the key features of the illusion, and motivate a hypothesis regarding a new "prose-to-lyrics" illusion that is based purely on text (i.e., no acoustic signal is provided) which we evaluate in both humans and LLMs (GPT-4; Achiam et al., 2023).

Our framework provides a principled account of the speech-to-song illusion that explicitly roots it in the learned statistics of speech and music, and allows for generalization across multiple modalities given suitable generative models. The paper proceeds as follows. We first review the empirical literature on the speech-to-song illusion, and the fundamentals of rational analysis in the context of a two-hypothesis decision problem. We then explicitly apply the latter to the speech-to-song problem and use it to motivate the textual ver-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

sion of the illusion. We then summarize the technical details of the corpus analysis and the human and LLM experiments, then present our results and distill their implications.

Background

Speech-to-Song Illusion

In its original form, the speech-to-song paradigm involves presenting participants with a spoken phrase (famously, "sometimes behave so strangely") and repeating it several times (Deutsch et al., 2011; Deutsch, 2019). In the pause between each repetition, participants report how the last utterance sounds to them from "exactly like speech" to "exactly like song" using a numeric Likert scale (Figure 1; Speech-to-Song). Deutsch et al. (2011) showed that this simple manipulation is sufficient to drive participants to perceive the speech excerpt as more like song than speech relative to the original (and identical) presentation. In a subsequent experiment, the authors asked participants to repeat the utterance after one and ten presentations, and found that in the former case the acoustic characteristics of the participants' responses (their fundamental frequency f_0 contour) resembled those of speech utterances, whereas in the latter case those of a tonal melody. The fact that stimuli in all repetitions are identical is significant as it allows for studying the internal processing of speech and music without relying on different stimuli.

Empirical research into the speech-to-song phenomenon has expanded substantially since the original study of Deutsch et al. (2011). First, Margulis et al. (2015) provided some evidence suggesting that the phenomenon may be affected by linguistic proficiency. Specifically, the authors showed that speech excerpts from languages that are hard to pronounce for native English speakers tended to transform more readily than those from languages that are easier to pronounce. Second, the increase of musicality of repeated sounds has been shown to generalize to non-speech sounds such as random tone sequences (Margulis & Simchy-Gross, 2016), environmental sounds like dripping water (Rowland et al., 2019), and animal sounds (Simchy-Gross & Margulis, 2018). Third, while repetition tends to make stimuli sound more musical on average, it has been shown that the effect at the level of individual stimuli is more nuanced, with some transforming to song effectively while others resisting transformation altogether (Tierney et al., 2018a). Specifically, Tierney et al. (2013) meticulously scanned audiobooks in English to find "song-like" sentences that show enhanced effect of repetition. They then contrasted these stimuli with similar utterances that, without repetition, have comparable speech vs. song ratings, but do not exhibit significant transformation when repeated.

Further research (Tierney et al., 2018a, 2018b) revealed that this interaction also occurs when participants are presented with only the fundamental frequency (f_0) contour of a spoken utterance. However, simple acoustic manipulations of the contour (such as changing the beat consistency, pitch slope, and melodic structure) that were supposed to enhance the effect did not influence musicality ratings as expected. These findings highlight the complexity of extracting acoustic correlates to predict whether a given phrase should transform under repetition (Falk et al., 2014; Tierney et al., 2018a).

Rational Analysis

Rational analysis is a cognitive modeling strategy that has been effectively applied across a wide array of topics including categorization (Anderson, 1990), causal induction (Griffiths & Tenenbaum, 2007), perception (Kersten & Yuille, 2003), and semantic memory (Griffiths, Steyvers, & Tenenbaum, 2007). The core idea behind rational analysis is to analyze human behavior from the perspective of the abstract computational problem that it attempts to solve and the optimal solution to that problem (Anderson, 1990). This is usually implemented as Bayesian inference whereby the posterior probabilities of different hypotheses $p(h|d)$ are estimated by inverting a generative model of the data d specified by some prior $p(h)$ and likelihood $p(d|h)$ using Bayes rule

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_h p(d|h)p(h)} \tag{1}$$

The prior captures inductive bias towards certain hypotheses independent of the data, the likelihood captures how well hypotheses are supported by the data, and the posterior integrates the two. The main challenge in applying Bayes rule is computing the denominator as it often requires summing over a large number of hypotheses. In the context of binary decision problems, however, in which an agent needs to decide between two hypotheses h_0 and h_1, a simpler formula can be derived by taking the ratio between the posteriors of the two hypotheses. This is known as the log-odds form

$$\log \frac{p(h_1|d)}{p(h_0|d)} = \log \frac{p(d|h_1)}{p(d|h_0)} + \log \frac{p(h_1)}{p(h_0)} \tag{2}$$

From here we see that the extent to which a given hypothesis is preferred over the other ultimately depends on the likelihood ratio of the data as well as the prior odds.

A Generative Account of Speech-to-Song

We now leverage the two-hypothesis Bayesian log-odds formula to formulate a rational model of the speech-to-song task. The data in this case is the presented utterance s, or its n-repetitions, which we will denote by s^n. The log-odds formula for the speech-to-song task then becomes

$$\log \frac{p(\text{song}|s^n)}{p(\text{speech}|s^n)} = \log \frac{p(s^n|\text{song})}{p(s^n|\text{speech})} + \log \frac{p(\text{song})}{p(\text{speech})} \tag{3}$$

Note that the log prior odds is just a constant independent of n, and by design we can set it to zero (a priori speech and song stimuli are equally likely). Thus, determining whether a sample is more speech-like or song-like becomes a matter of estimating the probability of the repeated stimulus s^n under generative models of speech and song. We hypothesize that differences in the strength of transformation for different sentences will follow from the statistics of those sentences under the different likelihood models. Equation (3) can also be

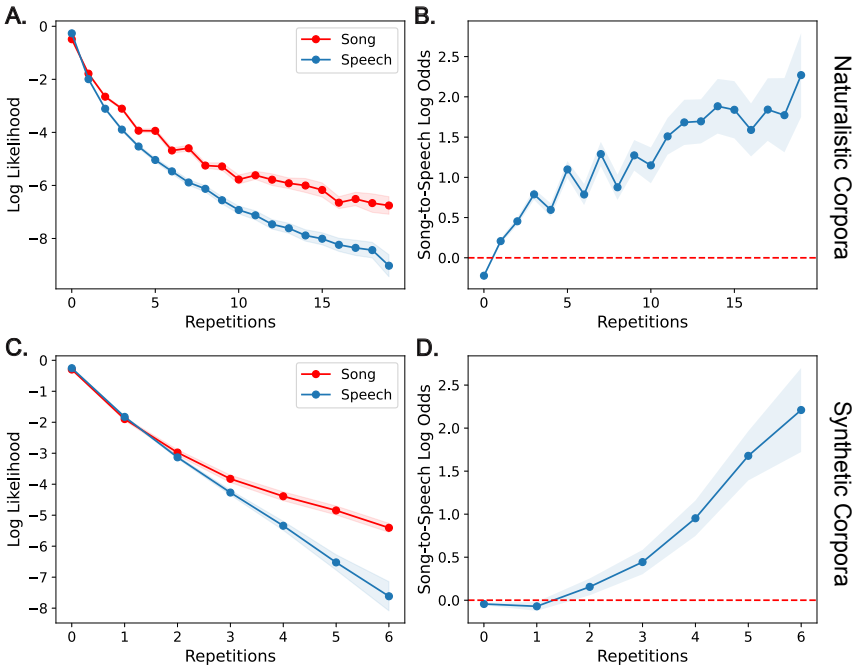

Figure 2: Aggregate repetition analysis. **A.** Log-likelihood curves for the naturalistic dataset as a function of the number of repetitions (aggregated across all words). **B.** Log-odds ratio for the naturalistic dataset (positive values favor song). **C.** and **D.** Similar to **A** and **B** for the synthetic (LLM-generated) dataset. Dashed red line indicates equal-likelihood threshold. Shaded area corresponds to 95% confidence intervals bootstrapped over documents.

used to predict the general effect of repetition by estimating $p(n|\text{speech})$ and $p(n|\text{song})$, i.e., the probability of observing an n-repetition in speech and song, respectively.

Finding generative audio models to estimate the probability of s^n is traditionally a hard problem (though recent advances in machine learning have changed this status; see Discussion). As an approximation, we can consider text corpora of speech and song transcripts. The idea here being that textual repetition in lyrics is indicative to some degree of repetition in melody (e.g., through refrains). By estimating the probability of sentences and their repetitions under such textual corpora we can evaluate whether their log-odds will exhibit the phenomenology of the illusion. Moreover, this treatment motivates the hypothesis that repetition should have a parallel effect in a purely textual variant of the illusion (Figure 1; Prose-to-Lyrics), as humans may also internalize similar statistical expectations in prose and lyrics. We will next present the modeling analysis and then proceed to the behavioral studies.

Modeling Methods

Text Preprocessing and Analysis

To construct repetition probabilities from text corpora we applied a bag-of-words approach. Specifically, we measured the repetition of words within each text document irrespective of their location within the document. This is a simplifying assumption that increases statistical power and is common to other textual analysis methods (e.g. topic modeling; Kherwa & Bansal, 2019). The processing steps were as follows. First, documents were tokenized, lemmatized, and cleaned from stop words and non-alphabetic entries using the `nltk` Python package (Bird, Klein, & Loper, 2009). Next, to create aggregate repetition probabilities per corpus, we counted the number of times each token was repeated within a given document, then created a histogram over the number of repetitions (i.e., the number of times a word was repeated n times within a document), normalized, and finally averaged over all documents. A similar procedure was applied for word-level repetition probabilities (including pronouns), by counting the number of documents in which a given word was repeated n times. Sentence-level repetition probabilities were constructed using a naïve Bayes approach (by multiplying the repetition probabilities of individual words). This is of course an approximation and can be generalized (see Discussion), but for the purpose of our theoretical analysis of short sentences constructed of common words it is sufficient. Specifically, the short sentences were constructed from the jointly most common words found in the naturalistic speech and song corpora (i.e., words that maximize the product of the number of documents from speech and song in which they appeared). These were useful as they allow for the construction of smooth word-level and sentence-level log-odds curves. Confidence intervals were constructed by bootstrapping over documents with replacement and 1,000 replicas.

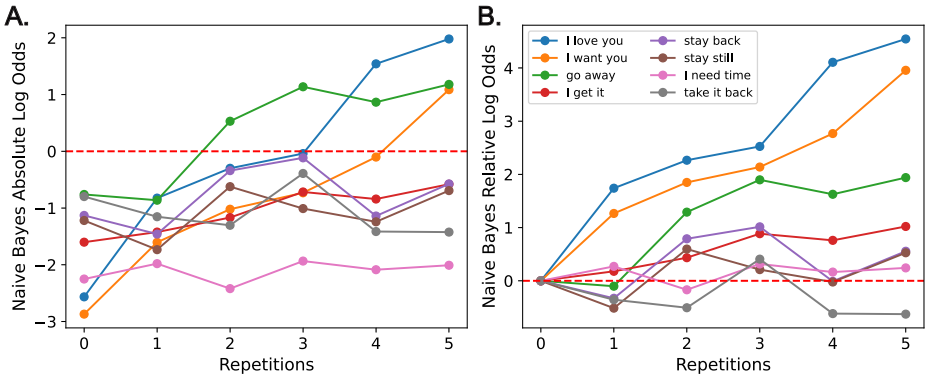

Figure 3: Log-odds profiles for common sentences constructed from the naturalistic speech and song corpora under a naïve Bayes approximation. **A.** Absolute log-odds profiles, dashed red line indicates equal-likelihood threshold, positive values favor song. **B.** Relative log-odds profiles, with the value at zero subtracted (indicated by the dashed red line).

Text Corpora for Speech and Song

For our prose and lyrics corpora we relied on two complementary approaches: one where we mine text from an ecologically valid ("naturalistic") source, but accept the fact that different documents will have varied lengths, and another where we synthesize text artificially using a large language model (GPT4) but have full control over the document length. For the naturalistic song corpus, we used a collection of 2,994 song lyrics extracted from songs on Shazam, a mobile app for song identification. The subset was extracted from a larger collection at random, whereby the lyrical language was detected using the FastText algorithm (Joulin et al., 2016). For the speech corpus, we used text documents from the Corpus of Contemporary American English (COCA; Davies, 2010), a large dataset constructed from a variety of sources such as magazines, blogs, and academic articles. Specifically, we focused on a set of 948 randomly sampled documents from the magazine subset of COCA. Note that while there were more song documents, these were naturally shorter, with a given document comprising 98 words on average as opposed to an average of 437 words for speech. To control for this variability, we used GPT4 to generate a synthetic set of 1,400 songs and 1,400 conversation transcripts, both constrained to approximately 20 lines per document, by repeatedly querying the model using the following prompt (with a temperature of 0.7 to encourage diversity in output): "Please write 10 [song lyrics|conversations] of length 20 lines each. Before each [song|conversation], write [Song|Conversation] and the number of the [song|conversation]". The average number of tokens was indeed more aligned in this case and amounted to 54 (95% CI: $[33, 76]$) for song and 49 (95% CI: $[32, 66]$) for speech. Data and code are provided at the following OSF link: `http://tinyurl.com/44dewc6p`.

Modeling Results

We begin with an empirical evaluation of the properties of the Bayes log-odds formula as applied to the textual speech and song corpora. Our goal here is to see if, despite its simplicity and pure reliance on text, the model is able to capture the qualitative phenomenology of the speech-to-song illusion. Figure 2 shows the log-likelihoods and log-odds for both speech and song as a function of general word repetition (naturalistic: Figure 2A-B, synthetic: Figure 2C-D). We see that for both cases, the log-likelihoods of repetition decrease with the number of repetitions. Crucially, while these initially overlap for song and speech, the tails of the former are much heavier, driving the log odds formula to favor song within two repetitions (three presentations). We also see that this difference in tails leads to a monotonic increase in the log-odds formula towards favoring song as a function of repetition. These results are consistent with the expectations for the speech-to-song phenomenology, namely, that repetition on average renders utterances more song-like (Deutsch et al., 2011), and increasingly so as a function of repetition (Tierney et al., 2013; Deutsch et al., 2011; Margulis & Simchy-Gross, 2016). Note that a richer model could capture ceiling effects by mapping the log-odds to an interval (e.g., via a sigmoid).

Next, we computed repetition log-odds for short sentences constructed from common words in the naturalistic corpus for which smooth profiles can be reliably estimated (see Modeling Methods). The resulting profiles are shown in Figure 3A. As is the case with Figure 2, we see a general trend of transformation from speech to song under repetition. Indeed, plotting the profiles relative to their value without repetition (i.e., collapsing the origin to zero; Figure 3B) reveals that seven out of eight sentences increased over their original value, and that the rate of change is different in different sentences, similar to findings from the literature (Tierney et al., 2013, 2018a, 2018b). For example, we see that intuitive sentences like "I love you" and "I want you" transform at a faster rate than more generic ones like "I get it" and "go away". Overall, this provides a principled explanation for the observed variation in phrase transformation rate based on natural statistics.

Having evaluated the qualitative predictions of the textual model, we are now ready to test whether these are realized quantitatively in the hypothesized prose-to-lyrics paradigm.

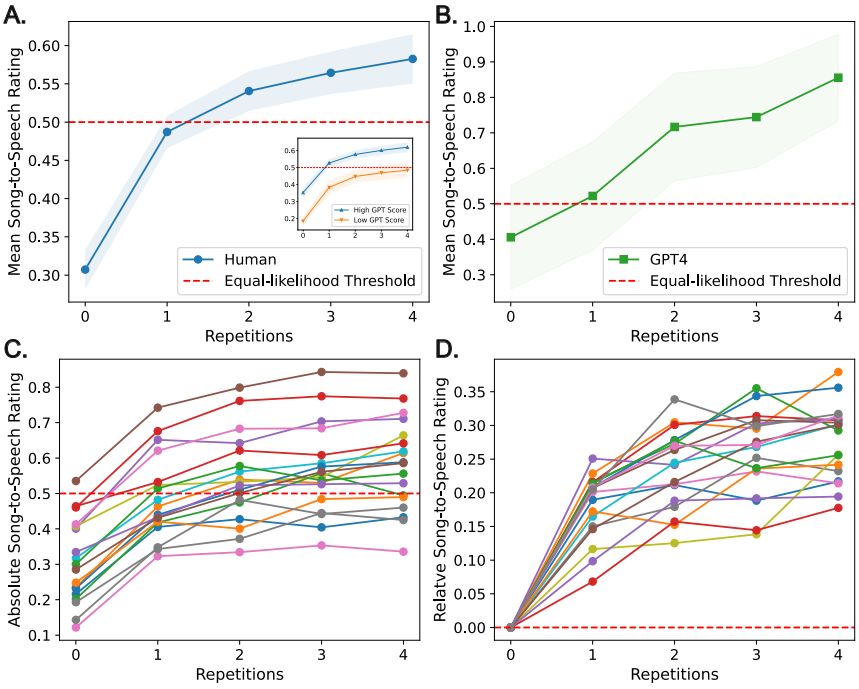

Figure 4: Behavioral results of the prose-to-lyrics experiments. **A.** Mean human ratings as a function of repetitions (rescaled to a 0-1 range via division by a factor of 6; dashed red line indicates speech-song equality threshold). Inset shows the data grouped based on a median split of the GPT4 scores at 4 repetitions. Shaded area indicates 95% confidence intervals (CIs) bootstrapped over participants. **B.** Mean rating as a function of repetition for the GPT4 control experiment. **C.** Mean sentence-level human ratings. **D.** Mean sentence-level human ratings relative to the value at 0 repetitions.

Behavioral Methods

Behavioral Paradigm

Humans For the prose-to-lyrics experiment, we recruited 120 UK participants from the online recruitment platform Prolific (`https://www.prolific.com/`). All participants provided informed consent prior to participation in accordance with the Max Planck Ethics Council (#2021_42). Upon providing informed consent, participants received the following instructions: 'In this study you will be presented with text transcripts of different audio recordings and your task is to decide based on those transcripts whether the original audio is more likely to have been a song or a speech excerpt. For each transcript you will have 7 response options ranging from 0 (definitely speech) to 6 (definitely song).' Participants then rated up to 20 randomly chosen stimuli and responded to the prompt: 'The following text transcript was extracted from an audio recording: ⟨transcript⟩. Based on this transcript alone, is the audio recording more likely to have been a speech or a song excerpt?' (Figure 1; Prose-to-Lyrics). Overall, there were 18 sentences (see Stimuli) each repeated 0 to 4 times (overall 90 items), and each rating item received an average of 63 ratings (range: 61 – 66). All items were shuffled and participants were randomly assigned items from this set.

GPT4 The LLM control experiment was constructed in a similar fashion. We used Azure's OpenAI API to query GPT4 (with temperature set to zero; version 0613) with the following prompt: "In what follows you will see a transcript of an audio recording and your task is to decide whether this transcript came from a song or a conversation. Please provide a numerical score between 0 and 1 where 0 is definitely conversation and 1 is definitely song. Do not provide any explanation or any other text. Transcript: ⟨transcript⟩. Answer:".

Stimuli

We used a set of 18 sentences based on the following sources. First, short sentences constructed from common words in the naturalistic corpora (see Modeling Methods). This set included, 'I love you', 'I get it', 'I want you', 'stay back', 'stay still', 'go away', 'take it back', and 'I need time'. Second, medium-length sentences that were randomly extracted from the speech corpus. These were 'how fast does a Zamboni go?', 'but once in a while you step on it from one end to the other', 'if it takes an extra minute to do it, that's fine', 'and wait 10 seconds or so between each bounce', 'I drilled a hole through the top of my rack', 'what kind of car would it be?', 'I grew up watching them', and 'it has really kept us growing'. Third, two additional sentences: 'sometimes behave so strangely' for its historical significance in the original paradigm, and an additional technical sentence 'I will head to the department store tomorrow'. Repeated versions were constructed by simply concatenating the sentences.

Behavioral Results

Figure 4 shows the behavioral results for the set of 18 sentences where each sentence was presented with 0-4 repetitions (see Behavioral Methods). Overall, the human average ratings over all sentences (Figure 4A) exhibited an excellent inter-rater reliability of $r = .96$ (95% CI: $[.94, .98]$) computed via a split-half method and bootstrapped over participants. Inspecting Figure 4A, we see that repetition steadily transformed the sentences from initially being significantly speech-like with mean rating of .31 (95% CI: $[.28, .33]$; with 0 being definitely speech and 1 being definitely song) to significantly being song-like with mean rating .58 (95% CI: $[.55, .61]$). A similar pattern was observed in the case of GPT4 ratings which performed an equivalent task (Figure 4B). We also found that the human and LLM data were significantly correlated with a Pearson correlation of $r = .64$ (95% CI: $[.60, .68]$). Moreover, splitting the human data into two groups based on a median split of the GPT4 scores at four repetitions (Figure 4A, inset) nicely separated the original curve into one that surpasses the 0.5 threshold and one that does not. These results show that we can replicate the effect of repetition in the speech-to-song illusion with our prose-to-lyrics illusion without directly playing an audio recording.

Next, inspecting the sentence-level human mean ratings (Figure 4C), we see that the aggregate trend of monotonic increase in song ratings as a function of repetitions is also reflected at the level of individual sentences. This becomes particularly clear when the value at zero repetitions is subtracted (Figure 4D). We see that not only song ratings increase for all sentences, but they also do so at different rates as found for the speech-to-song illusion (Tierney et al., 2013, 2018b, 2018a). To further highlight the analogy to the speech-to-song illusion findings (Tierney et al., 2013), we focus on a demonstrative subset of four sentences in Figure 5. We see that while the pair 'I love you' and 'sometimes behave so strangely' are not significantly separated at zero repetitions (.46, CI: $[.38, .54]$, and .46, CI: $[.40, .53]$, respectively), the ratings of the former increase at a significantly faster rate (at 4 repetitions: .77, CI: $[.71, .83]$, and .64, CI: $[.59, .70]$, respectively). Likewise, while the pair 'I drilled a hole through the top of my rack' and 'how fast does a Zamboni go?' were significantly speech-like at zero repetitions and not significantly separated (.22, CI: $[.16, .27]$, and .23, CI: $[.17, .29]$, respectively), only the latter transformed significantly to being song-like (.43, CI: $[.37, .50]$, and .59, CI: $[.52, .66]$, respectively). We confirmed this sentence-repetition interaction using a non-parametric test ($p < 0.001$ for both pairs). This provides clear evidence analogous to that found in the speech-to-song illusion (Tierney et al., 2013, 2018a, 2018b).

Discussion

By recasting the problem of deciding whether a stimulus is more like speech or song as rational statistical inference, we constructed a theoretical account of the speech-to-song illusion that not only provides principled explanations of its phenomenology based on the statistics of corpora, but also successfully predicts a new illusion that is purely textual.

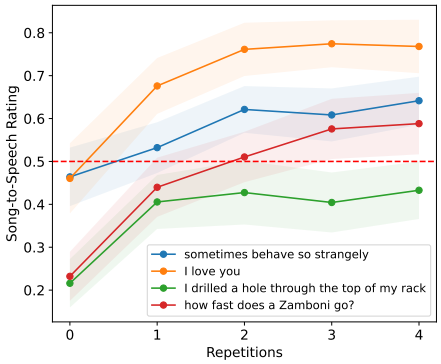

Figure 5: Prose-to-lyrics illusion for demonstrative sentences. Lines represent averages, shaded area represents 95% CIs.

Our framework opens up avenues for future investigation. First, a full treatment of the problem must use richer generative models of text and audio to estimate the probabilities of naturalistic stimuli and derive predictions (e.g. a naïve Bayes approach is likely to underestimate the required number of repetitions for the effect; Mullin, Norkey, Kodwani, Vitevitch, & Castro, 2021). For text, we could train or refine LLMs (e.g., Touvron et al., 2023) on lyrics and prose, and then estimate the probability of an entire (repeated) text conditioned on each subset. We could also use models that take into account phonology which can affect the illusion (Vitevitch, Ng, Hatley, & Castro, 2021). For audio, we could provide direct probability estimates of (repeated) utterances using modern audio transformer architectures such as the Music Transformer (Huang et al., 2018). This task can be made easier by referring back to the observation of Tierney et al. (2018a) that the speech-to-song effect can be replicated by providing only the f_0 contours of a sentence. This is significant because i) training a model only on f_0 contours requires significantly less data, and ii) it provides a strategy for analyzing non-speech sounds. Second, finding balanced transforming and non-transforming illusion pairs is a hard problem (Tierney et al., 2013). Our theoretical account could allow for the development of automated search algorithms for such pairs given suitable generative models. Third, being able to switch between generative models that are trained on different music and speech corpora could help develop principled hypotheses about the variation of this phenomenon across life span and culture (Margulis et al., 2015; Jaisin et al., 2016; Mullin et al., 2021). For example, would it be possible to find phrases that transform in one language but not in another? We hope to engage with these ideas in future work.

The speech-to-song illusion will continue to fascinate listeners for decades to come. Formalizing it as statistical inference whereby a perceived object provides evidence for an alternative, unconventional hypothesis serves as an exciting step towards a computational theory of perceptual illusions.

Acknowledgments

This work was supported by Microsoft Azure credits supplied to Princeton and by a Microsoft Foundation Models grant to TLG. The authors declare no competing interests.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047.

Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). Spectro-temporal acoustical markers differentiate speech from song across cultures. *bioRxiv*, 2023–01.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly.

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), e0198656.

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447–464.

Deutsch, D. (2003). *Phantom words and other curiosities*. Philomel Records.

Deutsch, D. (2019). *Musical illusions and phantom words: How music and speech unlock mysteries of the brain*. Oxford University Press.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, *129*(4), 2245–2252.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1491.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Herder, J. G. v. (2002). Treatise on the origin of language (1772). In M. N. Forster (Ed.), *Herder: Philosophical writings* (p. 65–164). Cambridge University Press.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., ... Eck, D. (2018). Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.

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

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, *13*(2), 150–158.

Kherwa, P., & Bansal, P. (2019). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, *7*(24).

Margulis, E. H., & Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music Perception: An Interdisciplinary Journal*, *33*(4), 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.

Mullin, H. A., Norkey, E. A., Kodwani, A., Vitevitch, M. S., & Castro, N. (2021). Does age affect perception of the speech-to-song illusion? *PloS one*, *16*(4), e0250042.

Ozaki, Y., Tierney, A., Pfordresher, P., Mcbride, J., Benetos, E., Proutskova, P., ... others (2022). Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [stage 2 registered report].

Rousseau, J.-J. (1782). *Collection complète des oeuvres de jj rousseau, citoyen de genève* (Vol. 3).

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 & Review*, *26*, 583–590.

Simchy-Gross, R., & Margulis, E. H. (2018). The sound-to-music illusion: Repetition can musicalize nonspeech sounds. *Music & Science*, *1*, 2059204317731992.

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.

Tierney, A., Patel, A. D., & Breen, M. (2018a). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, *147*(6), 888.

Tierney, A., Patel, A. D., & Breen, M. (2018b). Repetition enhances the musicality of speech and tone stimuli to similar degrees. *Music Perception: An Interdisciplinary Journal*, *35*(5), 573–578.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of experimental psychology: General*, *144*(2), e43.

Vitevitch, M. S., Ng, J. W., Hatley, E., & Castro, N. (2021). Phonological but not semantic influences on the speech-to-song illusion. *Quarterly Journal of Experimental Psychology*, *74*(4), 585–597.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.