I still hear a melody: investigating temporal dynamics of the Speech-to-Song Illusion

Groenveld, G.; Burgoyne, J.A.; Sadakata, M.

DOI
10.1007/s00426-018-1135-z

Publication date
2020

Document Version
Final published version

Published in
Psychological Research

Psychological Research (2020) 84:1451–1459
https://doi.org/10.1007/s00426-018-1135-z

ORIGINAL ARTICLE

I still hear a melody: investigating temporal dynamics of the Speech-to-Song Illusion

Gerben Groenveld[1] · John Ashley Burgoyne[1,2] · Makiko Sadakata[1,2,3]

Received: 1 June 2018 / Accepted: 10 December 2018 / Published online: 9 January 2019

Abstract

The Speech–to-Song Illusion (STS) refers to a dramatic shift in our perception of short speech fragments which, when repeatedly presented, may start to sound-like song. Anecdotally, once it is perceived as a song, it is difficult to unhear the melody of a speech fragment, and such temporal dynamics of the STS illusion has theoretical implications. The goal of the current study is to capture this temporal effect. In our experiment, speech fragments that initially did not elicit the STS illusion were manipulated to have increasingly stable F0 contours to strengthen the perceived 'song-likeness' of a fragment. Over the course of trials, the speech fragments with manipulated contours were repeatedly presented within blocks of decreasing, increasing, or random orders of F0 manipulations. Results showed that a presentation order where participants first heard the sentence with the maximum amount of F0 manipulations (decreasing condition) resulted in participants continuously giving higher overall song-like ratings than other presentation orders (increasing or random conditions). Our results thus capture the commonly reported phenomenon that it is hard to 'unhear' the illusion once a speech segment has been perceived as song.

Introduction

Judging whether someone is singing or speaking is usually an easy task. However, our judgement of vocalized sounds may "sometimes behave so strangely" (Deutsch, Henthorn, & Lapidis, 2011, p. 2246) that when a short fragment of speech is repeated, one's perception can transform from speech to song. This dramatic switch in perception, named the Speech-to-Song Illusion (STS, Deutsch et al., 2011), is a fascinating topic that offers a key to understand how our mental representation of language and music interact. Interestingly, listeners cannot easily escape from this illusion: once perception of speech fragment has transformed to a song, the melody tends to persist in one's mind and the original segment cannot be heard as pure speech anymore. While this is a common observation, and such temporal dynamics of the STS illusion could help us with considering the mechanism underlying the effect, it has not been empirically confirmed. In this study, we try to capture this temporal effect.

Up until now the academic discussion regarding the processing of speech and non-speech sounds has been divided into two contrasting theories: the domain-specific and the cue-specific models (Zatorre & Gandour, 2008). The former model states that speech and non-speech sounds are processed in separate mechanisms, and that each mechanism is dedicated exclusively to specific type of sounds. This is based on earlier findings pointing to neural independence between language and non-language domains. For example, a number of cases report that brain damage or dementia may affect speech but spare musical abilities or vice versa (Beatty, Zavadil, Bailly, & Rixen, 1988; Peretz & Morais, 1993). Some neuroimaging studies also demonstrated that speech sounds elicit different activation patterns of the brain areas than musical sounds (e.g., Binder, 2000; Tervaniemi et al., 1999, 2000; Zatorre, Meyer, Gjedde, & Evans, 1996). These agree with the idea of specialized mechanisms in the brain that are dedicated exclusively to speech processing and to music processing (e.g., Peretz & Coltheart, 2003).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00426-018-1135-z) contains supplementary material, which is available to authorized users.

✉ Makiko Sadakata
m.sadakata@uva.nl

[1] Musicology Department, University of Amsterdam, Amsterdam, The Netherlands

[2] Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

[3] Artificial Intelligence Department, Radboud University, Nijmegen, The Netherlands

The cue-specific model, in contrast, states that there can be an overlap in the processing of speech and non-speech sounds, especially that of common aspects of acoustic features. For example, the F0 contour of a sound can predict the grouping of auditory information in both domains (Patel, 2008). More recent studies indeed show that non-speech sounds, such as certain chord sequences (Koelsch et al., 2002), noise stimuli (Schonwiesner, Rübsamen, & Von Cramon, 2005) and tone sequences (Zatorre & Belin, 2001) induce activities in areas previously assumed to be solely responsible for speech processing. Furthermore, a number of studies reporting transfer of learned skills between domains support this line of approach (Patel et al., 2014; Roncaglia-Denissen, Roor, Chen, & Sadakata, 2016; Sadakata & Sekiyama, 2011). Deutsch et al. (2011) argues that mostly common neural pathways process speech and song but that neural networks dedicated to specific domains will produce the final percept. Most interestingly, the STS illusion provides a way to look into what conditions and acoustic information trigger the domain-specific neural network.

Interestingly, not all speech fragments transform to a song. In fact, the occurrence of STS is rather scarce among randomly selected speech fragments (Cornelissen, Sadakata, & Honing, 2016). Previous studies demonstrated the importance of stimulus acoustical cues to the illusion, such as rhythm and pitch patterns (Falk, Rathcke, & Dalla Bella, 2014). Especially how pitch patterns are formed in a speech fragment seems to provide the most reliable predictive factor of the illusion: speech fragments with longer stable portion of F0 contours tend to induce this illusion more often (Falk et al., 2014; Tierney, Dick, Deutsch & Sereno, 2013), possibly because stable tonal targets enhance the perception of discrete pitches, such as those typically found in music (Zatorre & Baum, 2012). A more recent study further demonstrated that the illusion is stronger when speech fragments include a F0 contour with pitch intervals closer to those of a Western diatonic scale, suggesting the importance of melodic structure (Tierney, Patel, & Breen, 2018). Next to these acoustical features, previous studies revealed the effect of listeners' experience: For experiencing the illusion, no special musical training is needed (Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015a). The illusion is more likely to occur with stimuli of unfamiliar languages for listeners (Margulis, Simchy-Gross, & Black, 2015) and it has been confirmed with listeners with various linguistic background, although a weaker effect was observed with native speakers of tonal language as compared to that of non-tonal languages (Jaisin et al., 2016).

It seems that the STS illusion embodies a dramatic shift in our perception of a phrase through which a musical reinterpretation of the phrase may settle in the mind for a quite long time. Patel (2008; 2011) argues that musical melody triggers a 'richer' response than linguistic prosody because it employs a larger set of perceptual meta-relations. The registration of aspects such as grouping structure, motivic similarity, pitch hierarchies and implied harmony, combined with possible relations between these aspects, is what makes our perception of musical melody to be psychologically distinct from perception of speech prosody. In line with this argument, the brain imaging results demonstrated greater activations in the area associated with pitch processing and auditory–motor integration while experiencing the transformation (Tierney et al., 2013). Such enhanced experience and activations may be contributing to the common observation regarding the robustness of melodic representation, contributing to the unidirectional temporal effect of the STS illusion.

The temporal effect predicts that one would perceive the stimuli more like a song after the STS illusion had occurred. The main goal of the current study is to capture this effect. As mentioned earlier, several papers have demonstrated that stable F0 contours resembling a Western tonal melody are the strongest cues for the occurrence of the STS Illusion (Falk et al., 2014; Tierney et al., 2018). We selected speech fragments that did not elicit the STS illusion on their own from our previous study (Cornelissen et al., 2016). These were then manipulated to have increasingly stable musical F0 contours to strengthen the 'song-likeness' of a fragment. All stimuli (of all different manipulation versions) were evaluated twice: one after a single presentation of the stimulus and another followed by a repeated presentation of the stimulus. The repeating presentation immediately followed the single presentation condition. This allowed us to study the original effect of repeating the same speech stimuli on the STS rating. Furthermore, the same speech fragments with manipulated contours were presented, consecutively, in decreasing, increasing, or random orders of F0 manipulations. In the decreasing order condition, participants first heard the stimuli with greatest F0 manipulations (single and repeat conditions), followed by stimuli with less F0 manipulations, while in the increasing order condition, they first heard the non-manipulated stimuli, followed by more and more manipulated stimuli. We expected that participants would continuously rate the stimuli as more 'song-like' in the decreased order condition where segments were first heard in the most melodic form than in the increased or random order conditions, where segments were first heard in their non-pitch manipulated versions or in a random order. With this setup, we tried to capture the effect that it is hard to unhear the melody of a speech once heard as song while controlling for a possible exposure effect.

Methods

Participants

Forty-two participants (18 male, 24 female) aged between 17 and 49 (mean: 25 years) with normal hearing participated in this study. The majority of participants were undergraduate and graduate students at the University of Amsterdam. This study was approved by the ethics committees of the Faculty of Humanities of the University of Amsterdam. Informed consent was obtained at the start of the experiment and music proficiency was determined after the experiment through a Gold-MSI, version 1.0 (Müllensiefen, Gingras, Musil, & Stewart, 2014). We included the general Gold-MSI score for the analysis. Two participants were excluded from final data analysis due to having misinterpreted instructions or dealing with medical issues.

Materials

Stimuli were derived from a study in which 259 short speech samples were tested for likeliness of inducing the STS illusion (Cornelissen et al., 2016). In this previous study, 137 participants gave perceived song-like ratings on a 11-scale slider during 7 repetitions of each segment. Based on the rating after the final repetition, 15 samples with the lowest song-like response (sentences mostly perceived as speech) were selected for the current experiment.

The 15 chosen samples consisted of sentences in German, English, Spanish, French and Dutch. Each sentence contained 4–8 words and had an average duration of 1843 ms (sd: 298 ms). As pointed out by previous studies (Falk et al., 2014; Tierney et al., 2018), stable tonal targets (flat pitch contours within syllables) and pitch contour resemblance to that of a Western tonal melody facilitate the occurrence of the STS illusion. We modified pitch contour enhancing such stimuli characteristics using Praat (Boersma, 2001).

First, based on syllable boundaries that were manually assigned, the average pitch of each syllable was calculated. The closest musical pitch per syllable was then calculated using a reference list featuring all musical pitches between 16.35 Hz (C0) and 7902.13 Hz (B#8) which corresponds to the general frequency limits of the human speaking voice. The selected musical frequencies were in turn compared to a list featuring all possible diatonic scales within an octave and each of the corresponding musical pitches, again ranging from any musical pitch between 16.35 and 7902.13 Hz after which the musical scale that would require the least amount of permutations to match the selection of frequencies in the sentence was selected. For an example of such calculations, see Table 1.

Table 1 Example of F0 stylization calculations for sentence "Son grand-père a commandé des escadres"

Syllable	Frequency (Hz)	Closest musical frequency (Hz) + tone	Closest diatonic scale frequency (Hz, B♭major)
Son	112.91	110–A2	110–A2
grand	115.67	116.54–B♭2	116.54–B♭2
père	134.80	138.59–D♭3	**130.81–C3**
a	111.57	110–A2	110–A2
com	111.51	110–A2	110–A2
man	107.20	110–A2	110–A2
dé	109.72	110–A2	110–A2
des	98.27	98–G2	98–G2
es	92.27	92.5–G♭2	**87.31–F2**
ca	113.47	116.54–B♭2	116.54–B♭2
dres	147.33	146.83–D3	146.83–D3

Frequency adjustments made to fit closest diatonic scale are printed bold

Having calculated the closest musical pitch for each syllable, F0 contours within each syllable were morphed towards the calculated nearest musical pitch in steps of 30%. A smoothing function was added to smoothen the transitions between different pitch values (e.g., going up or down evenly instead of changing abruptly). Based on these manipulations, the samples to be presented in the experiment contained 0% (base, non-transforming), 30%, 60% and 90% of F0 manipulations. We opted for 90% instead of 100% to avoid too mechanical rendition of fragments. A visualization of 0% and 100% contour manipulations can be seen in Fig. 1. A set of example stimuli can be found in Supplementary materials 1–4, including each step of contour manipulations (STS_0%, STS_30%, STS_60%, STS_90%).

Task and procedure

Participants were asked to rate 15 speech samples 2 times in four manipulation versions (0%, 30%, 60%, 90%), each after hearing them once and after 7 repetitions (170 ms ISI). Altogether, these eight trials constituted one block. Within a block, stimuli could be heard in increasing order of manipulations (0%, 30%, 60%, and 90%), decreasing order (90%, 60%, 30%, and 0%) or random order, further referred to as increasing, decreasing and random conditions, respectively. The rating of repeated stimuli presentation followed the single stimuli presentation at every stage before the program moved on to the next manipulation step. For example, the presentation order of increasing condition was "0% single, 0% repeat, 30% single, 30% repeat, 60% single, 60% repeat, 90% single, 90% repeat", while that of decreasing condition was "90% single, 90% repeat, 60% single, 60% repeat, 30% single, 30% repeat, 0% single, 0% repeat". Here, the random

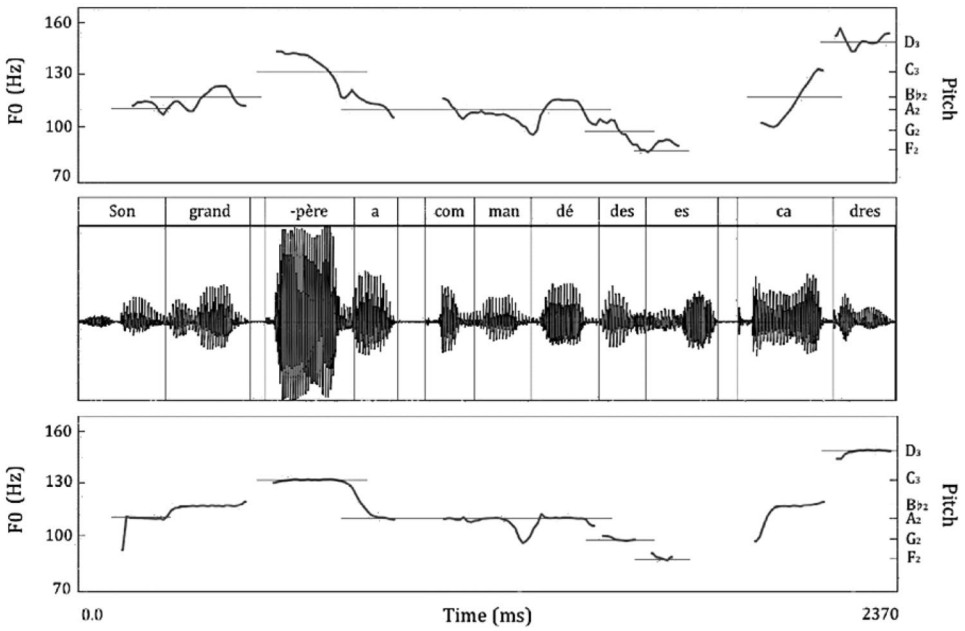

Fig. 1 Contour visualizations of a sentence before (top) and after (bottom) F0 stylizations (100%). Light gray lines show the frequencies corresponding to the scale of B-flat major

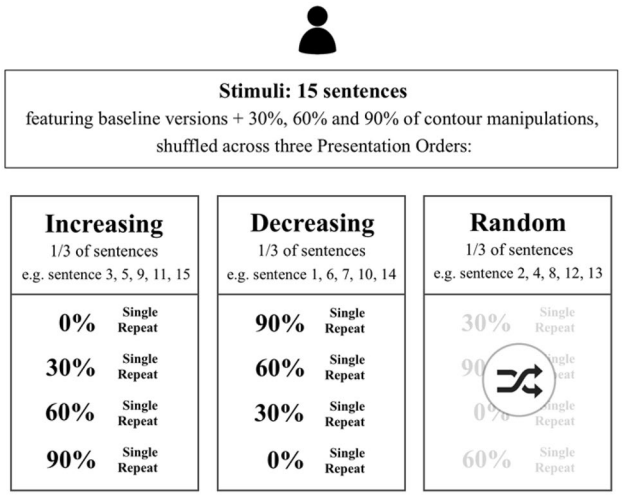

Fig. 2 An example of stimuli randomizations for presentation order conditions: increasing, decreasing and random. One block consisted of eight trials that are based on different presentation versions of the sample speech (single and repeated presentations for four manipulation versions)

condition provides an estimate of the exposure baseline (i.e., participants heard the stimuli of 0%, 30%, 60% and 90% manipulation at random order). The 15 speech samples were randomly allocated to one of the increasing, decreasing and random conditions and the presentation order of blocks was randomized. Figure 2 presents an example of how the 15 stimuli were presented.

Unlike many previous studies where a Likert scale was used for rating the STS illusion, only two rating options were provided in this study, namely "speech" or "song". We opted for this choice because each step in a Likert scale can be interpreted differently by participants. Ratings submitted for our analyses were average value of five trials (averaged across five sentences) per participant and we computed the statistics based on relatively large number of participants ($N = 40$). While this approach makes it difficult to directly compare values with other studies, we think that the interpretation is somewhat simpler, because higher values indicate simply more responses of song-like perception.

Participants were asked to use the left and right arrows on the keyboard to rate samples as 'Speech' or 'Song', hereafter we call this the STS rating. What each arrow referred to was randomized across participants to account for an index bias. Between every set of stimuli (all baseline + contour versions of a sentence), the participant could take a break and continue by pressing the right arrow button. The full test took about 25 min.

Results

To understand the effects of the three major within-subject variables—contour manipulation, presentation order, and stimulus repetition—we fit a sequence of nested mixed probit models on the STS ratings using the R package lme4 (version 1.1). We used polynomial contrasts to model linear, quadratic, and cubic effects of contour manipulation. We used treatment contrasts to check the effects of presentation order, setting random presentation order as the baseline. For checking the effect of stimulus repetition, we used a simple difference contrast: the mean for the repeated-presentation condition minus the

mean of the single-presentation condition. We also investigated possible between-subject effects of general music sophistication.

Table 2 presents the results. Model 1 contains only a random intercept per participant. Model 2 adds fixed effects for contour manipulation, presentation order, and stimulus repetition, including all interactions. Model 3 adds random slopes for the main effects. Model 4 adds linear and quadratic terms for general music sophistication and Model 5 adds interactions of these terms with stimulus repetition. In our initial analyses, Models 3–5 included random slopes not only for the linear contour manipulation contrast but also for the quadratic and cubic contrasts; because the variance of the quadratic and cubic contrasts was extremely small, we removed them to improve the stability of the model fit. Likewise, Model 5 originally included all second- and third-order interactions involving general music sophistication, but as only the interactions with stimulus repetition were significant, we dropped the others. Each of these simplifications improved AIC_c.

Overall, Model 3 (random slopes without general music sophistication) has the best AIC_c, although with ΔAIC_c of less than 4, Models 4 and 5 also have some empirical support. Figure 3 illustrates the model predictions for a median participant, based on Model 3, the best-fitting model. The strong linear effects of contour manipulation are clear, as well as the effects of stimulus repetition and the differences between the increasing and decreasing manipulation orders. In Models 2–5, the three-way contour manipulation × manipulation order × repeated presentation interaction is significant, $\chi^2(6) = 17.53$, $p = 0.008$, $\eta^2 = 0.004$; in Model 5, the two-way musical sophistication × repeated presentation interaction is also significant, $\chi^2(1) = 4.97$, $p = 0.026$, $\eta^2 < 0.001$. Looking at the contrasts in more detail, a common story emerges across all models. Participants were significantly more likely to perceive stimuli as song as contour manipulation increases (small effect), with no evidence for quadratic or cubic attenuation. Most interestingly, participants were significantly more likely to perceive stimuli as song when they heard contour manipulations in decreasing order (small effect). They also perceived the stimuli more as song after hearing them repeatedly (medium effect). Independent of these effects, there were very small but significant interactions, suggesting that when participants heard the contour manipulations in increasing order, the effect of contour manipulation itself is stronger (although slightly less so after stimuli are repeated), and when participants hear the contour manipulations in decreasing order, the effects of contour manipulation and stimulus repetition are weaker. This effect corresponds with different steepness of response patterns in Fig. 3. Overall, there is some evidence that musical sophistication reduces the effect of stimulus repetition (small effect).

Discussion

The current study aimed to capture the temporal dynamics of the STS illusion, namely, that it is difficult to unhear the melody once a speech fragment is perceived as song. By presenting speech stimuli featuring different amounts of contour manipulations in different presentation orders, we found that exposure to a repeated presentation of a melodic rendition of a speech segment (featuring syllables with a more stable pitch contour close to musical frequencies resembling a Western diatonic scale) significantly enhanced the perceived song-like rating of the segment. Stimuli were rated more as 'song' when heard in a decreasing order of contour manipulations (from song-like to baseline versions) than when heard in a random or increasing order of contour manipulations (from baseline to song-like versions). Our observation that the effect of contour manipulation itself was stronger for increasing than decreasing order among our participants also supports our hypothesis that having already heard a more melody-like version of a stimulus greatly facilitates the STS illusion. This is in line with the known robustness of melodic representation in the STS illusion, namely, that one cannot unhear a melody once perception of a speech fragment has transformed to music. The facilitation effects were observed even for the single presentation condition, supporting the common observation that the repetition is no longer needed to hear melody once one has heard melody (Deutsch et al., 2011).

The perceived song-like rating was always significantly greater for repeated conditions than for corresponding single listening conditions, confirming the general finding that repetition facilitates the STS illusion (Deutsch et al., 2011). Furthermore, exposure to stimuli featuring stable tonal targets (flat pitch contours within syllables) and a pitch contour resembling the contour of a Western tonal melody did increase the perceived song-like rating after repetitions, again confirming previous findings (Falk, et al., 2014; Tierney et al., 2018). As the contour manipulations were applied to stimuli in steps of 30%, the perceived song-like rating increased or decreased accordingly, depending on the presentation order. This effect was apparent across virtually all conditions of presentation mode and presentation order. The perceived song-like ratings of 90% stimuli for repeated presentation were all quite high and did not differ among three conditions, indicating a strong effect of pitch manipulation. In other words, the exposure to the stimulus did not play a big role here.

The right most panel in Fig. 3 shows that the STS ratings of single presentation at 90% manipulation in the decreasing order appears to be an outlier (lower rate of song perception as compared to the other order

Table 2 Fully *Y**-standardized probit estimates of fixed effects (top) and variance/correlation of random effects (bottom) for models of the Speech-to-Song Illusion

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
Fixed effects					
Intercept [random manip. order]	−0.45*** (0.07)	−0.54*** (0.07)	−0.58** (0.10)	−0.53* (0.23)	−0.64** (0.22)
Musical sophistication				0.09 (0.25)	−0.07 (0.24)
Musical sophistication^2				0.06 (0.27)	−0.08 (0.28)
Repeated presentation		0.32*** (0.03)	0.31*** (0.04)	0.31*** (0.04)	0.12 (0.10)
× Musical sophistication					−0.26* (0.12)
× Musical sophistication^2					−0.23 (0.13)
Contour manipulation		0.27*** (0.03)	0.28*** (0.04)	0.27*** (0.04)	0.27*** (0.04)
× Repeated presentation		0.04 (0.03)	0.05 (0.03)	0.05 (0.03)	0.05 (0.03)
Contour manipulation^2		−0.01 (0.03)	−0.02 (0.03)	−0.02 (0.03)	−0.02 (0.03)
× Repeated presentation		0.05 (0.03)	0.05 (0.03)	0.05 (0.03)	0.05 (0.03)
Contour manipulation^3		0.03 (0.03)	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)
× Repeated presentation		−0.01 (0.03)	0.00 (0.03)	0.00 (0.03)	0.00 (0.03)
Increasing manipulation order		−0.03 (0.02)	−0.03 (0.04)	−0.03 (0.04)	−0.03 (0.04)
× Repeated presentation		0.02 (0.02)	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
× Contour manipulation		0.12*** (0.02)	0.13*** (0.02)	0.13*** (0.02)	0.13*** (0.02)
× Repeated presentation		−0.04 (0.02)	−0.05* (0.02)	−0.05* (0.02)	−0.05* (0.02)
× Contour manipulation^2		−0.01 (0.02)	−0.02 (0.02)	−0.02 (0.02)	−0.02 (0.02)
× Repeated presentation		0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
× Contour Manipulation^3		0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)
× Repeated presentation		0.02 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
Decreasing manipulation order		0.15*** (0.02)	0.18*** (0.04)	0.18*** (0.04)	0.18*** (0.04)
× Repeated presentation		−0.04* (0.02)	−0.05* (0.02)	−0.05* (0.02)	−0.05** (0.02)
× Contour manipulation		−0.07*** (0.02)	−0.07*** (0.02)	−0.07*** (0.02)	−0.07*** (0.02)
× Repeated Presentation		0.05* (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.02)
× Contour manipulation^2		−0.04 (0.02)	−0.03 (0.02)	−0.03 (0.02)	−0.03 (0.02)
× Repeated presentation		0.02 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
× Contour manipulation^3		−0.02 (0.02)	−0.02 (0.02)	−0.02 (0.02)	−0.02 (0.02)
× Repeated presentation		0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
Random effects					
$\sigma^2_{\text{Intercept}}$	0.18	0.18	0.37	0.37	0.36
$\sigma^2_{\text{Repeated presentation}}$			0.02	0.02	0.02
$\sigma^2_{\text{Contour manipulation}}$			0.08	0.08	0.08
$\sigma^2_{\text{Increasing order}}$			0.04	0.04	0.04
$\sigma^2_{\text{Decreasing order}}$			0.05	0.05	0.05
$\rho_{\text{Intercept} \times \text{repeated presentation}}$			0.31	0.33	0.28
$\rho_{\text{Intercept} \times \text{contour manipulation}}$			−0.05	−0.04	−0.04
$\rho_{\text{Intercept} \times \text{increasing order}}$			−0.41	−0.42	−0.40
$\rho_{\text{Intercept} \times \text{decreasing order}}$			−0.82	−0.82	−0.82
$\rho_{\text{Repeated presentation} \times \text{contour manipulation}}$			−0.45	−0.45	−0.53
$\rho_{\text{Repeated presentation} \times \text{increasing order}}$			0.03	0.03	0.14
$\rho_{\text{Repeated presentation} \times \text{decreasing order}}$			−0.06	−0.06	0.00
$\rho_{\text{Contour manipulation} \times \text{increasing order}}$			−0.38	−0.37	−0.37
$\rho_{\text{Contour manipulation} \times \text{decreasing order}}$			−0.28	−0.28	−0.28
$\rho_{\text{Increasing order} \times \text{decreasing order}}$			0.46	0.47	0.46
Fit statistics					
R^2	0.17	0.43	0.54	0.54	0.54
ΔAIC_c	1029.41	222.71	0.00	3.92	3.14

Table 2 (continued)

Standard errors are in parentheses

*$p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

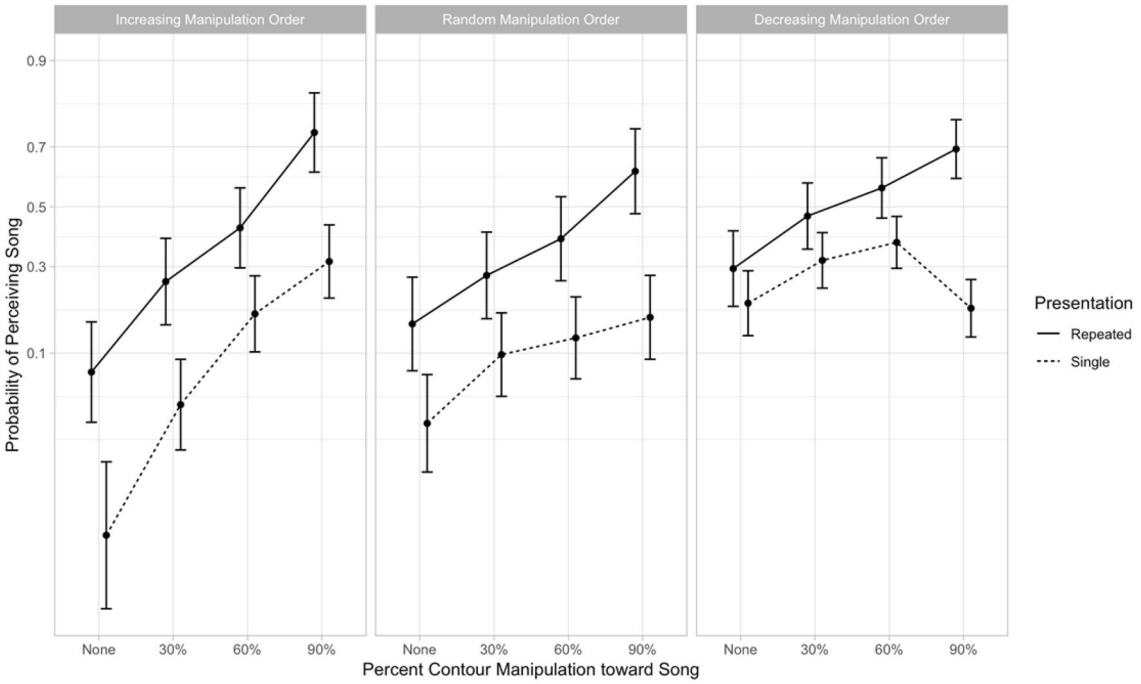

Fig. 3 Predicted probability of perceiving song for the median participant under Model 3. Error bars = 95% confidence intervals. There are strong effects of contour manipulation and presentation, which vary under different ordering conditions. The apparent outlier in the single presentation condition at 90% manipulation with decreasing manipulation order is most likely a side effect the experimental design

conditions). However, recall from the experiential design that this particular condition was always the first one that participants had heard the sentence in the decreasing order conditions. This may be supporting the idea that, despite of extreme pitch contour manipulation, participants' default listening strategy may be 'speech perception mode' (e.g., Margulis, 2014; Tierney et al., 2018). For similar reasons, the single presentation with 0% manipulation is slightly below (linear) trend in the increasing condition. This aspect of the experimental design is likely responsible for the presence of the three-way interaction.

Although it was not our primary research question, the observation that musical sophistication somewhat reduces the effect of stimulus repetition is an interesting finding. The original STS study demonstrated the illusion with musically trained individuals (Deutsch et al., 2011). The follow-up studies confirmed that no musical training is necessary but implicit musical knowledge is sufficient to experience the STS illusion (e.g., Falk et al., 2014; Vanden Bosch der Nederlanden et al., 2015a). The crucial difference between our study and the others lies in the way we model the effect of musical background. By analysing the effect using continuous scale of musical sophistication rather than splitting participants into groups, our study may have been able to pick up this small, somewhat surprising trend. Such individual differences in perception of the STS illusion should be explored further.

Of course, an intriguing question is why the STS takes place, and once occurred, why musical interpretations tend to stick to our mind. Patel (2008; 2011) argues that musical melody captures our interest or gets caught in our brain more than linguistic melody because it employs a larger set of meta-relations (e.g., pitch hierarchies, intervallic implications, implied harmony and event hierarchies). This is in line with evidence from an fMRI study demonstrating that the STS illusion triggers stronger responses than speech processing in networks associated with the musical pitch processing and auditory–motor integration, indicating that through exact repetition of a speech segment our attention may indeed shift to such perceptual meta-relations present in the stimulus (Tierney et al., 2013). The reason why musical melody sticks to our mind; thus may be the result of such a musical re-evaluation process engendering a richer set of perceptual relations.

Such temporal dynamics of the STS effect also has theoretical implications. Deutsch et al. (2011) argues that, in the STS illusion, repetition seems to boost an awareness of musical qualities in a speech segment. One way to elaborate this explanation is a satiation hypothesis: repetition may satiate the speech perception resources as found in another linguistic perceptual transformation experience, allowing listeners to shifting attention to other features in speech (Castro, Mendoza, Tampke, & Vitevitch, 2018; Margulis, 2013). One example is Verbal Transformation Effect (VTE), the phenomenon that our percept of a word transforms from one to others while the same word is continuously repeated (Warren & Gregory, 1958). However, as described by Tierney et al. (2018), not all features of the STS fit to this account. The temporal effect of the STS is one of them that distinguishes the STS from the VTE, while the STS illusion yields unidirectional transformation from speech to song, the transformation in VTE is unstable.

A drastic switch of interpretation when perceiving an event is, in fact, not uncommon. For example, the sine wave speech is another example that knowing original texts completely alters our listening mode of otherwise incomprehensible collection of sine wave composites (Remez, Rubin, Pisoni, & Carell, 1981). A similar example in visual perception would be the Gregory's camouflaged Dalmatian image, in which seeing a full-color image of a dog completely changes how our brain interprets otherwise seemingly random black–white image (Gregory, 1973). Together, these can be considered as a sign of our brain applying a top-down model (e.g., speech or a Dalmatian dog) to make sense of bottom-up sensory inputs. According to Reverse Hierarchy Theory (Hochstein & Ahissar, 2002; Ahissar & Hochstein, 2004), high-level representation tries to govern the interpretation at the initial stage of conscious perception. Once this is established, perceivers may attend to lower level representations as needed. Consequently, representation of top-down model may alter the way a perceiver interacts with lower level information. More specifically, it can inhibit or bias perception of bottom-up information. While initially proposed to account for visual perception, this theory has also been applied in auditory domain, and this is what exactly has found in the context of STS: Vanden Bosch der Nederlanden, Hannon and Snyder (2015b) demonstrated that song representation inhibits the sensitivity to pitch change.

Another promising account to consider is the predictive coding (PC, e.g., Clark et al., 2013). The PC formalizes how top-down predictions are continuously evaluated in comparison with inputs from lower level information layers to minimize the prediction error. We think the switch from one mode (speech) to another mode (song) can be interpreted in the light of this framework; because stable pitch contours and the exact repetition each strongly support the identity of song rather than that of speech, our system may consider the "song model" as more plausible than the "speech model" for some speech fragments. Robust song perception after experiencing the STS illusion could thus be showing that, once established, top-down models serve as a robust framework to guide input sensory information processing while continuously updating the probability. Although there are more things to consider, for example, how many levels of predictions contribute to this process, whether the simplistic dichotic competition between speech versus song models is realistic (perhaps not considering that perception of song does not exclude the perception of linguistic qualities), these theoretical frameworks are excellent tools for navigating our future study.

The Speech-to-Song Illusion shows an interesting case where the perception of vocalized sounds crosses the boundary between the domains of language and music. The illusion provides a valuable contribution in the recent discussion concerning the substrates of music and language processing: the domains of language and music are not dichotomies but rather forms of communication perception that may occasionally recruit shared or domain-specific circuitry (Zatorre & Grandour, 2008). As Deutsch et al. (2011) argue, the domains of speech and music may for a large part be processed by common neural pathways while certain song- or speech-specific circuitry is ultimately invoked to produce the final percept. Our results clearly suggest that the final percept is not only determined by processing of bottom-up acoustic features such as repetition and pitch contour characteristics, but also by top-down notion such as melodic expectation.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.08.011.

Beatty, W. W., Zavadil, K. D., Bailly, R. C., & Rixen, G. J. (1988). Preserved musical skill in a severely demented patient. *International Journal of Clinical Neuropsychology, 10*(4), 158–164.

Binder, J. R. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–347. https://doi.org/10.1097/AUD.0b013e31821473f7.

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*. https://doi.org/10.1371/journal.pone.0198656.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Cornelissen, B., Sadakata, M., & Honing, H. (2016). A classification approach to the speech to song transformation. In *International conference on music perception and cognition*, p. 386. San Francisco.

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. https://doi.org/10.1121/1.3562174.

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

Gregory, R. L. (1973). Eye and brain: The psychology of seeing. *British Medical Journal, 4*(5893), 682. https://doi.org/10.1136/bmj.4.5893.682-b.

Hochstein, S., & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791–804.

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*. https://doi.org/10.3389/fpsyg.2016.00662.

Koelsch, S., Gunter, T. C., Cramon, D.Y., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach speaks: A cortical "language-network" serves the processing of music. *NeuroImage, 17*(2), 956–966. https://doi.org/10.1016/S1053-8119(02)91154-7.

Margulis, E. H. (2013). Repetition and emotive communication in music versus speech. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00167.

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

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

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. (J. Snyder, Ed.). *PLoS One*. https://doi.org/10.1371/journal.pone.0089642.

Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195123753.001.0001.

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2011.00142.

Patel, A. D., Conard, N., Malina, M., Münzel, S., Brown, S., Jordania, J., Schmidt, M. (2014). The evolutionary biology of musical rhythm: Was Darwin wrong? *PLoS Biology, 12*(3), e1001821. https://doi.org/10.1371/journal.pbio.1001821.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6*(7), 688–691. https://doi.org/10.1038/nn1083.

Peretz, I., & Morais, J. (1993). Specificity for music. In F. Boller & J. Grahman (Eds.), *Handbook of neuropsychology* (vol 8, pp. 373–390). Amsterdam: Elsevier S.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carell, T. D. (1981). Speech perception without traditional speech cues. *Science*. https://doi.org/10.1126/science.7233191.

Roncaglia-Denissen, M. P., Roor, D. A., Chen, A., & Sadakata, M. (2016). The enhanced musical rhythmic perception in second language learners. *Frontiers in Human Neuroscience, 10*, 288. https://doi.org/10.3389/fnhum.2016.00288.

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica, 138*(1), 1–10ß. https://doi.org/10.1016/j.actpsy.2011.03.007.

Schonwiesner, M., Rübsamen, R., & Von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *European Journal of Neuroscience, 22*(6), 1521–1528. https://doi.org/10.1111/j.1460-9568.2005.04315.x.

Tervaniemi, M., Kujala, A., Alho, K., Virtanen, J., Ilmoniemi, R. J., & Näätänen, R. (1999). Functional specialization of the human auditory cortex in processing phonetic and musical sounds: A magnetoencephalographic (MEG) study. *NeuroImage, 9*(3), 330–336. https://doi.org/10.1006/nimg.1999.0405.

Tervaniemi, M., Medvedev, S. V., Alho, K., Pakhomov, S. V., Roudas, M. S., Van Zuijen, T. L., & Näätänen, R. (2000). Lateralized automatic auditory processing of phonetic versus musical information: A PET study. *Human Brain Mapping, 10*(2), 74–79. https://doi.org/10.1002/(SICI)1097-0193(200006)10:2%3C74::AID-HBM30%3E3.0.CO;2-2

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. (2018). Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000455.

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–e49. https://doi.org/10.1037/xge0000056.

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

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*. https://doi.org/10.2307/1420267.

Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune. *PLoS Biology, 10*(7), 5. https://doi.org/10.1371/journal.pbio.1001372.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex (New York, N.Y.: 1991), 11*(10), 946–953. https://doi.org/10.1093/cercor/11.10.946.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161.

Zatorre, R. J., Meyer, E., Gjedde, A., & Evans, A. C. (1996). PET studies of phonetic processing of speech: Review, replication, and reanalysis. *Cerebral Cortex, 6*(1), 21–30. https://doi.org/10.1093/cercor/6.1.21.