Music to Your Ears: Sentence Sonority and Listener Background Modulate the "Speech-to-Song Illusion"

Tamara Rathcke
University of Konstanz, Konstanz, Germany & Western Sydney University, Sydney, Australia & University of Kent, Canterbury, United Kingdom

Simone Falk
University of Montreal, Montreal, Canada & Université Sorbonne Nouvelle, Paris, France & International Laboratory for Brain, Music and Sound Research (BRAMS), Montreal, Canada

Simone Dalla Bella
International Laboratory for Brain, Music and Sound Research (BRAMS), Montreal, Canada & University of Economics and Human Sciences in Warsaw, Warsaw, Poland

Listeners usually have no difficulties telling the difference between speech and song. Yet when a spoken phrase is repeated several times, they often report a perceptual transformation that turns speech into song. There is a great deal of variability in the perception of the *speech-to-song illusion* (STS). It may result partly from linguistic properties of spoken phrases and be partly due to the individual processing difference of listeners exposed to STS. To date, existing evidence is insufficient to predict who is most likely to experience the transformation, and which sentences may be more conducive to the transformation once spoken repeatedly. The present study investigates these questions with French and English listeners, testing the hypothesis that the transformation is achieved by means of functional re-evaluation of phrasal prosody during repetition. Such prosodic re-analysis places demands on the phonological structure of sentences and language proficiency of listeners. Two experiments show that STS is facilitated in high-sonority sentences and in listeners' non-native languages and support the hypothesis that STS involves a switch between musical and linguistic perception modes.

Received: July 14, 2020, accepted March 21, 2021.

Key words: speech-to-song illusion, perceptual transformation, phonological sonority, L2-processing, L2-proficiency

The *speech-to-song illusion* is a perceptual phenomenon in which a spoken phrase shifts to being heard as sung by listeners after a series of repetitions. This transformation indicates a tight link between language and music and has attracted much research attention since its discovery (Deutsch, 1995). The transformation usually occurs during the third repetition of the phrase (Falk, Rathcke, & Dalla Bella, 2014) and is accompanied by a change in activation of the involved neural circuits that process spoken vs. musical signals, recruiting a network of areas associated with pitch extraction, song production, and auditory-motor integration (Tierney, Dick, Deutsch, & Sereno, 2012). Once transformed, the phrase often continues to be perceived as song, and its musical melody cannot be "unheard" (Groenveld, Burgoyne, & Sadakata, 2020). Despite having been the focus of several recent studies (e.g., Graber, Simchy-Gross, & Margulis, 2017; Groenveld et al., 2020; Jaisin, Suphanchaimat, Figueroa, & Warren, 2016; Tierney, Patel, & Breen, 2018), STS still poses many questions.

Not all sentences are equally likely to transform into song when repeated (Falk et al., 2014; Tierney et al., 2012). Tierney et al. (2012) report having discovered 24 high- and 24 low-transforming phrases after an exhaustive search through two large online libraries of audiobooks available in English. Alternatively, hypothesis-driven manipulations of spoken phrases provide an experimental tool for studying which properties induce or hinder STS (Falk et al., 2014). Both methodological approaches have so far provided converging evidence for the crucial role of the fundamental frequency (F0) that corresponds to the perceived pitch of the speaker's voice. Accordingly, stable local F0-trajectories provide a strong acoustic scaffold for STS to arise while increased local F0-dynamics tend to suppress the effect (see Figure 1). The importance of pitch track stability has recently been replicated with both naturally produced (Tierney et al., 2018) as well as manipulated (Groenveld et al., 2020) stimuli, and seems to be one of the most robust acoustic cues to STS. In contrast, pitch intervals that represent prominent scalar intervals in Western music (Cross, Howell, & West, 1983; Krumhansl, 2000) do not promote STS (Falk

Music Perception, volume 38, issue 5, pp. 499–508, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2021.38.5.499

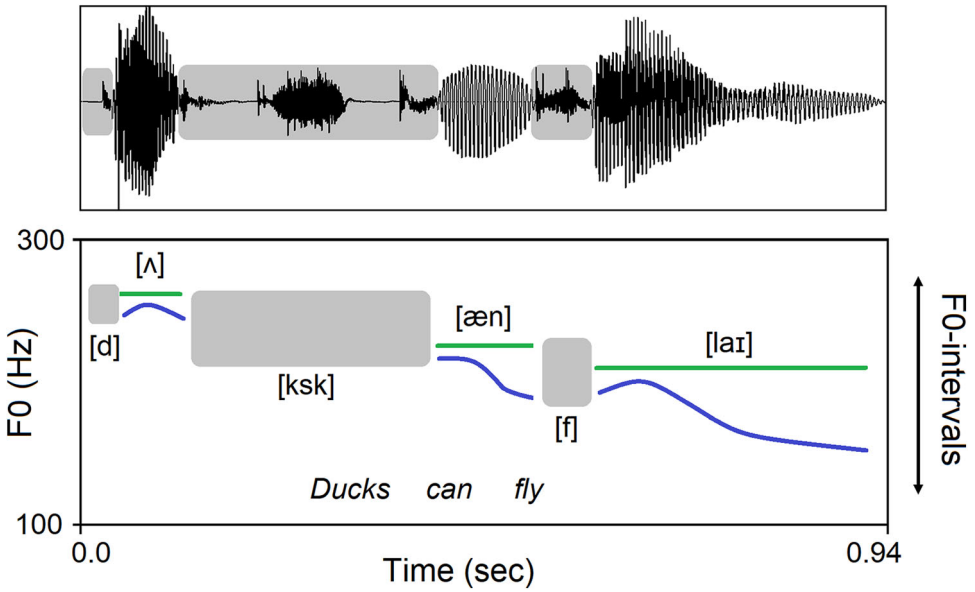

FIGURE 1. Waveform (top panel) and F0-trajectories (bottom panel) of the English sentence *"Ducks can fly"* spoken by a female. Stable F0-trajectories (shown in grey) indicate pitch patterns cueing STS, dynamic F0-trajectories (shown in black) indicate pitch patterns supressing STS. Grey squares highlight intervals of missing voicing and F0-information during the production of [d], [k s k], [f] (in *"Ducks can fly"*), in contrast to sonorant sections during the production of [ʌ], [ən], [laɪ].

et al., 2014, Figure 1). This finding has also been replicated in recent research (Tierney et al., 2018) and suggests that a musical melody can be established perceptually, without any strong acoustic cueing. As Deutsch, Henthorn, and Lapidis (2011, p. 2246) propose, "during the process of repetition, the pitches forming the phrase increase in perceptual salience, and . . . in addition they are perceptually distorted as to conform to a tonal melody." In support of this view, a recent study provided evidence that STS reduced the awareness and memory for pitch while enhancing those for duration (Graber et al., 2017).

Rhythmic properties of spoken sentences can also influence STS, although features conducive to the transformation do not rely on any regularity or isochrony either within the phrase (Falk et al., 2014, Tierney et al., 2018) or across repetitions (Falk et al., 2014; Margulis, Simchy-Gross, & Black, 2015). Instead, low-level timing variability that arises from groupings of speech sounds into intervocalic (as opposed to syllabic) intervals increases the likelihood and the ease of STS, possibly by supporting a metrical interpretation of spoken phrases (Falk et al., 2014).

We have previously hypothesized that repetition leads to a functional re-evaluation of prosodic properties of repeated spoken phrases whereby aspects relevant to speech processing dominate the perception initially and gradually give way to the percept of a musical melody (Falk et al., 2014). Similarly, Margulis (2013) proposes that a *speech perception mode* switches to a *music perception mode* during repetitions. Castro, Mendoza, Tampke, and Vitevitch (2018) explain STS within a connectionist framework of the node structure theory (MacKay, 1987) and assume that once lexical nodes have been saturated, the activation spreads to neighboring nodes that encode sound properties. Such theoretically grounded accounts of STS make it possible to advance the current understanding of STS in a hypothesis-driven way, which is the approach taken by the present study.

Research to date has primarily focused on acoustic properties of phrases that are looped to induce STS. It remains an open question whether or not an informed prediction can be made for a sentence based on its phonological properties alone. Assuming that STS relies on a melodic reanalysis (Deutsch et al., 2011, Falk et al., 2014) that is enhanced by high pitch (Groenveld et al., 2020; Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015), the phonological structure of spoken phrases ought to promote the transmission of pitch information in order to facilitate the transformation. Crucially, pitch perception hinges on the presence of vocal fold vibration and its acoustic correlate, the fundamental frequency (F0, Ladefoged & Maddieson, 1996), which

shape the phonological sonority of sentences. Sonority is often viewed on a scale from high (open vowels like [æ a ɑ]) to low (voiceless oral stops like [p t k], Clements, 1990) which primarily reflects the presence of vocal fold vibration and also the degree of vocal tract opening during sound production. Low-sonority phonemes (especially voiceless stops and fricatives) pose difficulties to the transmission of pitch information (see Figure 1). Hence, sentences with a high number of such phonemes (like voiceless stops or fricatives) are unlikely to support STS. In contrast, sentences containing many high-sonority sounds (like vowels, nasals, and approximants) can be expected to facilitate pitch extraction from the acoustic signal and thus promote STS. The *Sonority Hypothesis* is tested in Experiment 1.

Most of the unexplained variability in the perception of STS is, however, individual. Listeners' musicianship does not effectively predict the individual experience of STS as the effect can occur regardless of musical training (Vanden Bosch der Nederlanden et al. 2015), though musical aptitude tends to slightly increase the likelihood of STS arising (Falk et al., 2014). Listeners' musical aptitude may activate memory representations of music melodies and thus facilitate prosodic reanalysis of spoken phrases, though it is unlikely to play a role during the linguistic processing of phrases prior to such reanalysis.

Linguistic processing is evidently involved in the transformation. STS prevails in listeners whose native language uses pitch post-lexically (Deutsch et al., 2011; Falk et al., 2014; Jaisin et al., 2016) and is weak, if at all present, in listeners whose native language has lexical tone (Jaisin et al., 2016). Listeners of tonal languages tend to encode pitch patterns as having linguistic meaning (Bidelman & Lee, 2015), which may hinder their ability to experience STS in both their native (L1) and non-native (L2) language, even if the latter does not have lexical tone (Jaisin et al., 2016).

When exposed to repetitions of phrases from an unfamiliar language, listeners tend to have a stronger STS-experience in phrases that sound most foreign (or less pronounceable) to them, in contrast to phrases that sound more familiar (or more easily pronounceable, Margulis et al., 2015). This finding speaks to our main hypothesis that the linguistic processor ought to disengage from the analysis of incoming speech in order for the phrasal melody to be processed as singing (Castro et al., 2018; Falk et al., 2014; Margulis, 2013). Linguistic analyses are known to be cognitively costly, especially in L2 (Pérez, Hansen, & Bajo, 2019). Accordingly, the *Proficiency Hypothesis* of this study assumes that listeners' L2-mastery moderates their STS-experience in L2. The transformation is expected to be reduced in listeners with limited L2-skills who might take longer to extract the linguistic meaning of phrases, thus delaying or blocking prosodic reanalysis. In contrast, STS is likely to be facilitated in listeners with extensive L2-skills who will be faster at extracting linguistic meaning of sentences and reanalyzing phrasal prosody as singing. This hypothesis is tested in Experiment 2.

Method

MATERIALS

Twelve sentences were created in English and French (see Supplementary Materials at mp.ucpress.edu). The sentences varied in sonority (high vs. low) and were matched in length (4–14 syllables) and syntactic structure. A native female speaker of each target language read the sentences, paying attention to matching speech rate and pitch patterns across the two sonority conditions.

To ensure the success of the intended manipulation and to ascertain cross-linguistic comparability of the materials, we calculated a mean sonority index for each sentence. Each phoneme's location on the phonological sonority scale (Clements, 1990) was numerically coded from minimally 0 (for voiceless plosives) to maximally 9 (for open vowels). A mean sonority index was then calculated as an average across all sentence phonemes. Subsequent Welch two-sample *t*-tests revealed that the manipulation was successful in both French and English (see Table 1), whereas the subtle cross-linguistic differences between the high-sonority sets (5.62 in French vs. 5.88 in English) and low-sonority sets (3.93 in French vs. 4.30 in English) were not significant. Table 1 further compares how the intended sonority manipulation translates into its main acoustic-phonetic correlate, the duration of sonorous sounds (measured as percentage of the total sentence duration, % sonorous). On average, the duration of sounds that can carry pitch information made up less than half of the total duration of low-sonority sentences and about 80% of the total duration of high-sonority sentences.

To control for other factors that might influence STS, speech rate (in syllables per second) and pitch variability (in semitones) were also compared between the high-sonority and low-sonority sentences. To avoid measuring microprosodic F0-influences caused by intervening consonants (e.g., Hanson, 2009) and to ensure comparability across the two languages and sonority conditions, pitch variability was measured as the pitch change happening within a vowel. Accordingly, F0 was measured at 25% and 75% of each vowel.

TABLE 1. *Measurements of Phonological and Phonetic Properties of the Test Sentences*

	English			French		
	High-sonority	Low-sonority	*t-tests*	High-sonority	Low-sonority	*t-tests*
Sonority index	6.05	4.30	$t(7.9) = 6.6$, $p < .001$	5.62	3.93	$t(9.9) = 8.1$, $p < .001$
% Sonorous	84.88	46.44	$t(9.8) = 8.5$, $p < .001$	79.64	39.17	$t(7.5) = 8.8$, $p < .001$
Speech rate (syll/sec)	5.19	4.55	$t(10.0) = 1.3$, *ns*	5.54	4.91	$t(9.5) = 1.2$, *ns*
Pitch variability (st)	0.24	0.39	$t(84.8) = 0.8$, *ns*	0.11	0.39	$t(84.2) = 1.6$, *ns*

The F0-difference between the two measurement points was converted into semitones. Neither of the two extraneous factors (speech rate, pitch variability) differed significantly between the two experimental conditions (see Table 1) or between the two languages.

The test sentences were looped with eight repetitions, each separated by a 400 ms pause. Experiment 1 tested 12 stimuli in each language whereas Experiment 2 tested a subset of 10 stimuli per language (five sentences in each sonority condition).

PROCEDURE

English listeners were tested at the University of Kent, French listeners at the Sorbonne Nouvelle Paris-3 University. Prior to the experiment, participants filled in an online questionnaire that screened for amusia. The questionnaire further asked about music training, ongoing and past musical activities, the number of played instruments (which included singing) and the age at which participants took up musical training. A composite score of listener musicality (cf. Šturm & Volín, 2016) was derived from the questionnaire data.

Once in the lab, participants first rated individual test sentences on a scale from *1 (clearly speech)* to *8 (clearly song)*. These ratings established the baseline of perceived song-likeness of the experimental stimuli prior to repetition (Falk et al., 2014, Groenveld et al., 2020) and were followed by a series of distractor tasks. The session ended with the STS-test in which participants were instructed to listen to the looped sentences and indicate their STS-perception by pressing a button when (and only when) they experienced the transformation. They had to wait until the end of the loop without pressing any buttons if they did not experience STS. At the end of each trial, participants evaluated how song-like the sentence sounded to them after the last repetition, using the same Likert scale as in the baseline test. Experiment 1 was monolinguistic, i.e., listeners rated sentences in their native language only. Experiment 2 was cross-linguistic, i.e., listeners rated sentences in their L1 and L2. The order of the two languages was counterbalanced across individual sessions. A pair of good-quality headphones was used to present auditory stimuli. The study received approval from the ethics committee of the University of Kent. All listeners gave informed consent to participate in this research and received a small payment.

The above procedure obtained data on the speed and the likelihood of STS (Falk et al., 2014), along with the baseline and the strength of STS (e.g., Groenveld et al., 2020; Tierney et al., 2018).

LISTENERS

Forty English and forty French listeners (59 female, mean age = 27, range = 18–43) participated in Experiment 1. A different group of forty English and forty French listeners (59 female, mean age = 23, range = 18–42) took part in Experiment 2. Foreign language skills of the second group varied, and were assessed using a free online test by Education First. The test resulted in scores ranging from 0 (absolute beginner, A1) up to 100 (native-like proficiency, C2), according to the European Framework of Reference (Council of Europe, 2011). Attention was paid to counterbalancing the levels of L2-proficiency across the two language groups as far as possible, though, overall, L2-English skills in our French sample (mean: 88.9, or C2) were significantly higher, $t = 10.44$, $df = 77.8$, $p < .001$, than L2-French skills in our English sample (mean = 43.8, or B1). Nevertheless, individual variability spanned all proficiency levels in both groups of listeners.

Participants' musicality scores varied from 1 (*no musical training received*) to 26 (*high levels of music training and experience*), but there were no professional musicians among the four groups.

Results

Figure 2 displays "song-like" ratings of all sentences at baseline vs. after the exposure to repetitions in the two

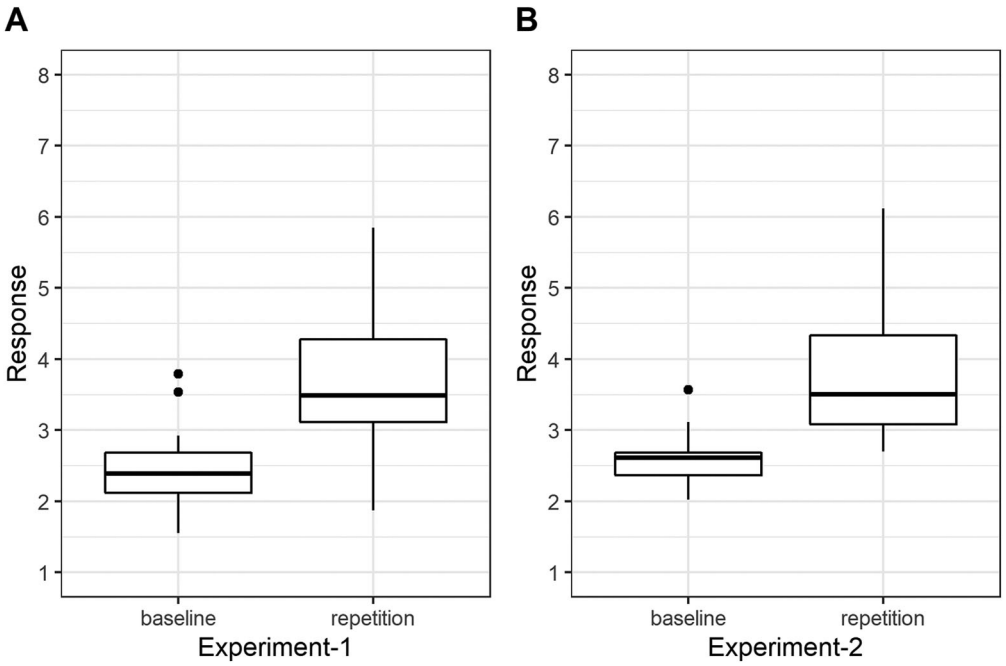

FIGURE 2. Ratings of test sentences on the 8-point scale (1 = *clearly speech*, 8 = *clearly song*) after a single exposure (baseline) vs. after eight repetitions. Responses from (mono-linguistic) Experiment 1 are plotted in panel A, those from (cross-linguistic) Experiment 2 in panel B.

experiments. Baseline responses clustered tightly at the lower end of the given 8-point scale. Wilcoxon signed-rank test confirmed that the test sentences were rated significantly more song-like after repetitions; Experiment 1: $V = 0$, $p < .001$; Experiment 2: $V = 0$, $p < .001$.

A series of mixed-effects models were subsequently fitted to the collected STS-data, to investigate:

- the *likelihood* of STS: the dependent variable measures whether or not (1/0) participants reported having perceived the transformation during repetitions of a given test sentence (binomial models),
- the *speed* of STS: the dependent variable reflects the repetition cycle (1–8) during which participants reported the transformation (ordinal models),
- the *strength* of STS: the dependent variable is based on the song-like ratings (1–8) of test sentences, collected after repetitions (ordinal models).

Listener and *sentence* were fitted as crossed random intercepts (Baayen, Davidson, & Bates, 2008). To control for variability in sentence length (Rowland, Kasdan, & Poeppel, 2019) and listener musicality (Falk et al., 2014), the number of syllables per sentence and individual musicality scores were included as covariates. The former was significant in some models, the latter in none. An RMarkdown file (see Supplementary Materials) outlines the analyses conducted in Rstudio (running R-version 4.0.3).

EXPERIMENT 1

Experiment 1 tested the effect of sentence sonority (high/low) on STS. Best-fit models showed the predicted effect on STS-likelihood, $x^2(1) = 7.40$, $p < .01$, speed, $x^2(1) = 8.90$, $p < .01$, and strength, $x^2(1) = 11.82$, $p < .001$, of the transformation. High-sonority sentences were more likely to transform, $z = 2.95$, $p < .01$ (Figure 3-A), and they did so one repetition cycle earlier than low-sonority sentences, $z = 3.19$, $p < .01$ (Figure 3-B). Moreover, high-sonority sentences sounded significantly more song-like after repetitions than their low-sonority counterparts, $z = 3.92$, $p < .001$ (Figure 3-C). We also checked the role of sentence language on STS-perception, but the French and the English listeners of Experiment 1 did not differ in any responses to the experimental stimuli of their native language.

EXPERIMENT 2

Experiment 2 investigated the role of the listener's language proficiency on their STS-experience. That is, the main factors of interest were *language* of the looped sentence (L1/L2), L2-score of the listener (numerical, 0–100), and their interaction. The best-fit model

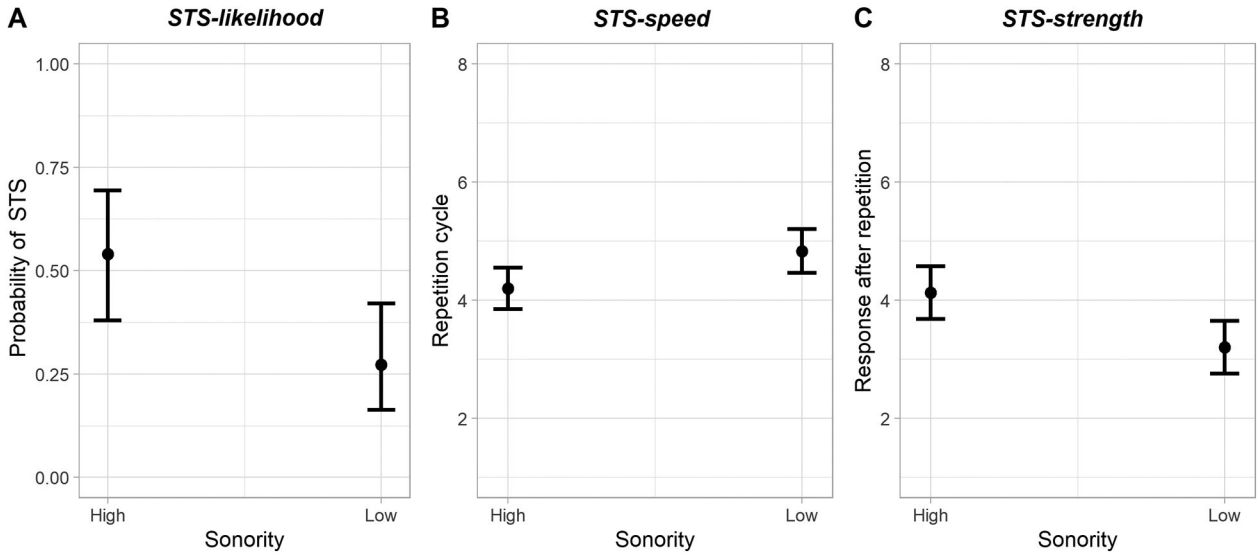

FIGURE 3. Effect of phrase sonority on the likelihood (A), speed (B, cycles 1-8) and strength (C) of STS. Responses after repetition (C) were collected on an 8-point Likert scale from 1 (*clearly speech*) to 8 (*clearly song*).

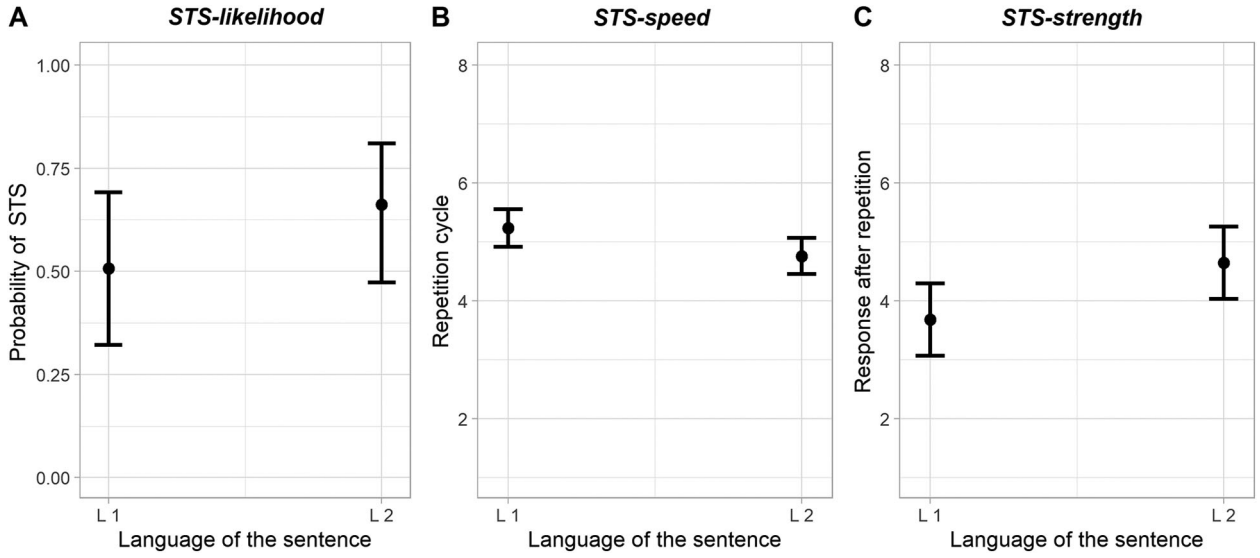

FIGURE 4 Main effect of language on the likelihood (A), speed (B, cycles 1-8) and strength (C) of STS. Responses after repetition (C) were collected on an 8-point Likert scale from 1 (*clearly speech*) to 8 (*clearly song*).

revealed a main effect of language of the sentence on its STS-likelihood, $x^2(1) = 22.33$, $p < .001$. Accordingly, all participants of the present study reported approximately 15% more transformations in their L2 than in their L1, $z = 4.78$, $p < .001$ (Figure 4-A). This effect held regardless of participants' proficiency levels in their L2. As far as the speed of STS was concerned, both language of the sentence, $x^2(1) = 17.01$, $p < .001$, and L2-scores of the listener, $x^2(1) = 6.19$, $p < .05$, affected the transformation (though not in interaction). Overall, STS was reported one repetition cycle earlier in L2 than in L1, $z = 4.12$, $p < .001$ (Figure 4-B) but was generally delayed in participants with a higher level of L2-skills, $z = 2.51$, $p < .05$ (Figure 5-A). In contrast, the strength of STS was significantly affected by the sentence language and listener L2-score in interaction, $x^2(1) = 8.90$, $p < .01$. The effect was rather subtle: while listeners with lower L2-scores tended to give slightly higher song-like

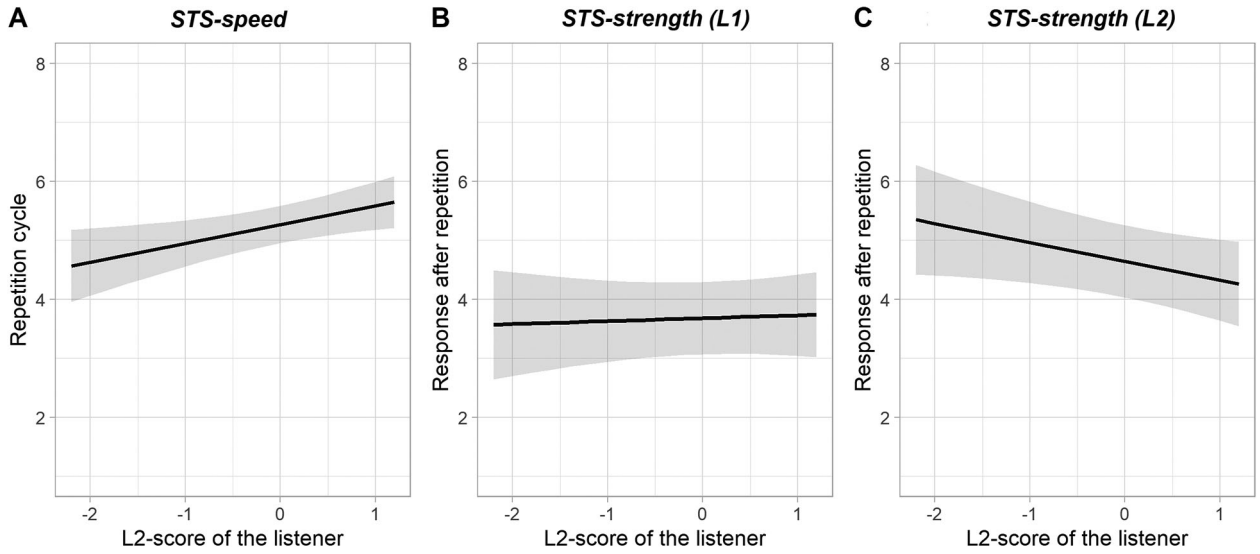

FIGURE 5. Effect of the listener's L2-proficiency on the speed (panel A, cycles 1-8) and the strength of STS in their native (panel B) and non-native (panel C) language. Responses after repetition (shown in panels B and C) were collected on an 8-point Likert scale from 1 (*clearly speech*) to 8 (*clearly song*). Note that L2-proficiency scores are centred around the group mean (0).

ratings to L2-sentences after repetition, $z = 1.88$, $p = 0.06$ (Figure 5-C), their L2-proficiency did not matter for the song-like ratings of L1-sentences, $z = 0.12$, *ns* (Figure 5-B). Overall, sentences heard in the listener's L2 were rated as sounding more song-like after repetition, $z = 12.91$, $p < .001$ (Figure 4-C).

Discussion

The aim of the present study was to test two aspects of a hypothesized mechanism that might give rise to STS (Castro et al., 2018; Falk et al., 2014; Margulis, 2013). The Sonority Hypothesis made a prediction for the susceptibility of sentences to STS based on their phonological structure while the Proficiency Hypothesis made a prediction for the listener's susceptibility to the transformation based on L2-language skills.

THE SONORITY HYPOTHESIS

Given the importance of pitch in STS (Falk et al., 2014; Groenveld et al., 2020; Tierney et al., 2018), the Sonority Hypothesis of Experiment 1 predicted that high-sonority sentences would facilitate the extraction of pitch information and thus promote STS. In contrast, low-sonority sentences were expected to inhibit STS as a general repetition effect that is known to bias perception of any acoustic signal toward the interpretation of musical structure (Rowland et al., 2019). These predictions were borne out in Experiment 1. On average, high-sonority sentences were twice as likely to induce STS when repeated, transformed one cycle earlier and sounded more song-like after repetition than their low-sonority counterparts of similar length and syntactic structure.

Highly sonorous sounds including vowels, nasals, and approximants have the ability to carry a tune because they are produced with an unobstructed vocal tract and a continuous vocal fold vibration that create resonance necessary for singing (cf. Ladefoged & Johnson, 2015). Hence, the facilitating effect of sonority for the perception of STS could stem from aspects of sound production that also play a crucial role in pitch transmission and perception (Ladefoged & Maddieson, 1996). However, phrases consisting exclusively of sonorants do not frequently occur in natural language (Rathcke, 2017), while in singing, sonority of underlying linguistic representations is typically enhanced by lengthening of vowels, i.e., by changing timing characteristics of speech acoustics (Eckardt, 1999). A deeper understanding of STS-foundations and mechanisms will benefit from future studies into potential interactions between timbral quality of varied sentence sonority (Clements, 1990) and the phonetics of resulting pitch patterns (cf. Allen & Oxenham, 2014; Caruso & Balaban, 2014).

Importantly, the likelihood of STS in high-sonority sentences of the present study was not as high as established in our previous research with pitch-manipulated stimuli (50% in Experiment 1 vs. 80% in Falk et al., 2014). This result suggests that properties of phrasal melody may be more central to the transformation (cf.

Deutsch et al., 2011; Falk et al., 2014; Tierney et al., 2018; Vanden Bosch der Nederlanden et al., 2015) than the phonological sonority that supports melodic reanalysis but does not actively promote a melodic interpretation. Nevertheless, Experiment 1 highlights that not all sentence-related features that facilitate STS derive from the acoustic make-up of the sentences.

THE PROFICIENCY HYPOTHESIS

Experiment 2 investigated STS-perception in 80 native listeners of English and French whose L2-ability in the other language varied from basic to advanced. According to the Proficiency Hypothesis, low-proficiency listeners would have a reduced STS-experience in their L2, since syntactic parsing, lexical access, and lexico-syntactic integration were delayed in L2 compared to L1 (Dufour & Kroll, 1995; Kilborn, 1992; Wartenburger et al., 2003). However, Experiment 2 shows that, overall, experiencing repeated speech in L2 strengthens STS and induces the transformation earlier and more frequently, regardless of listeners' L2-proficiency. The role of proficiency appears marginal in comparison, though listeners with fewer L2-skills do report higher STS-strength in their L2.

The Proficiency Hypothesis was based on the assumption that listeners would equally extract the linguistic message in their L1 and L2. However, L2-processing is not strictly automatic even in fluent bilinguals (Favreau & Segalowitz, 1983; Segalowitz, Segalowitz, & Wood, 1998). Given that Experiment 2 did not include assessment of L2-comprehension, there is a possibility that listeners (particularly those with low proficiency) did not attempt to process L2-phrases linguistically and experienced them as if they were spoken in a completely unfamiliar language, thus demonstrating the previously observed foreign-language effect (Jaisin et al., 2016; Margulis et al., 2015). Task demands are known to affect processing and comprehension of L2-speech (Kilborn, 1992; Tan & Foltz, 2020) as well as the perception and action more generally (Memelink & Hommel, 2007), and might have led listeners of this study to exclusively attend to the sound structure of the stimuli, bypassing other sources of linguistic information in the acoustic signal or engaging in a shallow encoding of L2 sentences. Experiment 2 will thus benefit from a replication design to include tests of lexico-syntactic integration and access in L2-listeners.

Overall, the results of Experiment 2 can be reconciled with the explanation that non-native listeners (especially those who have little experience in their L2) might be able to forego linguistic meaning extraction, focusing exclusively on the prosody of L2-messages and thus experiencing a stronger STS-effect. Importantly, the present findings indicate that the linguistic background of listeners contributes to STS more than their musical background, highlighting the need to further examine the involvement of linguistic processing in STS.

General Discussion and Conclusion

The present study enriched existing evidence on the workings of STS by documenting the effects of phonological sonority in the looped sentence and listener proficiency in the language of the spoken stimulus. These findings highlight the importance of pitch transmission for STS, regardless of the exact acoustic implementation of pitch relationships that have been extensively studied before (Deutsch et al., 2011; Falk et al., 2014; Tierney et al., 2018), and indicate a mediating effect of the language processor in STS (cf. Castro et al., 2018; Falk et al., 2014; Margulis, 2013).

At first glance, a strong involvement of linguistic processing in STS seems at odds with the idea of a "repetition-to-music" effect put forward in the studies demonstrating that a similar perceptual transformation can be induced in looped environmental sounds (Rowland et al., 2019, Simchy-Gross & Margulis, 2018). Indeed, sounds of water drops, ice cracks, wind noise or bee buzz and chicken cackle can be perceived as musical when repeated. The perceiver in all of these experiments is, however, the human listener whose processing of linguistic vs. environmental sounds or vocalisations of other species is known to differ (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Scott, Blank, Rosen, & Wise, 2000), potentially driven by the ability to categorise and assign specific meanings to the acoustic signal (Leech, Holt, Devlin, & Dick, 2009). The lack of semantic processing has been discussed as one of the main reasons why jumbling of segments in looped non-speech excerpts does not block the transformation to music in a similar way to how jumbling of syllables in looped sentences blocks STS (Simchy-Gross & Margulis, 2018). Assuming that lexico-syntactic processing in a listener's non-native language is shallow and task-dependent (Favreau & Segalowitz, 1983; Kilborn, 1992; Segalowitz et al., 1998; Tan & Foltz 2020), the L2-effects observed in Experiment 2 corroborate the idea that meaningfulness of the acoustic signal mediates the transformation. The "repetition-to-music" effect is therefore likely to be stronger in non-speech than in speech, as results of a previous study suggest (Rowland et al., 2019), though pertinent evidence for the meaning hypothesis is yet to be provided.

The role of acoustic properties of pitch in STS has been repeatedly discussed and well documented (Deutsch

et al., 2011; Falk et al., 2014; Tierney et al., 2018), while there are doubts that it matters as much for the "repetition-to-music" effect that might rely more heavily upon rhythmic processing (Rowland et al., 2019). Crucially, all environmental sounds that have been tested in previous research (Rowland et al., 2019, Simchy-Gross & Margulis, 2018) seem to have had measurable (or inducible) fundamental frequency. This resonates with the sonority effect in Experiment 1, demonstrating that an increased amount of transmittable pitch information fosters the transformation. The generalizability of the "repetition-to-music" effect is therefore yet to be attested with sounds whose acoustic properties are missing the fundamental and/or inhibit its induction like the noise of a radio static or rustling autumnal tree leaves.

To conclude, the present study demonstrates that STS links language to music in complex ways, involving a switch between musical and linguistic perception modes (Castro et al., 2018; Falk et al., 2014; Margulis, 2013) that is moderated by the linguistic (rather than musical, cf. Vanden Bosch der Nederlanden et al., 2015) background of listeners. The results have broader implications for the future study of the "repetition-to-music" effect as a general phenomenon that biases perception of acoustic signals toward the interpretation of musical structure upon repetition.

Author Note

We would like to thank our undergraduate research assistants Georgia Ann Carter and Katherine Willet at the University of Kent, Chloé Lehoucq and Sasha Lou Wegiera at the Sorbonne Nouvelle Paris-3 Université who helped with the data collection, and James Brand who helped with ggplot and R. We further thank two anonymous reviewers who helped to improve the manuscript. This research was supported by a Small Research Grant from the British Academy (SG152108) to the first author.

Correspondence concerning this article should be addressed to Tamara Rathcke, Fachbereich Linguistik, Universität Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany. E-mail: tamara.rathcke@uni-konstanz.de

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *Journal of the Acoustical Society of America, 135(3),* 1371–1379. https://doi.org/10.1121/1.4863269

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language, 59(4),* 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312. https://doi.org/10.1038/35002078

Bidelman, G., & Lee, C. (2015). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. *NeuroImage, 120,* 191–200. https://doi.org/10.1016/j.neuroimage.2015.06.087

Caruso, V. C., & Balaban, E. (2014). Pitch and timbre interfere when both are parametrically varied. *PloS ONE*, *9*(1), e87065. https://doi.org/10.1371/journal.pone.0087065

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

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* https://rm.coe.int/16802fc1bf

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech* (pp. 283–333). Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.017

Cross, I., Howell, P., & West, R. (1983). Preferences for scale structure in melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance, 9*(3), 444–460. https://doi.org/10.1037/0096-1523.9.3.444

Deutsch, D. (1995). *Musical illusions and paradoxes* [CD]. Philomel Records. www.philomel.com

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

Dufour, R., & Kroll, J. F. (1995). Matching words to concepts in two languages: A test of the concept mediation model of bilingual representation. *Memory and Cognition*, *23*(2), 166–180. https://doi.org/10.3758/BF03197219

Eckardt, F. (1999). *Singen und Sprechen im Vergleich artikulatorischer Bewegungen*. [Singing and speaking - articulatory movements compared]. THIASOS.

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

Favreau, M., & Segalowitz, N.S. (1983). Automatic and controlled processes in the first- and second-language reading of fluent bilinguals. *Memory and Cognition, 11,* 565–574. https://doi.org/10.3758/BF03198281

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

Groenveld, G., Burgoyne, J. A., & Sadakata, M. (2020). I still hear a melody: Investigating temporal dynamics of the speech-to-song illusion. *Psychological Research,* 84, 1451–1459. https://doi.org/10.1007/s00426-018-1135-z

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *Journal of the Acoustical Society of America, 125(1),* 425–441. https://doi.org/10.1121/1.3021306

Jaisin, K., Suphanchaimat, R., Figueroa, M.C., & Warren, J.D. (2016). The speech-to-song illusion is reduced in speakers of tonal (vs. non-tonal) languages. *Frontiers in Psychology 7,* 662. https://doi.org/10.3389/fpsyg.2016.00662

Kilborn, K. (1992). On-line integration of grammatical information in a second language. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 337–350). Elsevier. https://doi.org/10.1016/S0166-4115(08)61504-6

Krumhansl, C. L. (2000). Rhythm and pitch in music cognition. *Psychological Bulletin, 126,* 159–179. https://doi.org/10.1037/0033-2909.126.1.159

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages.* Blackwell.

Ladefoged, P., & Johnson, K. (2015). *A course in phonetics*. Cengage.

Leech, R., Holt, L. L., Devlin, J. T., & Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *Journal of Neuroscience, 29(16),* 5234–5239. https://doi.org/10.1523/JNEUROSCI.5758-08.2009

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

Margulis, E. H. (2013). *On repeat: How music plays the mind.* Oxford, UK: Oxford University Press.

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. https://doi.org/10.3389/fpsyg.2015.00048

Memelink J., & Hommel B. (2007). Tailoring perception and action to the task at hand. *European Journal of Cognitive Psychology, 18*(4), 579–592. https://doi.org/10.1080/09541440500423228

Pérez, A., Hansen, L., & Bajo, T. (2019). The nature of first and second language processing: The role of cognitive control and L2 proficiency during text-level comprehension. *Bilingualism: Language and Cognition, 22*(5), 930–948. https://doi.org/10.1017/S1366728918000846

Rathcke, T. (2017). How truncating are 'truncating languages'? Evidence from Russian and German. *Phonetica, 73,* 194–228. https://doi.org/10.1159/000444190

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

Scott, S. K., Blank, C. C., Rosen S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. Applied Psycholinguistics, 19*(1),* 53–67. https://doi.org/10.1017/S0142716400010572

Simchy-Gross, R., & Margulis, E. (2018). The sound-to-music illusion: Repetition can musicalize nonspeech sounds. *Music and Science, 1,* 1–6. https://doi.org/10.1177/2059204317731992

Šturm, P., & Volín, J. (2016). P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics, 55,* 38–52. https://doi.org/10.1016/j.wocn.2015.11.003

Tan, M., & Foltz, A. (2020). Task sensitivity in L2 English speakers' syntactic processing: evidence for Good-Enough processing in self-paced reading. *Frontiers in Psychology 11*, 575847. https://doi.org/10.3389/fpsyg.2020.575847

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2012). Speech versus song: Multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, *23*, 249–254. 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*, *147*(6), 888–904. https://doi.org/10.1037/xge0000455

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Every day musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology; General, 144,* e43–e49. https://doi.org/10.1037/xge0000056

Wartenburger, I., Heekeren, H. R., Abutalebi, J., Cappa, S. F., Villringer, A., Perani, D. (2003). Early setting of grammatical processing in the bilingual brain. *Neuron, 37*(1), 159–170. https://doi.org/10.1016/S0896-6273(02)01150-9