

Why Does Speech Sometimes Sound Like Song? Exploring the Role of Music-Related Priors in the “Speech-to-Song Illusion”

Tamara Rathcke¹ , Simone Falk^{2,3,4} and
Simone Dalla Bella^{3,4,5,6}

Abstract

The speech-to-song illusion is a perceptual effect emerging at the interplay of two cognitive domains, music and language. It arises upon repetitions of a spoken phrase that shifts to being perceived as song, and varies in the likelihood, ease, and vividness of its occurrence among individuals. A prevailing explanation of the illusion suggests that listeners' attention shifts to rhythm and melody of the phrase once their involvement with the linguistic meaning subsides. The present study tested this mechanism by manipulating meaning plausibility and structural complexity of French and English phrases and by obtaining measures of attentional and working memory capacity from 80 French and English listeners who were exposed to repetitions of sentences in their native language. The results show that the transformation was facilitated in listeners with fewer cognitive resources and in less plausible, more complex phrases, which is at odds with the previously proposed mechanism underpinning the speech-to-song illusion. The illusion-promoting effect of musical training was visible only in simple but not in complex phrases. We propose a new account of the perceptual transformation from speech to song as a cognitive effect arising from the accumulation of music-related priors in a linguistically ambiguous context of massed repetitions.

Keywords

Attentional resources, individual differences, meaning, musicality, the speech-to-song illusion, working memory capacity

Submission date: 19 October 2023; Acceptance date: 6 June 2024

Introduction

Our conscious experience of the world is a result of an external stimulation received by our senses and a processing effort made by our brain. As such, the process is prone to random error, though some errors appear more systematic, giving rise to what we might describe as perceptual illusions (McIntosh, 2022). The present article reports an empirical investigation of the speech-to-song illusion coined by Diana Deutsch (Deutsch, 1995; Deutsch et al., 2011; also referred to as the speech-to-song transformation, Falk et al., 2014; henceforth STS). This phenomenon describes an illusory transformation of spoken phrases which are perceived as singing upon repetition. Despite an extensive study of STS (e.g., Castro et al., 2018; Falk et al., 2014; Graber et al., 2017; Groenvelde et al., 2020; Jaisin et al., 2016; Rathcke et al., 2021b; Simchy-Gross & Margulis, 2018; Tierney et al., 2018; Tierney et al.,

¹ Department of Linguistics, University of Konstanz, Konstanz, Germany

² Faculté des arts et des sciences, Département de linguistique et de traduction, Université de Montréal, Montréal, Canada

³ International Laboratory for Brain, Music and Sound Research (BRAMS), University of Montréal, Montréal, Canada

⁴ Centre for Research on Brain, Language and Music (CRBLM), McGill University, Montréal, Canada

⁵ Department of Psychology, University of Montréal, Montréal, Canada

⁶ University of Economics and Human Sciences in Warsaw, Warsaw, Poland

Corresponding author:

Tamara Rathcke, Fachbereich Linguistik, Universität Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany.

Email: tamara.rathcke@uni-konstanz.de

Data Availability Statement included at the end of the article



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page

(<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

2021; Vanden Bosch der Nederlanden et al., 2015), the exact mechanism evoking STS is still not fully understood.

STS is one of the two auditory illusions that arise upon repetitions of speech. The first discovery of illusory transformations upon repetition was made with single words (Warren, 1961; Warren & Gregory, 1958). When a spoken word is repeated for a prolonged period of time, listeners tend to report that their perception of the word changes and alternates between auditory impressions of various, sometimes phonologically unrelated words, interspersed with percepts of the original input. Verbal transformations have been extensively studied since their discovery in 1958 (e.g., Basirat et al., 2012; Mackay et al., 1993; Natsoulas, 1965; Sato et al., 2006; Shoaf & Pitt, 2002; Warren, 1961; Warren & Warren, 1966). The results of this research demonstrate that the number of transformations can be highly variable across individuals (Warren, 1961; Warren & Warren, 1966), giving rise to 5–20 unique forms during a several-minute exposure (Shoaf & Pitt, 2002). The ability to experience verbal transformations emerges at the age of 6 and declines from the age of 60, suggesting that the illusory effect recruits a cognitive mechanism required for skilled listening to speech and is susceptible to maturational constraints and decay (Warren & Warren, 1966).

Previous research into verbal transformations has mostly focused on isolated words. Some experiments examined the effect of word length and demonstrated that polysyllabic words lead to a higher number of transformations than monosyllabic words (Kaminska et al., 2000; Shoaf & Pitt, 2002). However, the original experiments by Warren and colleagues (Warren, 1961; Warren & Gregory, 1958) also included short sentences like “Our ship has sailed.” or “Our side is right.”, which caused fewer transformations than isolated words. The observed patterns resulted in a general conclusion that “the greatest distortions occurred with phonetically simpler stimuli” (Warren, 1961, p. 255), paving the way to an extensive study of words rather than sentences in years to follow. An illusory perception that prevails when more complex speech stimuli (such as words forming phrases or sentences) are repeated was described as STS some 30 years later by Deutsch (1995). Apart from the primary difference in linguistic complexity, STS differs from verbal transformations in that the illusion affects only the prosodic shape of speech but not its segmental phonology (i.e., the perception of individual speech sounds constituting the phrase remains unaffected, Deutsch et al., 2011).

Prosody comprises melody, pitch, and rhythm of both spoken and written language (Frazier & Gibson, 2015; Shattuck-Hufnagel & Turk, 1996) and constitutes the best-known link between language and music (McMullen & Saffran, 2004; Patel, 2003, 2012). Not surprisingly, much research on STS has focused primarily on acoustic-prosodic properties of spoken utterances and their ability to induce the perceptual shift from speech to song (Deutsch et al., 2011; Falk et al., 2014; Falk et al., 2014; Groenvel-

et al., 2020; Tierney et al., 2018). Prosodic properties that promote STS include temporal stability of pitch tracks (Falk et al., 2014; Groenvelde et al., 2020; Tierney et al., 2018) and equalized duration of intervocalic intervals (as opposed to syllables, Falk et al., 2014). Moreover, sentence-inherent features that support the perception of pitch (Rathcke et al., 2021b) and individual characteristics that influence listeners’ ability to extract prosodic information from auditory signals (Tierney et al., 2021) both foster and enhance STS. Crucially, prosodic properties extend over large timescales that enable hierarchical relationships between neighboring structural units to be established (Beckman, 1996; Beckman et al., 2006; Beckman & Pierrehumbert, 1986). Such hierarchies are assumed to constitute mental representations of both language and music (e.g., Jackendoff, 2009; Jackendoff & Lerdahl, 2006; Patel, 2003, 2012). Accordingly, a hierarchical prosodic structure can hardly be derived from monosyllabic one-word utterances lacking alternations of strong and weak elements. Such utterances are thus very likely to give rise exclusively to verbal transformations rather than STS (Warren, 1961; Warren & Gregory, 1958), though even in case of verbal transformations, prosodic features such as (re-)grouping of phonemic constituents have also been hypothesized to shape the illusory effect (Basirat et al., 2012).

Repetition that is key to both STS and verbal transformations is also known to cause other auditory illusions. When non-linguistic sounds (like bee buzz, water droplets, chicken cackle or door noise) are played repeatedly to human listeners, they report perceiving musical excerpts as repetitions of the sound unfold in time (Rowland et al., 2019; Simchy-Gross & Margulis, 2018). This finding resonates with the observation that listeners tend to experience STS more readily and vividly when exposed to utterance repetitions spoken in a language that sounds foreign to them and is difficult for them to pronounce (Margulis et al., 2015). A “repetition-to-music” effect has been proposed and defined as a general perceptual tendency to induce musical attributes in any repetitive auditory signals (Simchy-Gross & Margulis, 2018). The proposal, however, overlooks the fact that massed repetitions of a single word in listeners’ native language(s) has not been previously reported to induce a musical percept (Basirat et al., 2012; Mackay et al., 1993; Natsoulas, 1965; Sato et al., 2006; Shoaf & Pitt, 2002; Warren, 1961; Warren & Warren, 1966). The “repetition-to-music” effect has so far been attested with non-linguistic sounds on relatively long timescales (2–5 s).¹ Moreover, these sounds had a measurable fundamental frequency, possibly enabling listeners to perceive pitch and to induce musical structure in the originally non-musical sound (cf. Rathcke et al., 2021b). In sum, a melody is never a note in isolation and requires time to evolve (Krumhansl, 2001; Krumhansl & Kessler, 1982).

Overall, existing theoretical accounts of STS have not yet determined if, and how, STS engages specific processes that are fundamentally different from the processes underpinning

verbal transformations (Warren, 1961; Warren & Gregory, 1958) or those involved in the domain-general “repetition-to-music” effect reported for nonspeech sounds (Rowland et al., 2019; Simchy-Gross & Margulis, 2018). A good point of departure to uncover the potential uniqueness of STS is the linguistic complexity of the auditory stimulus. While verbal transformations have been reported with isolated words, STS typically arises in phrases formed of several words (Deutsch, 1995; Deutsch et al., 2011; Falk et al., 2014; Jaisin et al., 2016; Margulis et al., 2015; Rathcke et al., 2021; Tierney et al., 2018; Vanden Bosch der Nederlanden et al., 2015). Such linguistic phrases have an internal structure that defines semantic relationships between phrasal constituents (e.g., “who did what to whom”), giving them their contextualized meaning that goes beyond purely lexical semantics (e.g., “cat”, “mouse”, “eat”; Stubbs, 2001). A linguistic utterance is thus much more than a list of isolated words (Stubbs, 2001). While this issue has been largely overlooked within current research on STS, many accounts of STS rest on the assumption that linguistic meaning is somehow involved in the perceptual switch between speech and song (e.g., Castro et al., 2018; Falk et al., 2014; Margulis, 2013; Margulis et al., 2015; Rathcke et al., 2021b; Tierney et al., 2021). The present study addresses the currently unresolved issue of how the switch from speech to song is moderated by the sentence-level meaning.

The aims of the present study are two-fold. First, we seek to clarify if, and how, sentence-level meaning influences STS. Second, we aim to test how individual listener traits that are involved in cognitive computations of the linguistic meaning might moderate the experience of STS which is known to be individually very variable (Rathcke et al., 2021b; Tierney et al., 2021). Starting from the idea shared by many existing accounts of STS that linguistic meaning has to somehow decay in the perception of listeners before they can experience sentence repetitions as being sung (e.g., Castro et al., 2018; Falk et al., 2014; Margulis, 2013; Rathcke et al., 2021b; Tierney et al., 2021), the present study tests the hypothesis that linguistic meanings that are difficult to extract from a complex phrasal structure will reduce STS. Two aspects of the process of the linguistic meaning extraction are manipulated, namely meaning plausibility and syntactic complexity. Semantically implausible (e.g., “My cat was eaten by a mouse.”) as well as syntactically complex (e.g., “The cotton clothing is made of grows in Mississippi.”²) sentences can cause difficulties during linguistic processing and comprehension (e.g., Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002). Accordingly, we predicted that implausible meanings and increased syntactic complexity would influence STS by interfering with the computation of linguistic meaning, and thus block, delay, and/or reduce the strength of STS-experience in listeners. In contrast to previous research that showed an increased tendency toward experiencing STS in a foreign or a non-native language of listeners (Margulis et al., 2015; Rathcke et al., 2021b), the

present study focuses on native listeners only. Sentence parsing and comprehension are arguably more automatic for language experts than novices (e.g., Ito & Pickering, 2021). Listening to someone talk and deliberately not understanding them is a difficult task when the language of the interaction is native to the listener, indicating that it may be harder to forego lexico-syntactic computations in a native (as compared to a non-native or a foreign) language.

The role of the listener background in STS has been extensively tested with regard to musical training and general musical aptitude, focusing on the song aspect of STS (e.g., Falk et al., 2014; Margulis et al., 2015; Tierney et al., 2013; Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015). However, listeners can also differ in terms of their ability to extract linguistic meanings from (complex) utterances of their native language(s) (e.g., Cummings, 2017; Tan & Foltz, 2020). They also differ in cognitive resources available to them for the computation of meaning and for language processing in general (Hannon & Daneman, 2001; Park & Reder, 2012). Given that lexico-syntactic processing is cognitively demanding and costly (e.g., Daneman & Merikle, 1996; Just et al., 1996; Just & Carpenter, 1992), the termination of semantic computations during repetitions may free up attentional and working memory resources (Tierney et al., 2021) which then can be used for a prosodic reanalysis leading to STS (Deutsch et al., 2011; Falk et al., 2014; Rathcke et al., 2021b; Tierney et al., 2021). The present study tests this hypothesis by investigating how individual working memory, attentional flexibility, and divided attention moderate listeners’ experience of STS.

Both attention and working memory are known to be capacity-limited cognitive resources involved in language processing (Baddeley, 2003; Kim et al., 2018; Kurland, 2011). Working memory plays an important role during sentence comprehension as it keeps the information about the ongoing syntactic parse and guides the understanding of the semantic relationships between syntactic constituents, helping to resolve potential ambiguities (MacDonald & Christiansen, 2002; Waters & Caplan, 1996). Similarly, attentional resources are involved in the assignment of syntactic roles to the constituents of a sentence (Tomlin, 1999) and are drawn upon in many tasks dealing with the semantic processing of both isolated words and complex sentences (Myachykov & Posner, 2005). Such resources have long been discussed to play an important role in a variety of illusions involving language (e.g., Hannon & Daneman, 2001; Park & Reder, 2012). Here, we hypothesized that listeners with a reduced attentional and/or working memory capacity would differ in their experience of STS, assuming that limited resources may compete during the auditory analysis of an incoming acoustic signal and therefore influence the activation of alternative mental representations. These individual effects are particularly likely to occur in conjunction with those lexico-syntactic features that affect semantic plausibility and syntactic complexity of sentences that are

repeated to create STS. Given that semantically implausible and syntactically complex sentences are more challenging to process (e.g., Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002), we expected that especially low-capacity listeners may have difficulties with the computation of linguistic meaning (Myachykov & Posner, 2005; Tomlin, 1999), and thus experience fewer, slower, or less vivid transformations. In addition, individually variable levels of musical training were measured and included as a control variable (e.g., Falk et al., 2014; Margulis et al., 2015; Tierney et al., 2013; Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015).

The two hypotheses were tested with 80 listeners whose cognitive resources were assessed using established batteries of working memory and attention (Wechsler, 1997; Zimmermann & Fimm, 2004). To increase the generalizability of the hypothesized effects, participants of the present study were recruited from two prosodically and structurally distinct languages – English and French. The participants were asked to listen to repetitions of sentences in their native language, and indicate if, when, and how strongly they experienced a shift from speech to song. The sentences varied in semantic plausibility and syntactic complexity, with similar lexico-syntactic manipulations across all stimuli of the two languages. We investigated how the sentence features, along with cognitive resources of listeners, may affect the likelihood and speed (Falk et al., 2014), as well as the perceived strength of STS (e.g., Groenveld et al., 2020; Margulis et al., 2015; Tierney et al., 2018).

Method

Participants

Forty English and forty French listeners (59 F) aged 18–43 (mean age: 27) volunteered to participate in this study. Prior to an experimental session, they filled in an online questionnaire that collected information about musical training and screened for amusia. Answers to a part of the questionnaire led to the calculation of an individual musicality index which was an aggregate of the scores given to the questions about years of musical training (from 0 to 19 in the sample), presence of a regular practice (0 for non-active and 1 for active participants), number of musical instruments (including singing, from 0 to 6 in the sample), age at the beginning of musical training (below the age of 10 coded as 2, from 10 up to 20 years coded as 1, above 20 years coded as 0). The derived index varied between 0 (no musical training received) to 26 (a relatively high level of musical training). This (or a comparable) way of capturing individual difference in musical training has been used in previous research, with compelling results (Rathcke et al. 2021a; Šturm & Volín, 2016). No professional musicians or singers participated in the experiment.

Three tests were chosen to measure individual cognitive abilities. The auditory working memory (WM) capacity

was tested using the forward and backward digit span test (WAIS-III, Wechsler, 1997) which measured individual storage and processing resources (Daneman & Carpenter, 1980). Additionally, three tests were selected from the Test Battery for Attentional Performance (TAP, Zimmermann & Fimm, 2004): divided attention, flexibility, and alertness. The TAP test of divided attention estimated listeners' ability to pay attention to two tasks simultaneously. Participants were asked to monitor for certain patterns involving either visual symbols or auditory tones and to quickly respond to both visual and auditory patterns. Here, the measure of interest was the number of errors participants made. Attentional flexibility estimated listeners' ability to quickly switch attention between two different tasks. Participants were asked to alternately respond to a letter and a number while both were present on the screen during each trial. Here, speed-accuracy trade-off characterized individual performance, with a compound measure derived from both reaction times and the total number of errors. Finally, alertness measured the general wakefulness in the presence of a stimulus ("phasic alertness"). Here, participants had to respond as quickly as possible to a visual stimulus appearing at randomly varying time intervals, with or without a preceding warning tone. The performance measure was calculated by subtracting reaction times to stimuli presented with a warning tone from reaction times to stimuli presented without a warning tone, divided by the median of total reaction times. The alertness index served as a control measure. Raw TAP and WM scores were transformed into percentiles of age-normed distributions (Wechsler, 1997; Zimmermann & Fimm, 2004). A correlation matrix of the listener characteristics is shown in Figure 1. None of the correlations between the individual measures was significant after the Bonferroni correction for multiple comparisons ($\alpha = .05/10$), which is in line with previous studies (e.g., Keye et al., 2009; Mall et al., 2014).

Materials

Two sets of sentence pairs varying in syntactic complexity of sentence-level structures and plausibility of sentence-level meanings were created in English and French. The first set (henceforth the "semantics set") consisted of syntactically simple sentences and manipulated the plausibility of lexical constituents. It contained 12 sentences contrasting plausible constituents (e.g., English: "The granma ate the lunch.") with implausible constituents (e.g., English: "The postbox ate the lunch."). The second set (henceforth the "syntax set") contained syntactically more complex sentences consisting of two verbal predicates and manipulated the mapping between syntactic constituents and prosodic breaks. It comprised 12 sentences in which the prosodic break location changed the sentence interpretability from plausible (e.g., English: "While the woman washed, the cat purred.") to implausible (e.g., English: "While the woman washed the cat, meowed."). The latter type of sentences introduces, or supports, the garden-path effect since

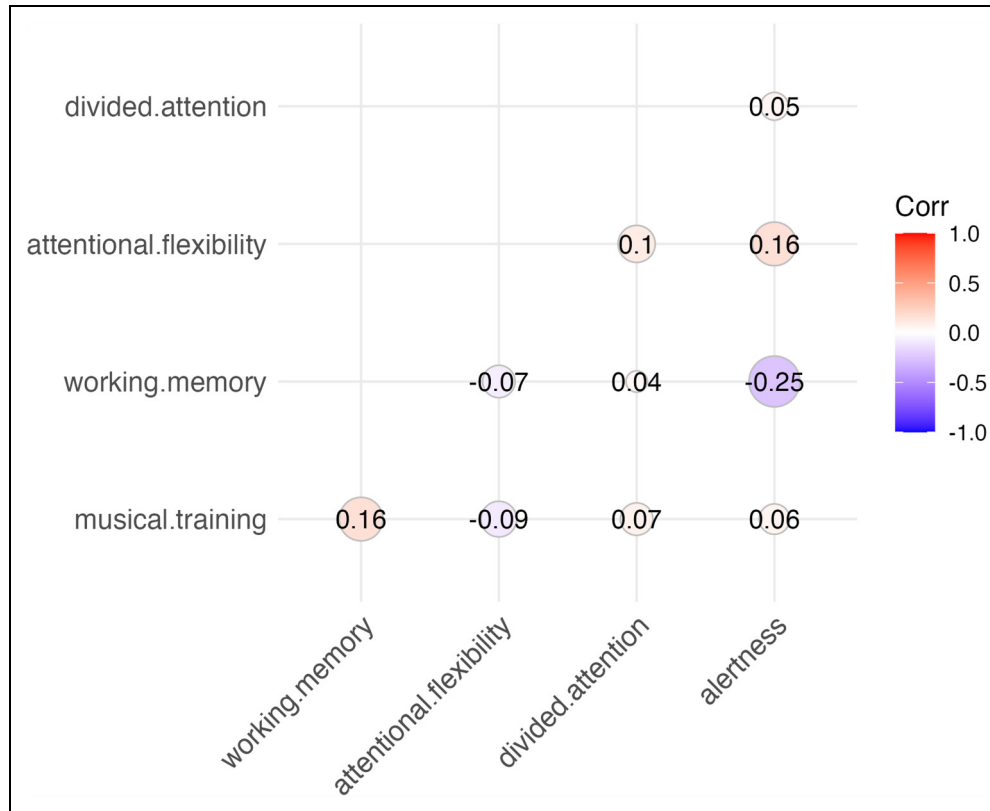


Figure 1. Correlation matrix displaying correlation coefficients for all pairs of the individual characteristics (including *alertness*, *divided attention*, *attentional flexibility*, *working memory capacity*, and *musical training*) measured in the study sample ($N = 80$).

the misplaced prosodic break invites incorrect syntactic parsing and results in the second verbal phrase lacking an argument (the subject). The prosodic break was expressed by means of an acoustic silence (with a constant duration of 200 ms across all stimuli). Minor changes in lexical choices (e.g., “purred”/ “meowed”) were implemented across pairs of the syntax set, to avoid monotony and to introduce some comparability to the semantics set. In the syntax set, the alternate items were taken from closely related semantic fields with the same syntactic roles. The full list of materials can be viewed on OSF (<https://osf.io/8d9mt>).

For each language, sentences of the two experimental conditions (plausible/ implausible) were matched in length (3–16 syllables) and syntactic structure. Attention was also paid to matching these aspects of the materials across the two languages wherever possible. By way of controlling for potential effects of phonology on STS (Rathcke et al., 2021b), each sentence was quantified in terms of its mean sonority. In the present materials, the sonority score varied between 4.0 and 5.35 (mean 4.68), reflecting the type of phonemes (e.g., high-sonority: approximants, nasals vs. low-sonority: stops, fricatives) that a sentence contained, and was balanced across the plausible/ implausible sentences as well as syntax/ semantics sets. Welch two-sample t-test confirmed that no systematic sonority differences existed in English ($t = .65$, $df = 18.19$, $p = .53$) and French ($t = 1.16$, $df = 17.33$, $p = .26$) sets.

To examine a potentially gradual effect of syntactic complexity across all materials, we additionally quantified the underlying structure of each sentence based on the following five aspects of its syntax: (1) the number of terminal syntactic nodes (i.e., the number of free morphemes); (2) the number of predicates (counting finite verbs, depending on whether or not a subordinate sentence was present); (3) the number of subordinate sentences; (4) the number of phrasal constituents (i.e., the number of arguments and adjuncts of the predicates); and (5) the probability of a garden-path effect (scored 0 in the “semantics” set, 1 or 2 in the “syntax” set, depending on the structure). A composite score of syntactic complexity was then derived. It varied between 6 and 19 (mean: 12.55) in the present materials, with higher scores reflecting higher levels of syntactic complexity. While syntax of any given sentence could be quantified and projected onto a continuum of complexity using the metric above, sentences of the syntax set were developed to contrast a relatively complex structure with a relatively simple structure of the semantics set. The difference is well reflected in the calculated scores. A Welch two-sample t-test confirmed that the syntax set had a significantly higher level of complexity than the semantics sets in English ($t = 13.42$, $df = 19.74$, $p < .001$) and in French ($t = 13.58$, $df = 21.53$, $p < .001$). We might expect STS-weakening to reflect either a relative difficulty of syntactic computations (captured on a scale) or an increased effort of integrating syntactic subordination into the sentence

meaning (captured in the syntax set, cf. Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002).

Two female speakers (undergraduate students at the University of Kent) were recruited to read the sentences in their native language. They were instructed to match speech rate and pitch patterns across the two experimental conditions (plausible/ implausible) and recorded in a soundproof booth of the Linguistics Lab at the University of Kent. Only the most satisfactory renditions of the sentence pairs were included in the experiment. Both speakers were re-recorded if the requirement of matching speech rate and pitch patterns across the two conditions was not met (the stimuli were prepared and perceptually evaluated by the first author). Timing properties and pitch trajectories of the recordings remained unmodified.

Given that prosody – especially pitch stability and syllable/vowel timing – plays an important role in STS (Falk et al., 2014; Groenveld et al., 2020; Tierney et al., 2018), Table 1 compares relevant prosodic features of the test sentences across all experimental conditions. The F0-measure focused on syllable nuclei only and excluded any abutting consonants, to avoid micro-prosodic perturbations of F0-contours (e.g., Hanson, 2009). To measure local F0, values were taken at the beginning (25%) and the end (75%) of each nucleus and converted into semitones to reflect pitch stability at the level of perceptually most salient, sonorous portions of sentences (Barnes et al., 2012; Rathcke, et al., 2021b). To measure global pitch stability within a sentence, we calculated coefficient of variability across F0-values measured at midpoint of all syllable nuclei using the formula: $CV_{F0} = \frac{SD_{Hz}}{mean_{Hz}}$. Rhythmic aspects of the test sentences included measurements of speech rate, duration of nucleus and syllable intervals. None of the acoustic parameters differed significantly between the experimental conditions (see Table 1), suggesting that prosody was highly unlikely to bias the results of the study in any significant way.

Previous work shows that STS tends to be experienced during the third repetition of a spoken sentence, with the number of reported illusions increasing after shorter rather than after longer inter-sentential pauses (Falk et al.,

2014). Accordingly, the test sentences were looped with eight repetitions in total, and the silent pause between repetitions was set to 400 ms – that is, the shortest pause from the experiments by Falk et al. (2014), yet twice as long as the duration of the intra-sentential break implemented in the syntax set. Examples of the stimuli can be found on OSF (<https://osf.io/8d9mt>).

Procedure

Experimental sessions took place in a sound-attenuated room at the University of Kent (experiments with English listeners) and the Sorbonne Nouvelle Paris-3 University (experiments with French listeners). Prior to the lab visit, the participants were emailed a copy of the consent form, along with web links to the online questionnaire that collected individual background information, with a request to fill in both. A lab session started with a baseline test that asked the participants to listen to the experimental sentences presented once and to judge each of them on a scale from 1 (clearly speech) to 8 (clearly song). The test tapped participants' intuitions about the nature of speech and singing as we did not provide any auditory examples for what sounds may constitute clear speech or clear song. The baseline test was then followed by individual measurements of the attentional resources using the TAP battery (Zimmermann & Fimm, 2004). We first tested divided attention, then flexibility and finally alertness. Following TAP, participants performed the WAIS-III auditory working memory test (Wechsler, 1997). In this test, a series of digits had to be repeated back to the experimenter, first in the same order and then in the reversed order of the original presentation. The presentation of the digits was slow-paced, approximately one digit per second (which is more taxing to the working memory than a fast-paced presentation). The complexity of digit series varied from 2- up to 9-digit lists. For each correctly repeated digit series, participants scored 1. They received 0 if they made mistakes in the digit order or identity. A failure to correctly repeat back digit series within two lists of the same complexity led to

Table 1. Comparisons of prosodic properties of the test sentences.

		Local F0 (st)	Global F0 (CV_{F0})	Speech rate (syllables/second)	Nucleus duration (in ms)	Syllable duration (in ms)
English	Plausible sentences	0.49	0.15	4.2	101.9	235.7
	Implausible sentences	0.39	0.15	4.3	97.6	236.2
	<i>t</i> -tests	$t(187) = 0.41, n.s.$	$t(22) = 0.10, n.s.$	$t(22) = 0.17, n.s.$	$t(201) = 0.52, n.s.$	$t(199) = 0.29, n.s.$
	Semantics set	0.38	0.14	4.4	101.1	227.8
	Syntax set	0.48	0.15	4.2	99.0	241.1
	<i>t</i> -tests	$t(187) = 0.43, n.s.$	$t(22) = 0.24, n.s.$	$t(22) = 0.62, n.s.$	$t(201) = 0.74, n.s.$	$t(199) = 0.03, n.s.$
French	Plausible sentences	0.22	0.11	5.2	82.3	192.8
	Implausible sentences	0.14	0.13	5.3	80.6	191.3
	<i>t</i> -tests	$t(230) = 0.46, n.s.$	$t(22) = 0.95, n.s.$	$t(22) = 0.20, n.s.$	$t(233) = 0.35, n.s.$	$t(234) = 0.67, n.s.$
	Semantics set	0.34	0.10	5.3	78.9	177.4
	Syntax set	0.06	0.13	4.9	83.2	202.8
	<i>t</i> -tests	$t(230) = 1.56, n.s.$	$t(22) = 1.93, n.s.$	$t(22) = 1.91, n.s.$	$t(233) = 0.27, n.s.$	$t(234) = 1.49, n.s.$

the termination of the test. The maximal score of WAIS-III was 30.

The STS test concluded the experimental session. In this test, we instructed the participants to listen to the looped sentences of their native language and indicate when (and only when) they experienced a transformation by pressing the return button. Participants had to wait until the end of the loop without pressing any button if they did not perceive any changes. At the end of each STS trial, we requested the participants to use the same scale as in the baseline test, and to indicate how song-like the sentence sounded to them after the eighth repetition. This procedure combined a previous approach of obtaining speed and frequency of STS (Falk et al., 2014; Rathcke et al., 2021b) with the experimental approaches that focused primarily on the strength of the subjective STS-experience (e.g., Groenveld et al., 2020; Tierney et al., 2018).

Overall, each individual session lasted 45–60 min. We conducted the WM aurally with an in-person experimenter and ran the remaining tests on a laptop computer using DMDX (Forster & Forster, 2003) and good-quality headphones for the auditory presentation of the stimuli.

Statistical Analyses

We conducted all statistical analyses in Rstudio (running R-version 4.1.2). Apart from the basic packages, we used several add-on libraries, including *ordinal* (Christensen, 2019), *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017). In all regression models below, the data were not aggregated (only NAs removed where appropriate). Wilcoxon signed-rank tests were run on the data aggregated by item.

We fitted mixed-effects models to the ordinal data (song-like ratings on the scale from 1 to 8; repetition cycles from 1 to 8 during which participants reported STS) and to the binomial data (whether or not (1/0) participants reported STS during a given loop). For all models, linguistic predictors of interest included *plausibility* (plausible/ implausible) and *syntactic complexity* (implemented either as a categorical factor of *sentence set* (semantics/ syntax) or as a continuous measure of variable complexity). Individual factors of interest included musical training, working memory, divided attention, and attentional flexibility. To control for potential effects of segmental phonology and sentence length (Rathcke et al., 2021b; Rowland et al., 2019), we included mean sonority index and number of syllables per sentence as covariates. All continuous factors (including syntactic complexity, musical training, working memory, divided attention, attentional flexibility, and sentence-specific covariates) were mean-centered. All models further included two crossed random intercepts: *listener* and *sentence*. Random slopes were only retained if the models converged. Model comparisons using the likelihood ratio test helped to determine the best model fit. To select the final model, we implemented a stepwise backward-fitting procedure. First, we defined all predicted main

effects and their interactions. Next, we reduced the model complexity by successively removing predictors that did not improve the model fit. Moreover, we checked for potential multi-collinearity issues among the predictors, by calculating the Variance Inflation Factor (VIF) of the best-fit models and checking it against established threshold values (Stine, 1995). Only converging best-fit models (whose VIFs were below the critical threshold of 5 for all significant predictors) are reported below.

Finally, we calculated Kendall's correlation coefficients τ_b to measure the strength and the direction of association between the three aspects of STS-experience (likelihood, speed, and strength of the transformation). For this test, the data were aggregated by item (using median ratings of song-likeness, median repetition cycle, and proportion of STS reported by all participants).

Results

Song-Like Ratings of the Test Sentences Before and After Repetition

Figure 2 displays “song-like” ratings of English and French sentences at baseline vs. post repetition. In both languages, baseline responses occupied the lower end of the Likert scale (median English: 2.5; median French: 3.2, with the highest median rating being 4 in both English and French) while responses to the same sentences after the exposure to eight repetitions clustered around the upper end of the Likert scale (median English: 5.0; median French: 5.3, with the highest median rating being 6 in both English and French). The slight numerical difference between the two languages was significant neither at baseline (Wilcoxon signed-rank, $W = 230.5$, n.s.) nor post repetition ($W = 201$, n.s.). In contrast, Wilcoxon signed-rank test confirmed that the sentences were rated significantly more song-like post repetition in both English ($V = 0$, $p < .001$) and French ($V = 0$, $p < .001$).

Likelihood of STS

Overall, the rate of reported transformations was relatively low in the present dataset, amounting to approximately 40%. We fitted a logistic mixed model (estimated using maximal likelihood and the BOBYQA optimizer) to predict the likelihood of experiencing STS by *sentence set* (semantics/ syntax), *plausibility* (plausible/ implausible), and their interaction; *syntactic complexity*, individual *musical training*, *working memory* and *attention* measures, and two-way interactions of the individual measures with *syntactic complexity*, *plausibility*, and *sentence set*. Number of syllables was fit as a covariate, to control for potential effect of sentence length (Rathcke et al., 2021b; Rowland et al., 2019). We also checked if listeners of the two languages differed in their perception of STS in these materials. Of these predictors, the best-fit model retained main effects of *plausibility* ($\chi^2(1) = 4.65$, $p < .05$) and

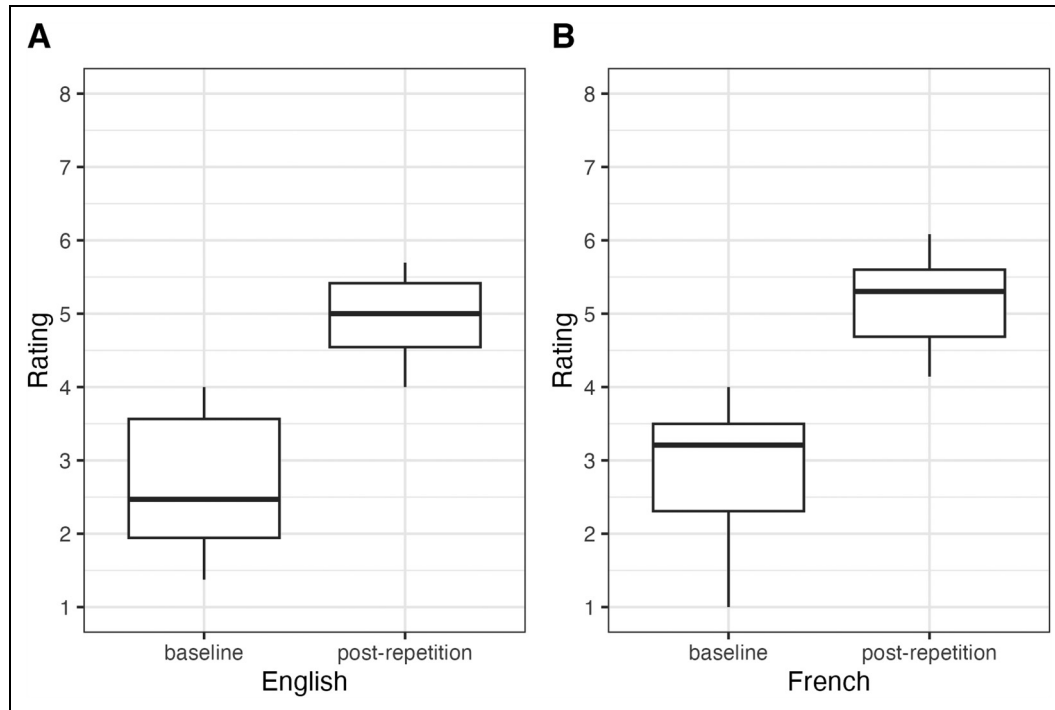


Figure 2. Perceptual ratings of the test sentences on the given Likert scale (1 = *clearly speech*, 8 = *clearly song*) after a single exposure (baseline) vs. after eight repetitions. Responses of the English participants are plotted in panel A, responses of the French participants in panel B.

flexibility ($\chi^2(1)=4.99$, $p<.05$), and two-way interactions of *syntactic complexity* with *working memory* ($\chi^2(1)=5.46$, $p<.05$) and with *musical training* ($\chi^2(1)=11.86$, $p<.001$). The two sentence sets did not differ in this regard.

Given the two-way interactions with the individual effects, we checked if musical training moderates the working memory finding but failed to find either a significant two-way interaction of *musical training* and individual *working memory* ($\chi^2(1)=.01$, n.s.) or a three-way interaction ($\chi^2(1)=.34$, n.s.). Moreover, there was no difference between English and French listeners in this task ($\chi^2(1)=.67$, n.s.). Among the control predictors, only *sentence length* ($\chi^2(1)=5.06$, $p<.05$) but not *sonority* ($\chi^2(1)=1.66$, n.s.) significantly improved model fit. Similarly, participants' *alertness* did not significantly influence their experience of STS (either alone: $\chi^2(1)=2.36$, n.s., or in interaction with syntactic complexity: $\chi^2(1)=.42$, n.s.).

Figure 3 displays the main effects of interest. Accordingly, all implausible sentences introduced STS more often than the plausible sentences of the present stimulus set ($\beta=.43$, $SE=.19$, $z=2.22$, $p<.05$, see Figure 3-A). Higher positive scores of attentional flexibility (measured as speed-accuracy trade-off, Zimmermann & Fimm, 2004) lowered the likelihood of individual STS-experience ($\beta=-.42$, $SE=.19$, $z=-2.05$, $p<.05$, see Figure 3-B). That is, participants with a tendency of being more accurate rather than being fast in the flexibility task showed lower STS in contrast to those who tended to respond faster at a cost of accuracy. Figure 4 shows the

interactions from the best-fit model. Both interactions involved syntactic complexity of the looped sentences and individual listener characteristics. Accordingly, syntactically more complex sentences led to a higher likelihood of STS-experience in listeners with lower working memory resources ($\beta=-.20$, $SE=.09$, $z=-2.37$, $p<.05$, see Figure 4-A). The promoting effect of musical training on STS (Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015) was observed exclusively in syntactically simpler, but not in syntactically complex, sentences used in the present study. That is, listeners with lower levels of musical training had lower likelihood of experiencing STS in syntactically simple sentences in contrast to listeners with a high level of musical training, who showed the opposite effect ($\beta=-.28$, $SE=.08$, $z=-3.52$, $p<.001$, see Figure 4-B). No difference between the listeners was observed in syntactically complex sentences.

Speed of STS

To examine how quickly participants experienced STS during repetitions, we fitted ordinal mixed-effects regressions to a subset of the data containing only those stimuli that induced STS and the accompanying information about the repetition cycle (1–8) during which participants reported to have experienced the transformation. The predictors of interest, again, included *sentence set* (semantics/syntax), *plausibility* (plausible/implausible), and their interaction; *syntactic complexity*, *musical training*,

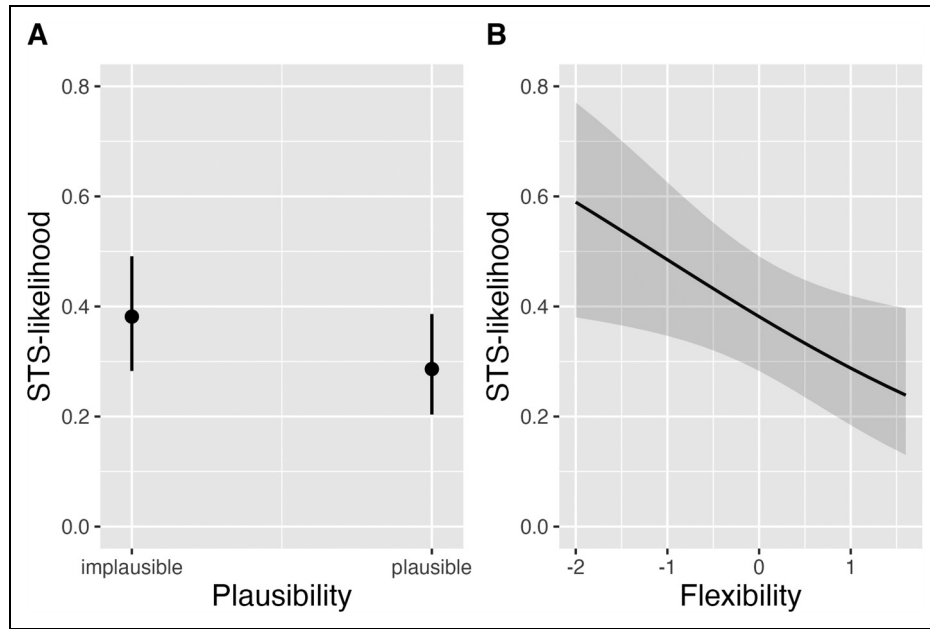


Figure 3. Model estimate plots obtained from the best-fit model of STS-likelihood, displaying the main effect of sentence plausibility (plausible vs. implausible sentences, panel A) and attentional flexibility (panel B, positive values of flexibility indicate that participants prioritized accuracy over speed of responses while negative values indicate a faster response at the cost of accuracy).

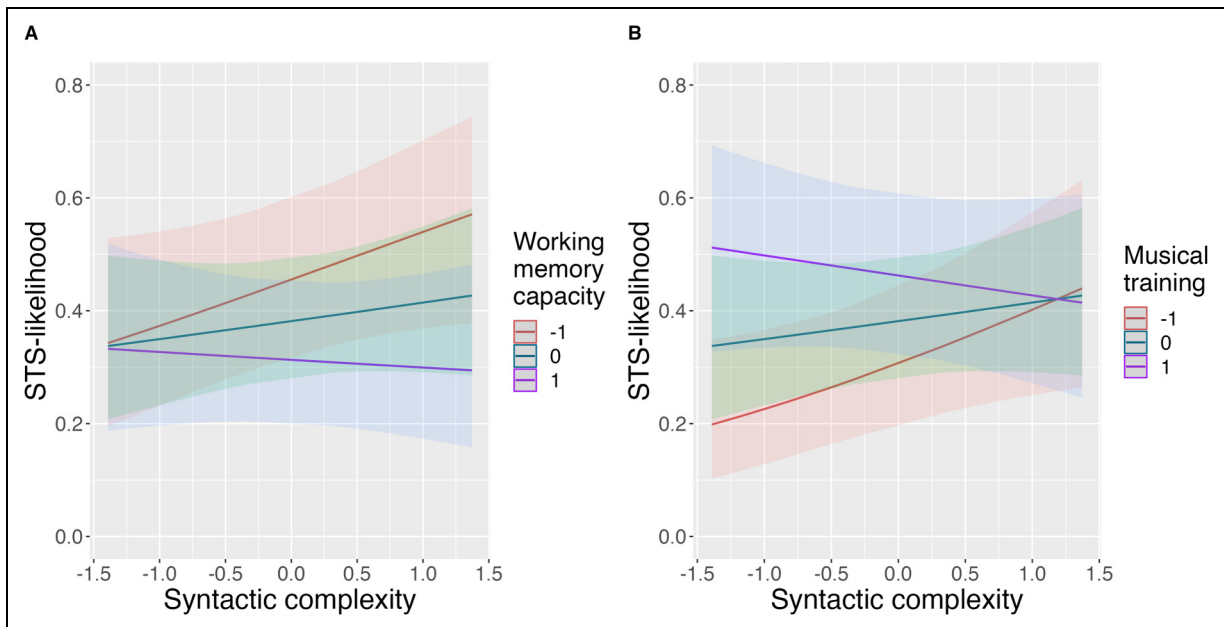


Figure 4. Model estimate plots obtained from the best-fit model of STS-likelihood, displaying the interactions of syntactic complexity with individual working memory capacity (panel A) and musical training (panel B). All values were scaled and centered around the mean ($= 0$). For the purpose of the visualization, listeners were grouped into high-scoring (1 standard deviation above the mean), average-scoring (i.e., close to the mean of the sample), and low-scoring (1 standard deviation below the mean).

working memory, attention, and two-way interactions of the individual measures with syntactic complexity, plausibility, and sentence set. The final best-fit model contained only one predictor, sentence set ($\chi^2(1)=7.32$, $p<.01$). The effect is shown in Figure 5. Accordingly, sentences from the semantics set transformed slightly later than sentences

from the syntax set that transformed earlier ($\beta = .41$, $SE = .15$, $z = 2.74$, $p < .01$). In contrast, syntactic complexity measured on a scale did not show a significant influence on the speed of STS ($\chi^2(1) = 1.77$, n.s.). There was no difference between French and English listeners in this regard ($\chi^2(1) = .56$, n.s.), nor did we find an impact of sentence

plausibility ($\chi^2(1) = .28$, n.s.). Similarly, none of the individual traits influenced STS-speed in significant ways.

This finding suggests that the ease of the STS-transformation does not reflect the relative ease of syntactic computations. The fact that the presence of syntactic subordination promotes STS rather than delaying it is at odds with the predictions of the linguistic meaning decay hypothesis that proposes a faster switch to the song percept in the absence of syntactic complexity (i.e., when the meaning of a sentence is easier to compute and is thus faster to decay).

Strength of STS

The last set of analyses examined the strength of the illusory experience. That is, how strongly song-like a given sentence sounded to the participants after the last repetition specifically in those trials that indeed showed the transformation. The ratings obtained on the Likert scale from 1 (clearly speech) to 8 (clearly song) were fitted to the predictors *sentence set* (semantics/ syntax), *plausibility* (plausible/ implausible), and their interaction; *syntactic complexity*, *musical training*, *working memory*, *attention*, and two-way interactions of the individual measures with *syntactic complexity*, *plausibility*, and *sentence set*. The best-fit model contained one predictor of interest, *plausibility* ($\chi^2(1) = 4.10$, $p < .05$), and the control variable *sentence length* ($\chi^2(1) = 6.27$, $p < .05$). The native language of listeners did not play a role in the strength of the effect ($\chi^2(1) =$

.11, n.s.). No individual listener characteristics or their interactions helped to significantly improve model fit.

The main effect of interest is plotted in Figure 6, which indicates that all implausible sentences (i.e., from both semantics and syntax sets) sounded significantly more song-like upon repetition than their plausible counterparts ($\beta = .27$, $SE = .13$, $z = 2.08$, $p < .05$). The effect of the control variable (not shown) was in full alignment with previous findings (Rathcke et al., 2021b; Rowland et al., 2019), with shorter sentences sounding significantly more song-like to listeners upon exposure to repetitions ($\beta = .07$, $SE = .03$, $z = 2.61$, $p < .05$).

Associations Between Likelihood, Speed, and Strength of STS

Correlation analyses showed that the three aspects of STS-experience were associated, at least to some extent. The strongest association was observed between the speed and the strength of the transformation ($\tau_b = -0.62$, $z = -5.17$, $p < .001$). The correlation indicated that those stimuli that were perceived more song-like upon repetitions also tended to transform sooner during the repetition cycles. The strength and the likelihood of the transformation were also correlated ($\tau_b = 0.30$, $z = 2.63$, $p < .01$), suggesting that higher transformation likelihood of a stimulus tended to be accompanied by a higher song-like rating upon repetition, though the correlation was relatively weak. In contrast, the speed and the likelihood of STS did not show a

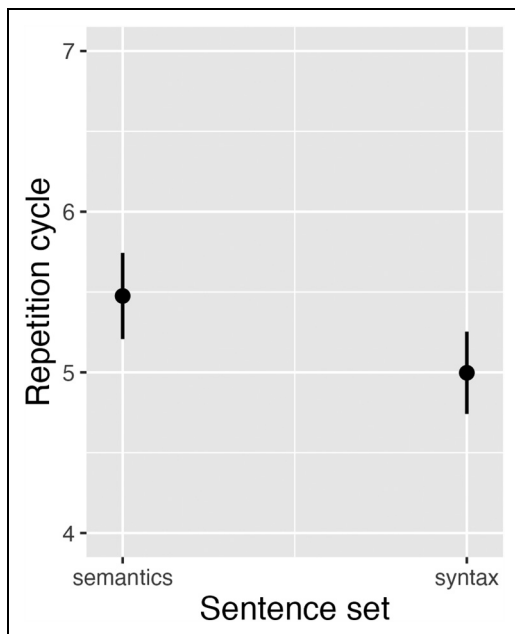


Figure 5. Model estimate plot obtained from the best-fit model of STS-speed, displaying the main effect of the sentence set. The speed of STS-experience during repetitions is plotted along the y-axis (the total number of repetition cycles was 8).

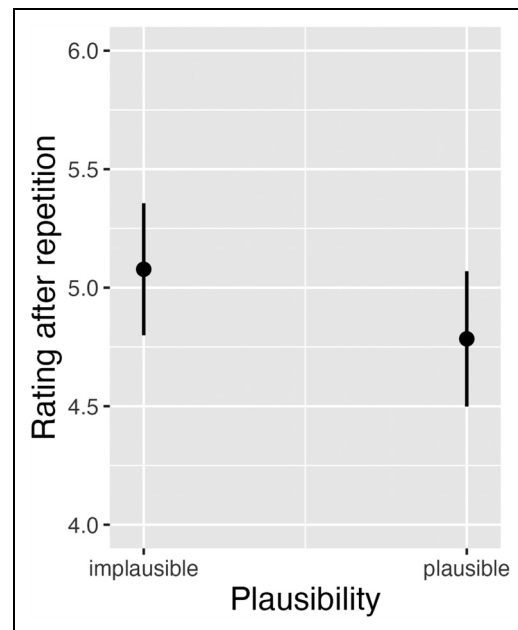


Figure 6. Model estimate plot obtained from the best-fit model of STS-strength, displaying the main effect of sentence plausibility. Responses after repetition were collected on an 8-point Likert scale from 1 (clearly speech) to 8 (clearly song).

statistically significant association following the Bonferroni correction ($\tau_b = -0.25$, $z = -2.27$, n.s.).

Discussion

The aim of the present study was to clarify several unresolved issues surrounding the cognitive underpinnings of the speech-to-song illusion (Deutsch, 1995). Even though the (dis)engagement of language processes that compute linguistic meaning has been frequently assumed to be the key prerequisite to an individual experience of STS (e.g., Castro et al., 2018; Falk et al., 2014; Jaisin et al., 2016; Margulis, 2013; Rathcke et al., 2021b; Tierney et al., 2021), previous studies did not provide sufficient evidence in support of this foundational assumption. The experiments presented here were conducted to fill this gap, by testing two specific aspects of STS: (1) the lexico-syntactic properties of the *sentences* that are repeated, and (2) the cognitive traits of the *listeners* who are exposed to sentence repetitions. The results obtained with French and English participants of the study did not differ, enforcing the generalizability of the findings discussed below.

The Contribution of Lexico-Syntactic Properties to STS

Guided by the idea that the transition from speech to song in STS is moderated by listeners' engagement with the linguistic meaning of the repeated sentence (e.g., Castro et al., 2018; Falk et al., 2014; Jaisin et al., 2016; Margulis, 2013; Rathcke et al., 2021b; Tierney et al., 2021), we tested two aspects of sentence-level meaning. In contrast to word-level meaning that is part of the mental lexicon storing the general knowledge about language (Aitchison, 2012), sentential meaning is computed on the fly (e.g., Culicover & Jackendoff, 2006). The sentence-level meaning derives not only from the meaning of individual words but also from the semantic roles assigned to the words because they occur in specific syntactic positions and reflect relationships between the constituents of a hierarchical structure (Chomsky, 1957; Hagoort, 2003). For example, the word "cat" can be the agent of a semantically plausible sentence "The cat chased a mouse in the house." or a semantically implausible sentence "The cat shot the mouse with a gun." The results of the present study demonstrate that sentences of the latter type that violate lexico-syntactic expectations are more conducive to STS.

Even though semantic implausibility tends to slow down linguistic processing in online comprehension tasks (e.g., Ferreira, 2003), it did not show the predicted effect of reducing the STS-experience. This result is possibly due to temporal differences between an experiment measuring processing cost of a lexico-semantic violation vs. the present study. The relatively short processing delay previously reported for implausible sentences in the relevant

literature (in the order of 300 ms, Ferreira, 2003; Patson & Warren, 2010) might have been obscured during sentence repetitions that are required for STS, interspersed with 400-ms long pauses. Similarly, implausible prosodic chunking of syntactically complex sentences (e.g., "While the woman washed the cat, meowed." vs. "While the woman washed, the cat meowed.") did not lead to a diminished STS-experience of the study participants as we had expected. Instead, implausible sentences of both the syntax and the semantics set transformed more often and sounded more song-like upon repetition, suggesting that online comprehension delays due to lexico-syntactic anomalies may be quickly resolved during repetitions. Moreover, the experimental task of the present study was not linguistic in nature. Listeners knew that their understanding of the sentences (and the depth of their lexico-syntactic processing) would not be examined at any point during the experiment. The task peculiarity (as compared to previous work on the processing of implausible sentences, Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002) may have encouraged listeners to forego in-depth computations of linguistic meaning from the lexico-syntactic structure (cf. Rathcke et al., 2021b).

Alternatively, listeners may have quickly assessed the meaning of sentences with lexico-syntactic violations as implausible, with meaning implausibility being the key driver of the effect. While lexico-syntactic implausibility may be both rare (Grice, 1989) and cognitively costly (Ferreira, 2003; Patson & Warren, 2010) in everyday language, it is far from uncommon in music. For example, text setting in lyrics obeys its own principles and often deviates from linguistic texts created for other purposes, especially with regards to prosodic segmentation and phrasing (Gordon et al., 2011; Liebling, 1908). Accordingly, prosodic breaks in lyrics do not always align with syntactic junctures between neighboring constituents. Rather, they serve the rhythmic structure that carries an expressive artistic effect and an emotional meaning. As a consequence, lyrics may frequently show enjambment, or incomplete syntax at the end of a line (Heller, 1977; van't Jagt et al., 2014). Principles and demands of the internal structure of lyrics can override many linguistic constraints, including the placement of lexical stress (Janda & Morgan, 1987) and even the properties of lexical tone (Schellenberg, 2009).

Interestingly, both meaning and structure also influence verbal transformations that arise upon repetitions of single words (Warren, 1961; Warren & Gregory, 1958). Meaningful words tend to transform slower than meaningless nonce-words or structurally illicit pseudo-words, with the overall number of unique forms being higher in nonce- and pseudo-words than in real words (Natsoulas, 1965; Shoaf & Pitt, 2002). Such parallels between the two perceptual repetition effects highlight a shared origin of the two illusions, as previously identified in the connectionist account of STS (Castro et al., 2018). A certain level of (syntactic) implausibility can be frequently observed in

STS-stimuli used in previous research. For example, the original phrase by Deutsch (1995) that introduced STS as a perceptual phenomenon – “sometimes behave so strangely” – is syntactically incomplete as it misses the subject (“they”) and cannot be fully parsed (e.g., Hagoort, 2003; Van Gompel & Pickering, 2007). Similarly, the database of high-transforming STS-stimuli created by Tierney et al. (2013, 2018, 2021) contain syntactic omissions of one or two constituents that also prohibit a complete syntactic parse (e.g., “here is no less”, “gave the houses”, “somehow I can get”), possibly enhancing the STS-effect.

Moreover, violations of the semantic plausibility constraints examined in the present study (Ferreira, 2003; Ferreira et al., 2002) are commonplace in (song) lyrics (Pattison, 1991). Interestingly, an example of such semantically implausible phrase of the famous song lyrics “Excuse me while I kiss the sky” (from “Purple Haze” by Jimmy Hendrix, 1967) is also an example of the most frequently misheard lines in song lyrics, being predominantly perceived as “Excuse me while I kiss this guy” and giving rise to the title of an archive of misheard song lyrics (<https://www.kissthisguy.com/>, Kentner, 2015). While displaying a segmentation error typical of mondegreens (Kentner, 2015), this misperception further demonstrates a perceptual repair of a semantically implausible phrase (cf. Beck et al., 2014) and is in line with many other illusory phenomena involving speech (Warren, 1970; Warren & Obusek, 1971; Warren & Sherman, 1974).

Adult listeners are likely to have gathered a lifetime’s worth of experience with lexico-syntactic discrepancies between speech and song that stem from the different structural and functional demands placed on spoken vocalizations by the linguistic vs. the poetic system. Casual music listeners with no formal training can implicitly acquire a wealth of knowledge about different aspects of music prevalent in their culture (Bigand & Poulin-Charronnat, 2006). Such everyday music exposure and implicit knowledge of musical structure have indeed been argued to shape STS in musically untrained listeners (Falk et al., 2014; Vanden Bosch der Nederlanden et al., 2015). A growing body of research (in both language and music) provides compelling evidence for the incredible ability of the human mind to extract underlying regularities from auditory signals without directed attention and to acquire substantial knowledge about such regularities without an explicit instruction (e.g., Bigand & Poulin-Charronnat, 2006; Oh et al., 2020; Rohrmeier & Rebuschat, 2012). The present findings further corroborate this idea and extend existing evidence to include implicit knowledge of lexico-syntactic structure in speech vs. songs.

Previous experience and knowledge are at heart of perceptual phenomena that heavily involve cognitive processes, and in which prior knowledge frames the experience of an ambiguous sensory input by computing a percept consistent with the priors (cf. Gregory, 1997, 2009; McIntosh, 2022; Stocker & Simoncelli, 2006). Many well-described illusions involving speech – for

example, phoneme restoration (Warren, 1970; Warren & Gregory, 1958) or the McGurk effect (McGurk & MacDonald, 1976) – testify to speech perception being a constructive process and involving top-down biases. It is based on an active, adaptive engagement with the sensory input and as such, is shaped by listeners’ priors that arise from previous experience and knowledge (Davis & Johnsruide, 2007). Given the often fragmentary and noisy character of speech signals, top-down inferences enable listeners to efficiently deal with an impoverished signal (Pressnitzer et al., 2018). As argued in previous research on verbal transformations (Kaminska et al., 2000; Warren, 1983), massed, verbatim repetitions of speech create a situation of ambiguity for listeners, given the absence of any contextual information and an abnormal stream of invariant acoustic input. The invariance of speech acoustics during sentence repetitions is a prerequisite for STS (Deutsch et al., 2011; Vanden Bosch der Nederlanden et al., 2015), yet natural speech lacks such invariance as the same word spoken twice is never acoustically identical (cf. Perkell & Klatt, 2014). In contrast to speech, natural music has a more readily observable tendency toward repetition (e.g., of pitch and timing patterns), with present-day listeners being highly accustomed to repetition due to the ubiquity of recorded and synthesized pieces (Margulis, 2013). In addition, music of Western (and, to some extent, other) cultures maintains precise definitions of notes, intervals, and the corresponding pitch frequencies that lay foundations to melodies (Krumhansl, 2001). Such prior experience with musical signals and their difference from spoken language may help listeners frame the ambiguity of massed, acoustically invariant repetitions that are required to elicit STS and thus bias the perceptual interpretation of such linguistically ambiguous input towards singing.

The role of previous experience and other top-down processes that shape perception when sensory input is ambiguous, degraded, or distorted in some way (Gregory, 1997, 2009; McIntosh, 2022; Stocker & Simoncelli, 2006) might be an important – and so far, overlooked – aspect giving rise to STS. In previous discussions of STS, the linguistic ambiguity of an acoustically invariant stream of sentence repetitions has not been identified as a potential source of the illusory effect. In contrast, an existing account of verbal transformations recognized this mechanism early on (Kaminska et al., 2000; Warren, 1983). We therefore suggest that STS, like verbal transformations, is best understood as a cognitive effect, or an auditory illusion that relies on cognitive priors. This account of STS predicts that experiential and individual priors will play a crucial role in the emergence of STS, similar to other ambiguous encounters of spoken language. It is in full alignment with existing evidence on the role of music- vs. speech-related acoustic priors in STS (Deutsch et al., 2011; Falk et al., 2014; Groenveld et al., 2020; Tierney et al., 2018). The present results further reinforce this account by showing that lexico-syntactic features of a spoken phrase can also serve as such priors and introduce a perceptual

bias toward speech or song, depending on lexico-syntactic structures typically encountered in the expressive systems of language vs. music (cf. Fedorenko et al., 2020). The results of the present study thus extend previous findings on the signal-specific priors that can influence STS, by adding an indicator of the structure-specific priors (see also Rathcke et al., 2021b). Overall, the results of the present study provide compelling evidence that linguistic structures that convey meaning through lexical or syntactic choices play an important role in STS. They can influence not only the likelihood, but also the speed and the strength of STS. Implausible meanings that violate linguistic expectations (Grice, 1989) have been found to facilitate STS, with a specially notable effect of implausibility on the vividness of the song-like experience.

The Role of Listener-Specific Characteristics in STS

Viewing STS as an auditory illusion under the influence of prior knowledge paves the way to new hypotheses. Specifically, the cognitive account of STS predicts that listener experience and traits will affect the individually experienced transformation. Only a few individual effects have been studied and documented to date. Tierney et al. (2021) provide evidence that an increased musical aptitude (e.g., a greater skill in beat and tonality perception, a better selective attention to pitch) may lead listeners to experience STS more vividly. A heightened ability to detect musical attributes present in spoken phrases seems to bias the perception toward song upon repetition (Deutsch et al., 2011), though it is unlikely to constitute the only prerequisite of STS given that listeners with only casual music exposure can also experience the effect (Vanden Bosch der Nederlanden et al., 2015). Another possibility, not incompatible, is that variability in general cognitive capacities that are not music-specific may interact with the emergence of STS. The present study tested the relation between individual STS-experience and variability in selected cognitive functions that were hypothesized to play a role in STS (cf. Falk et al., 2014; Rathcke et al., 2021b; Tierney et al., 2021). As predicted, cognitive traits of individual listeners contributed to their experience of STS. The present experiments showed that both attentional flexibility and working memory (but not divided attention) influence the individual tendency to experience STS when exposed to massed repetitions of spoken phrases. While attentional flexibility had a general impact on STS, the effect of working memory capacity was specific to sentences with high syntactic complexity.

Attentional flexibility measures the ability to willfully alternate the focus of attention between different sources of incoming information (e.g., Calcott & Berkman, 2014; Zimmermann & Fimm, 2004). In the present experimental task, participants who traded speed for accuracy were more likely to experience STS. In other words, lower attentional flexibility (reflected in fast but incorrect responses) was associated with an increased likelihood of an individual

STS-experience and was otherwise independent of any linguistic properties manipulated in the experimental stimuli. This finding suggests that the involvement of attention during STS is unrelated to the integration of semantic and syntactic cues necessary for language processing as we had originally proposed (cf. Myachykov & Posner, 2005; Tomlin, 1999). Rather, the present effect of attentional flexibility might be indicative of an individually variable ability to focus exclusively on what the listener expects to be the primarily relevant cue for the task at hand and to ignore other information concomitantly present in the sensory input (e.g., Guinote, 2007; Posner & Petersen, 1990; Rothbart & Posner, 2001). Under this explanation, listeners with an attentional tendency to privilege accuracy at the expense of speed (i.e., slow-speed, high-accuracy listeners) would maintain the veridical linguistic percept by attending to all aspects of the incoming linguistic input while listeners with lower attentional flexibility (i.e., high-speed, low-accuracy listeners) might be more likely to shift their attention from lexico-syntactic properties of a phrase to its prosodic shape and would thus be more prone to a musical reinterpretation of the phrase (Falk et al., 2014; Tierney et al., 2021). More work is required to further corroborate the present finding. If the current explanation is correct, similar results would be expected for individuals scoring high vs. low in focused attention, inhibitory control, and attentional flexibility tasks (e.g., Posner & Petersen, 1990; Rothbart & Posner, 2001; Tiego et al., 2018; Yantis & Johnston, 1990; Zimmermann & Fimm, 2004).

Working memory also influenced STS, but only in interaction with syntactic complexity of looped sentences. Specifically, individual capacity limitations seem to only play a role in syntactically complex but not syntactically simple sentences. Listeners with better working memory performance were unaffected by the variable levels of syntactic complexity in the materials as their STS-rates remained largely constant across all sentences they listened to. In contrast, listeners with lower working memory performance tended to report more STS in syntactically more complex sentences. The result is at odds with the frequently discussed hypothesis that the switch from speech to song is mediated by linguistic meaning decay or satiation of a semantic processing unit (e.g., Castro et al., 2018; Falk et al., 2014; Margulis, 2013; Rathcke et al., 2021b; Tierney et al., 2021). Being easy to parse syntactically and to process semantically, sentences with a simpler structure do not require increased processing resources (Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002). We would therefore expect them to be more readily reinterpreted as musical upon repetition by all listeners regardless of their working memory capacity (which was not the case).

There is some evidence to indicate that a reduced working memory capacity impinges upon syntactic parsing (King & Just, 1991), possibly making listeners more attentive to prosody. As Speer et al. (1996) suggest,

speech prosody provides an initial structure in working memory where utterances are maintained until linguistic analysis has taken place and the comprehension process is complete. If comprehension becomes more vulnerable and prone to error under capacity limitations (cf. Myachykov & Posner, 2005; Tomlin, 1999), low-capacity listeners may rely more heavily on the initial stages of linguistic analysis and therefore on prosody (cf. King & Just, 1991; Kjølgaard & Speer, 1999; Speer et al., 1996). In general, sentence-level prosody is a key factor in parsing syntactic constituency in many languages of the world (Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996). Sensitivity to prosodic detail is among the main developmental predictors of the syntactic mastery in both native and non-native language learners (Goad et al., 2003; Hawthorne & Gerken, 2014; Morgan & Demuth, 2014; Tremblay et al., 2016). Moreover, foreign language learners are known to frequently forego syntactic analyses and attend primarily to prosody when processing speech in their non-native language(s) (Harley, 2000; Harley et al., 1995), which is also in line with the finding of an enhanced STS-effect in listeners' non-native languages (Margulis et al., 2015; Rathcke et al., 2021b). In the present experiment, only STS-likelihood (i.e., the likelihood to report an illusion) shows an effect of individual working memory capacity, in interaction with syntactic complexity. This result suggests that, given the requirements of the STS-task, listeners with a limited working memory capacity do not fully engage in the syntactic analysis of increasingly complex sentences, rather than starting and abandoning the computations prematurely which would have influenced STS-speed instead of STS-likelihood.

In sum, both findings of the present study confirm that STS is moderated by capacity-limited cognitive resources of attention and working memory (Falk et al., 2014; Rathcke et al., 2021b; Tierney et al., 2021). However, the hypothesized mechanisms supporting the involvement of basic cognition in STS have to be revised. The present results cast doubt on the idea that a reduction (or the termination) of lexico-syntactic computations during language processing simply frees up cognitive resources that then become available for the extraction of melodic and rhythmic characteristics of speech and their reassessment in terms of a musical structure (Tierney et al., 2021). If this were the case, high-capacity listeners and syntactically simple sentences would show enhanced STS. Instead, the illusion prevails in low-capacity listeners and syntactically complex sentences that are more difficult to process (Christianson et al., 2010; Ferreira, 2003; Ferreira et al., 2002). As an alternative to be more extensively addressed in future research, we propose that the availability of cognitive resources for language processing influences either listeners' reliance on prosody (King & Just, 1991; Kjølgaard & Speer, 1999; Speer et al., 1996) or their attentional shifts to prosody (cf. Guinote, 2007), which in turn moderates the individual likelihood of experiencing STS.

The present study adds to the steadily growing body of evidence that STS crucially hinges on speech prosody in

languages that do not use lexical tone to encode word meanings (Jaisin et al., 2016). Both the acoustic-prosodic shape of the phrase repeated (Deutsch et al., 2011; Falk et al., 2014; Rathcke et al., 2021b; Tierney et al., 2018) and the individually variable ability to attend to the prosodic properties of the phrase (Falk et al., 2014; Tierney et al., 2021) seem to unleash the perceptual effect. As far as listener-specific characteristics are concerned, the cognitive priors of STS discussed here and the musical priors discussed in previous research (Falk et al., 2014; Tierney et al., 2021) are orthogonal in that higher importance of prosody during linguistic analyses and a stronger bias toward interpreting prosodic relations as musical can be seen as stemming from different aspects of individual experience and arising at different stages of the hypothesized switch from a language percept to a music percept (e.g., Castro et al., 2018; Falk et al., 2014; Margulis, 2013; Rathcke et al., 2021b; Tierney et al., 2021). A potential interplay of these traits and their impact on the individual experience of STS is an empirical question worth pursuing in future studies of the illusion.

As a first step toward understanding such interplay, the results of the present study provide a more nuanced view on the role of listener musicality (as reflected by the amount of musical training) in STS. Accordingly, individually variable levels of musical training bring about different experiences of STS exclusively in sentences of reduced syntactic complexity. The STS-promoting effect of listener musicality disappears when linguistic structure gains in complexity. In a way, the results can also be interpreted as being indicative of a stable STS-effect in musically trained participants in contrast to a more variable STS-effect in musically less versed participants. As mentioned above, previous research often relied on syntactically simple, short phrases (e.g., “snags and sandbars”, “people in the neighborhood”, Tierney et al., 2013; 2018; 2021) which have likely foregrounded the effect of listener musicality at the expense of the contribution of linguistic structure to STS, obscuring the interaction of the linguistic and musical perception modes. This interaction is, however, key to the understanding of the perceptual foundations of STS. As the results of the present study suggest, the interaction may only be uncovered if sentence-level linguistic structure forms an important part of the theoretical account of STS, but is not amenable to empirical observation with word-level structures (cf. Castro et al., 2018).

Multiple Aspects of an Illusory Experience

The present study highlights that there are three different aspects to an individual experience of a cognitive effect such as STS: (1) whether or not a listener experiences the transformation, or the *likelihood* of an illusory effect; (2) the ease with which the experience arises in their perception, or the *speed* of an illusory occurrence; and (3) how vividly they experience it, or the *strength* of an illusory experience. Most of the present findings derive from

(some) listeners' (in)ability to experience STS in certain sentences. That is, these results are concerned with the STS-likelihood. Such findings can be easily overlooked in experimental approaches that divide linguistic materials into transforming vs. non-transforming *a-priori* (e.g., Tierney et al., 2013; 2018; 2021). Moreover, the transformation likelihood seems to be rather weakly – if at all – correlated with either strength or vividness of STS, signifying the central importance of this aspect of STS.

Previous research into STS has rarely distinguished between the three aspects of an illusory experience. It is not unlikely that an apparent disagreement regarding the role of listeners' musical background in STS stems from the fact that some studies focused on the likelihood of STS, not finding an effect of musical training (Rathcke et al., 2021b) while others examined the vividness of STS, providing evidence in favor of musical skill (Falk et al., 2014; Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015). Like many other auditory illusions involving speech, STS can be considered a cognitive effect that arises while an ambiguous sensory stimulation (cf. Warren, 1983) is supplemented with prior knowledge in order to arrive at an unambiguous perceptual interpretation (cf. McIntosh, 2022). In case of musically experienced or apt listeners, such perceptual priors might include rich tonal representations that bias their perceptual responses towards a more vivid impression of singing than the one observed in musically less experienced or apt listeners (Deutsch et al., 2011; Falk et al., 2014; Tierney et al., 2021). It thus seems plausible that specifically the ease and the strength of a song percept is shaped by an individually variable musicality, rather than their likelihood of experiencing STS *per se*.

Future studies of STS will benefit from a theoretically informed approach to testing those factors that may have a specific influence on each aspect of the illusion. Given that illusions such as STS are shaped by prior knowledge (McIntosh, 2022; Stocker & Simoncelli, 2006), it appears particularly informative to ask the question what cognitive traits determine an individual auditory experience – its occurrence, vividness, and ease. As seen in the present study, different factors may exert an independent influence on each aspect of STS, highlighting an intricate complexity of illusory phenomena involving cognition.

Conclusions

The speech-to-song illusion (STS, Deutsch, 1995; Deutsch et al., 2011) describes a striking perceptual experience of spoken phrases transforming into singing upon repetition and signifies one of the most complex auditory illusions connecting two phenomena of human cognition: language and music. The present results place STS among perceptual illusions that result from a process of integrating sensory evidence with individual experience and context-based expectations (McIntosh, 2022). Accordingly, massed repetitions of speech necessary for STS may create a context

that frequently occurs in music (Margulis, 2013; Rowland et al., 2019; Vanden Bosch der Nederlanden et al., 2015) but is unlikely and ambiguous in language (Warren, 1983), thus setting a context-based bias toward song perception. An evidence-based bias toward song is added if auditory input contains acoustic or structural features known from song and music, like lexico-syntactic implausibility or stable melodies (Falk et al., 2014; Tierney et al., 2018), regular intervocalic intervals (Falk et al., 2014), and prolonged periods of vocal sonority (Rathcke et al., 2021b). An individual bias ignites the transformation. The latter bias can arise from individual reliance on speech prosody for complex linguistic analyses or attentional shift to prosody due to capacity limitations as argued in the present study (Goad et al., 2003; Hawthorne & Gerken, 2014; King & Just, 1991; Morgan & Demuth, 2014; Speer et al., 1996; Tremblay et al., 2016) or from an individual musicality as argued in previous work (Falk et al., 2014; Tierney et al., 2021). Overall, we propose that the linguistic meaning decay is unlikely to be the key mechanism driving the perceptual transformation from speech to song. Rather, it arises when experience-based and cognitive priors combine and set a strong bias for the interpretation of a contextually ambiguous phrase as musical. Each prior may influence one specific aspect of STS, be it the likelihood, the ease, or the vividness of the effect, though more work is needed to further advance the understanding of the individual aspect of STS. At this point, it is also unclear if the effect of the priors may be additive or non-linear, and to what extent. Multiple interactions of linguistic structure and individual listeners traits are, however, suggestive of a non-linear complexity among the involved priors.

In conclusion, STS resembles other illusory phenomena involving cognitive processes in that it testifies to auditory perception being an active, malleable process of an individual mind's engagement with sensory input, whose goal is to establish the most probable percept in noisy, ambiguous environments (Davis & Johnsrude, 2007; Pressnitzer et al., 2018; Stocker & Simoncelli, 2006). As such, the process can be strongly influenced by previous experience and is likely to be subject to maturational constraints (Warren & Warren, 1966). While adult listeners of all ages have been reported to perceive the transformation from speech to song upon repetition (Mullin et al., 2021), it is unclear when children start being able to experience STS. Even though the music network seems to develop rapidly within the first few months of life (Dehaene-Lambertz et al., 2010), the ability to categorize songs as such matures with age and experience (Vanden Bosch der Nederlanden et al., 2023). The present account of STS predicts that child listeners' susceptibility to STS would evolve along with their musical enculturation, mirroring their ability to draw a distinction between speech and song. The cognitive view discussed here, thus, opens new avenues for future research on the perceptual phenomenon of the speech-to-song illusion.

Acknowledgements

We would like to thank our undergraduate research assistants Georgia Ann Carter and Katherine Willet at the University of Kent, Chlo   Lehoucq and Sasha Lou Wegiera at the Sorbonne Nouvelle Paris-3 Universit   who helped with the data collection. This research was supported by a Small Research Grant from the British Academy (SG152108) to the first author.

Action Editor

Ian Cross, University of Cambridge, Department of Music.

Peer Review

Emily Graber, Allegheny College.

Christina Vanden Bosch der Nederlanden, University of Toronto Mississauga, Psychology.

Author Contributions

TR and SF researched literature and conceived the study. SDB contributed to the study design. TR was responsible for gaining ethical approval, experimental set-up, and data analysis. TR wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of Conflicting Interests

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

Ethical Approval

The study received approval from the ethics committee of the University of Kent, UK (Reference number: 0031516; date of approval: 21/09/2015). All listeners gave an informed consent to participate in this research and were remunerated.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Tamara Rathcke  <https://orcid.org/0000-0002-4831-7387>

Data Availability Statement

The data that support the findings of the present study can be made available from the corresponding author, upon reasonable request.

Supplemental Material

Supplemental material for this article is available online (<https://osf.io/8d9mt/>).

Notes

1. Note that a monosyllabic word can be about 0.25 seconds (or even shorter, like “the”, “a”, or “an”), depending on the phonological structure of the word.
2. The sentence should be read (and is only meaningful if read) as ‘*The cotton / clothing is made of / grows in Mississippi*’, though the first parse of the sentence tends to be read as ‘*The cotton clothing is made of / grows in Mississippi*’ which causes short-

term comprehension issues (Ferreira, 2003). This example of a garden-path sentence contains an embedded phrase that increases the syntactic complexity of the sentence (Ferreira et al., 2001).

References

- Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10), 829–839. <https://doi.org/10.1038/nrn1201>
- Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, 3(2). <https://doi.org/10.1515/lp-2012-0017>
- Basirat, A., Schwartz, J-L, & Sato, M. (2012). Perceptuo-motor interactions in the perceptual organization of speech: Evidence from the verbal transformation effect. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367, 965–976. <https://doi.org/10.1098/rstb.2011.0374>
- Bates, D., M  chler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beck, C., Kardatzki, B., & Ethofer, T. (2014). Mondegreens and Soramimi as a Method to Induce Misperceptions of Speech Content - Influence of Familiarity, Wittiness, and Language Competence. *PLoS ONE*, 9(1), e84667. <https://doi.org/10.1371/journal.pone.0084667>
- Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes*, 11(1–2), 17–68. <https://doi.org/10.1080/016909696387213>
- Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2006). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (Vol. 1). Oxford University Press.
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–309. <https://doi.org/10.1017/S095267570000066X>
- Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, 100(1), 100–130. <https://doi.org/10.1016/j.cognition.2005.11.007>
- Bosker, H. R. (2018). Putting Laurel and Yanny in context. *The Journal of the Acoustical Society of America*, 144(6). <https://doi.org/10.1121/1.5070144>
- Calcott, R. D., & Berkman, E. T. (2014). Attentional flexibility during approach and avoidance motivational states: The role of context in shifts of attentional breadth. *Journal of Experimental Psychology: General*, 143(3), 1393–1408. <https://doi.org/10.1037/a0035060>
- Calef, R. S., Calef, R. A., Kesecker, M. P., & Burwell, R. (1974). Verbal transformations of “stabilized” taboo and neutral words. *Perceptual and Motor Skills*, 38(1), 177–178. <https://doi.org/10.2466/pms.1974.38.1.177>

- Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1493), 1001–1010. <https://doi.org/10.1098/rstb.2007.2155>
- 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>
- Chomsky, N. (1957). *Syntactic structures*. Mouton & Co.
- Christensen, R. H. B. (2019). *ordial – Regression Models for Ordinal Data: R Package version 2019.12-10*.
- Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2), 538–544. <https://doi.org/10.1037/a0018027>
- Culicover, P. W., & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, 10, 413–418. <https://doi.org/10.1016/j.tics.2006.07.007>
- Cummings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4), 659–678. <https://doi.org/10.1017/S1366728916000675>
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466. [https://doi.org/10.1016/S0022-5371\(80\)90312-6](https://doi.org/10.1016/S0022-5371(80)90312-6)
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 3(4), 422–433. <https://doi.org/10.3758/BF03214546>
- Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1–2), 132–147. <https://doi.org/10.1016/j.heares.2007.01.014>
- Dehaene-Lambertz, G., Montavont, A., Jobert, A., Alliol, L., Dubois, J., Hertz-Pannier, L., & Dehaene, S. (2010). Language or music, mother or Mozart? Structural and environmental influences on infants' language networks. *Brain and Language*, 114(2), 53–65. <https://doi.org/10.1016/j.bandl.2009.09.003>
- Deutsch, D. (1995). *Musical illusions and paradoxes*. Philomel. Retrieved from <https://scholar.google.com/citations?user=5ssmvbyaaaaj&hl=de&oi=sra>
- 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>
- Evans, C. R., & Kitson, A. (1967). An experimental investigation of the relation between the “familiarity” of a word and the number of changes in its perception which occur with repeated presentation as a “stabilized” auditory image. *National Physical Laboratory Audio Reports*, 36.
- Falk, S. (2011). Temporal variability and stability in infant-directed sung speech: Evidence for language-specific patterns. *Language and Speech*, 54(2), 167–180. <https://doi.org/10.1177/0023830910397490>
- Falk, S., Fasolo, M., Genovese, G., Romero-Lauro, L., & Franco, F. (2021). Sing for me, mama! Infants' discrimination of novel vowels in song. *Infancy*, 26(2), 248–270. <https://doi.org/10.1111/infa.12387>
- 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/a003685>
- Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203, <https://doi.org/10.1016/j.cognition.2020.104348>
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203. [https://doi.org/10.1016/S0010-0285\(03\)00005-7](https://doi.org/10.1016/S0010-0285(03)00005-7)
- Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11–15. <https://doi.org/10.1111/1467-8721.00158>
- Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30(1), 3–20. <https://doi.org/10.1023/A:1005290706460>
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1), 116–124. <https://doi.org/10.3758/BF03195503>
- Frazier, L., & Gibson, E. (Eds.) (2015). *SpringerLink Bücher: Vol. 46. Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor*. Springer. <https://doi.org/10.1007/978-3-319-12961-7>
- Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, 462(7272), 502–504. <https://doi.org/10.1038/nature08572>
- Goad, H., White, L., & Steele, J. (2003). Missing inflection in L2 acquisition: Defective syntax or LI-constrained prosodic representations? *Canadian Journal of Linguistics/Revue Canadienne De Linguistique*, 48(3–4), 243–263. <https://doi.org/10.1017/S0008413100000669>
- Gordon, R. L., Magne, C. L., & Large, E. W. (2011). EEG correlates of song prosody: A new look at the relationship between linguistic and musical rhythm. *Frontiers in Psychology*, 2, 352. <https://doi.org/10.3389/fpsyg.2011.00352>
- 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. *The Journal of the Acoustical Society of America*, 142(6), 3593. <https://doi.org/10.1121/1.5016806>
- Gregory, R. L. (1997). Visual illusions classified. *Trends in Cognitive Sciences*, 1(5), 190–194. [https://doi.org/10.1016/S1364-6613\(97\)01060-7](https://doi.org/10.1016/S1364-6613(97)01060-7)
- Gregory, R. L. (2009). *Seeing through illusions*. Oxford University Press.
- Grice, P. (1989). *Studies in the way of words*. Harvard University Press.
- 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(5), 1451–1459. <https://doi.org/10.1007/s00426-018-1135-z>
- Guinote, A. (2007). Power affects basic cognition: Increased attentional inhibition and flexibility. *Journal of Experimental Social*

- Psychology*, 43(5), 685–697. <https://doi.org/10.1016/j.jesp.2006.06.008>
- Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, 15(6), 883–899. <https://doi.org/10.1162/089892903322370807>
- Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual- difference perspective. *Memory & Cognition*, 29, 449–461. <https://doi.org/10.3758/BF03196396>
- 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>
- Harley, B. (2000). Listening strategies in ESL: Do age and L1 make a difference? *TESOL Quarterly*, 34(4), 769–777. <https://doi.org/10.2307/3587790>
- Harley, B., Howard, J., & Hart, D. (1995). Second language processing at different ages: Do younger learners pay more attention to prosodic cues to sentence structure? *Language Learning*, 45(1), 43–71. <https://doi.org/10.1111/j.1467-1770.1995.tb00962.x>
- Hawthorne, K., & Gerken, L. (2014). From pauses to clauses: Prosody facilitates learning of syntactic constituency. *Cognition*, 133(2), 420–428. <https://doi.org/10.1016/j.cognition.2014.07.013>
- Heller, J. R. (1977). Enjambment as a metrical force in romantic conversation poems. *Poetics*, 6(1), 15–25. [https://doi.org/10.1016/0304-422X\(77\)90018-3](https://doi.org/10.1016/0304-422X(77)90018-3)
- Ito, A., & Pickering, M. J. (2021). Automaticity and prediction in non-native language comprehension. *Prediction in Second Language Processing and Learning*, 26–46. <https://doi.org/10.1075/bpa.12.02ito>
- Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception*, 26(3), 195–204. <https://doi.org/10.1525/mp.2009.26.3.195>
- Jackendoff, R., & Lerdahl, F. (2006). The capacity for music: What is it, and what's special about it? *Cognition*, 100(1), 33–72. <https://doi.org/10.1016/j.cognition.2005.11.005>
- 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>
- Jakobovits, L. A., & Lambert, W. E. (1964). Stimulus-characteristics as determinants of semantic changes with repeated presentation. *The American Journal of Psychology*, 77(1), 84–92. <https://doi.org/10.2307/1419274>
- Janda, R. D., & Morgan, T. A. (1987). El acento dislocadó – pues cantadó – castellanó: On explaining stress-shift in song-texts from Spanish (and certain other romance languages). In *Advances in romance linguistics* (pp. 151–170). De Gruyter. <https://doi.org/10.1515/9783112420140-010>
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149. <https://doi.org/10.1037/0033-295x.99.1.122>
- Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, 274(5284), 114–116. <https://doi.org/10.1126/science.274.5284.114>
- Kaminska, Z., Pool, M., & Mayer, P. (2000). Verbal transformation: Habituation or spreading activation? *Brain and Language*, 71(2), 285–298. <https://doi.org/10.1006/brln.1999.2181>
- Kentner, G. (2015). Rhythmic segmentation in auditory illusions – Evidence from cross-linguistic mondegreens. *Proceedings of the International Congress of Phonetic Sciences*.
- Keye, D., Wilhelm, O., Oberauer, K., et al. (2009). Individual differences in conflict-monitoring: Testing means and covariance hypothesis about the Simon and the Eriksen flanker task. *Psychological Research Psychologische Forschung*, 73, 762–776. <https://doi.org/10.1007/s00426-008-0188-9>
- Kim, A. E., Oines, L., & Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: An ERP investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(3), 406–420. <https://doi.org/10.1037/xlm0000457>
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580–602. [https://doi.org/10.1016/0749-596X\(91\)90027-H](https://doi.org/10.1016/0749-596X(91)90027-H)
- Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, 40(2), 153–194. <https://doi.org/10.1006/jmla.1998.2620>
- Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195148367.001.0001>
- Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4), 334–368. <https://doi.org/10.1037/0033-295X.89.4.334>
- Kuhn, G., Amlani, A. A., & Rensink, R. A. (2008). Towards a science of magic. *Trends in Cognitive Sciences*, 12(9), 349–354. <https://doi.org/10.1016/j.tics.2008.05.008>
- Kurland, J. (2011). The role that attention plays in language processing. *Perspectives on Neuropsychology and Neurogenic Speech and Language Disorders*, 21(2), 47–54. <https://doi.org/10.1044/nnsld21.2.47>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Liebling, E. (1908). Phrasing. In *The American history and encyclopedia of music* (Vol. 1, pp. 267–282). Irving Squire.
- MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1), 35–54. <https://doi.org/10.1037/0033-295x.109.1.35>
- Mackay, D. G. (1988). The organization of perception and action. A theory for language and other cognitive skills. *The Italian Journal of Neurological Sciences*, 9(3). <https://doi.org/10.1007/BF02334060>

- Mackay, D. G., Wulf, G., Yin, C., & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language*, 32(5), 624–646. <https://doi.org/10.1006/jmla.1993.1032>
- Mall, J. T., Morey, C. C., Wolff, M. J., et al. (2014). Visual selective attention is equally functional for individuals with low and high working memory capacity: Evidence from accuracy and eye movements. *Attention, Perception, and Psychophysics*, 76, 1998–2014. <https://doi.org/10.3758/s13414-013-0610-2>
- Margulis, E. H. (2013). *On repeat: How music plays the mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199990825.001.0001>
- 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>
- Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1), 1–71. [https://doi.org/10.1016/0010-0277\(80\)90015-3](https://doi.org/10.1016/0010-0277(80)90015-3)
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. <https://doi.org/10.1038/264746a0>
- McIntosh, R. D. (2022). Perceptual illusions. In S. Della Sala (Ed.), *Encyclopedia of behavioral neuroscience* (pp. 588–596). Elsevier. <https://doi.org/10.1016/B978-0-12-819641-0.00006-2>
- McMullen, E., & Saffran, J. R. (2004). Music and language: A developmental comparison. *Music Perception: An Interdisciplinary Journal*, 21(3), 289–311. <https://doi.org/10.1525/mp.2004.21.3.289>
- Meyerhoff, H. S., Gehrer, N. A., Merz, S., & Frings, C. (2022). The beep-speed illusion: Non-spatial tones increase perceived speed of visual objects in a forced-choice paradigm. *Cognition*, 219, <https://doi.org/10.1016/j.cognition.2021.104978>
- Morgan, J. L., & Demuth, K. (2014). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Taylor and Francis.
- Mullin, H. A. C., 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). <https://doi.org/10.1371/journal.pone.0250042>
- Myachykov, A., & Posner, M. I. (2005). Attention in language. *Neurobiology of Attention*, 324–329. <https://doi.org/10.1016/B978-012375731-9/50057-4>
- Mykhalonok, M. (2020). Music and prosody: Suprasegmental features of reggaeton songs. *ATeM Archiv für Textmusikforschung*, 4(1), 1–19. https://doi.org/10.15203/ATeM_2019_1.07
- Natsoulas, T. (1965). A study of the verbal-transformation effect. *The American Journal of Psychology*, 78(2), 257–263. <https://doi.org/10.2307/1420498>
- Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Foris Publications. xiv + 327. <https://doi.org/10.1017/S0952675700002219>
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234. [https://doi.org/10.1016/0010-0277\(94\)90043-4](https://doi.org/10.1016/0010-0277(94)90043-4)
- O’Callaghan, C. (2011). Lessons from beyond vision (sounds and audition). *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 153(1), 143–160. <https://doi.org/10.1007/s11098-010-9652-7>
- Oh, Y., Todd, S., Beckner, C., et al. (2020). Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific Reports*, 10, 22318. <https://doi.org/10.1038/s41598-020-78810-4>
- Park, H., & Reder, L. M. (2012). Moses illusion. In *Cognitive illusions* (pp. 287–304). Psychology Press. <https://doi.org/10.4324/9780203720615>
- Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, 6(7), 674–681. <https://doi.org/10.1038/nn1082>
- Patel, A. D. (2012). Language, music, and the brain: A resource-sharing framework. In P. Rebuschat, M. Rohmeier, J. A. Hawkins, & I. Cross (Eds.), *Language and music as cognitive systems* (pp. 204–223). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199553426.003.0022>
- Patson, N. D., & Warren, T. (2010). Eye movements when reading implausible sentences: Investigating potential structural influences on semantic integration. *Q J Exp Psychol (Hove)*, 63(8). <https://doi.org/10.1080/17470210903380822>
- Pattison, P. (1991). *Songwriting: Essential guide to lyric form and structure: Tools and techniques for writing better lyrics*. Hal Leonard Corporation.
- Perkell, J. S., & Klatt, D. H. (Eds.). (2014). *Invariance and variability in speech processes*. Psychology Press. <https://doi.org/10.4324/9781315802350>
- Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, 4(1–3), 75–95. [https://doi.org/10.1016/0167-6393\(85\)90037-8](https://doi.org/10.1016/0167-6393(85)90037-8)
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25–42. <https://doi.org/10.1146/annurev.ne.13.030190.000325>
- Pressnitzer, D., Graves, J., Chambers, C., de Gardelle, V., & Egré, P. (2018). Auditory perception: Laurel and Yanny together at last. *Current Biology : CB*, 28(13), R739–R741. <https://doi.org/10.1016/j.cub.2018.06.002>
- Pressnitzer, D., & Hupé, J-M (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology: CB*, 16(13), 1351–1357. <https://doi.org/10.1016/j.cub.2006.05.054>
- Rathcke, T., Falk, S., & Dalla Bella, S. (2021b). Music to your ears: Sentences sonority and listener background modulate the “speech-to-song illusion”. *Music Perception*, 38(5), 499–508. <https://doi.org/10.1525/mp.2021.38.5.499>
- Rathcke, T., Lin, C.-Y., Falk, S., & Bella, S. D. (2021a). Tapping into linguistic rhythm. *Laboratory Phonology*, 12(1), Article 1. <https://doi.org/10.5334/labphon.248>
- Rensink, R. A., & Kuhn, G. (2015). A framework for using magic to study the mind. *Frontiers in Psychology*, 5, 1508. <https://doi.org/10.3389/fpsyg.2014.01508>

- Rohrmeier, M., & Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics in Cognitive Science*, 4(4), 525–553. <https://doi.org/10.1111/j.1756-8765.2012.01223.x>
- Rothbart, M. K., & Posner, M. I. (2001). Mechanism and variation in the development of attentional networks. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 353–363). MIT Press.
- 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(2), 583–590. <https://doi.org/10.3758/s13423-018-1527-5>
- Sato, M., Schwartz, J.-L., Abry, C., Cathiard, M.-A., & Loevenbruck, H. (2006). Multistable syllables as enacted percepts: A source of an asymmetric bias in the verbal transformation effect. *Perception & Psychophysics*, 68(3), 458–474. <https://doi.org/10.3758/BF03193690>
- Schellenberg, M. (2009). Singing in a tone language: Shona. In *Selected proceedings of the 39th annual conference on African linguistics* (pp. 137–144). Cascadia Proceedings Project.
- Severance, E., & Washburn, M. F. (1907). The loss of associative power in words after long fixation. *The American Journal of Psychology*, 18(2), 182. <https://doi.org/10.2307/1412411>
- Shahin, A. J., Bishop, C. W., & Miller, L. M. (2009). Neural mechanisms for illusory filling-in of degraded speech. *NeuroImage*, 44(3), 1133–1143. <https://doi.org/10.1016/j.neuroimage.2008.09.045>
- Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25(2), 193–247. <https://doi.org/10.1007/BF01708572>
- Shoaf, L. C., & Pitt, M. A. (2002). Does node stability underlie the verbal transformation effect? A test of node structure theory. *Perception & Psychophysics*, 64(5), 795–803. <https://doi.org/10.3758/BF03194746>
- Simchy-Gross, R., & Margulis, E. H. (2018). The sound-to-music illusion: Repetition can musicalize nonspeech sounds. *Music and Science*, 1(1–6). <https://doi.org/10.1177/2059204317731992>
- Snyder, K. A., Calef, R. S., Choban, M. C., & Geller, E. S. (1992). Frequency of verbal transformations as a function of word-presentation styles. *Bulletin of the Psychonomic Society*, 30(5), 363–364. <https://doi.org/10.3758/BF03334089>
- Snyder, K. A., Calef, R. S., Choban, M. C., & Geller, E. S. (1993). Effects of word repetition and presentation rate on the frequency of verbal transformations: Support for habituation. *Bulletin of the Psychonomic Society*, 31(2), 91–93. <https://doi.org/10.3758/BF03334148>
- Speer, S. R., Kjølgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, 25(2), 249–271. <https://doi.org/10.1007/BF01708573>
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.
- Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *American Statistician*, 49, 53–56. <https://doi.org/10.1080/00031305.1995.10476113>
- Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9, 578–585. <https://doi.org/10.1038/nn1669>
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell publishers.
- 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>
- Sundberg, J. (1989). Synthesis of singing by rule. In M. V. Mathews & J. R. Pierce (Eds.), *Current directions in computer music research* (pp. 45–55). MIT Press.
- 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>
- Tiego, J., Testa, R., Bellgrove, M. A., Pantelis, C., & Whittle, S. (2018). A hierarchical model of inhibitory control. *Frontiers in Psychology*, 9, 1339. <https://doi.org/10.3389/fpsyg.2018.01339>
- 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*, 147(6), 888–904. <https://doi.org/10.1037/xge0000455>
- Tierney, A., Patel, A. D., Jasmin, K., & Breen, M. (2021). Individual differences in perception of the speech-to-song illusion are linked to musical aptitude but not musical training. *Journal of Experimental Psychology: Human Perception and Performance*, 47(12), 1681–1697. <https://doi.org/10.1037/xhp0000968>
- Tomlin, R. S. (1999). Mapping conceptual representations into linguistic representations: The role of attention in grammar. In E. Pederson & J. Nuyts (Eds.), *Language, culture, and cognition: Vol. 1. Language and conceptualization* (pp. 162–189). Cambridge University Press. <https://doi.org/10.1017/CBO9781139086677.007>
- Townsend, D. J. (1983). Thematic processing in sentences and texts. *Cognition*, 13(2), 223–261. [https://doi.org/10.1016/0010-0277\(83\)90023-9](https://doi.org/10.1016/0010-0277(83)90023-9)
- Tremblay, A., Broersma, M., Coughlin, C. E., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology*, 7, 985. <https://doi.org/10.3389/fpsyg.2016.00985>
- Vanden Bosch der Nederlanden, Christina, M., Xin, Q., Sarah, S., Prakhara, S., Jessica, A. G., Marc, F. J., & Erin, E. H. (2023). Developmental changes in the categorization of speech and song. *Developmental Science*, 26(5), e13346. <https://doi.org/10.1111/desc.13346>
- 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>
- Van Gompel, R. P., & Pickering, M. J. (2007). Syntactic parsing. In M. G. Gaskell & G. Altmann (Eds.), *The Oxford Handbook of Psycholinguistics* (pp. 289–307). Oxford University Press.
- van't Jagt, R. K., Hoeks, J. C., Dorleijn, G. J., & Hendriks, P. (2014). Look before you leap: How enjambment affects the processing of poetry. *Scientific Study of Literature*, 4(1), 3–24. <https://doi.org/10.1075/ssol.4.1.01jag>
- Verschuure, J., & Brocaar, M. P. (1983). Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. *Perception & Psychophysics*, 33(3), 232–240. <https://doi.org/10.3758/BF03202859>
- 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, 585–597. <https://doi.org/10.1177/1747021820969144>
- Warren, R. M. (1961). Illusory changes of distinct speech upon repetition – the verbal transformation effect. *British Journal of Psychology (London, England : 1953)*, 52, 249–258. <https://doi.org/10.1111/j.2044-8295.1961.tb00787.x>
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science (New York, N.Y.)*, 167(3917), 392–393. <https://doi.org/10.1126/science.167.3917.392>
- Warren, R. M. (1983). Auditory illusions and their relation to mechanisms normally enhancing accuracy of perception. *Journal of the Audio Engineering Society*, 31(9), 623–629. <https://aes2.org/publications/elibrary-page/?id=4557>
- Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, 71(3), 612–613. <https://doi.org/10.2307/1420267>
- Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception and Psychophysics*, 9, 358–362. <https://doi.org/10.3758/BF03212667>
- Warren, R. M., & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. *Perception and Psychophysics*, 16, 150–156. <https://doi.org/10.3758/BF03203268>
- Warren, R. M., & Warren, R. P. (1966). A comparison of speech perception in childhood, maturity, and old age by means of the verbal transformation effect. *Journal of Verbal Learning and Verbal Behavior*, 5, 142–146. [https://doi.org/10.1016/S0022-5371\(66\)80007-5](https://doi.org/10.1016/S0022-5371(66)80007-5)
- Waters, G., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, 1003(4). <https://doi.org/10.1037/0033-295X.103.4.761>
- Wechsler, D. (1997). *Wechsler adult intelligence scale, Third Edition*. Psychological Corporation. <https://doi.org/10.1037/t49755-000>
- Yantis, S., & Johnston, J. C. (1990). On the locus of visual selection: Evidence from focused attention tasks. *Journal of Experimental Psychology: Human Perception and Performance*, 16(1), 135–149. <https://doi.org/10.1037/0096-1523.16.1.135>
- Zimmermann, P., & Fimm, B. (2004). A test battery for attentional performance. In M. Leclercq & P. Zimmermann (Eds.), *Applied neuropsychology of attention: Theory, diagnosis and rehabilitation* (pp. 124–165). Taylor and Francis. <https://doi.org/10.4324/9780203307014-12>