Illusory transformation from speech to song[a)]

Diana Deutsch,[b)] Trevor Henthorn, and Rachael Lapidis
Department of Psychology, University of California, San Diego, La Jolla, California 92093

(Received 8 December 2010; revised 12 February 2011; accepted 14 February 2011)

An illusion is explored in which a spoken phrase is perceptually transformed to sound like song rather than speech, simply by repeating it several times over. In experiment I, subjects listened to ten presentations of the phrase and judged how it sounded on a five-point scale with endpoints marked "exactly like speech" and "exactly like singing." The initial and final presentations of the phrase were identical. When the intervening presentations were also identical, judgments moved solidly from speech to song. However, this did not occur when the intervening phrases were transposed slightly or when the syllables were presented in jumbled orderings. In experiment II, the phrase was presented either once or ten times, and subjects repeated it back as they finally heard it. Following one presentation, the subjects repeated the phrase back as speech; however, following ten presentations they repeated it back as song. The pitch values of the subjects' renditions following ten presentations were closer to those of the original spoken phrase than were the pitch values following a single presentation. Furthermore, the renditions following ten presentations were even closer to a hypothesized representation in terms of a simple tonal melody than they were to the original spoken phrase. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3562174]

PACS number(s): 43.75.Cd, 43.75.Rs [NHF] Pages: 2245–2252

I. INTRODUCTION

There has recently been an upsurge of interest in relationships between music and speech, particularly in how these two forms of communication are processed by the auditory system (cf. Zatorre *et al*., 2002; Koelsch *et al*., 2002; Koelsch and Siebel, 2005; Zatorre and Gandour, 2007; Schon *et al*., 2004; Peretz and Coltheart, 2003; Patel, 2008; Hyde *et al*., 2009; Deutsch, 2010). In exploring this issue, it is generally assumed that whether a phrase is heard as spoken or sung depends on its acoustical characteristics. Speech consists of frequency glides that are often steep, and of rapid amplitude and frequency transitions. In contrast, song consists largely of discrete pitches that are sustained over relatively long durations and that tend to follow each other in small steps. At the phenomenological level, speech appears as a succession of rapid changes in timbre, which are interpreted as consonants and vowels, and in which pitch contours are only broadly defined (at least in nontone languages). In contrast, song is heard primarily as a succession of well-defined musical notes (though also with consonants and vowels) and these are combined to form well-defined pitch relationships and rhythmic patterns. The dichotomy between the physical characteristics of speech and nonspeech is not clearcut, however. It has been found that certain nonspeech sounds can be interpreted as speech as a consequence of training (Remez *et al*., 1981; Mottonen *et al*., 2006) or when they are placed in verbal contexts (Shtyrov *et al*., 2005).

The widespread view that speech and music can be defined in terms of their acoustical properties is reflected in studies that explore their perceptual characteristics and neurological underpinnings. For speech, researchers have focused on features such as fast formant transitions and voice onset time (Diehl *et al*., 2004), while for music, researchers have examined such issues as the processing of pitch sequences, musical instrument timbres, and rhythmic patterns (Stewart *et al*., 2006).

The use of signals with different physical characteristics is necessary for studying music and speech taken independently. However, when differences are found in the ways in which they are processed, these could be either due to the differences in the signals employed or due to the processing of these signals by different neural pathways (Zatorre and Gandour, 2007). In contrast, this paper describes and explores an illusion in which a spoken phrase is perceptually transformed so as to be heard as sung rather than spoken. The illusion occurs without altering the signal in any way, without training, and without any context provided by other sounds, but simply as a result of repeating the phrase several times over. Research on this illusion therefore provides insights into differences in the processing of speech and music, without the complication of invoking different signal parameters or different contexts.

The illusion was first published as a demonstration on the compact disc by Deutsch (2003). Here, a spoken sentence is presented, followed repeatedly by a phrase that had been embedded in it. Most people hear the repeated phrase transform into a sung melody, generally as notated in Fig. 1. This paper describes the first formal exploration of the illusion and presents a discussion of its possible underlying bases.

The study consisted of two experiments. Experiment I explored certain constraints governing the illusion, using a

[a)]Portions of this work were presented at a meeting of the Acoustical Society of America, Miami, November 2008, as Deutsch, D., Lapidis, R., and Henthorn, T. (2008). "The speech-to-song illusion," J. Acoust. Soc. Am. **124,** 2471.

[b)]Author to whom correspondence should be addressed. Electronic mail: ddeutsch@ucsd.edu

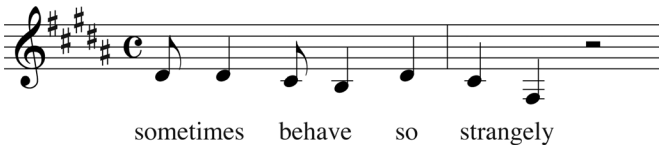

FIG. 1. The spoken phrase, as it appears to be sung. From Deutsch (2003).

rating task as the measure. The illusion was found to occur when the repeated presentations of the spoken phrase were exact replicas of the original one. However, when on repetition, the phrase was transposed slightly or the syllables were jumbled, the illusion did not occur. In experiment II, the characteristics of this perceptual transformation were explored in detail by having subjects repeat back the phrase exactly as they had heard it, both following a single repetition and following ten repetitions. It is hypothesized that during the process of repetition, the pitches forming the phrase increase in perceptual salience and that in addition they are perceptually distorted so as to conform to a tonal melody. The findings from the experiment provide evidence in favor of this hypothesis. Finally, the hypothesized neurological underpinnings of this illusion are explored and its general implications for relationships between speech and music are discussed.

II. EXPERIMENT I

A. Method

1. Subjects

Fifty-four subjects with at least 5 yr of musical training participated in the experiment and were paid for their services. They were divided into three groups of 18 subjects each, with each group serving in one condition. The subjects in the first group (three males and 15 females) were of average age 21.7 yr (range, 18–33 yr) and with an average of 10.2 yr of musical training (range, 6–14 yr). Those in the second group (four males and 14 females) were of average age 22.4 yr (range, 18–29 yr) and with an average of 10.6 yr (range, 6–15 yr) of musical training. Those in the third group (three males and 15 females) were of average age 20.3 yr (range 18–28 yr) and with an average of 10.0 yr (range 6–14 yr) of musical training. None of the subjects had perfect pitch. All had normal hearing in the range of 250 Hz−6 kHz, as determined by audiometric testing, and all were naïve concerning the purpose of the experiment and the nature of the illusion.

2. Stimulus patterns and procedure

The experiment was carried out in a quiet room. The stimulus patterns were derived from the sentence on track 22 of the compact disc by Deutsch (2003). The sentence states "*The sounds as they appear to you are not only different from those that are really present, but they sometimes behave so strangely as to seem quite impossible*." In all conditions, this sentence was presented, followed by a pause of 2300 ms in duration and then by ten presentations of the embedded phrase "*sometimes behave so strangely*," which were separated pauses of 2300 ms in duration. During the pause following each presentation, the subjects judged how the phrase had sounded on a five-point scale with endpoints 1 and 5 marked "exactly like speech" and "exactly like singing."

In all conditions, the initial and final presentations of the phrase were untransformed; however, the phrases in the intervening presentations varied depending on the condition: In the *untransformed* condition, the intervening phrases were also untransformed. In the *transposed* condition, the intervening phrases were transposed slightly, while the formant frequencies were preserved. The degree of transposition on each of the intervening presentations, given in the order of presentation, was +2/3 semitone; −1 1/3 semitone; +1 1/3 semitone; −2/3 semitone; +1 1/3 semitone; −1 1/3 semitone; +2/3 semitone; and −2/3 semitone. In the *jumbled* condition, the intervening phrases were untransposed, but they were presented in jumbled orderings. The phrase consisted of seven syllables (1 = "*some*;" 2 = "*times*;" 3 = "*be*;" 4 = "*have*;" 5 = "*so*;" 6 = "*strange*;" and 7 = "*ly*"), and in the intervening repetitions, the orderings of the syllables, given in the order of presentation, were 6, 4, 3, 2, 5, 7, 1; 7, 5, 4, 1, 3, 2, 6; 1, 3, 5, 7, 6, 2, 4; 3, 6, 2, 5, 7, 1, 4; 2, 6, 1, 7, 4, 3, 5; 4, 7, 1, 3, 5, 2, 6; 6, 1,5, 3, 2, 4, 7; and 2, 5, 4, 3, 7, 1, 6.

Finally, all subjects filled out a questionnaire that enquired into their age and musical training.

3. Instrumentation and software

The original sentence was recorded from track 22 of the Deutsch (2003) compact disc onto a Power Mac G5 computer, and it was saved as an AIF file at a sampling frequency of 44.1 kHz. The software package BIAS PEAK PRO Version 4.01 was employed to create the stimuli in all conditions and also to create the jumbled orderings in the *jumbled* condition. The software package PRAAT Version 4.5.06 (Boersma and Weenink, 2006) was employed to create the transpositions in the *transposed* condition, using the pitch-synchronous overlap-and-add method. The reconstituted signals were then recorded onto compact disc. They were played to subjects on a Denon DCD-815 compact disc player, the output of which was passed through a Mackie CR 1604-VLZ mixer and presented via two Dynaudio BM15A loudspeakers at a level of approximately 70 dB sound pressure level (SPL) at the subjects' ears.

B. Results

Figure 2 shows the average ratings of the phrase on the initial and final presentations and under each of the three conditions. It can be seen that the phrase was perceived as speech on the initial presentation; however, the way it was perceived on the final presentation depended on the nature of the intervening presentations. When these were untransformed, the phrase on the final presentation was heard as song. However, when the intervening presentations were transposed, judgments on the final presentation remained as speech, though displaced slightly towards song.

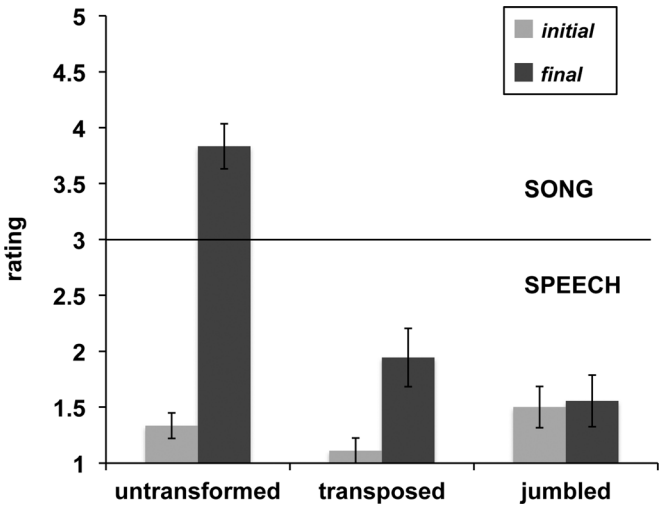

FIG. 2. Average ratings of the spoken phrase on the initial and final presentations, under the three conditions: *untransformed*, *transposed*, and *jumbled*. Subjects rated the phrase on a five-point scale with endpoints 1 and 5 marked exactly like speech, and exactly like singing, respectively.

When the syllables in the intervening presentations were presented in jumbled orderings, the phrase on the final presentation was heard as speech.

To make statistical comparison between the judgments under the different conditions, a 2×3 analysis of variance (ANOVA) was performed, with presentation (*initial* and *final*) as a within-subjects factor, and condition (*untransformed*, *transposed*, and *jumbled*) as a between-subjects factor. The overall difference between the initial and final presentations was highly significant [$F(1, 51) = 62.817$; $p < 0.001$], the effect of condition was highly significant [$F(2, 51) = 16.965$; $p < 0.001$], and the interaction between presentation and condition was highly significant [$F(2, 51) = 25.593$; $p < 0.001$].

Given these findings, two further ANOVAs were performed, in which judgments in the three conditions were compared for the initial and final presentations taken separately. For the initial presentation, judgments in the different conditions did not differ significantly [$F(2, 51) = 1.912$; $p > 0.05$]. However, for the final presentation, the effect of type of intervening phrase was highly significant [$F(2, 51) = 27.317$, $p < 0.001$]. Post hoc comparisons were therefore made taking the final presentation alone. It was found that judgments in the *untransformed* condition were significantly different from those in the *transposed* condition ($p < 0.001$) and were also significantly different from those in the *jumbled* condition ($p < 0.001$). The difference between judgments in the *transposed* and *jumbled* conditions was nonsignificant ($p > 0.05$).

C. Discussion

In this experiment, it was found that for a group of subjects who were naïve concerning the purpose of the experiment and who had been selected only on the basis of having had at least 5 yr of musical training, the repeated presentation of a spoken phrase caused it to be heard as sung rather than spoken. However, this perceptual transformation did not occur when, during the intervening presentations, the phrase was transposed slightly or the syllables were presented in jumbled orderings. The illusion could not, therefore, have been due to repetition of the pitch contour of the phrase or even repetition of the exact melodic intervals, since these were preserved under transposition. Further, since the perceptual transformation did not occur when the intervening patterns were transposed leaving the timing of the signal unaltered, it could not have been due to the repetition of the exact timing of the phrase. In addition, since the perceptual transformation did not occur when the syllables were presented in jumbled orderings, it could not have been due to the exact repetition of the unordered set of syllables. The illusion therefore appears to require repetition of the untransposed set of syllables, presented in the same ordering.

Experiment II explored this transformation effect in further detail, by employing a production task. The embedded phrase was presented either once or ten times following the complete sentence, and the subjects were asked to repeat it back exactly as they had most recently heard it. Differences in the subjects' renditions under these two conditions were analyzed.

The experiment was motivated by two hypotheses. First, in contrast to song, the pitch characteristics of speech are rarely salient perceptually, and one striking characteristic of the present illusion is that the perceived pitch salience of the syllables increases substantially through repetition. It was therefore hypothesized that following repeated listening to the phrase, this perceptual increase in pitch salience would result in renditions whose pitches would be closer to the original spoken phrase, and with more inter-subject consistency. Second, it was hypothesized that once the syllables were heard as forming salient pitches, they would also be perceptually distorted so as to be in accordance with a plausible melodic representation; specifically, it was hypothesized that the pitches produced by the subjects would be as notated in Fig. 1.

III. EXPERIMENT II

A. Method

1. Subjects

Thirty-one female subjects participated in the experiment and were paid for their services. They were divided into three groups. The first group consisted of 11 subjects, of average age 23.8 yr (range, 19–35 yr), and with an average of 11.2 yr (range, 5–15 yr) of musical training. Before participating in the experiment, they had listened to the sentence followed by the repeating phrase and had all reported that they heard the phrase as having been transformed into song. The second group also consisted of 11 subjects and were of average age 18.9 yr (range, 18–20 yr) and with an average of 8.9 yr (range, 6–12 yr) of musical training. They had not been presented with the stimulus pattern before participating in the experiment. The third group consisted of nine subjects, of average age 19.2 yr (range, 18–22 yr), and with an

average of 8.0 yr (range, 4–11 yr) of musical training. None of the subjects had perfect pitch. All had normal hearing in the range of 250 Hz−6 kHz, as determined by audiometric testing, and all were naïve concerning the purpose of the experiment and the nature of the illusion.

2. Experimental conditions and procedure

There were four conditions in the experiment. In the *repeat speech* condition, the stimulus pattern was as in experiment 1, so that the embedded phrase *sometimes behave so strangely* was presented ten times in succession, except that the pauses between repeated presentations were 780 ms in duration. The *nonrepeat speech* condition was identical to the *repeat speech* condition, except that the embedded phrase was presented only once. In the *nonrepeat song* condition, the stimulus pattern consisted of a recording of a single rendition of the embedded phrase sung by one of the authors (R.L.) as she had heard it following multiple presentations. In all three conditions, the subjects were asked to listen to the stimulus pattern and then to repeat it back three or four times exactly as they had heard it; the second rendition of the phrase was then extracted for analysis. The first group of subjects served in the *repeat speech* condition and the second group served in the *nonrepeat speech* condition, followed immediately by the *nonrepeat song* condition.

Finally, in the *evaluation* condition, the 22 renditions of the spoken phrase, which were taken from the utterances of the 11 subjects in the *repeat speech* condition and the 11 subjects in the *nonrepeat speech* condition, were presented to the third group of subjects. The phrases were presented in random order and were separated by 8-s pauses. During the pause following each presentation, the subject indicated on a response sheet whether the phrase sounded as speech or as song.

3. Instrumentation and software

The subjects produced their renditions individually in a quiet room. The instrumentation used to deliver the stimulus patterns was identical to that in experiment I. The subjects' vocalizations were recorded onto an Edirol R-1 24 bit recorder at a sampling frequency of 44.1 kHz. The recordings were made using an AKG C 1000 S microphone placed roughly 8 in. from the subject's mouth. The sound files were transferred to an iMac computer, where they were saved as AIF files at a sampling frequency of 44.1 K. Then from each sound file, the second rendition of the phrase was extracted, saved as a separate sound file, and normalized for amplitude using the software package BIAS PEAK PRO Version 5.2. F0 estimates of the subject's vocalizations were then obtained at 5 ms intervals using the software package PRAAT Version 5.0.09 (autocorrelation method). Then for each sound file, the F0 estimates were averaged along the musical scale; that is, along a log frequency continuum, so producing an average F0 for the phrase. In addition, each phrase was segmented into the seven syllables (*some*, *times*, *be*, *have*, *so*, *strange*, and *ly*), and the F0 estimates were averaged over each syllable separately.

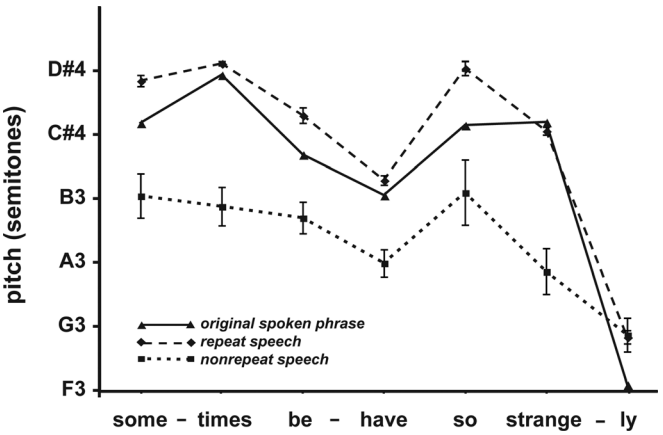

FIG. 3. Triangles show the average F0 of each syllable in the original spoken phrase. Diamonds show the average F0 of each syllable, averaged over the renditions of all subjects in the *repeat speech* condition. Squares show the average F0 of each syllable, averaged over the renditions of all subjects in the *nonrepeat speech* condition. F3 = 174.6 Hz; G3 = 196.0 Hz; A3 = 220 Hz; B3 = 246.9 Hz; C#4 = 277.2 Hz; and D#4 = 311.1 Hz.

B. Results

The judgments made in the *evaluation* condition showed that renditions in the *nonrepeat speech* condition were heard as spoken, while those in the *repeat speech* condition were heard as sung. Specifically, this was true for 97.5% of the 198 judgments that were made. This result is as expected from the findings from experiment I, in which subjects judged the initial presentation of the original spoken phrase as spoken and the final presentation as sung.

Detailed analyses of the pitch patterns in the renditions were undertaken in order to characterize the changes that resulted from repeated exposure to the original spoken phrase. Figure 3 displays the average F0s of all the syllables in the original spoken phrase, together with those averaged over all renditions in the *repeat speech* condition and in the *nonrepeat speech* condition. As further illustration, Fig. 4 displays the pitch tracings of the original spoken phrase, together with those from four subjects in the *repeat speech* condition and four subjects in the *nonrepeat speech* condition. These pitch tracings are representative of those produced by all subjects in each condition.

Two findings emerged that were predicted from the hypothesis that pitch salience increased as a result of repetition. These showed that the average pitch for the phrase as a whole, and also for each syllable taken separately, was more consistent across subjects and closer to the original spoken phrase, in the *repeat speech* condition than in the *nonrepeat speech* condition.

First, the across-subjects variance in average F0 was considerably lower for renditions in the *repeat speech* condition than in the *nonrepeat speech* condition. Taking the average F0s for renditions of the entire phrase, this difference in variance was highly significant statistically [$F(10, 10) = 5.62$, $p < 0.01$]. This pattern held when comparisons were made for each syllable taken separately: for *some*, $F(10, 10) = 19.72$, $p < 0.0001$; for *times*, $F(10, 10) = 69.22$, $p < 0.0001$; for *be*, $F(10, 10) = 6.2$, $p < 0.01$; for *have*, $F(10, 10) = 9.71$, $p < 0.001$; for *so*, $F(10, 10) = 22.68$, $p < 0.0001$;

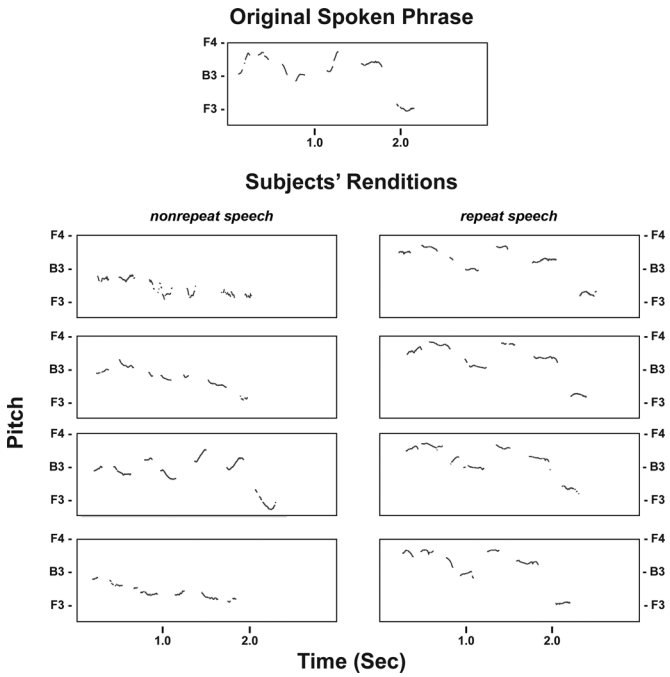

FIG. 4. Pitch tracings of the original spoken phrase, together with those from four representative subjects in the *repeat speech* condition, and from four representative subjects in the *nonrepeat speech* condition. F3 = 174.6 Hz; B3 = 246.9 Hz; and F4 = 349.2 Hz.

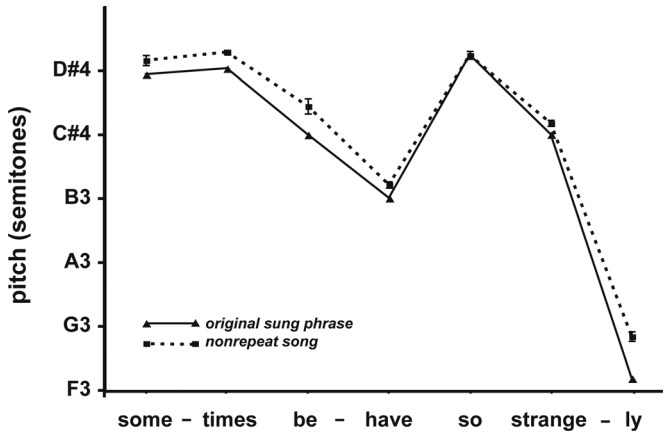

FIG. 5. Triangles show the average F0s of each syllable in the original sung phrase. Squares show the average F0 of each syllable, averaged over the renditions of all subjects in the *nonrepeat song* condition. F3 = 174.6 Hz; G3 = 196.0 Hz; A3 = 220 Hz; B3 = 246.9 Hz; C#4 = 277.2 Hz; and D#4 = 311.1 Hz.

for *strange*, $F(10, 10) = 35.76$, $p < 0.0001$; and for *ly*, $F(10, 10) = 12.71$, $p < 0.001$.

Second, the average F0s for renditions in the *repeat speech* condition were found to be considerably closer to the average F0 of the original spoken phrase compared with those in the *nonrepeat speech* condition. To evaluate this effect statistically, for each subject a difference score was obtained between the average F0 of her rendition of the phrase and that of the original spoken phrase. Using an independent samples t-test assuming unequal sample variances, the difference scores were found to be significantly lower for renditions following ten presentations than for those following a single presentation [$t(13.34) = -4.03$, $p < 0.01$]. This pattern held for all syllables taken individually except for the last one: for *some* $t(11.01) = 5.37$, $p < 0.001$; for *times*, $t(10.29) = 7.68$, $p < 0.0001$; for *be*, $t(13.14) = 6.46$, $p < 0.0001$; for *have*, $t(12.04) = 6.28$, $p < 0.0001$; for *so*, $t(10.88) = 4.23$, $p < 0.01$; for *strange*, $t(10.73) = 6.07$, $p < 0.001$; and for *ly*, there was a nonsignificant [$t(11.19) = 1.34$, $p = 0.23$] trend in the same direction.

To ensure that the differences between conditions found here were not due to a simple effect of repetition, comparison was made between renditions in the *nonrepeat speech* condition and the *nonrepeat song* condition, in which the stimulus pattern was also presented to the subjects only once. Figure 5 displays the average F0s of all syllables in the original sung phrase, together with those averaged over all renditions in the *nonrepeat song* condition. It can be seen that the subjects' renditions in this condition corresponded closely to each other and also to the original sung phrase. As further illustration, Fig. 6 displays the pitch tracings from the original sung phrase, together with those from four representative subjects in the *nonrepeat song* condition. (The tracings were taken from those subjects whose tracings in the *nonrepeat speech* condition are shown in Fig. 4.) It can be seen that the renditions in the *nonrepeat song* condition were more consistent across subjects and considerably closer to the original sung phrase than were those in the *nonrepeat speech* condition in relation to the original spoken phrase.

Two types of statistical comparison were made between renditions in these two conditions. First, it was found that the across-subjects variance in average F0 was considerably lower for renditions in the *nonrepeat song* condition than in the *nonrepeat speech* condition. Taking the average F0s for renditions of the entire phrase, the difference in variance was highly significant statistically [$F(10, 10) = 7.39$, $p < 0.01$]. This pattern held for all the syllables taken separately: for *some* $F(10, 10) = 19.89$, $p < 0.0001$; for *times* $F(10, 10) = 97.66$, $p < 0.0001$; for *be* $F(10, 10) = 5.31$, $p < 0.01$; for *have* $F(10, 10) = 21.63$, $p < 0.0001$; for *so* $F(10, 10) = 60.51$, $p < 0.0001$; for *strange* $F(10, 10) = 66.06$, $p < 0.0001$; and for *ly* $F(10, 10) = 17.74$, $p < 0.0001$.

Second, for each subject a difference score was obtained in the *nonrepeat song* condition, taking the difference between the average F0 of her rendition of the entire phrase and that of the original sung phrase. Using a correlated samples t-test, the difference scores in the *nonrepeat song* condition were found to be significantly lower than those in the *nonrepeat speech* condition [$t(10) = 3.31$, $p < 0.01$]. The same pattern held for each of the seven syllables taken individually, with the exception of the last one: for *some* $t(10) = -4.16$, $p < 0.01$; for *times* $t(10) = -7.56$, $p < 0.0001$; for *be* $t(10) = -5.69$, $p < 0.001$; for *have* $t(10) = -6.12$, $p < 0.001$; for *so* $t(10) = -2.41$, $p < 0.05$; for *strange* $t(10) = -6.6$, $p < 0.0001$; and for *ly* there was a nonsignificant trend in the same direction, $t(10) = -1.37$, $p = 0.200$.

The above findings are in accordance with the hypothesis that repeated listening to the original spoken phrase causes its pitches to be heard more saliently, and in this respect more appropriately to song than to speech.

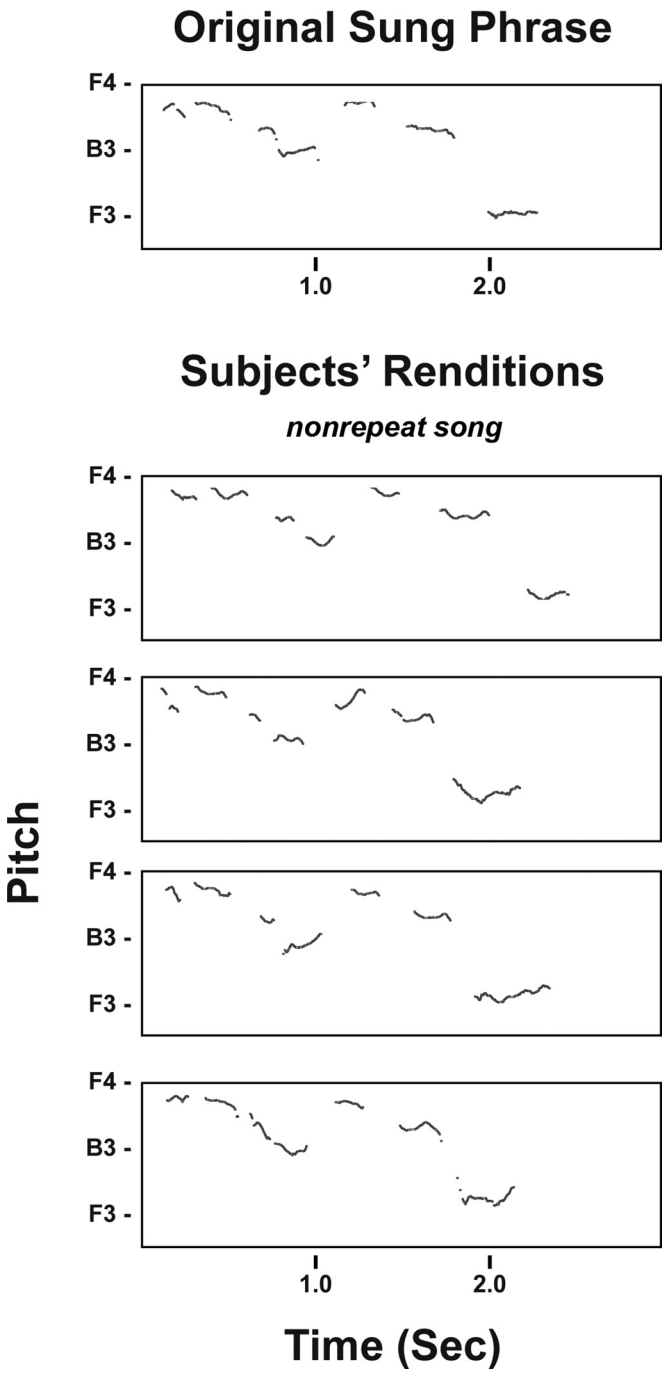

FIG. 6. Pitch tracings from the original sung phrase, together with those from four representative subjects in the *nonrepeat song* condition. These were the same subjects as whose pitch tracings, taken from renditions in the *nonrepeat speech* condition, are shown in Fig. 4. F3 = 174.6 Hz; B3 = 246.9 Hz; and F4 = 349.2 Hz.

Renditions in the *repeat speech* condition were considerably closer to the original and had considerably closer inter-subject agreement than were renditions in the *nonrepeat speech* condition. However, renditions in the *nonrepeat song* condition were very close to the original, with strong inter-subject agreement.

We now turn to the prediction that once the syllables constituting the original phrase were heard as forming salient pitches, these would also be perceptually distorted so as to conform to a tonal melody. Specifically, it was predicted that the subjects' renditions of the spoken phrase in the *repeat speech* condition would correspond to the pattern of pitches notated in Fig. 1 and so would correspond to the sequence of intervals (in semitones) of 0, −2, −2, +4, −2, and −7. It was therefore hypothesized that the sequence of intervals formed by the subjects' renditions of the spoken phrase in the *repeat speech* condition would be more consistent with this melodic representation than with the sequence of intervals formed by the original spoken phrase.

To test this hypothesis, the six melodic intervals formed by the original spoken phrase were calculated, as were those produced by the subjects' renditions in the *repeat speech* condition and the *nonrepeat speech* condition, taking the average F0 of each syllable as the basic measure. Then for each condition two sets of difference scores were calculated for each of the six intervals: (a) between the interval produced by the subject and that in the original spoken phrase, and (b) between the interval produced by the subject and that based on the hypothesized melodic representation. These two sets of difference scores were then compared statistically.

The results are shown in Table I. It can be seen that, as expected, the difference scores for renditions in the *nonrepeat speech* condition did not differ significantly for the two types of comparison. However, the renditions in the *repeat speech* condition were significantly closer to the hypothesized melodic representation than they were to the original spoken phrase. Specifically, in the *repeat speech* condition, for each of the six intervals, the difference between the subjects' renditions and the hypothesized melodic representation was smaller than between the subjects' renditions and the original spoken phrase. This difference between the two types of comparison was statistically significant ($p < 0.016$, one-tailed, on a binomial test). This result is in accordance with the hypothesis that the subjects' renditions following repeated listening to the spoken phrase would be heavily

TABLE I. Difference scores, averaged across subjects, between the intervals produced by the subjects' renditions and (a) produced by the original spoken phrase, and (b) based on the hypothesized melodic representation.

	Average difference (semitones)	
	(a) From original spoken phrase	(b) From melodic representation
Repeat speech condition		
some to *times*	0.89	0.75
times to *be*	0.63	0.41
be to *have*	0.68	0.43
have to *so*	1.42	0.55
so to *strange*	2.04	0.51
strange to *ly*	1.41	0.72
Nonrepeat speech condition		
some to *times*	1.85	1.53
times to *be*	1.63	1.31
be to *have*	0.91	1.20
have to *so*	2.16	2.53
so to *strange*	2.35	1.68
strange to *ly*	5.06	4.06

TABLE II. Difference scores, averaged across subjects, between the intervals produced by the subjects and (a) produced by the original sung phrase and (b) based on the hypothesized melodic representation.

	Average difference (semitones)	
	(a) From original sung phrase	(b) From melodic representation
Nonrepeat song condition		
some to *times*	0.46	0.52
times to *be*	0.40	0.43
be to *have*	0.35	0.35
have to *so*	0.39	0.43
so to *strange*	0.40	0.31
strange to *ly*	0.82	0.36

influenced by the hypothesized perceptual representation in terms of a tonal melody.

Table II shows the results of the same calculations that were carried out based on the original sung phrase, i.e., in the *nonrepeat song* condition. It can be seen that the difference scores were here very small and that they did not differ depending on whether comparison was made with the original sung phrase or with the hypothesized melodic representation. This result is expected on the assumption that the original sung phrase was itself strongly influenced by the melodic representation that had been constructed by the singer.

IV. DISCUSSION

In considering the possible basis for this transformation effect, we note that the vowel components of speech are composed of harmonic series, so that one might expect their pitches to be clearly perceived even in the absence of repetition. Yet in contrast to song, the pitch characteristics of speech are rarely salient perceptually. We can therefore hypothesize that in listening to the normal flow of speech, the neural circuitry underlying pitch salience is somewhat inhibited, perhaps to enable the listener to focus more on other characteristics of the speech stream that are essential to meaning, i.e., consonants and vowels. We can also hypothesize that exact repetition of the phrase causes this circuitry to become disinhibited, with the result that the salience of the perceived pitches is enhanced. Concerning the brain regions that might be involved here, brain imaging studies have identified a bilateral region in the temporal lobe, anterolateral to primary auditory cortex, which responds preferentially to sounds of high pitch salience (Patterson *et al.*, 2002; Penagos *et al.*, 2004; Schneider *et al.*, 2005). This leads to the prediction that, as a result of repeated listening to this phrase, activation in these regions would be enhanced.

The process underlying the perceptual transformation of the spoken phrase into a well-formed tonal melody must necessarily be a complex one, involving several levels of abstraction (Deutsch, 1999). At the lowest level, the formation of melodic intervals occurs (Deutsch, 1969; Demany and Ramos, 2005). This process involves regions of the temporal lobe that are further removed from the primary auditory cortex, with emphasis on the right hemisphere (Patterson *et al.*, 2002; Hyde *et al.*, 2008; Stewart *et al.*, 2008). Furthermore, in order for listeners to perceive the transformed phrase as a tonal melody, they must draw on their long term memories for familiar music. This entails projecting the pitch information onto overlearned scales (Burns, 1999) and invoking further rule-based characteristics of our tonal system (Deutsch, 1999; Deutsch and Feroe, 1981; Lerdahl and Jackendoff, 1983; Krumhansl, 1990; Lerdahl, 2001) and so requires the processing of musical syntax. Further brain regions must therefore also be involved in the perception of this phrase once it is perceived as song. Brain imaging studies have shown that regions in the frontal lobe in both hemispheres, particularly Broca's area and its homologue, are involved in processing musical syntax (Patel *et al.*, 1998; Maess *et al.*, 2001; Janata *et al.*, 2002; Koelsch *et al.*, 2002; Koelsch and Siebel, 2005). Furthermore, regions in the parietal lobe, particularly in the left supramarginal gyrus, have been found to be involved in the short term memory for musical tones (Schmithorst and Holland, 2003; Vines *et al.*, 2006; Koelsch *et al.*, 2009), as have other cortical regions, such as the superior temporal gyrus (Janata *et al.*, 2002; Koelsch *et al.*, 2002; Schmithorst and Holland, 2003; Warrier and Zatorre, 2004). We can therefore hypothesize that when the phrase is perceived as sung, there would be further activation in these regions also.

It should be pointed out that the subjects in the present study had all received musical training. It has been found that musically trained listeners are more sensitive to pitch structures than untrained listeners (Schneider *et al.*, 2002; Thompson *et al.*, 2004; Magne *et al.*, 2006; Musacchia *et al.*, 2007; Wong *et al.*, 2007; Kraus *et al.*, 2009; Hyde *et al.*, 2009), so it is possible that musically untrained subjects may not produce results as clear as those obtained in the present experiment.

Finally, the present findings have implications for general theories concerning the substrates of music and speech perception. As reviewed by Diehl *et al.* (2004) and Zatorre and Gandour (2007), much research in this area has been motivated by two competing theories. The domain-specific theory assumes that the sounds of speech and of music are each processed by a system that is dedicated specifically to processing these sounds and that excludes other sounds (Liberman and Mattingly, 1985; Peretz and Coltheart, 2003). The cue-based theory assumes instead that whether a stimulus is processed as speech, music, or some other sound depends on its acoustic characteristics and that it is unnecessary to posit special-purpose mechanisms for processing speech and music (Diehl *et al.*, 2004; Zatorre *et al.*, 2002). The present findings cannot be accommodated by a strong version of either theory, since the identical stimulus pattern is perceived convincingly as speech under some conditions and as music under others. It is proposed instead (similarly to the proposal advanced by Zatorre and Gandour, 2007) that speech and music are processed to a large extent by common neural pathways, but that certain circuitries that are specific either to speech or to music are ultimately invoked to produce the final percept.

V. CONCLUSION

In conclusion, we have described and explored a new perceptual transformation effect, in which a spoken phrase comes to be heard as sung rather than spoken, simply as a result of repetition. This effect is not just one of interpretation, since listeners upon hearing several repetitions of the phrase sing it back with the pitches distorted so as to give rise to a well-formed melody. Further research is needed to characterize the features of a spoken phrase that are necessary to produce this illusion, to document its neural underpinnings, and to understand why it occurs. Such research should provide useful information concerning the brain mechanisms underlying speech and song.

ACKNOWLEDGMENTS

We thank Adam Tierney, David Huber, and Julian Parris for discussions and Stefan Koelsch and an anonymous reviewer for helpful comments on an earlier draft of this manuscript.

Boersma, P., and Weenink, D. (**2006**). "Praat: Doing phonetics by computer (version 4.5.06)," http://www.praat.org/ (Last viewed December 8, 2010).

Burns, E. M. (**1999**). "Intervals, scales, and tuning," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic Press, New York), pp. 215–258.

Demany, L., and Ramos, C. (**2005**). "On the binding of successive sounds: Perceiving shifts in nonperceived pitches," J. Acoust. Soc. Am. **117**, 833–841.

Deutsch, D. (**2010**). "Speaking in tones," Sci. Am. Mind 21, 36–43.

Deutsch, D. (**2003**). *Phantom Words, and Other Curiosities* (Philomel Records, La Jolla) (compact disc; Track 22).

Deutsch, D. (**1999**). "Processing of pitch combinations," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic Press, New York), pp. 349–412.

Deutsch, D. (**1969**). "Music recognition," Psychol. Rev. **76**, 300–309.

Deutsch, D., and Feroe, J. (**1981**). "The internal representation of pitch sequences in tonal music," Psychol. Rev. **88**, 503–522.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Hyde, K. L., Peretz, I., and Zatorre, R. J. (**2008**). "Evidence for the role of the right auditory cortex in fine pitch resolution," Neuropsychologia **46**, 632–639.

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., and Schlaug, G. (**2009**). "Musical training shapes structural brain development," J. Neurosci. **29**, 3019–3025.

Janata, P., Birk, J. L., Horn, J. D. V., Leman, M., Tillmann, B., and Bharucha, J. J. (**2002**). "The cortical topography of tonal structures underlying Western music" Science **298**, 2167–2170.

Koelsch, S., Gunter, T. C., von Cramon, D. Y., Zysset, S., Lohmann, G., and Friederici, A. D. (**2002**). "Bach speaks: A cortical 'language-network' serves the processing of music," Neuroimage **17**, 956–966.

Koelsch, S., and Siebel, W.A. (**2005**). "Towards a neural basis of music perception," Trends Cogn. Sci. **9**, 578–584.

Koelsch, S., Schulze, K., Sammler, D., Fritz, T., Muller, K., and Gruber, O. (**2009**). "Functional architecture of verbal and tonal working memory: An fMRI study," Hum. Brain Mapp. **30**, 859–873.

Kraus, N., Skoe, E., Parbery-Clark, A., and Ashley, R. (**2009**). "Experience-induced malleability in neural encoding of *pitch, timbre and timing*," Ann. N. Y. Acad. Sci. **1169**, 543–557.

Krumhansl, C. L. (**1990**). *Cognitive Foundations of Musical Pitch* (Oxford University Press, New York), pp 1–318.

Lerdahl, F. (**2001**). *Tonal Pitch Space* (Oxford University Press, New York), pp. 1–411.

Lerdahl, F., and Jackendoff, R. (**1983**). *A Generative Theory of Tonal Music* (MIT Press, Cambridge, MA), pp. 1–368.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.

Maess, B., Koelsch, S., Gunter, T. C., and Friederici, A. D. (**2001**). "Musical syntax is processed in Broca's area: An MEG study," Nat. Neurosci. **4**, 540–545.

Magne, C., Schön, D., and Besson, M. (**2006**). "Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches," J. Cogn. Neurosci. **18**, 199–211.

Mottonen, R., Calvert, G. A., Jaaskelainen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., and Sams, M. (**2006**). "Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus," Neuroimage **30**, 563–569.

Musacchia, G., Sams, M., Skoe, E., and Kraus, N. (**2007**). "Musicians have enhanced subcortical auditory and audiovisual processing of speech and music," Proc. Natl. Acad. Sci. U.S.A. **104**, 15894–15898.

Patel, A. D., (**2008**). *Music, Language, and the Brain* (Oxford University Press, Oxford), pp. 1–513.

Patel, A. D., Peretz, I., Tramo, M. J., and Lebreque, R. (**1998**). "Processing prosodic and musical patterns: A neuropsychological investigation," Brain Lang. **61**, 123–144.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., and Griffiths, T. D. (**2002**). "The processing of temporal pitch and melody information in auditory cortex," Neuron **36**, 767–776.

Penagos, H., Melcher, J. R., and Oxenham, A. J. (**2004**). "A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging," J. Neurosci. **24**, 6810–6815.

Peretz, I., and Coltheart, M. (**2003**). "Modularity of music processing," Nat. Neurosci. **6**, 688–691.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.

Schmithorst, V. J., and Holland, S. K. (**2003**). "The effect of musical training on music processing: A functional magnetic resonance imaging study in humans," Neurosci. Lett. **348**, 65–68.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., and Rupp, A. (**2002**). "Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians," Nat. Neurosci. **5**, 688–694.

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Dosch, H. G., Bleeck, S., Stippich, C., and Rupp, A. (**2005**). "Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference," Nat. Neurosci. **8**, 1241–1247.

Schon, D., Magne, C., and Besson, M. (**2004**). "The music of speech: Music training facilitates pitch processing in both music and language," Psychophysiology **41**, 341–349.

Shtyrov, Y., Pihko, E., and Pulvermuller, F. (**2005**). "Determinants of dominance: Is language laterality explained by physical or linguistic features of speech?" Neuroimage **27**, 37–47.

Stewart, L., Von Kriegstein, K., Warren, J. D., and Griffiths, T. D. (**2006**). "Music and the brain: Disorders of musical listening," Brain **129**, 2533–2553

Stewart, L., Overath, T., Warren, J. D., Foxton, J. M., and Griffiths, T. D. (**2008**). "fMRI evidence for a cortical hierarchy of pitch pattern processing," PLoS ONE **3**, e1470.

Thompson, W. F., Schellenberg, E.G. and Husain, G. (**2004**). "Decoding speech prosody: Do music lessons help?" Emotion **4**, 46–64.

Vines, B. W., Schnider N. M., and Schlaug, G. (**2006**). "Testing for causality with transcranial direct current stimulation: Pitch memory and the left supramarginal gyrus," NeuroReport **17**, 1047–1050.

Warrier, C. M., and Zatorre, R. J. (**2004**). "Right temporal cortex is critical for utilization of melodic contextual cues in a pitch constancy task," Brain **127**, 1616–1625.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (**2007**). "Musical experience shapes human brainstem encoding of linguistic pitch patterns," Nat. Neurosci. **10**, 420–422.

Zatorre, R. J., and Gandour, J. T. (**2007**). "Neural specializations for speech and pitch: Moving beyond the dichotomies," Philos. Trans. R. Soc. London Ser. B **2161**, 1–18.

Zatorre, R. J., Belin, P., and Penhune, V. B. (**2002**). "Structure and function of auditory cortex: Music and speech," Trends Cogn. Sci. **6**, 37–46.