Cognition 254 (2025) 105933

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Speech-to-song transformation in perception and production

Yan Chen[a], Adam Tierney[b], Peter Q. Pfordresher[a,*]

[a] Department of Psychology, University at Buffalo, State University of New York, Buffalo, USA
[b] Department of Psychological Sciences, Birkbeck, University of London, London, UK

ARTICLE INFO

Keywords:
Speech-to-song transformation
Speech and music
Perception and production

ABSTRACT

The speech-to-song transformation is an illusion in which certain spoken phrases are perceived as more song-like after being repeated several times. The present study addresses whether this perceptual transformation leads to a corresponding change in how accurately participants imitate pitch/time patterns in speech. We used illusion-inducing (illusion stimuli) and non-inducing (control stimuli) spoken phrases as stimuli. In each trial, one stimulus was presented eight times in succession. Participants were asked to reproduce the phrase and rate how music-like the phrase sounded after the first and final (eighth) repetitions. The ratings of illusion stimuli reflected more song-like perception after the final repetition than the first repetition, but the ratings of control stimuli did not change over repetitions. The results from imitative production mirrored the perceptual effects: pitch matching of illusion stimuli improved from the first to the final repetition, but pitch matching of control stimuli did not improve. These findings suggest a consistent pattern of speech-to-song transformation in both perception and production, suggesting that distinctions between music and language may be more malleable than originally thought both in perception and production.

1. Introduction

Music and language are commonly considered clearly separable cognitive domains (Peretz & Coltheart, 2003), a distinction that may extend to production (Peretz, 2009). However, recent evidence suggests that the dividing line between speech and song can be modified by context, at least in perception. Some spoken phrases can transform perceptually from speech to song after being repeated several times (Deutsch et al., 2011), an effect referred to as an illusory *speech-to-song transformation,* that has been widely replicated (Castro et al., 2018; Deutsch et al., 2011; Falk et al., 2014; Jaisin et al., 2016; Margulis et al., 2015; Tierney et al., 2013; Tierney et al., 2018; Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015; Vitevitch et al., 2021). We report evidence that the experience of this transformation yields effects on production similar to those found for the imitation of naturally occurring speech versus song (Mantell & Pfordresher, 2013). These results complement claims that perception and action share common representations (e.g., Hommel, 2015; Wilson et al., 2005).

The initial report on the transformation from speech to song (Deutsch et al., 2011) included a study that addressed the association between perceptual transformation and changes in vocal production. One group of participants reproduced an illusion-inducing phrase after hearing it once, while the other group reproduced the same phrase after hearing it ten times. Participants who heard the phrase ten times reproduced the pitch values more accurately than those who heard it once, suggesting that the perceptual speech-to-song transformation facilitates imitative production. However, there were several limitations of this study. The authors only used one spoken phrase that was expected to induce the illusion, but no control phrases that fail to produce the illusion. The effect of the speech-to-song transformation could therefore not be disentangled from the effect of stimulus repetition. Furthermore, comparisons between performances following the first and final repetitions were based on different groups of participants. Finally, only musicians with at least five years of musical training participated, making it unclear whether results generalize to musically untrained individuals.

To investigate whether a transformation from speech to song in perception leads to a transformation from speech-like imitation to song-like imitation, we conducted a study using stimuli drawn from Tierney et al. (2013). That study identified stimuli that are likely to yield an illusory transformation (illusion stimuli) and others that do not (control stimuli). These two stimulus categories allow us to separate effects based on the speech-to-song transformation from basic effects of repetition. In the present experiment, participants heard each phrase 8 times. After the

[*] Corresponding author at: University at Buffalo, 362 Park Hall, Buffalo, NY 14260, United States of America.
E-mail address: pqp@buffalo.edu (P.Q. Pfordresher).

https://doi.org/10.1016/j.cognition.2024.105933
Received 19 January 2024; Received in revised form 20 August 2024; Accepted 21 August 2024

first and last repetition, participants were asked to vocally reproduce (imitate) the phrase, and then rate the phrase on a speech/song continuum. By comparing the accuracy of imitative production with perceptual ratings, we can investigate whether sensorimotor interaction accompanies the illusory perceptual transformation from speech to song. In general, the pitch patterns of song are imitated more accurately than those of speech (Mantell & Pfordresher, 2013; Pfordresher et al., 2022); however, it remains unclear whether this song advantage is driven by perception of a stimulus as song versus speech or by the acoustic characteristics which separate song and speech. Here we predicted that the speech-to-song transformation, found in illusion but not control stimuli, would be associated with a commensurate increase in pitch imitation accuracy from the first to last repetition for the illusion but not control stimuli, indicating that perceiving a stimulus as song leads to an enhanced ability to imitate its pitch.

2. Methods

2.1. Subjects

40 participants (20 female and 20 male) from the University at Buffalo subject pool participated in exchange for course credit. The average age of participants was 18.8 years (ranging from 18 to 23). Their average years of instrumental training was 3.35 (ranging from 0 to 15), and their average years of vocal training was 0.93 (ranging from 0 to 10). Twenty five of the subjects had at least one year of instrumental training and eleven of them had at least one year of vocal training. All subjects were native English speakers. Participants were excluded if they reported a medically diagnosed hearing disorder or disorder of vocal motor control. The procedure was approved by the Institutional Review Board of the University at Buffalo, and verbal informed consent was obtained from each participant.

2.2. Stimuli

Stimuli were short phrases selected from the illusion and control stimuli from Tierney et al. (2013), described earlier. Because the participants in our study spoke with an American English accent, only short phrases spoken with this accent were selected from the original stimulus set to avoid the potential challenge of imitating an unfamiliar accent. 12 illusion stimuli and 12 control stimuli were included on this basis. Acoustic differences across the subset we used mirrored those found for the entire original sample, as detailed in the Supplementary Information document. The phrases were spoken by three different talkers, with equal contributions of talker to each group of stimuli. We also included four filler stimuli from Tierney et al. (2018), in which the same talker first repeated a spoken phrase four times and then sung the same phrase another four times at the same rate and with similar pitches. Filler stimuli guard help prevent participants from shifting their ratings from “speech” to “song” based on mere repetition, by including trial sequences associated with changes in the acoustical signal. These stimuli also served as a check to make sure that subjects were paying attention to the speech to song changes.

Every talker in the original stimulus set used a male-gendered voice. This posed a problem for our production study given that prior work suggests that stimuli are imitated more accurately when they fall into participants’ vocal range (Pfordresher & Brown, 2007; Price, 2000; Welch, 1979). Therefore, we generated a new matched set of stimuli more suitable to female gendered voices by shifting the fundamental frequency one octave upward and adjusting formant frequencies in MATLAB (MathWorks, Inc., Natick, MA). Both male and female stimuli are available online (https://osf.io/j5xms/). Follow-up analyses of the results reported below showed no effect of male versus female stimuli on the strength of the speech-to-song transformation (no significant 3-way interaction between gender, stimulus type, and repetition). We therefore aggregated across vocal genders for sake of simplicity and maximizing statistical power.

2.3. Procedure

The experimenter interacted with participants via Zoom (Zoom Video Communications, San Jose, CA). Prior to the beginning of the experiment, participants were instructed to sit in a quiet place and to use headphones if possible. The experimenter checked the ambient noise level in the participant’s recording area during each session, and the experiment was rescheduled if the noise level was deemed too high. Specifically, the experimenter made sure no other media was playing in the background and the participant would not be disturbed by other people during the session. The experiment was run on an online data collection platform (Findingfive.com) and comprised three sections: speech-to-song task, pitch imitation, and questionnaires (Fig. 1).

2.4. Speech-to-song task

In the speech-to-song task, one of the spoken phrases was repeated eight times during each trial. After the first and eighth repetitions, participants were instructed to record themselves reproducing the phrase (production) and then provide a rating (perception) indicating the extent to which the phrase sounded like speech or song. A rating of 1 indicated that the phrase is completely speech-like, while a rating of 10 indicated that the phrase is completely song-like. Participants were instructed to start the recording by clicking a record icon and to stop the recording by clicking a square icon on the platform. From the second to the seventh repetitions, participants were only required to listen to the phrases, and the time interval between each repetition was 100 milliseconds. A practice trial was given at the beginning of the session to confirm that participants understood the instruction and that their microphones were working properly. Once the practice trial was successful, participants proceeded to the actual experiment. The illusion, control and filler stimuli were randomly intermingled across trials.

2.5. Pitch imitation task

After completing the speech-to-song task, participants were assessed on their abilities to match pitch using a subset of trials from the Seattle Singing Accuracy Protocol (SSAP; Demorest et al., 2015; Pfordresher & Demorest, 2020). On each trial, a single tone was presented for one second, and participants were asked to imitate the pitch after hearing the tone. The instruction for recording was the same as that in the speech-to-song task. Each tone used a human vocal timbre, matched to the vocal gender of the participant. This task utilized two different sets of tones spanning a musical perfect 5th (7 semitones). Specifically, 5 tones with voices in a typical male timbre and range were used for male participants (f_0: C3, D3, E3, F3, G3), whereas 5 tones with voices in a typical female timbre and range were used for female participants (f_0: C4, D4, E4, F4, G4). Each of the five tones was presented twice, and the tones for two successive trials were different.

2.6. Questionnaire

At the end of the experiment, the participants were asked to fill out a short questionnaire about their language and musical background.

2.7. Data analysis for vocal imitation

To evaluate the performance of imitation, the recordings of imitation were compared to the corresponding stimuli. During the preprocessing stage, recordings were eliminated before further analysis if the number of syllables in the recording did not match the number of syllables in the corresponding stimulus (7.6 % of the trials were removed).

The remaining audio files (including imitations and stimuli) were processed using MATLAB scripts following Mantell & Pfordresher

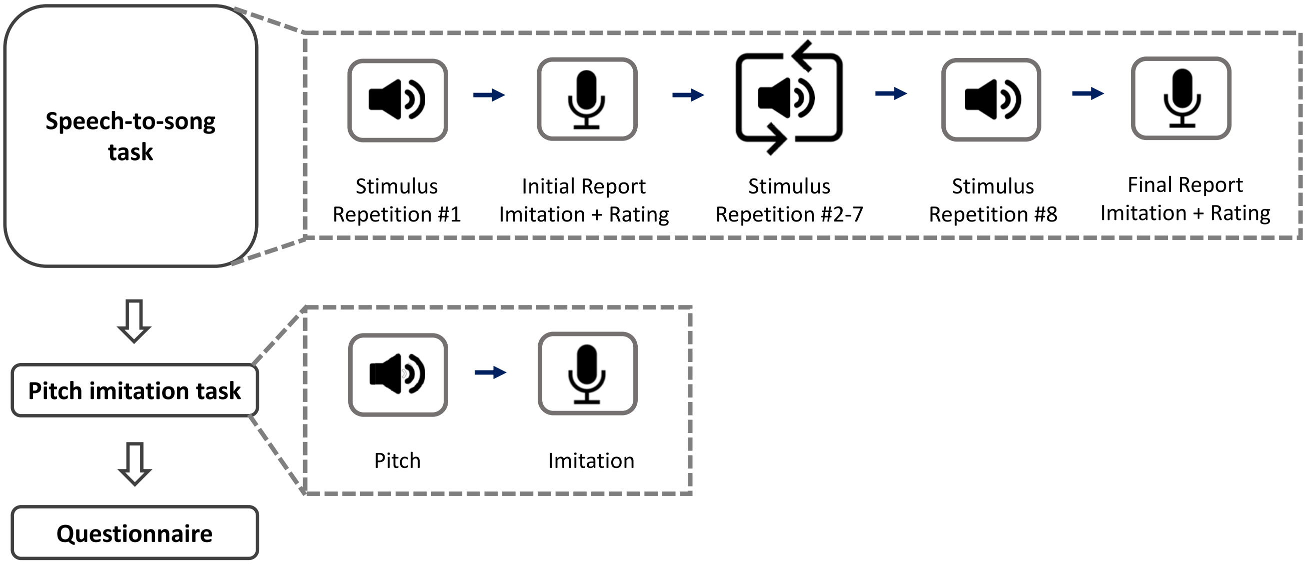

Fig. 1. Experimental procedure.
Note. Arrows indicate ordering of tasks across time. Dashed boxes indicate flow of tasks within single trials, where the speaker icon indicates perception and the microphone indicates production.

(2013). Fundamental frequencies (f_o) were extracted at each time point using the Matlab function Yin (De Cheveigné & Kawahara, 2002), resulting in a vector of f_o values sampled at an interval of 25 ms. All values were converted from Hz to cents, where 100 cents equal 1 semitone, based on a referent frequency of 440 Hz. Next, paired recordings of target stimuli and the corresponding imitations were equated for duration. This was done by resampling the target f_o vector so that its length matched the length of f_o vector from the imitation.[1]

Two measures were used to assess the pitch accuracy of imitation based on temporally aligned f_o vectors. *Absolute pitch error* is the mean absolute difference between the target and imitation vectors across all duration-matched samples in a trial. *Pitch correlation* is the Pearson correlation between matched imitation and target samples in a trial, and measures how closely the pattern of change in the imitated pitch trajectory corresponds to the pattern of change in the target (i.e., relative pitch).

Statistical analyses were performed with a 2 (Stimulus Type: illusion versus control stimuli) x 2 (Repetition: first versus final repetitions) repeated measures ANOVA. Prior work suggested the increase in rating should be found only for the illusion stimuli and not for the control stimuli (Tierney et al., 2018). Therefore, planned contrasts were also conducted with independent samples *t*-tests between the first and final repetitions. All statistical decisions were made with $\alpha = 0.05$.

3. Results

3.1. Perceptual ratings of speech versus song

The average initial and final ratings for control and illusion stimuli are displayed in Fig. 2A. Difference scores between the initial and final ratings for control and illusion stimuli are shown in Fig. 2B. The ANOVA yielded a significant main effect of Stimulus Type, $F(1, 39) = 53.54$, $p < .001$, $\eta_p^2 = 0.58$, indicating that the illusion stimuli ($M = 3.91$, $SD = 1.69$) were rated as more song-like compared to the control stimuli ($M = 2.78$, $SD = 1.44$) across repetitions. There was also a significant main effect of Repetition, $F(1, 39) = 34.30$, $p < .001$, $\eta_p^2 = 0.47$, indicating that ratings increased with repetition ($M_{\text{first}} = 3.03$, $SD_{\text{first}} = 1.45$; $M_{\text{final}} = 3.66$, $SD_{\text{final}} = 1.81$). In addition, there was a significant Stimulus Type x Repetition interaction, $F(1, 39) = 35.70$, $p < .001$, $\eta_p^2 = 0.48$. Planned contrast analyses revealed that for illusion stimuli the mean final ratings ($M = 4.48$, $SD = 1.68$) were significantly higher than the mean initial ratings ($M = 3.35$, $SD = 1.52$; $t(39) = 7.03$, $p < .001$), while for control stimuli there was no significant difference between the mean initial ($M = 2.71$, $SD = 1.31$) and final ratings ($M = 2.84$, $SD = 1.57$; $t(39) = 1.19$, $p = .12$). These results suggest a perceptual transformation from speech to song for the illusion stimuli but not for the control stimuli, replicating the findings of Tierney et al. (2018).

3.2. Absolute pitch error for imitations

Fig. 3A displays the mean absolute pitch errors of the phrase imitations across four Stimulus Type x Repetition conditions. Fig. 3B shows the differences in mean absolute errors across repetitions. There was a significant main effect of Stimulus Type, $F(1, 39) = 67.08$, $p < .001$, $\eta_p^2 = 0.63$, indicating that the absolute pitch errors for illusion stimuli ($M = 295.06$, $SD = 105.85$) were lower than for control stimuli ($M = 395.58$, $SD = 102.21$). There was also a main effect of Repetition, $F(1, 39) = 11.12$, $p = .002$, $\eta_p^2 = 0.22$, indicating that the absolute pitch errors decreased from the first to the final repetition ($M_{\text{first}} = 353.08$, $SD_{\text{first}} = 110.91$, $M_{\text{final}} = 337.56$, $SD_{\text{final}} = 119.77$). The Stimulus Type x Repetition interaction was not significant, $F(1, 39) = 1.70$, $p = .20$, $\eta_p^2 = 0.04$. However, planned contrasts indicated that the absolute pitch errors decreased significantly from the first to the final repetition for illusion stimuli, $t(39) = 3.61$, $p = .005$, but not for the control stimuli, $t(39) = 0.97$, $p = .175$.

3.3. Pitch correlation for imitations

Fig. 4A shows pitch correlations of the phrase imitations across four Stimulus Type x Repetition conditions. Fig. 4B displays the differences in pitch correlation scores across repetitions. The ANOVA revealed a significant main effect of Stimulus Type, $F(1, 39) = 59.59$, $p < .001$, $\eta_p^2 = 0.60$, indicating that pitch correlation scores for illusion stimuli ($M = 0.45$, $SD = 0.18$) were greater than for control stimuli ($M = 0.29$, $SD = 0.17$). There was also a main effect of Repetition, $F(1, 39) = 7.30$, $p = .010$, $\eta_p^2 = 0.16$, indicating an increase in pitch correlation scores from

[1] The Supplementary Information document reports analyses that address possible effects of the alignment process, which ultimately did not change the interpretation of results reported here.

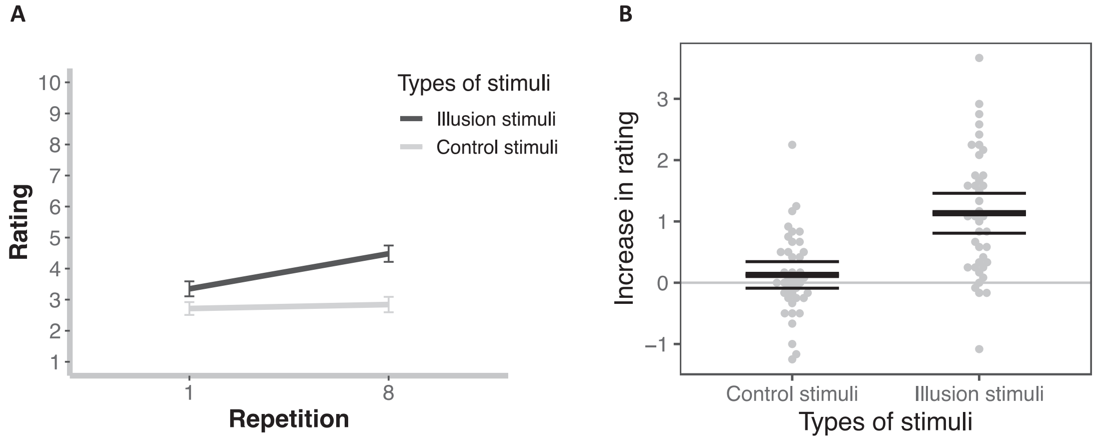

Fig. 2. Ratings for the first and final repetitions.

Note. A: Mean ratings for the 1st and 8th repetitions averaged across participants for illusion stimuli (black line) and control stimuli (gray line). Error bars represent one standard error of the mean. B: Swarm charts displaying the differences in ratings (final ratings minus initial ratings) for control and illusion stimuli. Dark horizontal lines in each panel represent means surrounded by 95 % confidence intervals, and each dot represents the mean difference score for a single participant.

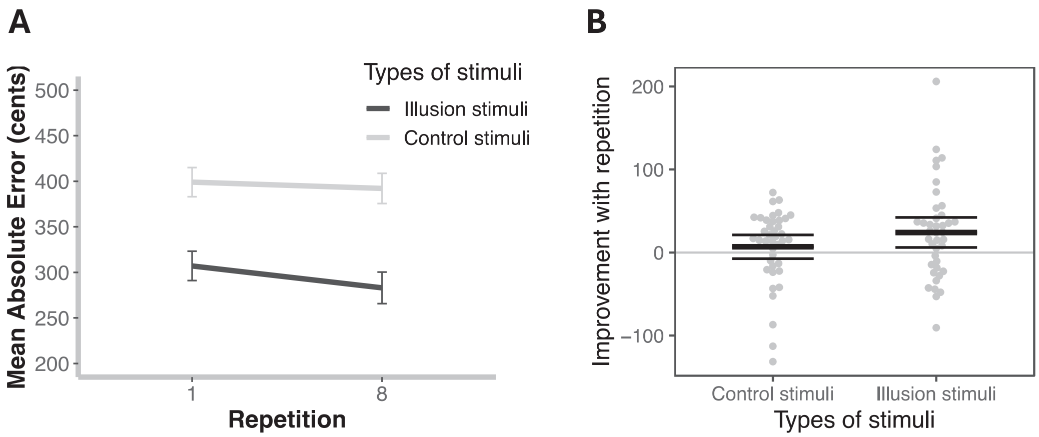

Fig. 3. Absolute pitch error.

Note. A: Absolute pitch error (in cents) for the 1st and 8th repetitions averaged across participants for illusion stimuli (black line) and control stimuli (gray line). Error bars represent one standard error of the mean. B: Swarm charts displaying differences in mean absolute errors across repetitions (Absolute pitch errors of initial recordings minus absolute pitch errors of final recordings). Dark horizontal lines in each panel represent means surrounded by 95 % confidence intervals, and each dot represents the mean difference score for a single participant.

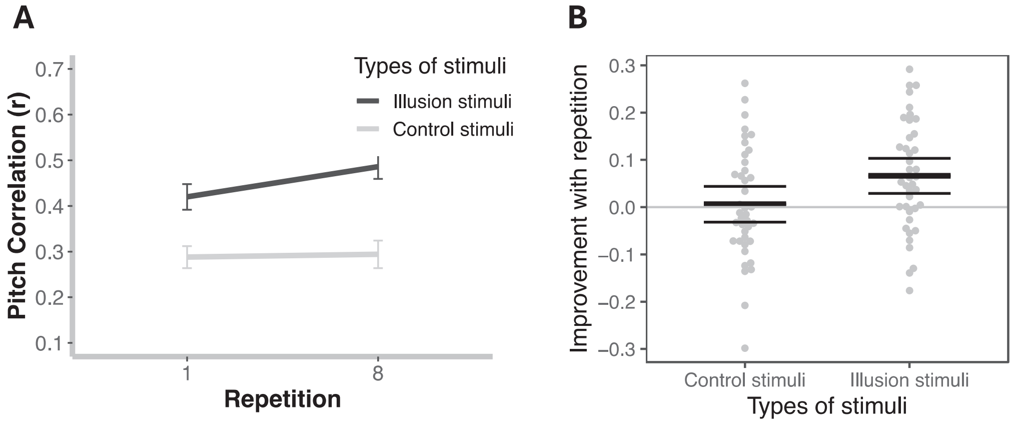

Fig. 4. Pitch correlation.

Note. A: Mean pitch correlations for the 1st and 8th repetitions averaged across participants for illusion stimuli (black line) and control stimuli (gray line). Error bars represent one standard error of the mean. B: Swarm charts displaying differences in pitch correlation across repetitions (pitch correlation of final recordings minus pitch correlation of initial recordings). Dark horizontal lines in each panel represent means surrounded by 95 % confidence intervals and each dot represents the mean difference score for a single participant.

the first to the final repetition ($M_{\text{first}} = 0.35$, $SD_{\text{first}} = 0.18$; $M_{\text{final}} = 0.39$, $SD_{\text{final}} = 0.20$). Additionally, there was a significant Stimulus Type x Repetition interaction, $F(1, 39) = 5.49$, $p = .024$, $\eta_p^2 = 0.12$, suggesting the improvement in pitch correlation scores was greater for illusion stimuli than for control stimuli. This interpretation was supported by planned contrasts analyses, in that pitch correlations increased significantly from the first to the final repetition for illusion stimuli, $t(39) = 3.62$, $p < .001$, but not for control stimuli, $t(39) = 0.32$, $p = .373$.

3.4. Correlational analyses

We next assessed the association between perception and production on a more granular level using correlational analyses, looking at the association between the degree of change from repetitions one to eight in perceptual ratings, with the commensurate degree of change in imitation accuracy (via each measure). When parameterized by individual participant (i.e., each data point is the mean score of a participant across items), the association between perception and production was not significant for either measure of production. However, both of these associations were significant when correlations were parameterized by item (cf. Tierney et al., 2018, their Fig. 3), for pitch deviation (Fig. 5A), $r(22) = 0.53$, $p = .004$, for pitch correlation (Fig. 5B), $r(22) = 0.42$, $p = .021$. We also evaluated whether any demographic variables related to instrumental or vocal training correlated with change in perceptual ratings or imitative performance among illusion stimuli. The only significant association we found was between years of instrumental training and improvement in production measured via pitch correlations, $r(39) = 0.34$, $p = .04$, suggesting that participants with more years of training exhibited a larger effect of repetition within illusion-generating stimuli with respect to tracking relative pitch.

4. Discussion

This study reports a replication of the perceptual speech-to-song transformation and an extension of this effect to the accuracy with which pitch contours in speech are imitated. These results suggest that the illusory transformation found in perception also exerts an effect on sensorimotor associations that influence vocal-motor planning. The present study is thus consistent with frameworks advocating for the integration of perception and action (e.g., Hommel et al., 2001; MacKay, 1987; Pfordresher, 2019; Pickering & Garrod, 2013). For example, in a previous neuroimaging study, Tierney et al. (2013) showed that the perception of the speech-to-song transformation is linked to increased activation in a motor region associated with vocalization. The significance of the present effect is that the associations found here are based on phenomenological aspects of perception (i.e., perceiving a stimulus as more representative of language or music), beyond effects related to acoustic structure or practice. These results also suggest that the advantage in imitating song over speech (e.g., Mantell & Pfordresher, 2013) may not simply reflect differences in acoustic features across domains.[2] Taken together, certain acoustical parameters may lend flexibility to certain acoustical parameters, such that manipulations like repetition (used here) can cause items to vary phenomenologically between song and speech. The fact that these phenomenological changes affect production is the novel contribution here.

Correlational analyses suggested an association between perception and production at the item level. Furthermore, the magnitude of the effect of repetition on perceptual ratings scaled with the magnitude of the effect on production, for both measures of imitation accuracy. These effects suggest that items within the two stimulus categories reported here fall on a perceptual continuum between speech and song which is additionally associated with graded effects on pitch production. The continuum is largely defined by acoustic variables such as pitch stability and rhythmic regularity (see Supplementary Results for more analyses of these variables). Other correlational analyses, however, did not yield robust results. In particular, correlations based on individual differences (where the regression is parameterized by participant rather than by item), were not significant. The difference between group-level and individual-level association may reflect the combination of shared versus unshared factors that contribute to perception during production. For instance, various models predict that different factors contribute to perception used for explicit decision making (such as a speech versus song categorization tasks) as opposed to the more implicit role perception has in our imitative production task (cf., Hutchins & Moreno, 2013; Loui, 2015). Following Tierney et al. (2018), we suggest that the graded transformation effect across items in both tasks follows from listeners' ability to detect music-like features in speech, whereas individual differences are based on additional task-specific features such as response biases and internal category boundaries, for perception, and vocal motor control, for production.

The current results also add to previous studies that have explored the influence of musical background on the speech-to-song transformation. Like Vanden Bosch der Nederlanden et al. (2015), we found that the magnitude of the perceptual transformation effect was not significantly correlated with years of musical training. This is analogous to other research showing that individual differences in musical background (Tierney et al., 2021; Vanden Bosch der Nederlanden et al., 2015) and tonal language background (Kachlicka et al., 2024) are not significantly correlated with differences in the magnitude of the transformation. However, years of training did predict the magnitude of the transformation effect on the accuracy of relative pitch in imitation (viz. the pitch correlation). This distinction suggests a subtle differentiation between perception and production in which musical training may influence sensorimotor integration of pitch perception and production. This finding should be interpreted with caution because we did not correct for multiple comparisons here.

One possible explanation for the improved imitation for illusion-generating stimuli after repetition is that listeners engaged in tonal encoding of pitches when they experienced the perceptual shift, a conclusion suggested by Deutsch et al. (2011). Tonal encoding is associated with greater precision of pitch processing for with music as opposed to speech (Patel, 2011, 2014; Zatorre et al., 2002) and may be a hallmark of music-specific neural processing (Peretz & Coltheart, 2003). This explanation is also consistent with the fact that the illusion stimuli are more open to tonal encoding based on having more stable pitches and pitches that more closely approximate Western tonal scales than the control stimuli. However, post-hoc analyses of produced pitch (suggested by an anonymous reviewer) did not support this explanation. In fact, produced pitches were less consistent with Western tonal hierarchies after 8 repetitions than after the first repetition of a phrase, and this tendency was found for both illusion and control stimuli (there was no interaction with stimulus type). Details on this analysis can be found in the Supplementary Information document. Thus, improved pitch matching after the speech-to-song transformation may not reflect tonal encoding based on Western prototypes but instead may reflect upweighting of pitch precision.

The motivation for this study was based in part on previous evidence for an advantage in imitating sung pitch patterns in comparison to patterns of pitch used in speech (a.k.a. the song advantage, Mantell & Pfordresher, 2013; Pfordresher et al., 2022). The present study offered a new opportunity to determine whether song associations that are independent of acoustic structure can lead to changes in performance akin to the song advantage. The fact that pitch imitation can improve simply based on the phenomenology of perception, beyond effects related to acoustic structure, is surprising in the context of previous research. It is important to note, however, that effects on production here are not directly analogous to those seen in other studies that contrasted stimuli

[2] This holds even if one considers repetition to be an acoustic feature (a possibility that an anonymous reviewer proposed) given that repetition led to improved imitation for illusion but not control stimuli.

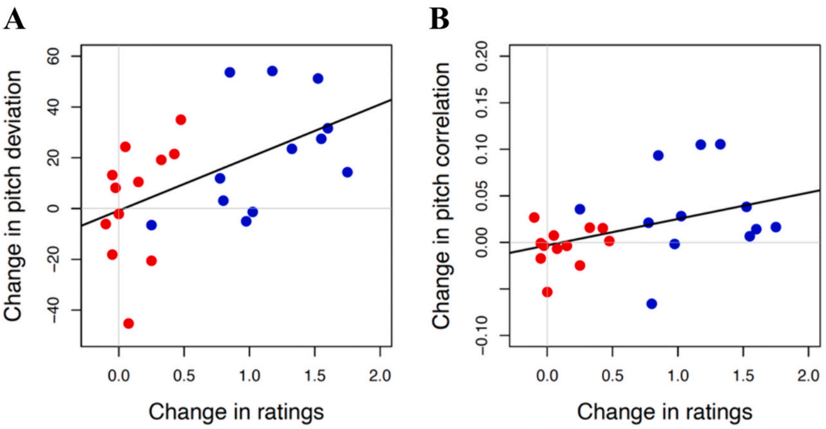

Fig. 5. Associations between perception and production.

Note. All values reflect change from repetitions 1 to 8, averaged across all participants for a given stimulus item. Terms in the differences are arranged so that positive values reflect increased "song-like" ratings or improved imitative production and light gray lines highlight zero crossings. Red dots denote control stimuli and blue dots are illusion stimuli. Panels differ with respect to the Y-variable, A: change in pitch deviation scores, B: change in pitch correlation scores.

with different acoustic structures. First, the effect magnitude seen here is subtler than what has been found elsewhere. Here, we found that pitch deviations for illusion stimuli improved by approximately 20 cents from the first to the final repetition, whereas the song advantage in other studies is nearly 80 cents (Pfordresher, 2022, Table 1). Second, whereas the song advantage found earlier tends to be more strongly associated with absolute than relative pitch deviations, the opposite was found here given the presence of a significant Stimulus Type x Repetition interaction for pitch correlations but not pitch error.

In closing, our study presents a novel finding that speech-to-song transformation in perception is associated with related changes to the accuracy of imitative production. Future research could explore the role of pitch perception in the speech-to-song transformation, providing a better understanding of the perception-action loop associated with this phenomenon.

CRediT authorship contribution statement

Yan Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adam Tierney:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Peter Q. Pfordresher:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Data availability

Data and stimuli are available at https://osf.io/j5xms/.

Acknowledgement

We thank Andreane Syra Cruz and Bidya Dahal for assistance with data collection. We thank Renee Branche, Emma Boiano, Jianghan Ren, and Shreeja Yarabolu for assistance with data processing. This work was funded in part by NSF grant BCS-1848930.

Appendix A. Supplementary data

Supplementary results for this article can be found online at https://doi.org/10.1016/j.cognition.2024.105933.

References

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), Article e0198656. https://doi.org/10.1371/journal.pone.0198656

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917–1930. https://pubs.aip.org/asa/jasa/article-abstract/111/4/1917/547221/YIN-a-fundamental-frequency-estimator-for-speech?redirectedFrom=fulltext.

Demorest, S. M., Pfordresher, P. Q., Bella, S. D., Hutchins, S., Loui, P., Rutkowski, J., & Welch, G. F. (2015). Methodological perspectives on singing accuracy: An introduction to the special issue on singing accuracy (part 2). *Music Perception: An Interdisciplinary Journal, 32*(3), 266–271.

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

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

Hommel, B. (2015). The theory of event coding (TEC) as embodied-cognition framework. *Frontiers in Psychology, 6*, 1318. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4554939/pdf/fpsyg-06-01318.pdf.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849–937.

Hutchins, S., & Moreno, S. (2013). The linked dual representation model of vocal perception and production [Hypothesis & Theory]. *Frontiers in Psychology, 4*, 825. https://doi.org/10.3389/fpsyg.2013.00825

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

Kachlicka, M., Patel, A. D., Liu, F., & Tierney, A. (2024). Weighting of cues to categorization of song versus speech in tone-language and non-tone-language speakers. *Cognition, 246*, Article 105757. https://doi.org/10.1016/j.cognition.2024.105757

Loui, P. (2015). A dual-stream neuroanatomy of singing. *Music Perception, 32*(3), 232–241. https://doi.org/10.1525/mp.2015.32.3.232

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

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of song and speech. *Cognition, 127*(2), 177–202. https://doi.org/10.1016/j.cognition.2012.12.008

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

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis [Hypothesis & Theory]. *Frontiers in Psychology, 2*, 142. https://doi.org/10.3389/fpsyg.2011.00142

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research, 308*, 98–108. https://doi.org/10.1016/j.heares.2013.08.011

Peretz, I. (2009). Music, language and modularity framed in action. *Psychologica Belgica, 49*(2–3), 157–175.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6*(7), 688–691. https://www.nature.com/articles/nn1083.

Pfordresher, P. Q. (2019). *Sound and action in music performance.* Academic Press.

Pfordresher, P. Q. (2022). A reversal of the song advantage in vocal pitch imitation. *JASA Express Letters, 2*(3), Article 034401. https://doi.org/10.1121/10.0009729

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of" tone deafness". *Music Perception, 25*(2), 95–115. https://doi.org/10.1525/mp.2007.25.2.95

Pfordresher, P. Q., & Demorest, S. M. (2020). Construction and validation of the Seattle singing accuracy protocol (SSAP): An automated online measure of singing accuracy. In *The Routledge companion to interdisciplinary studies in singing* (pp. 322–333). Routledge.

Pfordresher, P. Q., Mantell, J. T., & Pruitt, T. A. (2022). Effects of intention in the imitation of sung and spoken pitch. *Psychological Research, 86*(3), 792–807. https://doi.org/10.1007/s00426-021-01527-0

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 1–64.

Price, H. E. (2000). Interval matching by undergraduate nonmusic majors. *Journal of Research in Music Education, 48*(4), 360–372. https://doi.org/10.2307/3345369

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.

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

Vitevitch, M. S., Ng, J. W., Hatley, E., & Castro, N. (2021). Phonological but not semantic influences on the speech-to-song illusion. *Quarterly Journal of Experimental Psychology, 74*(4), 585–597. https://doi.org/10.1177/1747021820969144

Welch, G. F. (1979). Vocal range and poor pitch singing. *Psychology of Music, 7*(2), 13–31. https://doi.org/10.1177/030573567972002

Wilson, A. D., Collins, D. R., & Bingham, G. P. (2005). Perceptual coupling in rhythmic movement coordination: Stable perception leads to stable action. *Experimental Brain Research, 164*, 517–528. https://link.springer.com/article/10.1007/s00221-005-2272-3.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6*(1), 37–46. http://www.elsevier.com/inca/publications/store/6/0/0/3/5/6/index.htt.