Acoustic and musical foundations of the speech/song illusion

Adam Tierney,[*1] Aniruddh Patel[#2], Mara Breen[^3]

[*]*Department of Psychological Sciences, Birkbeck, University of London, United Kingdom*
[#]*Department of Psychology, Tufts University, United States*
[^]*Department of Psychology and Education, Mount Holyoke College, United States*
[1]a.tierney@bbk.ac.uk, [2]a.patel@tufts.edu, [3]mbreen@mtholyoke.edu,

ABSTRACT

In the 'speech-to-song illusion', certain spoken phrases sound like song when isolated from context and played repeatedly. Previous work has shown that this perceptual transformation occurs more readily for some phrases than others, suggesting that the switch from speech to song perception depends in part on certain cues. We conducted three experiments to explore how stimulus characteristics affect the illusion. In Experiment 1, we presented 32 participants with a corpus of 24 spoken phrases which become more song-like when repeated and 24 spoken phrases which continue to sound like speech when repeated. After each of 8 repetitions participants rated the extent to which the phrase sounded like song versus speech. Regression modeling indicated that an increase in song perception between the first and eighth repetitions was predicted by a) greater stability of the pitches within syllables, b) a better fit of average syllable pitches to a Bayesian model of melodic structure, and c) less variability in beat timing, as extracted by a beat-tracking algorithm. To investigate whether pitch characteristics play a causal role in the speech-to-song transformation, we elicited ratings of the stimuli from Experiment 1 after manipulating them to have larger pitch movements within syllables (Experiment 2, n = 27) or to have average pitches of syllables which resulted in poorer melodic structure (Experiment 3, n = 31). Larger pitch movements within syllables did not decrease the size of the illusion compared to Experiment 1; however, the illusion was significantly weaker when the intervals between pitches were altered. These results suggest that the strength of the illusion is determined more by pitch relationships between than within syllables, such that phrases with pitch relationships between spoken syllables that resemble those of Western tonal music are more likely to perceptually transform than those that do not.

I. INTRODUCTION

Music can take on a wide variety of forms, even within Western culture. Musical genres are marked by large differences in the timbral, rhythmic, melodic, and harmonic patterns upon which composers and musicians draw. Given this variety of musical subcultures, one might expect there to be little agreement across the general population as to the qualities that cause a sound sequence to be perceived as more or less musical.

However, there exist certain spoken recordings that listeners tend to perceive as sounding like song when isolated from context and repeated (Deutsch 2011). Across subjects, this transformation is stronger for some recordings than others (Tierney et al. 2013), and musicians and non-musicians agree as to which examples do and do not transform (Vanden Bosch der Nederlanden et al. 2015a, 2015b). The existence of this illusion suggests that there are certain cues on which listeners tend to rely when judging the musicality of a linguistic phrase, and that these cues are relatively unaffected by musical experience. The fact that repetition is necessary for the speech-song transformation to take place also suggests that a certain amount of time is necessary for the detection of at least a subset of these cues.

What, then, are the minimal characteristics which need to be present in order for a linguistic phrase to sound musical? One explanation—the acoustic cue theory--is that these characteristics are primarily acoustic in nature. For example, the speech-to-song transformation takes place more often for stimuli that have relatively flat pitches within syllables (Tierney et al. 2013, Falk et al. 2014). This may facilitate the categorization of pitches into pitch classes, enabling listeners to perceive the pitch sequence as a melody—with the result that the pitches that listeners perceive are distorted from the original pitches of the sequence (Vanden Bosch der Nederlanden et al., 2015b). Thus, listeners may tend to hear any sequence of flat pitches as musical. Supporting this account is the fact that random sequences of pitches are rated as more musical if they have been repeated (Margulis and Simchy-Gross 2016). Furthermore, repetition may be necessary for the illusion to take place because exact repetition causes speech perception resources to become satiated, allowing musical perception to take over. This would explain why the speech-song transformation is stronger for languages that are more difficult for a listener to pronounce (Margulis et al. 2015).

A second possibility—the musical cue theory—is that basic acoustic cues such as flat pitches within syllables are necessary but not sufficient for the perception of a spoken sentence as song. According to this account the speech-to-song illusion would only occur for spoken sentences which feature both acoustic pre-requisites such as pitch flatness and musical cues matching certain basic characteristics of Western music. For example, a sequence featuring relatively flat pitches but an abundance of tritone intervals may be unlikely to be perceived as music, because tritone intervals are rare in Western music. This theory offers an alternate (but not exclusive) explanation for the necessity of repetition for eliciting the illusion, as a certain amount of time may be necessary for the detection of some or all of these musical cues. This theory suggests not only that listeners across the general population can make relatively sophisticated musical judgments (Bigand and Poulin-Charronnat 2006), but that they can make these judgments about non-musical stimuli.

Here we tested the musical cue theory of the speech-song illusion using the speech-song corpus first reported by Tierney et al. (2013). This corpus consists of 24 stimuli which listeners perceive as song after repetition (Illusion stimuli) and 24 stimuli which listeners perceive as speech after repetition (Control stimuli). We repeated each stimulus eight times and asked a new set of listeners with a relatively small amount of musical training to rate how much the stimulus sounded like

song after each repetition. We measured the musical characteristics of each stimulus using computational models of melodic structure (Temperley 2007) and musical beat structure (Ellis 2007). Our prediction was that these musical characteristics would explain additional variance in the extent to which each stimulus transformed into song with repetition, even after syllable pitch flatness was accounted for. Furthermore, we predicted that musical characteristics would correlate with the change in song ratings due to repetition but not with song ratings after a single repetition.

II. Experiment 1

A. Methods

1) Participants. 32 participants were tested (14 male). The mean age was 33.7 (standard deviation 9.4). The mean amount of musical training was 1.9 years (standard deviation 3.8).

2) Stimuli. Stimuli consisted of the 24 Illusion stimuli and 24 Control stimuli described in Tierney et al. (2013). These were short phrases (mean 6.6 (sd 1.5) syllables) extracted from audiobooks.

3) Procedures. Participants were recruited via Mechanical Turk, a website which enables the recruitment of participants for internet-based work of various kinds. This study was conducted online using the Ibex paradigm for internet-based psycholinguistics experiments (http://spellout.net/ibexfarm). Participants were told that they would hear a series of spoken phrases, each repeated eight times. After each presentation they were given three seconds to indicate, on a scale from 1 to 10, how much the stimuli sounded like song versus like speech. Judgments could be made either by clicking on boxes which contained the numbers or by pressing the number keys. Order of presentation of the stimuli was randomized. Participants also heard, interspersed throughout the test, four "catch" trials in which the stimulus actually changed from a spoken phrase to a sung phrase between the fourth and fifth repetitions. Data from participants whose judgments of the sung portions of the catch trials were not at least 1.5 points higher than their judgments of the spoken portions were excluded from analysis. This constraint resulted in the exclusion of 8 participants from Experiment 1, 7 participants from Experiment 2, and 8 participants from Experiment 3. Excluded participants are not included in the subject counts of demographic descriptions within the Participants section for each Experiment.

4) Analysis. To minimize the influence of inter-subject differences on baseline ratings, ratings were normalized prior to analysis with the following procedure: each subject's mean rating across all repetitions for all stimuli was subtracted from each data point for that subject. Data were then averaged across stimuli within a single subject. This generated, for each subject, average scores for each repetition for Illusion and Control stimuli. A repeated measures ANOVA with two within-subjects factors (condition, two levels; repetition, eight levels) was then run; an interaction between repetition and condition was predicted, indicating that the Illusion stimuli transformed to a greater extent than the Control stimuli after repetition.

To investigate the stimulus factors contributing to the speech-song illusion, for each stimulus initial and final ratings were averaged across all subjects. The initial rating and the rating change between the first and last repetition were then calculated and correlated with several stimulus characteristics. First, we measured the average pitch flatness within syllables by manually marking the onset and offset of each syllable, tracking the pitch contour using autocorrelation in Praat, calculating the change in pitch between each time point (in fractions of a semitone), and then dividing by the length of the syllable. Thus, pitch flatness was measured in semitones per second.

Second, we determined the best fit of the pitches in each sequence to a diatonic key using a Bayesian key-fitting algorithm (Temperley, 2007) which evaluates the extent to which pitch sequences fit characteristics of melodies from western tonal music by assessing them along a number of dimensions, including the extent to which the distribution of interval sizes fits the interval size distribution of Western vocal music and fit to tonal key profiles.

The model considers four sets of probabilities—*key profile, central pitch profile, range profile,* and *proximity profile*—which are all empirically generated from the Essen Folksong Collection, a corpus of 6217 European folk songs. The key profile is a vector of 12 values indicating the probability of occurrence for each of the 12 scale tones in a melody from a specific key, normalized to sum to 1. For example, on average 18.4% of the notes in a melody in a Major key are scale degree 1 (e.g., C in C Major), while only 0.1% are scale degree #1 (e.g., C# in C Major). The central pitch profile (c) is a normal distribution of pitches represented as integers (C4 = 60) with a mean of 68 (Ab4) and variance of 13.2, which captures the fact that melodies in the Essen corpus are centered within a specific pitch range. The range profile is a normal distribution with a mean of the first note of the melody, and variance of 29, which captures the fact that melodies in the Essen corpus are constrained in their range. The proximity profile is a normal distribution with a mean of the previous note, and variance of 8.69, which captures the fact that melodies in the Essen corpus have relatively small intervals between adjacent notes. The final parameter of the model is the *RPK profile*, which is calculated at each new note as the product of the key profile, range profile, and proximity profile. The inclusion of the RPK profile captures the fact that specific notes are more probable after some tones than others.

Calculating the probability of each of the 24 (major and minor) diatonic keys given a set of notes is done using the equation below:

$$P(pitch\ sequence) = \sum_{k,c}\left(P(k)P(c)\prod_{n}RPK_n\right)$$

P(k) is the probability of any key (k) being chosen (higher for major than minor keys), *P(c)* is the probability of a central pitch being chosen, and RPK_n is the RPK profile value for all pitches of the melody given the key, central pitch, and prior pitch for each note. We defined melodic structure as the best fit of each sequence to the key (k) that maximized key fit in the equation.

Finally, we used a computer algorithm designed to find beat times in music (Ellis 2007) to determine the location of each

beat, and then we calculated the standard deviation of inter-beat times to measure beat regularity.

The beat tracking algorithm works as follows. First, it divides a sound sequence into 40 equally-spaced Mel frequency bands, and extracts the onset envelope for each band by taking the first derivative across time. These 40 onset envelopes are then averaged, giving a one-dimensional vector of onset strength across time. Next, the global tempo of the sequence is estimated by taking the autocorrelation of the onset envelope, then multiplying the result by a Gaussian weighting function. Since we did not have a strong a priori hypothesis for the tempo of each sequence, the center of the Gaussian weighting function was set at 120 BPM, and the standard deviation was set at 1.5 octaves. The peak of the weighted autocorrelation function is then chosen as the global tempo of the sequence. Finally, beat onset times are chosen using a dynamic programming algorithm which maximizes both the onset strength at chosen beat onset times and the fit between intervals between beats and the global tempo. A variable called "tightness" sets the relative weighting of onset strength and fit to the global tempo; this value was set to 100, which allowed a moderate degree of variation from the target tempo.

The beat times chosen by this algorithm tend to correspond to onsets of stressed syllables, but can also appear at other times (including silences) provided that there is enough evidence for beat continuation from surrounding stressed syllable onsets. This algorithm permits non-isochronous beats, and is therefore ideal for extracting beat times from speech, despite the absence of metronomic regularity (Schultz et al. 2015). As this procedure was only possible when phrases contained at least three identifiable beats, we were only able to calculate beat variability for 42 of the 48 stimuli. As a result, correlational and regression analyses were only run on these 42 stimuli.

B. Results

Song ratings increased with repetition across all stimuli (main effect of repetition, $F(1.5, 48.0) = 47.9$, $p < 0.001$). However, this increase was larger for the Illusion stimuli (interaction between condition and repetition, $F(1.5, 46.8) = 62.7$, $p < 0.001$). Moreover, song ratings were greater overall for the Illusion stimuli compared to the Control stimuli (main effect of condition, $F(1, 31) = 83.6$, $p < 0.001$). (See Figure 1, black lines, for a visual display of song ratings across repetitions for Illusion and Control stimuli in Experiment 1.)

Table 1 displays correlations between initial ratings and rating changes, and the three stimulus attributes. After a single repetition, subjects' reported song perception was only correlated with beat variability. However, rating change was correlated with pitch flatness, melodic structure, and beat variability.

Table 1. Correlation between song ratings and stimulus characteristics. Bolded cells indicate significance at $p < 0.05$ (Bonferroni corrected).

r-values	Initial rating	Rating change
Pitch flatness	0.151	**0.588**
Melodic structure	0.141	**0.541**
Beat variability	**0.456**	**0.402**

We used hierarchical linear regression to determine whether pitch flatness, melodic structure, and beat variability contributed independent variance to song ratings. By itself, pitch flatness predicted 34.6% of variance in song ratings ($\Delta R^2 = 0.346$, $F = 21.2$, $p < 0.001$). Adding melodic structure increased the variance predicted to 45.2% ($\Delta R^2 = 0.105$, $F = 7.5$, $p < 0.01$). Adding beat variability increased the variance predicted to 55.6% ($\Delta R^2 = 0.105$, $F = 7.5$, $p < 0.01$).

C. Discussion

As reported previously (Tierney et al. 2013), stimuli for which the speech-song transformation was stronger had flatter pitch contours within syllables. However, this characteristic was not sufficient to fully explain why some stimuli transformed more than others. Adding musical cues, namely melodic structure and beat variability, improved the model. This result suggests that listeners who are relatively musically inexperienced rely on melodic and rhythmic structure when judging the musical qualities of spoken phrases.

We also found, contrary to our prediction, that musical beat variability correlated not just with the increase in song perception with repetition but also with song ratings after a single repetition. Melodic structure and pitch flatness, however, only correlated with song ratings after repetition. This result suggests that the rhythmic aspects important for the perception of song can be perceived immediately, while the melodic aspects take time to extract. This finding could provide a partial explanation for why repetition is necessary for the speech-song illusion to take place, but does not rule out the possibility that satiation of speech perception resources is playing an additional role.

Although these results suggest that syllable pitch flatness, melodic structure, and beat variability all influence speech-song transformation, our correlational design was unable to show that these factors played a causal role. To begin to investigate this issue we ran two follow-up experiments in which syllable pitch flatness (Experiment 2) and melodic structure (Experiment 3) were experimentally varied. We predicted that the removal of either of these cues would diminish the speech-song effect.

III. Experiment 2

D. Methods

1) Participants. 27 participants were tested (17 male). The mean age was 33.1 years (standard deviation 7.4). The mean amount of musical training was 1.0 years (standard deviation 1.6).

2) Stimuli. Illusion and Control stimuli from Experiment 1 were altered for Experiment 2. First, the ratio of pitch variability within syllables in the Control stimuli and Illusion stimuli was measured. On average pitch variability within syllables was 1.49 times greater for Control stimuli than for Illusion stimuli. This ratio was then used to alter the pitch contours of syllables (using Praat) such that pitch movements within syllables for Illusion stimuli were multiplied by 1.49, while pitch movements within syllables for Control stimuli were divided by 1.49. This process switched the pitch flatness of Control and Illusion stimuli while maintaining other characteristics such as beat variability and melodic structure.

3) Procedures. Same as Experiment 1.

4) Analysis. To determine whether the pitch flatness manipulation altered song perception ratings, data from Experiment 1 and Experiment 2 were compared using a repeated measures ANOVA with two within-subject factors (condition, two levels; repetition, eight levels) and one between-subject factor (experiment).

E. Results

Similar to the data from Experiment 1, across all stimuli for both experiments there was an increase in song ratings with repetition (main effect of repetition, $F(1.4, 81.9) = 82.2$, $p < 0.001$), although this increase was larger for the Illusion stimuli (interaction between condition and repetition, $F(1.6, 88.6) = 99.9$, $p < 0.001$). Song ratings were also greater overall for the Illusion stimuli compared to the Control stimuli (main effect of condition, $F(1,57) = 133.9$, $p < 0.001$). However, the pitch flatness manipulation had no measurable effect on song perception ratings (no interaction between experiment and repetition, $F(1,57) = 1.0$, $p > 0.1$; no interaction between experiment, repetition, and condition, $F(1.6, 88.6) = 0.56$, $p > 0.1$). See Figure 1 for a visual comparison between song ratings from Experiment 1 and from Experiment 2.

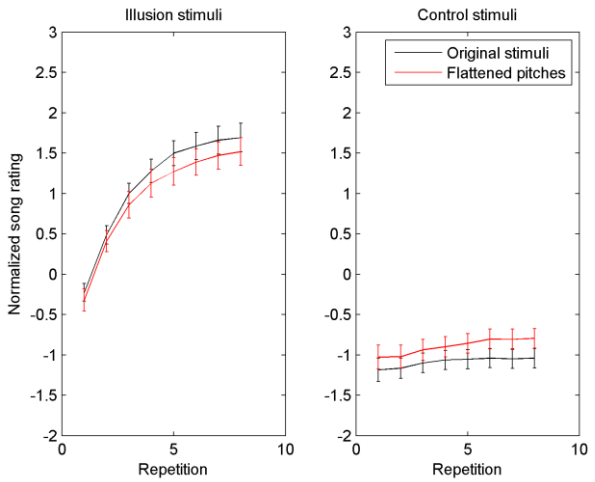

Figure 1. Song ratings for Illusion and Control stimuli, unaltered (black line) and with increased pitch variation for the Illusion stimuli and decreased pitch variation for the Control stimuli (red line).

F. Discussion

Contrary to our predictions, switching the syllable pitch flatness characteristics of the Illusion and Control stimuli did not affect the magnitude of the speech-song transformation. Falk et al. (2014), on the other hand, found that increasing tonal target stability boosted the speech-song transformation; however, our pitch manipulations in this study were much smaller than those used by Falk et al. (2014). While these results do not rule out a role for syllable pitch flatness entirely, they do suggest that this factor does not play a major role in distinguishing transforming from non-transforming stimuli in this particular corpus.

IV. Experiment 3

G. Methods

1) Participants. 31 participants were tested (21 male). The mean age was 35.5 years (standard deviation 7.5). The mean amount of musical training was 1.1 years (standard deviation 2.3).

2) Stimuli. Illusion and Control stimuli from Experiment 1 were altered for Experiment 3 using a Monte Carlo approach with 250 iterations across each of the 48 stimuli. For each iteration the pitch of each syllable was randomly shifted to between 3 semitones below and 3 semitones above its original value. Temperley's (2007) algorithm was then used to determine which of the 250 randomizations resulted in the worst fit to Western melodic structure, and this randomization was used to construct the final stimulus. Note that this manipulation does not affect the pitch flatness within syllables.

3) Procedures. Same as Experiment 1.

4) Analysis. To determine whether the melodic structure manipulation altered song perception ratings, data from Experiment 1 and Experiment 3 were compared using a repeated measures ANOVA with two within-subject factors (condition, two levels; repetition, eight levels) and one between-subject factor (experiment).

H. Results

Similar to the data from Experiment 1, across all stimuli for both experiments there was an increase in song ratings with repetition (main effect of repetition, $F(1.4, 86.0) = 81.6$, $p < 0.001$), although this increase was larger for the Illusion stimuli (interaction between condition and repetition, $F(1.5, 91.2) = 105.6$, $p < 0.001$). There was also a tendency for the Illusion stimuli to sound more song-like overall (main effect of condition, $F(1, 62) = 161.9$, $p < 0.001$). Importantly, however, the melodic structure manipulation changed song perception ratings: the repetition effect for Illusion stimuli was larger for the original stimuli than for the altered stimuli (3-way interaction between repetition, condition, and experiment, $F(1.5, 91.2) = 6.1$, $p < 0.01$). The melodic structure manipulation also had different overall effects for the two classes of stimuli, decreasing song perception ratings for the Illusion stimuli but increasing song perception ratings for the Control stimuli (interaction between condition and experiment, ($F(1, 62) = 6.3$, $p < 0.05$). See Figure 2 for a visual comparison between song ratings from Experiment 1 and from Experiment 3.

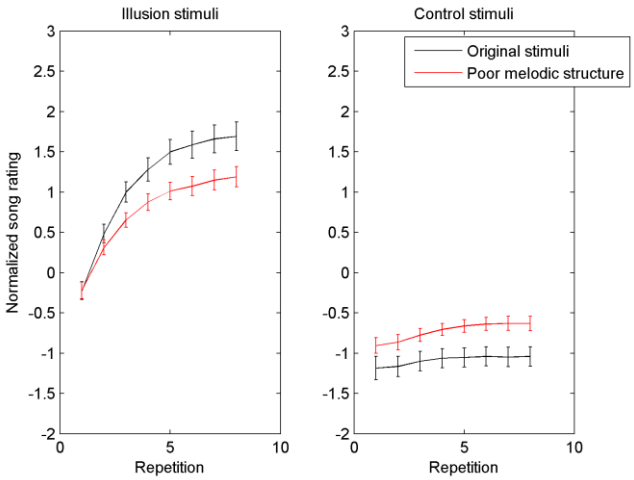

Figure 2. Song ratings for Illusion and Control stimuli, unaltered (black line) and altered to poorly fit a Bayesian model of melodic structure (red line).

I. Discussion

As predicted, forcing the stimuli to more poorly fit a model of melodic structure diminished the magnitude of the speech-song transformation. These results suggest that melodic structure plays a causal role in the speech-song transformation. However, initial song perception ratings were unaffected by the melodic structure manipulation. This, along with the results of Experiment 1, further supports the idea that the increase in song perception with repetition is due in part to a gradual extraction of melodic information from the sequence. Unexpectedly, the melodic fit manipulation also increased song perception for the Control stimuli. We do not currently have a theoretical framework for this finding and so further investigation is needed to pinpoint the source of this effect.

V. General Discussion

We found that within-syllable pitch contour flatness, melodic structure, and beat variability predicted the magnitude of the speech-song transformation. Of these characteristics, only beat variability predicted song ratings after a single repetition. This result suggests that the melodic aspects of a sound sequence take time to extract and could explain why the speech-song illusion requires repetition, as well as why the repetition of random tone sequences increases judgments of their musicality (Margulis and Simchy-Gross 2016). The rhythmic aspects of a sound sequence, on the other hand, may be immediately accessible. If so, one prediction is that a non-tonal rhythmic sequence may not increase in musicality with repetition.

Altering within-syllable pitch flatness did not modulate the speech-song illusion. However, decreasing the extent to which a sequence fit a Bayesian model of melodic structure decreased the intensity of the speech-song illusion. It is possible that the correlation between syllable pitch flatness and speech-song transformation is driven by a third variable. For example, recordings with more overall pitch movement may have both less flat pitch contours within syllables and larger pitch intervals across syllables, and this second factor may be a more important cue for speech-song transformation. Future work in which the rhythmic properties of the stimuli are altered could determine whether beat variability plays a causal role in the speech-song illusion. We predict that increasing beat variability will diminish both initial song perception ratings and the increase in song perception with repetition.

As a whole, our results suggest that musical characteristics of stimuli such as melodic structure and beat variability may be more important than acoustic characteristics such as pitch variability within syllables in determining the strength of the speech-song illusion. If true, this may enable the creation of stimuli that are closely matched on acoustic characteristics, differing only in those musical characteristics necessary for eliciting the speech-song illusion. Such stimuli would be ideal for comparing the neural correlates and perceptual consequences of speech and music perception. Furthermore, our results add to the growing body of work demonstrating that musical sophistication is widespread in the general population (Bigand and Poulin-Charronnat 2006). Indeed, these findings suggest that listeners not only possess sophisticated musical knowledge, they can apply this knowledge to judge the musicality of sound sequences that were never intended to be heard as music.

REFERENCES

Bigand, D., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal training. *Cognition, 100,* 100-130.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America, 129,* 2245-2252.

Ellis, D. (2007). Beat tracking by dynamic programming. *Journal of New Music Research, 36,* 51-60.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1491-1506.

Margulis, E., Simchy-Gross, R., & Black, J. (2015). Pronunciation difficulty, temporal regularity, and the speech-to-song illusion. *Frontiers in Psychology, 6,* 48.

Margulis, E., & Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music Perception, 33,* 509-514.

Schultz, B., O'Brien, I., Phillips, N., McFarland, D., Titone, D., & Palmer, C. (2015). *Applied Psycholinguistics.* DOI: 10.1017/S0142716415000545.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Vanden Bosch der Nederlanden, C., Hannon, E., & Snyder, J. (2015a). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology: General, 144,* e43-e49.

Vanden Bosch der Nederlanden, C., Hannon, E., & Snyder, J. (2015b). Finding the music of speech: musical knowledge influences pitch processing in speech. *Cognition, 143,* 135-140.