Cerebral Cortex February 2013;23:249–254
doi:10.1093/cercor/bhs003
Advance Access publication February 7, 2012

Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion

Adam Tierney[1], Fred Dick[2], Diana Deutsch[3] and Marty Sereno[2]

[1]Department of Communication Sciences and Disorders, Auditory Neuroscience Laboratory, Northwestern University, Evanston, IL 60208, USA, [2]Birkbeck-UCL Centre for NeuroImaging, London WC1H 0AP, UK and [3]Department of Psychology, University of California, San Diego, La Jolla, CA 92023, USA

Address correspondence to Dr Adam Tierney, Department of Communication Sciences and Disorders, Auditory Neuroscience Laboratory, Northwestern University, 2-233, Frances Searle Building, 2240 Campus Drive, Evanston, IL 60208. Email: AdamTierney@gmail.com.

It is normally obvious to listeners whether a human vocalization is intended to be heard as speech or song. However, the 2 signals are remarkably similar acoustically. A naturally occurring boundary case between speech and song has been discovered where a spoken phrase sounds as if it were sung when isolated and repeated. In the present study, an extensive search of audiobooks uncovered additional similar examples, which were contrasted with samples from the same corpus that do not sound like song, despite containing clear prosodic pitch contours. Using functional magnetic resonance imaging, we show that hearing these 2 closely matched stimuli is not associated with differences in response of early auditory areas. Rather, we find that a network of 8 regions, including the anterior superior temporal gyrus (STG) just anterior to Heschl's gyrus and the right midposterior STG, respond more strongly to speech perceived as song than to mere speech. This network overlaps a number of areas previously associated with pitch extraction and song production, confirming that phrases originally intended to be heard as speech can, under certain circumstances, be heard as song. Our results suggest that song processing compared with speech processing makes increased demands on pitch processing and auditory–motor integration.

Keywords: fMRI, music, pitch, song, speech

Introduction

When listening to musicals or opera it is trivial for listeners to determine when a singer suddenly switches from speaking to singing. However, the acoustical differences between song and speech are subtle. Both consist of connected words produced with a relatively smooth fundamental frequency contour. Both are divided up into phrases that often correspond to a breath group. The ends of phrases in both speech and song are marked by final lengthening (Klatt 1975; Fant et al. 1991; Vaissière 1991; Venditti and van Santen 1998; Sundberg 2000) and final lowering (Liberman and Pierrehumbert 1984; Connell and Ladd 1990; Herman 1996; Huron 1996; Prieto and Shih 1996; Truckenbrodt 2004). Some of the few acoustic differences between song and speech are the more isochronous rhythm and greater fundamental frequency stability within each syllable in song (Gerhard 2003; Lindblom and Sundberg 2007).

Speech and song are prototypical examples of language and music, domains that have been used as a test bed for theories of high-level brain organization. Results from neuropsychological and neuroimaging studies have often been used to argue for domain-specific representations and cortical processing modules (Peretz and Coltheart 2003; Peretz and Zatorre 2005). However, such claims are bedeviled by the difficulty of matching linguistic and musical stimuli across a variety of perceptual and cognitive dimensions due to differences in mechanisms of sound generation, segmental properties, and semantic interpretation. This problem exists even when comparing speech with song.

A naturally occurring boundary case between speech and song was discovered by Deutsch (2003) and (Deutsch et al. 2011)—an ambiguous phrase that in context sounds like speech but that when isolated and repeated sounds as if it were being sung. The results of Deutsch et al. (2011) suggest several hypotheses regarding the neural correlates of the percept of song. The authors found that when subjects listened to the illusory stimulus only once and were asked to repeat what they heard the subjects spoke the phrase back and produced fundamental frequencies that were, on average, markedly lower than those of the original stimulus. If, however, the stimulus was repeated, the subjects sang the phrase back, producing fundamental frequencies that were both closer to those of the original recording and corresponded more closely to musical intervals. These results imply that song perception may entail both an increase in the salience of the fundamental frequencies making up a perceived phrase and a perceptual transformation of the fundamental frequencies, matching them to expected statistical characteristics of music (such as a predominance of intervals that are multiples of a semitone). As a result, we would expect song perception to be accompanied by an increase in response in regions associated with pitch salience (e.g., the area encompassing and just anterior to lateral Heschl's gyrus, Patterson et al. 2002; Warren et al. 2003; Brown et al. 2004; Puschmann et al. 2010) and working memory for pitch (e.g., the supramarginal gyrus [SMG], Gaab et al. 2003, 2006; Vines et al. 2006). (Pitch is the perceptual correlate of sound periodicity and typically corresponds to the fundamental frequency of a periodic sound.)

By an exhaustive search through an audiobook library, we discovered 24 spoken phrases that, when excised and repeated, are perceived as song. Each song phrase was matched with a control speech phrase spoken by the same speaker and containing the same number of syllables that continued to be perceived as speech when excised and repeated. Using this stimulus set, we were able to make a particularly closely controlled comparison between perception of language and music (cf. Jeffries et al. 2003; Callan et al. 2006; Schön et al. 2010), analogous to comparisons using sine wave speech and nonspeech (Möttönen et al. 2006). A number of acoustical parameters were measured for these 2 groups of perceptually chosen stimuli. The stimuli were then used in a functional magnetic resonance imaging (fMRI) experiment to determine

which brain regions responded differentially and in common to phrases perceived as speech and song.

Materials and Methods

Materials

All stimuli were collected with permission from audiobooks (via audiobooksforfree.com and librivox.org). Twenty-four spoken phrases that do, on average, sound like song when removed from context and repeated and twenty-four that do not were found. Both speech and song stimuli were taken from the same 3 male talkers in the same proportions. All phrases were taken from passages that were intended to be heard as speech by the readers rather than as song. The semantic content of these phrases is listed in Supplementary Table S1. Speech and song stimuli were closely matched on a number of acoustic measures. Speech phrases had an average duration of 1305 ms, while song phrases had an average duration of 1431 ms. Song and speech stimuli were matched for syllable length. The average syllable rates, in syllables per second, of the song and speech stimuli were 5.13 and 5.00, respectively. Song phrases had an average median fundamental frequency of 141.75 Hz, while speech phrases had an average median fundamental frequency of 134.83 Hz. Song and speech phrases did not significantly differ along any of these dimensions except that of average median fundamental frequency according to the Mann–Whitney U test ($P < 0.05$). This difference is, however, extremely small—less than one semitone—and is therefore unlikely to be the source of the difference in the way in which the 2 sets of stimuli are perceived or any differences in brain response found.

To ensure that people do, on average, hear the song stimuli as song when repeated and the speech stimuli as speech when repeated 15 subjects in a pilot study were asked to listen to 8 repetitions of each stimulus and indicate whether what they heard more closely resembled speech or song. On average, song stimuli were heard more as song than as speech by 12.67 subjects (standard deviation [SD] 1.52), while speech stimuli were heard more as song than as speech by only 2.75 subjects (SD 1.62). The mean hit rate, therefore, was 0.845, while the mean false alarm rate was 0.183. fMRI subjects were given the same test prior to being scanned; to ensure that only subjects who actually heard the illusion were tested, only subjects whose classifications agreed with the predetermined song/speech classifications of at least 80% of the presented stimuli were scanned.

To ensure that the "song" and "speech" stimuli were matched on phonetic content, we calculated frequency counts of every phoneme appearing in the 2 classes of stimuli; the resulting raw data are shown in Supplementary Table S2. A chi-square test was used to determine whether the 2 phoneme distributions were significantly different; they were not (chi-square = 53.0, $P = 0.4$). Moreover, to ensure that the song and speech stimuli were matched in semantic content, we tabulated the frequency counts of different parts of speech in the 2 classes of stimuli. The resulting raw data are shown in Supplementary Table S3. A chi-square test showed that the 2 distributions were not significantly different (chi-square = 10.7, $P = 0.1$).

To examine the degree of fundamental frequency stability within syllables, the onset and offset of each syllable were manually marked by one of the authors while viewing the spectrograms in Praat. The fundamental frequency contour of each phrase was then extracted using Praat. Within each syllable, the sum of the absolute value of the fundamental frequency distances (in semitones) between each time point (one every 10 ms) was computed. This value was then divided by the number of time points and multiplied by 100, giving the average fundamental frequency change over the course of the syllable in semitones per second.

To examine the possibility that the song stimuli contain more regularly spaced stressed syllables, which could give rise to the impression that they sound more rhythmic than the speech stimuli, we marked each syllable as stressed or unstressed using the CMU Pronouncing Dictionary, including both primary and secondary stresses. The onset of the syllable, for the purpose of measuring rhythm, was defined as the beginning of the vowel, which was marked in Praat. For any phrase with at least 3 stressed syllables, we measured the SD of the durations of the intervals between the onsets in order to determine the extent to which stressed intervals were produced in a regular rhythm.

Participants

Fourteen subjects (mean age 30.85 (9.18) years, 9 female) with normal hearing and no history of neurological disorders participated in the fMRI study. All subjects were monolingual speakers of English. Subjects came from a wide variety of musical backgrounds: the subject with the least amount of musical experience had played an instrument for only 1 year, while the most experienced subject had 37 years of musical experience. On average, the subjects had 12.9 (9) years of musical experience. All subjects gave informed written consent. Each subject underwent four 512-s functional scans, at least one structural scan, and an alignment scan (if the structural scan was collected in a different session than the functional scan). No subjects were excluded because of excessive head motion.

During functional imaging, stimuli were presented in 16-s blocks. During each block, a single phrase was repeated with a 0.5-s interstimulus interval as many times as possible within 16 s. Speech, song, and silence blocks were presented in pseudorandom order. Subjects were asked to listen carefully to each phrase and mentally note whether the phrase sounded like speech or like song. Subjects were not, however, asked to explicitly respond to the stimuli in any way. Stimuli were delivered via CONFON headphones and a Denon amplifier.

Data were acquired with a 1.5-T Avanto magnetic resonance imaging scanner (Siemens) using a 12-channel head coil. This magnet is rather quiet for an MRI scanner—82 dB—and this, along with the passive damping offered by the CONFON headphones, allowed us to use a continuous scanning protocol with minimal distracting acoustic interference. MR slices were 3.2 mm thick, with an in-plane resolution of 3.2 × 3.2 mm. A single scan took 516 s, with 258 single-shot echo planar imaging images per slice (time repetition [TR] = 2 s, 24 slices, time echo [TE] = 39 ms, flip angle = 90°, PACE). Stimulus presentation began after the first 4 TRs, which were discarded, because recovered longitudinal magnetization only reaches a steady state after multiple RF pulses. Each of the 14 subjects underwent 4 functional scans.

Each subject's cortical surface was reconstructed from at least one separate structural scan (T_1-weighted magnetization prepared rapid gradient echo, 1 × 1 × 1 mm, 176 slices, TR = 2730 ms, TE = 3.57 ms, flip angle = 7°). Surface reconstruction was performed using FreeSurfer (Dale et al. 1999; Fischl et al. 1999). If more than one structural image was collected for a given subject, the 2 images were registered using AFNI 3dvolreg and averaged prior to surface reconstruction.

The functional scans were registered to the higher resolution structural scan in native space using a boundary-based registration method (Greve and Fischl 2009; bbregister). Statistical analysis was carried out using AFNI (Cox 1996). After being concatenated, the functional scans were motion corrected using AFNI's 3dvolreg function. Images were registered to the middle of the last functional run (the closest to the T_1-weighted alignment scan). This registration results in 6 estimated motion parameters.

Blood oxygen level–dependent responses were analyzed using AFNI's 3dDeconvolve and 3dREMLfit, which uses multiple linear regression to estimate the extent to which each voxel's time series data fits the predicted hemodynamic response, which was generated by convolving the experimental design with a 2-parameter gamma function. A quadratic polynomial was used to model the baseline, and boundaries between concatenated scans were respected (hemodynamic responses were not allowed to cross them). We carried out 3 general linear model tests: song versus baseline, speech versus baseline, and song versus speech.

Group averaging was performed by sampling first level individual subject 3D parameter estimates onto each individual's folded cortical surface, performing 10 steps of surface-based smoothing (equivalent to a surface-based kernel of ~3 mm full-width at half-maximum), inflating each subject's reconstructed surface to a sphere, and then aligning it to the FreeSurfer spherical atlas average using a vertex-by-vertex measure of sulcal depth. After calculating second level

group statistics, results were then painted back onto a single subject's surface for viewing.

To correct for multiple comparisons, we used surface-based cluster-size exclusion for the group-averaged data with an initial surface-vertexwise threshold of $P < 0.005$. Any brain response within a cluster containing fewer vertices was excluded. AFNI's ALPHASIM (Ward 2000), adapted for use with cortical surface-based statistics (Hagler et al. 2006), was used to ensure that these thresholds resulted in corrected P values of <0.05 across the cortex.

Results

Song stimuli were marked by a higher degree of fundamental frequency stability within syllables than speech stimuli. The mean rate of fundamental frequency change within song syllables was 27.6 (25.3) semitones per second, while the mean rate of fundamental frequency change within speech syllables was 40 (23.1) semitones per second: t-test, $P < 0.0001$. Song stimuli contained a slight trend for stressed syllables to be separated by more regular intervals. The average SD of interstress intervals for song stimuli was 0.0995 (0.0539) and for speech stimuli was 0.1232 (0.0554). This trend was, however, not significant according to Student's t-test ($P = 0.1734$).

The song stimuli versus nonstimulus baseline comparison elicited bilateral responses in the superior temporal plane, the precentral and postcentral gyri, the middle temporal gyrus (MTG), and the inferior frontal gyrus (IFG), as well as a response on the left SMG (Fig. 1). The speech stimuli versus nonstimulus baseline comparison elicited bilateral responses in the superior temporal plane and the postcentral gyri, as well as responses in the left inferior frontal cortex, left precentral gyrus (PCG), and right MTG.

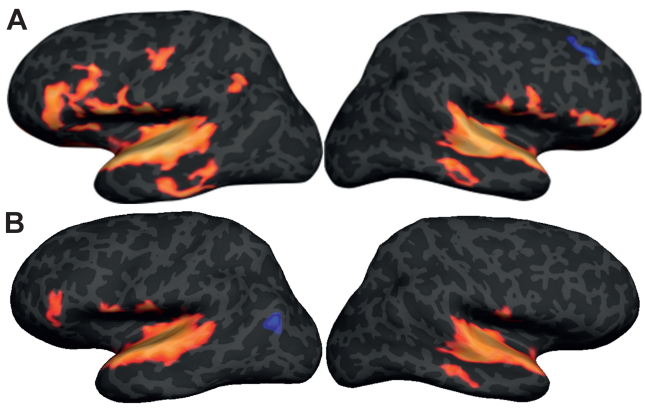

Figure 1. (*A*) Between-subjects surface-based average showing greater response for song versus nonstimulus baseline. (*B*) Between-subjects surface-based average showing greater response for speech versus nonstimulus baseline.

Figure 2. Between-subjects surface-based average showing greater response for song versus speech stimuli.

The intersubject surface-based average song versus speech subtraction is shown in Figure 2. In both hemispheres, there was increased response associated with song perception along the anterior temporal plane considerably anterior to Heschl's gyrus and on a small patch of cortex on the MTG. In addition, there was extensive response on the left SMG, the right posterior superior temporal gyrus (STG), and the left IFG. Finally, there was significant response on the inferior PCG in the right hemisphere but only a small equivalent on the left. (Talairach coordinates of the areas showing greater response in the song condition, as compared with the speech condition, are listed in Supplementary Table S4.) No areas were found to be significantly more activated by the speech stimuli than by the song stimuli.

Discussion

A closely controlled set of naturally occurring song-like versus speech-like stimuli was used to uncover areas involved in processing specifically musical pitch. Both of these stimuli evoked robust responses in auditory language-associated areas. Since both song and speech stimuli consisted of speech, they both contained similar stretches of prosodic contours that were matched for syllable duration, syllable rate, and utterance duration. Both stimuli consisted of semantically meaningful short phrases.

The song stimuli were characterized by slightly more stable fundamental frequency contours within syllables. At first glance, it might appear plausible that this acoustic difference could be driving the differences in brain response we find between conditions. Consistently, however, many previous studies have shown that stimuli with a greater amount of pitch variation lead to increased response in auditory cortex, particularly in lateral Heschl's gyrus and just anterior along the planum polare (Griffiths et al. 2001; Zatorre and Belin 2001; Patterson et al. 2002; Brown et al. 2004). We, on the other hand, found that no regions were more highly responsive to speech than to song. Moreover, increased response for song did not lead to an increase in response in primary auditory cortex, confirming that the differences in the patterns of brain response elicited by the 2 classes of stimuli were not the result of simple low-level acoustic differences. Instead, we argue that both the song percept and the difference in brain response between conditions were driven by higher level musical regularities inherent in the song stimuli, regularities that cannot be detected unless fundamental frequency contours are sufficiently stable within syllables. Further evidence for this view is supplied by Deutsch et al. (2011), who found that the perception of a repeated spoken phrase as song does not take place if the order of the syllables is randomized, suggesting that the effect cannot stem from the acoustic characteristics of individual syllables but instead must find its source in some acoustic regularity spanning at least several syllables. According to this view, therefore, the presence of stable fundamental frequency contours within syllables may be necessary, but not sufficient, for the elicitation of a song percept.

Five foci were more responsive to song than to speech (Fig. 2) and were significantly more responsive in the song condition than the "silence" condition (scanner noise). These foci include areas on the anterior superior temporal gyrus bilaterally (aSTG), right midposterior superior temporal gyrus

(pSTG), right lateral PCG, MTG bilaterally, left SMG, and left IFG. One additional song versus speech focus on the left SMG did not appear in the song versus OFF condition. These 7 foci can be broadly divided into areas likely involved in pitch processing (aSTG, pSTG, SMG, and MTG) and areas involved in vocalization and auditory–motor integration (PCG, SMG, and IFG). The most important implications of our results are as follows. First, our results therefore suggest that one important difference between song and speech processing is the increased demands that song processing makes on both pitch processing and motor processing. Moreover, as we do not find any areas to be more highly responsive to speech than to song, our results imply that, overall, song perception makes greater demands on neural resources than does speech perception. Finally, our pattern of results is strikingly similar to that found by previous studies of song perception and production. Studies of song perception have found increased response in aSTG, pSTG, MTG, and SMG (Gelfand and Bookheimer 2003; Hickok et al. 2003; Callan et al. 2006; Schön et al. 2010), while studies of song production have found increased response in pSTG, PCG, MTG, SMG, and IFG (Riecker et al. 2000; Jeffries et al. 2003; Brown et al. 2004; Callan et al. 2006; Özdemir et al. 2006). Therefore, our results provide neural evidence that the ambiguous speech/song stimuli reported by Deutsch (2003) truly result in song perception, a claim that was previously supported only by behavioral data (Deutsch et al. 2011). The following paragraphs discuss each responsive area in greater detail.

Our main contrast was linked to a strong bilateral response in an aSTG region just anterior to lateral Heschl's gyrus (focus A in Fig. 2). Lateral Heschl's gyrus has long been associated with pitch perception (e.g., Patterson et al. 2002; Warren et al. 2003); recent experiments manipulating pitch salience have not, however, consistently replicated this finding. For example, Hall and Plack (2009) found that only iterated ripple noise stimuli—but not Huggins pitch and a pure tone in noise—led to increased response in lateral Heschl's gyrus, while Puschmann et al. (2010) found that increased pitch salience in Huggins pitch, binaural band pitch, and pure tones in noise were all associated with increased response in lateral Heschl's gyrus.

Our focus was slightly more anterior, in an area that has been shown to be responsive to sequences of changing pitches (Patterson et al. 2002; Brown et al. 2004) and sequences of changing pure tones (Zatorre and Belin 2001). As noted earlier, however, our song stimuli actually featured a smaller degree of fundamental frequency variation within syllables than did the speech stimuli. Rather than detecting simple pitch changes, therefore, this region may be responsible for the detection of pitch patterns extending across multiple different notes or syllables. This region may not be purely "musical"; a recent study (Dick et al. 2011) showed that in the left hemisphere, an anterior STG focus (slightly posterior to focus A) showed sensitivity to subjects' relative expertise in perceiving and producing a given sound class. Here, fMRI responses in actors were greater for speech than violin music, but in violinists, fMRI responses were greater for violin music than for speech.

Our contrast was also linked to response in a midposterior STG area in the right hemisphere (focus B). This region is responsive to speech sounds (Binder et al. 2000) and shows increased response when the number of channels used to noise-vocode speech stimuli is increased, thereby increasing intelligibility (Scott et al. 2006). A difference in response was also found in this area bilaterally to 2 musical sounds that differed in timbre (Menon et al. 2002). Several studies have, however, found larger response in the right hemisphere than in the left hemisphere in this area when musical stimuli are compared with speech stimuli (Gelfand and Bookheimer 2003; Schön et al. 2010) or when subjects are asked to sing rather than speak (Callan et al. 2006; Özdemir et al. 2006). Therefore, our results, combined with previous findings, suggest that in both hemispheres, this area may be responsible for processing a variety of complex sounds, but in the right hemisphere, it may be biased toward pitch processing, as opposed to spectral processing.

Our contrast was also linked to response in 2 adjacent inferior parietal areas, one of which (focus E) responded in our song versus silence contrast and the other of which (focus F) did not. Together, these areas reach from the anterior-most part of the SMG (at the border of the Sylvian fissure) to the posterior-most part at the intersection with the intraparietal sulcus. These foci have been associated with working memory for pitch. Gaab et al. (2003), for example, found that the left SMG responded when subjects performed a short-term pitch memory task, while Gaab et al. (2006) found that subjects who were better able to learn a pitch memory task showed a larger increase in response in the left SMG while performing the same task. Moreover, Vines et al. (2006) found that stimulating the left SMG with transcranial magnetic stimulation led to a decrease in performance on a pitch memory task. Our subjects reported hearing the fundamental frequencies in the song stimuli as being drawn from diatonic scales. Performing this transformation may draw upon short-term memory for pitch, as subjects would need to hold pitches in short-term memory in order for the set of pitches as a whole to be matched to the closest diatonic scale model and perceptually distorted. Our finding of an increase in left SMG response supports this hypothesis. Overall, our finding of an increase in response in the left SMG, the aSTG bilaterally, and the right pSTG, all of which have been repeatedly linked to pitch processing, strongly suggests that song perception compared with speech perception puts greater demands on the neural resources underlying pitch perception.

The left SMG has also been linked to vocal auditory–motor integration. The left SMG responds when subjects are asked to both perceive and produce speech and song (Hickok et al. 2003) and when singers and speakers are presented with altered feedback (Hashimoto and Sakai 2003; Toyomura et al. 2007; Zarate and Zatorre 2008). It has been suggested that this area performs auditory–motor integration, closing an auditory–motor feedback loop (Hickok and Sakai 2007). Moreover, this area may be specifically tuned to vocal auditory–motor integration—when skilled pianists were asked to listen to novel melodies and imagine either humming or playing them on a keyboard, the area was more active for the humming than for the playing condition (Pa and Hickok 2008).

Our main contrast also led to bilateral increased response in the lateral PCG (focus C), which extended much more medially in the right hemisphere. This area lies within the cortical region thought to contain mouth motor and/or premotor representations. It remains somewhat unclear, however, where the motor representations of the different parts of the vocal apparatus are located, both with respect to each other and with respect to structural landmarks, as only in recent years has this issue begun to be investigated using brain imaging techniques

(Brown et al. 2007; Olthoff et al. 2009; Takai et al. 2010). While it is safe to say, therefore, that this area is a motor area containing a representation of some part of the mouth/vocal apparatus, exactly which part is represented remains somewhat unclear (with possibilities including the facial, labial, pharyngeal, laryngeal, jaw, and tongue muscles). An increase in response in this region may indicate that song perception is linked to an increased tendency to covertly vocalize along with stimuli, perhaps indicating a stronger auditory–motor coupling than that found during speech perception.

Previous studies directly comparing song and speech production have found conflicting results regarding whether the resulting response in primary motor cortex is bilateral or biased toward one of the hemispheres (Riecker et al. 2000; Brown et al. 2004; Özdemir et al. 2006). Özdemir et al. (2006) note that hemispheric lateralization of song and speech has only been found when covert tasks have been used, while overt song- and speech-production tasks have led to bilateral responses; our findings fit this pattern.

Our contrast was also linked to response in a region in the left IFG (focus G). Increased response in the IFG has been found when subjects perform song production tasks, both covert (Riecker et al. 2000) and overt (Özdemir et al. 2006). Our finding of increased response in this region, along with the lateral PCG and SMG—all of which have been repeatedly shown to be involved in vocal motor production—strongly suggests that song perception, in the absence of any explicit task, is linked to an increased demand on motor processing. This increased reliance on motor processing during song perception, as opposed to speech perception, may result from subjects covertly initiating synchronized movement to the stimuli that were perceived as musical. This possibility is consistent with the interpretation of the dorsal auditory pathway advanced by Warren et al. (2005), who suggested that a pathway leading from the posterior superior temporal plane to frontal areas is responsible for preparing motor responses to incoming auditory information.

Finally, our contrast was also linked to bilateral response in a region in the MTG (focus D). Although there is no consensus regarding the function of this region, it has been shown to respond bilaterally in several other studies using auditory stimuli, including speech syllables (Jäncke et al. 2002) and noise-vocoded speech (Warren et al. 2006). Increased response has also been reported in this area for sung words as compared with spoken words (Schön et al. 2010), for spoken words as compared with environmental sounds (Dick et al. 2007), and for song production as compared with speech production (Jeffries et al. 2003). Interestingly, in a possible macaque monkey homologue of this region, Barnes and Pandya (1992) showed interdigitated auditory and visual inputs (e.g., their Fig. 4).

We interpret our findings as indicating the neural substrates of song perception. An alternative interpretation of our results is that they are driven in part by the resolution of auditory ambiguity, as the song stimuli are more perceptually ambiguous than the speech stimuli. While the possibility that auditory ambiguity is influencing our results cannot be entirely ruled out, previous work comparing ambiguous auditory stimuli with acoustically matched unambiguous auditory stimuli has revealed response patterns that only slightly overlap with our results. For example, Benson et al. (2006) presented subjects with sine wave speech containing phonetic content and acoustically matched nonspeech stimuli. The 2 types of stimuli, therefore, differed in both ambiguity and presence of phonetic content. Sine wave speech stimuli led to a bilateral response in 2 areas on the superior temporal sulcus overlapping with the MTG. This overlaps somewhat with the MTG response we find to song stimuli. The responses we find in this area may, therefore, be driven by auditory ambiguity, but our findings elsewhere more likely result from song perception.

In summary, we delineated a cerebral network responsible for processing song using 2 sets of naturalistic spoken phrases matched on a number of acoustic dimensions; subjects reported that one set was heard as song when repeated, while the other was heard as speech. Despite minimal acoustic differences between the 2 sets of stimuli, a network of brain regions associated with the perception and production of pitch sequences showed greater response when subjects listened to the song stimuli, as compared with the speech stimuli. This network consisted of areas responsible for the detection of complex pitch patterns and areas responsible for vocal motor processing. Our results delineate a potential neural substrate for the perceptual transformation of speech into song.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/

Funding

This work was supported by the National Institute of Mental Health at the National Institutes of Health (grant number RO1-MH-081990).

Notes

We would like to thank all of the subjects who participated in this study. *Conflict of Interest* : None declared.

References

Barnes C, Pandya D. 1992. Efferent cortical connections of multimodal cortex of the superior temporal sulcus in the rhesus monkey. J Comp Neurol. 318:222–244.

Benson R, Richardson M, Whalen D, Lai S. 2006. Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. Neuroimage. 31:342–353.

Binder J, Frost J, Hammeke T, Bellgowan P, Springer J, Kaufman J, Possing E. 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex. 10:512.

Brown S, Martinez M, Hodges D, Fox P, Parsons L. 2004. The song system of the human brain. Cogn Brain Res. 20:363–375.

Brown S, Ngan E, Liotti M. 2007. A larynx area in the human motor cortex. Cereb Cortex. 18:837–845.

Callan D, Tsytsarev V, Hanakawa T, Callan A, Katsuhara M, Fukuyama H, Turner R. 2006. Song and speech: brain regions involved with perception and covert production. Neuroimage. 31:1327–1342.

Connell B, Ladd D. 1990. Aspects of pitch realization in Yoruba. Phonology. 7:1–29.

Cox R. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 29:162–173.

Dale A, Fischl B, Sereno M. 1999. Cortical surface-based analysis I: segmentation and surface reconstruction. Neuroimage. 9:179–194.

Deutsch D. 2003. Phantom words, and other curiosities. La Jolla (CA): Philomel Records.

Deutsch D, Henthorn T, Lapidis R. 2011. Illusory transformation from speech to song. J Acoust Soc Am. 129:2245–2252.

Dick F, Lee H, Nusbaum H, Price C. 2011. Auditory-motor expertise alters "speech selectivity" in professional musicians and actors. Cereb Cortex. 21:938–948.
Dick F, Saygin A, Galati G, Pitzalis S, Bentrovato S, D'Amico S, Wilson S, Bates E, Pizzamiglio L. 2007. What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: evidence from functional magnetic resonance imaging and lesion data. J Cogn Neurosci. 19:799–816.
Fant G, Kruckenberg A, Nord L. 1991. Prosodic and segmental speaker variations. Speech Commun. 10:521–531.
Fischl B, Sereno M, Dale A. 1999. Cortical surface-based analysis II: inflation, flattening, and a surface-based coordinate system. Neuroimage. 9:195–207.
Gaab N, Gaser C, Schlaug G. 2006. Improvement-related functional plasticity following pitch memory training. Neuroimage. 31:255–263.
Gaab N, Gaser C, Zaehle T, Jancke L, Schlaug G. 2003. Functional anatomy of pitch memory—an fMRI study with sparse temporal sampling. Neuroimage. 19:1417–1426.
Gelfand J, Bookheimer S. 2003. Dissociating neural mechanisms of temporal sequencing and processing phonemes. Neuron. 38:831–842.
Gerhard D. 2003. Computationally measurable differences between speech and song [PhD thesis]. Department of Computer Science, Simon Fraser University.
Greve D, Fischl B. 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage. 48:63–72.
Griffiths T, Uppenkamp S, Johnsrude I, Josephs O, Patterson R. 2001. Encoding of the temporal regularity of sound in the human brainstem. Nat Neurosci. 4:633–637.
Hagler D, Saygin A, Sereno M. 2006. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. NeuroImage. 33:1093–1103.
Hall D, Plack C. 2009. Pitch processing sites in the human auditory brain. Cereb Cortex. 19:576–585.
Hashimoto Y, Sakai K. 2003. Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. Hum Brain Mapp. 20:22–28.
Herman R. 1996. Final lowering in Kipare. Phonology. 13:171–196.
Hickok G, Buchsbaum B, Humphries C, Muftuler T. 2003. Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J Cogn Neurosci. 15:673–682.
Hickok G, Poeppel D. 2007. The cortical organization of speech processing. Nat Rev Neurosci. 8:393–402.
Huron D. 1996. The melodic arch in Western folksongs. Comput Musicol. 10:3–23.
Jäncke L, Wustenberg T, Scheich H, Heinze H-J. 2002. Phonetic perception and the temporal cortex. Neuroimage. 15:733–746.
Jeffries K, Fritz J, Braun A. 2003. Words in melody: an H215O PET study of brain activation during singing and speaking. Neuroreport. 14:745–749.
Klatt D. 1975. Vowel lengthening is syntactically determined in a connected discourse. J Phon. 3:129–140.
Liberman M, Pierrehumbert J. 1984. Intonational invariance under changes in pitch range and length. In: Aronoff M, Oehrle R, editors. Language, sound, structure: studies in phonology presented to Morris Halle by his teacher and students. Cambridge. (MA): MIT Press. p. 157–223.
Lindblom B, Sundberg J. 2007. The human voice in speech and singing. In: Rossing T, editor. Springer handbook of acoustics. New York: Springer Verlag. p. 669–706.
Menon V, Levitin D, Smith B, Lembke A, Krasnow B, Glazer D, Glover G, McAdams S. 2002. Neural correlates of timbre change in harmonic sounds. Neuroimage. 17:1742–1754.
Möttönen R, Calvert G, Jääskeläinen I, Matthews P, Thesen T, Tuomainen J, Sams M. 2006. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage. 30:563–569.
Olthoff A, Baudewig J, Kruse E, Dechent P. 2009. Cortical sensorimotor control in vocalization: a functional magnetic resonance imaging study. Laryngoscope. 118:2091–2096.
Özdemir E, Norton A, Schlaug G. 2006. Shared and distinct neural correlates of singing and speaking. Neuroimage. 33:628–635.
Pa J, Hickok G. 2008. A parietal-temporal sensory-motor integration area for the human vocal tract: evidence from an fMRI study of skilled musicians. Neuropsychologia. 46:362–368.
Patterson R, Uppenkamp S, Johnsrude I, Griffiths T. 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron. 36:767–776.
Peretz I, Coltheart M. 2003. Modularity of music processing. Nat Neurosci. 6:688–691.
Peretz I, Zatorre R. 2005. Brain organization for music processing. Annu Rev Psychol. 56:89–114.
Prieto P, Shih C. 1996. Pitch downtrend in Spanish. J Phon. 24:445–473.
Puschmann S, Uppenkamp S, Kollmeier B, Thiel C. 2010. Dichotic pitch activates pitch processing centre in Heschl's gyrus. Neuroimage. 49:1641–1649.
Riecker A, Ackermann H, Wildgruber D, Dogil G, Grodd W. 2000. Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. Neuroreport. 11:1997–2001.
Schön D, Gordon R, Campagne A, Magne C, Artésano C, Anton J, Besson M. 2010. Similar cerebral networks in language, music and song perception. Neuroimage. 51:450–461.
Scott S, Rosen S, Lang H, Wise R. 2006. Neural correlates of intelligibility in speech investigated with noise vocoded speech—a positron emission tomography study. J Acoust Soc Am. 120:1075–1083.
Sundberg J. 2000. Emotive transforms. Phonetica. 57:95–112.
Takai O, Brown S, Liotti M. 2010. Representation of the speech effectors in the human motor cortex: somatotopy or overlap? Brain Lang. 113:39–44.
Toyomura A, Koyama S, Miyamaoto T, Terao A, Omori T, Murohashi H, Kuriki S. 2007. Neural correlates of auditory feedback control in human. Neuroscience. 146:499–503.
Truckenbrodt H. 2004. Final lowering in non-final position. J Phonetics. 32:313–348.
Vaissière J. 1991. Rhythm, accentuation, and final lengthening in French. In: Sundberg J, Nord L, Carson R, editors. Music, language, and brain: Wenner-Gren International Symposium Series, Stockholm (Sweden): Macmillan. p. 108–120.
Venditti J, van Santen J. 1998. Modeling vowel duration for Japanese text-to-speech synthesis. Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia.
Vines B, Schnider N, Schlaug G. 2006. Testing for causality with transcranial direct current stimulation: pitch memory and the left supramarginal gyrus. Neuroreport. 17:1047–1050.
Ward B. 2000. Deconvolution analysis of fMRI time series data. AFNI 3dDeconvolve documentation, Medical College of Wisconsin. http://afni.nimh.nih.gov/pub/dist/doc/manual/Deconvolvem.pdf.
Warren J, Scott S, Price C, Griffiths T. 2006. Human brain mechanisms for the early analysis of voices. NeuroImage. 31:1389–1397.
Warren J, Uppenkamp S, Patterson R, Griffiths T. 2003. Separating pitch chroma and pitch height in the human brain. Proc Natl Acad Sci U S A. 100:10038–10042.
Warren J, Wise R, Warren J. 2005. Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends Neurosci. 28:636–643.
Zarate J, Zatorre R. 2008. Experience-dependent neural substrates involved in vocal pitch regulation during singing. Neuroimage. 40:1871–1887.
Zatorre R, Belin P. 2001. Spectral and temporal processing in human auditory cortex. Cereb Cortex. 11:946–953.