Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept

CrossMark

Mark Hymers [a,*], Garreth Prendergast [a,d], Can Liu [b,1], Anja Schulze [b,1], Michellie L. Young [b,1], Stephen J. Wastling [c], Gareth J. Barker [c], Rebecca E. Millman [a]

[a] York Neuroimaging Centre, University of York, York Science Park, YO10 5NY, United Kingdom
[b] Department of Psychology, University of York, YO10 5DD, United Kingdom
[c] Institute of Psychiatry, King's College London, SE5 8AF, United Kingdom
[d] Audiology and Deafness Group, School of Psychological Sciences, University of Manchester, Manchester, M13 9PL, UK

ARTICLE INFO

ABSTRACT

The issue of whether human perception of speech and song recruits integrated or dissociated neural systems is contentious. This issue is difficult to address directly since these stimulus classes differ in their physical attributes. We therefore used a compelling illusion (Deutsch et al. 2011) in which acoustically identical auditory stimuli are perceived as either speech or song. Deutsch's illusion was used in a functional MRI experiment to provide a direct, within-subject investigation of the brain regions involved in the perceptual transformation from speech into song, independent of the physical characteristics of the presented stimuli. An overall differential effect resulting from the perception of song compared with that of speech was revealed in right midposterior superior temporal sulcus/right middle temporal gyrus. A left frontotemporal network, previously implicated in higher-level cognitive analyses of music and speech, was found to co-vary with a behavioural measure of the subjective vividness of the illusion, and this effect was driven by the illusory transformation. These findings provide evidence that illusory song perception is instantiated by a network of brain regions that are predominantly shared with the speech perception network.

Introduction

Perceiving language and music constitutes two of the highest level cognitive skills evident in humans. The concept that the hierarchy of syntactic structures found in language and music result in shared perceptual representations (e.g. Koelsch et al., 2002; Patel, 2003) contrasts with the idea that such stimuli are perceived using entirely disparate neural mechanisms (e.g. Peretz and Coltheart, 2003; Rogalsky et al., 2011), whilst others propose a more emergent functional architecture (Zatorre et al., 2002). Song is a well-known example of a stimulus category which evokes both linguistic and musical perception and therefore provides an avenue with which to explore the relationship between these perceptual systems.

There is currently debate regarding the extent to which the representations of melody and lyrics are integrated or segregated during the perception of song. This issue has been examined in a wide range of experiments including integration of memory for melody and lyrics of songs (Serafine, 1984; Serafine et al., 1986), neurophysiological

changes resulting from semantic and harmonic incongruities in familiar music (Besson et al., 1998; Bonnel et al., 2001), fMRI repetition suppression induced by listening to unfamiliar lyrics and tunes (Sammler et al., 2010) and modulations of BOLD response to changes in words, pitch and rhythm for both spoken and sung stimuli (Merrill et al., 2012).

Existing fMRI studies have implicated an extensive network of brain regions which show larger BOLD responses to the perception of sung stimuli as compared to speech stimuli, including bilateral anterior superior temporal gyrus (STG), superior temporal sulcus (STS), middle temporal gyrus (MTG), Heschl's gyrus (HG), planum temporale (PT) and superior frontal gyrus (SFG) as well as left inferior frontal gyrus (IFG), left pre-motor cortex (PMC) and left orbitofrontal cortex (Callan et al., 2006; Schön et al., 2010).

The question of whether speech and song recruit shared or distinct neural systems remains a contentious and controversial topic which is difficult to address directly, since linguistic and musical stimuli differ in their physical attributes. Even when the same syllable is spoken or sung significant differences in the physical properties of the spoken and sung syllable are apparent, such as the minimal and maximal fundamental frequency (F0) and amplitude variation (e.g. Angenstein et al., 2012). Physical differences between spoken and sung stimuli have introduced potential low-level confounds in previous studies designed

* Corresponding author.
   E-mail address: mark.hymers@ynic.york.ac.uk (M. Hymers).
[1] Student author.

to examine the dissociation and/or integration of speech and song perception.

Deutsch et al. (2011) demonstrated an auditory illusion in which identical auditory stimuli may be perceived as either speech or song. Deutsch's speech-to-song illusion is achieved simply through repetition of a spoken phrase. When the spoken phrase was heard for the first time, participants rated the stimulus as speech-like. Following several repetitions of the same spoken phrase, the perception of the stimulus changed and participants rated the stimulus as song-like. The perceptual transformation did not occur if the pitch of the spoken phrase was transposed, or the order of the syllables in the spoken phrase was changed during the repetition phase of the experiment. As identical stimuli can be perceived as both speech and song, Deutsch's speech-to-song illusion provides an elegant solution to controlling auditory confounds, i.e. physical differences in speech and musical stimuli.

Tierney et al. (2013) carried out an fMRI study in which they contrasted neural activity when listeners were presented with song-like and speech-like stimuli. However, rather than using identical stimuli (i.e. Deutsch's illusion in its original form), different spoken phrases were used as song- and speech-like stimuli based upon prior behavioural judgements. Using this approach, they reported BOLD changes within bilateral anterior STG, bilateral MTG, right posterior STG, left IFG and right-lateralised activity in the inferior pre-central gyrus. In contrast, in the current fMRI study, we exploited the power of Deutsch's speech-to-song illusion and employed *physically identical* stimuli that could be perceived as either speech or song. By contrasting brain regions responsive to the percept of the same stimulus as speech-like or song-like, this approach provides a direct, within-subject investigation of the integration or dissociation of neuronal activity involved in differentially perceiving speech and song. As the stimuli are physically identical in the present study, we predict that our approach should show differences in regions of higher-level auditory cortex (e.g. anterior/posterior STG, STS and MTG) as well as higher-order, heteromodal regions including left IFG and left PMC when comparing the perception of speech and illusory song.

Materials and methods

Participants

Thirty-one native English-speaking, right-handed adults gave full informed consent to participate in the study. Before taking part in the main experiment, all participants were screened for normal hearing and absence of amusia in a double-walled sound-attenuating booth. Participants who had absolute thresholds better than 20 dB HL for octave frequencies from 250 to 8000 Hz in both ears progressed to the main experiment. Four participants did not meet this requirement. Participants were also screened using a relevant subset of the Montreal Battery for the Evaluation of Amusia (MBEA: Peretz et al., 2003). One participant did not meet this requirement. As part of the MBEA, participants were asked about the number of years of formal musical training they had received. The average number of years of formal musical training was 3.3 years (range 0–16 years) in this participant group. Of the twenty-five participants who took part in the fMRI study, 15 participants had some formal musical training and 10 participants had received no formal musical training.

Twenty-six participants (mean age 22.6 years, SD 4.0 years; 8 female) were therefore entered into the main experiment. One initial pilot subject was discarded due to technical problems with data acquisition. All data from the remaining 25 participants were analysed. Participants were not paid for taking part in the experiment. The project was approved by the Research Governance Committee, York Neuroimaging Centre, University of York and conformed to the guidelines given in the Declaration of Helsinki.

Stimuli

Auditory stimuli for the main experiment were drawn from the Institute of Electrical and Electronics Engineers sentence lists (Rothauser et al., 1969). Thirty sentences were identified which contained fragments of 4–6 syllables (mean duration 2.37 s, range 1.92 to 2.83 s) — for example "in the red hot sun". The extracted sentence fragments were used as stimuli. The experiment layout was based around 30 "trial-sets". The layout of each of these individual trial-sets can be seen in Fig. 1. Each trial-set consisted of three pre-presentations of a stimulus, a repetition phase based around the same stimulus and three post-presentations of the same stimulus. Each trial-set used a single stimulus from the pool of 30 fragments and each participant heard each fragment in only one trial-set. The two conditions within the experiment were termed *untransformed* and *jumbled*. The difference between the two conditions occurred only during the repetition phase of the stimulus presentation — during the pre-repetition and post-repetition phases the stimuli were always presented in their original, unmodified form (see Fig. 1). In the *untransformed* condition, the repetition phase consisted of presenting the unprocessed fragment ten times, i.e. the number of repetitions shown to cause the perceptual transformation from speech to song (Deutsch et al., 2011). This was to ensure that in the post-repetition phase, the illusory transformation had already taken place. For the *jumbled* condition the Praat software (Boersma and Weenink, 2013) was used to divide each sentence fragment into individual syllables. Five-millisecond logarithmic ramps were applied to the start and end of individual syllables which were then recombined into a *jumbled* fragment as described in Deutsch et al. (2011). The repetition phase in the *jumbled* condition then consisted of the presentation of 10 sentence fragments with different syllable orderings. No perceptual transformation was predicted to occur in the *jumbled* condition. Each participant was presented with 15 trial-sets for the *jumbled* condition and 15 trial-sets for the *untransformed* condition. The experiment was performed over three scanning blocks — each of which contained 5 *jumbled* and 5 *untransformed* trial-sets. The order of the presentation of *jumbled* and *untransformed* trial-sets within the blocks was pseudo-randomised.

In order to further minimise the difference between the *untransformed* and *jumbled* conditions, the 30 stimuli were chosen from the sentence battery such that 15 pairs of stimuli approximately matched for content were derived. As an example, for the sentence fragment "in the red hot sun", the paired fragment was "in the hot June sun". It should be noted that the exact content of the fragments was irrelevant as only trials in which identical sentence fragments were presented were contrasted with each other in the fMRI analysis. Thus there were two sets of 15 stimuli. For each participant, one of these sets was assigned to the *untransformed* condition and the other half to the *jumbled* condition. The assignment of stimulus sets to either the *untransformed* or the *jumbled* condition was counterbalanced across subjects. This pairing counterbalancing was an extra step to minimise any potential differences between conditions.

fMRI procedure

The noise generated by MR scanners poses serious problems to researchers who wish to carry out auditory fMRI experiments (e.g. Gaab et al., 2007a, 2007b). To alleviate some of these issues, data were acquired using Interleaved Silent Steady-State Imaging (ISSS) (Schwarzbauer et al., 2006). This method of fMRI data acquisition differs from traditional sparse imaging in that even during the quiet periods, the slice-select gradient and radio-frequency excitation pulses are applied in the normal way. However, frequency-encoding, phase-encoding and data acquisition do not take place during the quiet periods. This method allows for the acquisition of multiple temporal volumes after a quiet period, without the necessity of modelling T1 saturation effects, and has been shown to be more sensitive than