# UC San Diego

## UC San Diego Electronic Theses and Dissertations

**Title**

The structure and perception of human vocalizations

**Permalink**

https://escholarship.org/uc/item/34g0616c

**Author**

Tierney, Adam Taylor

**Publication Date**

2010

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

# The Structure and Perception of Human Vocalizations

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Cognitive Science

by

Adam Taylor Tierney

Committee in charge:

Sarah Creel, Chair
Martin Sereno, Co-Chair
Andrea Chiba
Diana Deutsch
Marta Kutas
Aniruddh Patel

2010

The Dissertation of Adam Taylor Tierney is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

---

---

---

---

---

Co-Chair

---

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

Signature Page.....iii

Table of Contents.....iv

List of Abbreviations.....vi

List of Figures.....vii

List of Tables.....x

Acknowledgements.....xi

Curriculum Vitae.....xiii

Abstract of the Dissertation.....xvi

Introduction.....1

Chapter 1. Shared Pitch Patterns in Human Song and Birdsong.....37

1.1 Introduction.....38

1.2 Methods.....42

1.3 Results.....50

1.4 Discussion.....62

Chapter 2. Pitch Patterns in Human Speech and Instrumental Music.....68

2.1 Introduction.....69

2.2 Methods.....72

2.3 Results.....74

2.4 Discussion.....80

Chapter 3. Tonotopic Mapping of Human Auditory Cortex Using Meaningful Stimuli.....85

3.1 Introduction.....86

3.2 Methods.....90

3.3 Results..... 96
3.4 Discussion.....115
Chapter 4. Neural Correlates of Perceiving Speech and Song..... 125
4.1 Introduction..... 126
4.2 Methods..... 136
4.3 Results.....143
4.4 Discussion.....147
Chapter 5. Conclusions.....161

## LIST OF ABBREVIATIONS

EEG: electroencephalogram

fMRI: functional magnetic resonance imaging

HG: Heschl's gyrus

MRI: magnetic resonance imaging

PET: positron emission tomography

STS: superior temporal sulcus

STG: superior temporal gyrus

LIST OF FIGURES

Figure 1.1. Pitch encoding method for human and bird song..... 47

Figure 1.2. Histogram of length in number of notes of human songs, human song phrases, and birdsongs..... 49

Figure 1.3. Final lengthening in human song and birdsong.....50

Figure 1.4. Final lengthening in German and Chinese song..... 51

Figure 1.5. Skip-reversal patterns in human song and birdsong..... 52

Figure 1.6. Skip-reversal patterns in German and Chinese song..... 52

Figure 1.7. In human song and birdsong, skips that are followed by reversals fall further from the mean than skips that are followed by continuations..... 54

Figure 1.8. In Chinese song and German song, skips that are followed by reversals fall further from the mean than skips that are followed by continuations.....55

Figure 1.9. Average pitch contour of human song phrases and birdsong individual notes, both unaltered and detrended..... 56

Figure 1.10. Average pitch contour of German and Chinese song phrases, both unaltered and detrended..... 57

Figure 1.11. Classification of pitch contours in human song phrases and birdsong individual notes.....58

Figure 1.12. Classification of pitch contours in German and Chinese song phrases..... 59

Figure 1.13. Tendency of bird songs and human songs to contain smaller pitch intervals than one would expect based solely on the pitch values used in a given song or song phrase..... 62

Figure 2.1. Skip-reversal patterns in instrumental music and speech..... 75

Figure 2.2. In instrumental music, skips that are followed by reversals fall further from the mean than skips that are followed by continuations.....76

Figure 2.3. In speech, skips that are followed by reversals fall further from the mean than skips that are followed by continuations..... 77

Figure 2.4. Final lengthening in instrumental music and speech.....78

Figure 2.5. Final lengthening in English, French, Yoruba, and Japanese..... 79

Figure 2.6. Average note contours for speech and instrumental music.....80

Figure 3.1. Spectrogram of the stimulus used for a single 8-minute, 32-second down-ramped tonotopy scan.....91

Figure 3.2. Average pitch contour of affective voice stimuli..... 92

Figure 3.3. Pitch contour classification of affective voice stimuli.....93

Figure 3.4. Representative example of the pitch contour of a single affective voice stimulus.....93

Figure 3.5. Tonotopic activation in the right hemisphere of subject 1.....101

Figure 3.6. Tonotopic activation in the left hemisphere of subject 1.....102

Figure 3.7. Tonotopic activation in the right hemisphere of subject 2.....103

Figure 3.8. Tonotopic activation in the left hemisphere of subject 2.....104

Figure 3.9. Tonotopic activation in the right hemisphere of subject 3..... 105

Figure 3.10. Tonotopic activation in the left hemisphere of subject 3..... 106

Figure 3.11. Tonotopic activation in the right hemisphere of subject 4..... 107

Figure 3.12. Tonotopic activation in the left hemisphere of subject 4..... 108

Figure 3.13. Tonotopic activation in the right hemisphere of subject 5.....109

Figure 3.14. Tonotopic activation in the left hemisphere of subject 5.....110

Figure 3.15. Tonotopic activation in the right hemisphere of subject 6..... 111

Figure 3.16. Tonotopic activation in the left hemisphere of subject 6..... 112

Figure 3.17. Tonotopy mapping results of left hemisphere between-subjects average..... 113

Figure 3.18. Tonotopy mapping results of right-hemisphere between-subjects average..... 114

Figure 4.1. Pitch contour of single examples of “speech” and “song” illusion stimuli.....137

Figure 4.2. Speech rate, in syllables per second, of “speech” and “song” stimuli..... 138

Figure 4.3. Mean pitch of “speech” and “song” stimuli..... 139

Figure 4.4. Percent of vowels marked as glides by the prosogram for “speech” and “song” stimuli.....140

Figure 4.5. Number of people who reported hearing song for the “song” and “speech” stimuli.....141

Figure 4.6. Examples of single-subject fMRI block design data, activation elicited by “song” stimuli subtracted from activation elicited by “speech” stimuli..... 144

Figure 4.7. Between-subjects average, activation in left hemisphere elicited by “song” stimuli subtracted from activation elicited by “speech” stimuli, $p < 0.005$ uncorrected..... 145

Figure 4.8. Between-subjects average, activation in right hemisphere elicited by “song” stimuli subtracted from activation elicited by “speech” stimuli, $p < 0.005$ uncorrected.....146

Figure 4.9. Comparison of single-subject data from Chapters 3 and 4.....147

LIST OF TABLES

Table 1.1. Family, common name, and species name of each bird whose song we studied.....43

Table 1.2. Number of phrases and songs taken from each country of origin for the folksongs contained in the analyzed database.....47

ACKNOWLEDGEMENTS

There are a great many people who have provided me with support, advice, and friendship over the years, without whom I could not have done the work presented here. The following list is but a small sample:

Marty Sereno taught me the bulk of what I know about neuroscience, and his boundless enthusiasm for science continually inspired me to learn more.

Aniruddh Patel supplied an endless stream of crystal clear hypotheses for testing, and gave my work and my development as a scientist a great deal of attention and direction.

Diana Deutsch helped immensely with the process of making and testing the stimuli used in chapter 4, and was in general exceedingly generous with her time and insight.

My collaborator Fred Dick, for providing guidance and advice on innumerable occasions; Flavia Filimon, from whom I learned a large portion of my knowledge of fMRI methods; Sarah Creel, for providing advice and support whenever needed; Kaitlyn Means, who has been an inexhaustible source of help and support; My family, especially my parents William and Mary and my brother Ryan.

Chapter 1, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this

manuscript: Frank Russo and Aniruddh Patel.

Chapter 2, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-author as collaborator on this manuscript: Aniruddh Patel.

Chapter 3, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Martin Sereno and Fred Dick.

Chapter 4, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Diana Deutsch and Martin Sereno.

CURRICULUM VITAE

## Education

Ph.D., Cognitive Science, University of California, San Diego, 2010

M.S., Cognitive Science, University of California, San Diego, 2006

B.A., Cognitive Science and Spanish, B.S., Psychology, Indiana University, Bloomington, 2004

## Publications

Tierney, A., Dick, F., and Sereno, M. (in preparation). Tonotopic mapping of human auditory cortex using meaningful stimuli.

Tierney, A., Deutsch, D., and Sereno, M. (in prepatation). Neural correlates of perceiving speech and song.

Tierney, A., Russo, F., and Patel, A. (in preparation). Shared pitch patterns in human song and birdsong.

Tierney, A., Bergeson, T., and Pisoni, D. (2008). "Effects of early musical experience on auditory sequence memory." Empirical Musicology Review *3*, 178-186.

Clopper, C., Pisoni, D., and Tierney, A. (2006). "Effects of open-set and closed-set demands on spoken word recognition." *Journal of the American Academy of Audiology* 17: 331-349.

**Conference Abstracts**

Tierney, A., Dick, F., and Sereno, M. (2009). "Representation of frequency and pitch in superior temporal cortex." Poster presented at the Society for Neuroscience meeting, Chicago, Illinois.

Tierney, A., Russo, F., and Patel, A. (2009). "Statistical regularities of human melodies reflect perceptual-motor constraints: evidence from comparisons with bird songs." Talk presented at the Society for Music Perception and Cognition meeting, Indianapolis, Indiana.

Tierney, A., Russo, F., and Patel, A. (2008). "Empirical comparisons of pitch patterns in music, speech, and birdsong." *J. Acoust. Soc. Am.,* 123(5):3721.

Tierney, A., and Sereno, M. (2007). "Topographic mapping of second formant in superior temporal cortex." Poster presented at the Cognitive Neuroscience Society Meeting, New York NY.

Tierney, A., and Pisoni, D. (2005). "The effects of early music experience on verbal skill." Poster presented at the Society of Music Perception and Cognition Meeting, San Diego, CA.

Clopper, CG, Tierney AT, Pisoni DB. (2003). "Effects of response format on speech intelligibility in noise: Results obtained from open-set, closed-set, and delayed response tasks." *J. Acoust. Soc. Am.,* 13(4,2):2254.

## Honors and Awards

Center for Research in Language Fellow, University of California, San Diego 2008-2009, 2006-2007.

Department Travel Award to attend Society for Neuroscience meeting (UCSD, Department of Cognitive Science) 2009.

Department Travel Award to attend Acoustical Society of America meeting (UCSD, Department of Cognitive Science) 2008.

Superior Teaching Award, University of California, San Diego 2008.

Department Travel Award to attend Cognitive Neuroscience Society meeting (UCSD, Department of Cognitive Science) 2007.

Chancellor's and Dean's Fellow, University of California, San Diego 2004-2006.

Wells Scholarship recipient, Indiana University, Bloomington 2000-2004.

## Professional Experience

2004-2010 Graduate Research Assistant, Professor Martin I. Sereno, Department of Cognitive Science, University of California, San Diego, CA.

2006-2010 Graduate Research Assistant, Dr. Aniruddh Patel, The Neurosciences Institute, La Jolla, CA.

2007-2010 Graduate Research Assistant, Professor Diana Deutsch, Department of Psychology, University of California, San Diego, CA.

2003-2004 Research Assistant, Professor David Pisoni, Department of Psychology, Indiana University, Bloomington, IN.

ABSTRACT OF THE DISSERTATION

The Structure and Perception of Human Vocalizations

by

Adam Taylor Tierney

Doctor of Philosophy in Cognitive Science

University of California, San Diego, 2010

Professor Sarah Creel, Chair
Professor Martin Sereno, Co-chair

Music takes on a variety of forms in different cultures, but several cross-cultural regularities exist. Large jumps in pitch, for example, tend to be followed by reversals in pitch direction, phrase-final notes tend to be lengthened, and pitch contours of song phrases tend to be arch-shaped. I

show that these patterns exist in a corpus of European and Chinese folksongs and in a phylogenetically diverse selection of bird songs. As birds and humans diverged from a common ancestor more than 200 million years ago, but share similar song-production mechanisms, the simplest explanation for the existence of similar patterns in the songs of both animals is that they are a consequence of motor constraints. I also show that these patterns exist in instrumental composed music and speech.

If these patterns are a consequence of motor constraints, the auditory system has had ample time to adapt to their presence. Simple stimuli following these patterns could, then, be useful tools for studying the auditory system. I demonstrate the viability of this approach by using emotional non-verbal vocalizations, simple stimuli with arch-like pitch contours, to tonotopically map human auditory cortex using fMRI. I demonstrate, for the first time, tonotopic organization in the superior temporal sulcus, an area commonly activated by voice stimuli.

Speech and song are acoustically highly similar, a fact emphasized by my finding that they share several patterns by virtue of being produced by the same vocal production system. In fact, as Diana Deutsch has shown, it is possible to find fragments of recorded speech that, when repeated out of context, suddenly begin to sound like song. I collected twenty-four additional examples of this phenomenon from audio books, along with twenty-four control stimuli spoken by the same speakers. In an imaging study using these

two sets of stimuli I show that perceiving stimuli as song rather than as speech is linked to an increase in activation in a number of areas involved with pitch processing. It appears, therefore, that one of the main differences between the auditory processing of speech and song may lie in the degree of attention directed to pitch information.

# Introduction

The prosody and music of different cultures differ unendingly in the specific rhythmic and pitch patterns they use and the way they use these patterns to communicate. Nevertheless, researchers have noticed a few regularities across musical traditions and languages. In Chapters 1 and 2 I argue that these regularities are due to motor constraints, as evidenced by the fact that they are all found in the song of birds—animals phylogenetically distant from humans, but which produce vocalizations using a similar system. I also show that these patterns are found in human song, human speech, and human instrumental music. Chapters 3 and 4 examine two consequences of this finding. First, the human auditory system may be specialized for processing stimuli that follow these simple patterns, as they have likely been characteristic of animal vocalizations for hundreds of millions of years. Therefore, stimuli following these patterns are useful tools for probing the structure and function of the auditory system, which I illustrate by using band-passed emotional vocalizations to tonotopically map human auditory cortex. Second, given the acoustic similarities between speech and song, it is possible to find ambiguous stimuli that can be perceived as either speech or song, a perceptual difference that is linked to a few rather simple differences in brain activation, suggesting that speech and song are processed by largely overlapping neural resources.

**Characteristic patterns of music across cultures**

Composers and music theorists have for decades discussed the presence of two patterns found throughout western classical music: the tendency for large jumps in pitch to be followed by pitch movements in the opposite direction (or "reversals"), and the tendency for the pitch contours of phrases to follow an arch shape. These two patterns have only been empirically validated rather recently; the skip-reversal pattern was shown in a large corpus of (mostly Western) folk songs by von Hippel and Huron (2000), and the arch contour was found in song phrases from the same corpus by Huron (1996). Several studies have also found that classical pianists (Repp 1990, Penel and Drake 2004) and a single opera singer (Sundberg 2000) lengthen notes just prior to the ends of musical phrases. The presence of these regularities does not seem to be lost on listeners; infants can take advantage of the existence of phrase-final lowering and lengthening in order to detect the phrase structure of music (Jusczyk and Krumhansl 1993).

In speech, the last syllable before a phrase boundary or the end of an utterance is lengthened, an effect which has been demonstrated in a wide variety of languages, including English (Klatt 1975), Japanese (Venditti and van Santen 1998), Swedish (Fant et al. 1991), and French (Vaissiere 1991). The end of declarative sentences in many languages is also marked by a rapid decline in pitch, including English (Liberman and Pierrehumbert 1984), Spanish (Prieto et al. 1996), German (Truckenbrodt 2004), Kipare (Herman

1996), and Yoruba (Connell and Ladd 1990). This final decrease in pitch is in addition to a constant decrease in pitch throughout an utterance known as declination, which is found nearly universally in declarative sentences as well (Bolinger 1978). Infants can make use of the presence of phrase-final lengthening and lowering in order to detect the phrase structure of spoken language (Jusczyk et al. 1992).

The few universal patterns that have been found in music and speech are remarkably similar, particularly those marking the boundaries between phrases. At the present, unfortunately, the source of these similarities can only be speculated upon, as despite the intriguing similarities between the two domains essentially no direct comparative work has been done. One possibility is that the tendency to produce these patterns is somehow directly specified genetically, perhaps because they help communicate with the listener. Both the melodic arch and phrase-final lengthening, for example, could conceivably help listeners detect structural boundaries. In this dissertation, however, I will argue that all of the above-mentioned patterns stem from motor constraints.

As an aside, I should note that the theory that these patterns are initially caused by motor constraints does not imply that speakers and musicians cannot take advantage of the regularities and use them for communicative purposes. In fact, both musicians and speakers seem to exaggerate these effects when necessary. Musicians, for example, produce more final lengthening when asked to play expressively (Penel and Drake 2004), while

speakers are able to produce final lengthening in order to disambiguate syntactically ambiguous sentences, a phenomenon that has been demonstrated in both professional speakers (Scott 1982, Price et al. 1991) and randomly selected subjects (Schafer et al. 2000). Moreover, motor constraints should not be seen as dictating what speakers and musicians *must* do in a particular context; instead, they simply specify which patterns are easy to produce and which are more difficult. Speakers are perfectly able to resist these effects, and might do so in order to convey, for example, the fact that they are not done talking or to invite a conversation partner to complete an intonational phrase by answering a question. This could account for the near-universal pairing of questions with rising intonational contours (though certain kinds of questions, such as wh-questions, are often associated with falling contours) (Bolinger 1978).

**Motor components of speech production**

There are three main parts to the vocal instrument (Sundberg 1977): the lungs act as the power source, driving the oscillation of the vocal folds, and then the resulting sound is filtered through the pharyngeal and nasal cavities, which act as a resonator. Each of these parts can be linked to a different motor constraint and, in turn, to one of the three patterns described above. The lungs, for example, drive the sub-glottal pressure that causes the vocal folds to vibrate, but as an utterance within a breath group proceeds, keeping this pressure constant becomes more and more difficult. Moreover, it takes time

for sub-glottal pressure to reach its steady state at the beginning of an utterance, and for sub-glottal pressure to fall at the end of an utterance. As a result, since speakers do not wait for sub-glottal pressure to reach its maximum before beginning an utterance, over the course of a spoken phrase, subglottal pressure sharply rises, gradually declines, and then sharply falls (Slifka 2006). As sub-glottal pressure and fundamental frequency have been shown to be correlated in human speech (Beckman et al. 1996) and song (Lindblom and Sundberg 2007), these changes in pressure could potentially lead to an arched average pitch contour.

The vocal folds are only able to produce pitches from a bounded distribution, and pitches closer to the center of this distribution are easier to produce than pitches closer to the edge. This causes the normalized distribution of pitches within the phrases of folk songs to be normally distributed (von Hippel 2000). This simple fact could give rise to the skip-reversal pattern, as large jumps in pitch are more likely to land closer to the edge of the distribution, and all else being equal pitches closer to the center of the distribution are more likely to be produced than pitches closer to the edge. This interpretation is supported by the fact that only skips departing from the median or crossing the median are more likely to be followed by reversals, while skips approaching the median are instead more likely to be followed by continuations (von Hippel and Huron 2000).

Finally, over the course of a spoken or sung phrase, the articulators need to be rapidly moved to change the resonant frequencies of the vocal tract as different speech sounds are produced. It may be easier for speakers to slow this movement down before stopping, rather than coming to an abrupt halt, and this slowing would result in final lengthening (Myers and Hansen 2007). This explanation is consistent with the finding that final lengthening is greater for speakers who naturally speak at a faster rate (Smith 2000) or who have been asked to speak at an increased tempo (Smith 2002).

Many of these motor constraint hypotheses are not entirely new. As mentioned above, for example, Myers and Hansen (2007) suggested that final lengthening is a consequence of articulatory slowing near the end of an utterance, a possibility also discussed by Klatt (1975). Penel and Drake (2004) consider a motor constraint explanation for final lengthening in music, though they conclude that it is primarily, instead, a communicatory phenomenon. Liberman and Pierrehumbert (1984) mention the possibility that final lowering could be attributed to changes in subglottal pressure, but wave this explanation aside due to a lack of evidence for the changes in pressure over time that would be necessary to account for a sudden final drop in pitch. Finally, Gussenhoven (2002) and Lindblom and Sundberg (2007) suggest that declination is a consequence of the gradual drop in subglottal pressure throughout an utterance, but do not extend this explanation to encompass final lowering.

Although the possibility that some of these regularities found between cultures are due to motor constraints has been discussed for decades, there is little to no direct evidence that this is the case. What could be counted as evidence for motor constraints as a source of these patterns? One could argue that the remarkable similarity between the acoustic correlates of phrase boundaries in music and speech is evidence for motor constraints shared between the two domains. Recent research, however, has shown that rhythmic and pitch patterns in speech can have an effect on the structure of music through statistical learning. Patel and Daniele (2003) found that British musical themes and spoken phrases both exhibit high pairwise durational variability, whereas French music and speech exhibit low durational variability (that is, notes of similar duration tend to occur together). This result was replicated by Huron and Ollen (2003) in a much larger corpus. Moreover, French speech and music have low pitch interval variability, while English speech and music have higher interval variability (Patel et al. 2006). These results suggest that similarities between speech and music cannot necessarily be taken as evidence for shared motor constraints—composers and musicians could be learning the patterns from speech and incorporating them into their music.

**Investigating a motor basis for constraints on song form**

An alternate way to bring evidence to bear on the source of these regularities is to see if they're also found in birdsong. Although birds differ

from humans in many obvious ways—size, average pitch of songs, ability to fly—their vocal apparatus is broadly similar to the human vocal apparatus, consisting of a power source (the lungs) hooked up to an oscillator (two syrinxes), producing a sound that is filtered through a resonating cavity. One would expect to find the same skip-reversal pattern in birdsong, as the syrinxes are only able to produce a limited range of pitches, and pitches near the center of that range should be easier to produce than pitches closer to the edge. One would expect to find final lengthening in birdsong as well, as birds that produce pure-tonal songs actively modify the resonating characteristics of their vocal tract by opening and closing their beaks, which filters out the higher harmonics (Nowicki 1987, Podos et al. 1995, Goller et al. 2004). This movement can be quite rapid (Westneat et al. 1993), and as a result birds may need to slow this movement down before halting entirely at the end of a song. Finally, as bird song is powered by the lungs, one would expect to find an arch-shaped contour (as well as declination) associated with the breath group. However, since birds are able to breathe in between notes (Hartley and Suthers 1989, Wild et al. 1989), for birds the breath group would be a single note.

In Chapter 1, I show that these three patterns—final lengthening, arch-shaped contours, and skips being more likely to be followed by reversals—are found in both bird song and human song. This is evidence that these patterns are, in fact, due to motor constraints, in which case one would expect to find them in speech as well, and I show in Chapter 2 that this is the case. This is the

first demonstration that speech is characterized not just by final lowering, but also by an initial rise in pitch. (Vaissiere (1995) suggested that arch contours underlie spoken phrases, but did not present any evidence for their existence.) In Chapter 2 I find all three patterns present in composed instrumental music as well.

Arch-shaped contours and final lengthening are, therefore, patterns characterizing human vocalizations and communication systems of many kinds—for example, I show in Chapter 3 that non-verbal emotional cries also predominantly follow the contour of the declining arch. The fact that bird songs also exhibit these patterns suggests that they are the consequence of shared motor constraints and, therefore, have been present in the vertebrate acoustic environment for hundreds of millions of years. Moreover, these simple characteristics indicate that a sound is being produced by an animal, and more likely than not a human attempting to communicate; as a result, any sound exhibiting these patterns should be very interesting to the human auditory system. It's likely, therefore, that the auditory system has somehow adapted to process sounds that follow final lengthening and the melodic arch.

**Neural processing of auditory signals**

The exact nature of this specialization is not yet clear. Possibilities range from greater temporal and frequency resolution to easier sound source segregation. Whatever its exact nature, it is likely that larger amounts of

auditory cortex are recruited for the processing of stimuli that follow these patterns. As a result, simple stimuli following these patterns are an ideal tool for probing the structure of the auditory system, as they could be simple enough to be easily controlled and manipulated but nonetheless able to activate cortex sensitive to human vocalizations. In Chapter 3 I demonstrate the viability of this approach by using band-passed human emotional vocalizations to tonotopically map human auditory cortex. It is widely thought that the human superior temporal sulcus (STS) is not organized tonotopically, as previous mapping studies have shown tonotopic organization to be confined to the superior temporal gyrus and the temporal plane. However, many studies have shown the STS to be more strongly activated by human voices than by acoustically closely matched non-vocal stimuli. As a result, it's possible that previous studies may not have been able to show tonotopic organization in these areas because the stimuli used were overly abstract and simple, being largely composed of pure tones and noise bands.

Using fMRI, Binder et al. (2000) found that the STS, bilaterally, became more active when subjects listened to words, as opposed to tones or white noise. This activation extended throughout a long stretch of the STS, inferior to the parts of the planum temporale and STG that have since been tentatively identified as human core and belt cortex. These areas may, therefore, correspond to monkey parabelt cortex. Unfortunately, the contrasts used in this study were rather crude; words, tones, and white noise differ along so

many dimensions that it is impossible to determine what was driving the increase in activation, be it complexity, ecological validity, or a specific adaptation to process human vocalizations. Scott et al. (2000) compared normal speech to three types of altered speech: noise-vocoded speech, which is intelligible after some training (though it contains no intonational information); spectrally-rotated speech, which is not intelligible but contains phonetic features; and rotated noise-vocoded speech, which is not intelligible and does not contain phonetic features. Activation in the left posterior STS was found to be linked to the presence of phonetic features, while left anterior STS was activated only by intelligible signals. No significant right STS activation was found for any of the contrasts tested.

Thus, not only des speech processing in the STS appear to be left-lateralized, but it seems the left STS can be divided up into sub-regions based on the intelligibility of the signals that activate it. This finding was replicated by Narain et al. (2003), who used noise-vocoded and spectrally rotated speech stimuli and found that processing of intelligible versus unintelligible speech was linked to activation in the mid and anterior STS. Liebenthal et al. (2005) investigated the effect of speech intelligibility on STS activation by presenting tokens from a phonemic consonant-vowel syllable continuum from /ba/ to /da/. There was also an analogous nonphonemic continuum; the anchor points of this continuum were constructed by spectrally inverting the first formants of the anchor points of the phonemic continuum. The resulting sounds were not

legal phonemes in American English and were not categorically perceived by subjects. Activation associated with both categories included the posterior STG, bilaterally, and areas in the frontal and parietal lobes. Increased activation in the phonemic condition was observed predominantly in the anterior and middle portions of the left STS.

Yet another way to manipulate the intelligibility of speech signals is to use noise-vocoding with different numbers of channels, rather than the single-channel noise-vocoding used by Scott et al (2000) and Narain et al (2003). The more channels that are used when a signal is noise-vocoded, the more intelligible the resulting signal. Scott et al. (2006) found that the number of channels used in noise-vocoded speech stimuli correlated significantly with activation in the left anterior STS. (Control stimuli were constructed by spectrally rotating the different noise-vocoded stimuli to control for signal complexity.) Warren et al. (2006) presented subjects with blocks of noise-vocoded speech stimuli in which the speaker and the number of channels used either changed or remained constant. When the intelligibility of the signal changed during a block, the STS showed increased activation bilaterally, but when the voice source changed, increased activation was limited to posterior bilateral STS.

Although speech signals, for the most part, appear to preferentially activate the left STS, listening to non-speech vocal sounds appears to activate the STS bilaterally. Belin et al. (2002) used fMRI to compare brain activation

in response to vocal sounds both linguistic and nonlinguistic (laughter, emotional cries) to activation in response to a wide variety of non-vocal sounds, including environmental sounds, animal sounds, mechanical sounds, and musical sounds. Vocal sounds, both linguistic and non-linguistic, led to increased bilateral activation in several different areas along the superior temporal sulcus, both posterior and anterior to Heschl's gyrus. The maximum voice sensitivity was seen in the anterior part of the STS, with the strongest difference on the right side. Furthermore, Belin and Zatorre (2003) found that the same area in the right anterior STS showed adaptation to constant speaker identity, as opposed to constant syllable identity.

To summarize, the STS appears to be more strongly activated by stimuli more complex than simple pure tones or noise bands. Moreover, there is some evidence that the right anterior STS processes information specific to particular speakers. The left anterior STS, on the other hand, is preferentially activated by intelligible speech signals, while researchers have suggested that posterior STS areas perform more domain-general feature-extraction. Recent evidence (Petkov et al. 2008) suggests that anterior auditory cortex in macaques may also contain a "voice region" preferring macaque vocalizations to heterospecific animal vocalizations and natural sounds. Unlike the human voice region, the monkey voice region does not appear to reside inside the STS, and is instead on the surface of the STG, suggesting that the voice region may have migrated in one of the two primates after their evolutionary split.

**Investigating tonotopy in the STS**

Based on prior research, therefore, the ideal stimulus for searching for topographic maps in the STS would be a stimulus more complex than simple pure tones or noise bands. It would be a human vocalization that, even when filtered so that only very high or very low frequencies remain, would be not only identifiable as a human vocalization, but also in some sense semantically meaningful. The Montreal Affective Voices (Belin et al. 2008), a set of non-linguistic vocalizations produced by actors expressing eight different emotional categories, fulfill all of these requirements. These voices predominantly feature descending, arch-shaped pitch contours; this feature may help these stimuli to sound voice-like, even when they are filtered such that only very high or very low frequencies remain. In Chapter 3 I show that using these stimuli, a 32-channel send-receive coil, and a set of headphones with very good frequency response at a wide range of frequencies, it is possible to demonstrate tonotopic organization in the STS of the right hemisphere.

The finding that speech and song share several statistical patterns underscores the very close acoustic similarities between these two forms of human communication. Both consist of speech sounds formed by the movement of the articulators, under which lies a pitch contour. Both are divided up into phrases that often (but not always) correspond to a breath group. The phrases of both are set apart by final lengthening and arch-shaped contours. In fact, at the level of a single phrase, there are only a few acoustic

differences between song and speech: 1) song features isochronous rhythm, a feature that has never been found to be present in human speech. 2) The song of professional singers is marked by a higher degree of pitch stability during each syllable (Lindblom and Sundberg 2007; this is almost certainly far less true for untrained singers). It is this stability within a syllable that allows listeners to assign heard pitches to scale steps. 3) In song, intervals between pitches tend to be closer to exact multiples of a semitone than in speech, and finally, 4) as a result of scale structure, 2-semitone intervals are much more common in song than in speech.

**Music vs. speech: a processing difference?**

Given these rather minimal differences between speech and song, it's surprising that everyday speech is not normally heard as song—occasionally natural speech must contain flat pitch contours, for example, if only by chance. In fact, Deutch et al. (2008) have demonstrated an example of a spoken phrase that, heard in its original context, sounds like speech, but that sounds as if it were sung when it is removed from context and repeated. In Chapter 4 I demonstrate that it is possible to find several more examples of these ambiguous speech/song stimuli, along with phrases spoken by the same speakers that do not sound like song when removed from context and repeated.

These ambiguous stimuli are useful tools for studying the neural correlates of the perception of speech and song, as they should differ from other randomly selected phrases spoken by the same speakers in only those acoustic dimensions both necessary and sufficient for causing stimuli to sound like song, rather than speech. As mentioned above, there is no a priori reason to believe that these acoustic dimensions should include basic properties of the stimuli such as speech rate, f0 range, mean f0, etc., and in fact analysis of the stimuli found that they did not differ on any of these dimensions. Instead, they differed only in the degree of pitch variation within a syllable and the amount of durational variability existing between stressed syllables. As a result, by using these stimuli in an fMRI study, it should be possible to isolate neural correlates of the perception of speech and song without the data being contaminated by irrelevant differences between the two categories of stimuli.

## Is music processing specialized in the brain? Evidence from brain damage

In the last couple of decades, there has been a great deal of research and speculation centered on the neural resources underpinning speech and song and, in particular, the extent to which they draw on overlapping resources. Some scholars, citing among other sources of evidence cases of impairment in music processing following brain damage in the absence of verbal impairment (and vice versa), have argued that music and speech are processed in almost entirely separate domain-specific modules (Peretz and Coltheart 2003). On the

other hand, some scholars have argued that music and speech are subserved in large part by shared domain-general processes (see, for example, Patel 2008). In this dissertation I will present evidence for an exceedingly high degree of overlap in the neural correlates of music and speech processing. How, then, to reconcile this finding with the evidence that amusia can be present without aphasia, and vice versa?

To answer this question, we need to examine these cases of music and speech impairment following brain damage in detail. All of the cases described below have been cited as evidence for the modularity of music processing (Marin and Perry 1999, Peretz and Coltheart 2003, Peretz and Zatorre 2005). In large part, however, these case studies did not examine processing of speech information contained in intonational contours—and in those cases that did, prosody processing was nearly always found to be impaired when music processing was impaired, or spared when music processing was spared. The overall picture that emerges from these studies is that music perception and prosody perception rely on, at the very least, somewhat overlapping processes.

I will first examine cases that have been described as amusia without aphasia. Peretz (1993), for example, described a patient with a moderate difficulty in detecting intervals, severe difficulty perceiving tonality, and normal speech comprehension. The patient performed worse than controls on a short-term pitch memory task as well. This pattern of impairments could potentially be simply caused by a deficit in the ability to represent fine-grained

pitch information, especially over time. This would likely lead to a disruption of prosodic processing as well, but the authors did not examine the patient's prosody perception. On the other hand, Peretz (1994) later studied a patient with deficits on melody naming and musical instrument naming who had no problems comprehending spoken words. The patient was, however, impaired on a test of the ability to use intonational information to resolve syntactic ambiguities, was less able than control subjects to match intonational contours with labels (such as "declarative" and "question"), and impaired on identification of speaking voices. The authors argue that their results support modularity of music, as prosodic processing was not impaired to the same degree as music perception, but as music contains much finer-grained pitch information than is found in intonational contours, it's likely that prosody perception is simply less affected by impairments in the ability to make fine pitch discriminations.

Griffiths (1997) examined a patient who had problems recognizing familiar melodies, but who did not display any clear signs of aphasia. The patient also did not have any problems processing affective prosody—that is, they were able to distinguish between speech that was meant to convey different emotions. That is not, however, conclusive evidence that prosodic processing was completely unimpaired in this patient, as emotion is conveyed in part by cues such as rate and intensity. Moreover, the emotional cues contained in pitch consist of overall pitch range and mean pitch levels, and

therefore a deficit in, for example, the ability to detect changes in pitch direction should not affect the perception of affective prosody (Juslin and Laukka 2003). The patient also had a deficit on a task in which they heard two three-tone sequences, and were asked if the second tone differed between the sequences. The patient was, however, only impaired on this task when it was presented at a fast rate, suggesting that the music deficit was due to an impairment in tone sequencing or temporal resolution of pitch information, rather than domain-specific processing.

Piccirilli et al. (2000) examined a patient with impaired melody recognition, tonal memory, and music production, but who did not suffer from any problems with spoken word comprehension. The patient also did not suffer from problems with affective prosody perception. The patient's impaired tonal memory, however, leaves open the possibility that the music deficits were due to a lower-level domain-general impairment. Finally, Steinke et al. (2001) examined a patient with impairments in melody recognition, singing and rhythm reproduction. The patient exhibited impairments on several pitch-discrimination and rhythm-discrimination tasks, and compared to control subjects was impaired on prosody processing as well (it is unclear from the paper, unfortunately, what the exact nature of the prosody processing test was.)

Of the studies described above, therefore, all those who tested their patients' ability to perceive the information contained in the intonational

contours of speech found that they were impaired. Moreover, every one of these studies found that their patients suffered from impairments in basic, domain-general pitch processing that could underlie the music deficits also found.

I will now turn to the cases of aphasia without amusia. Metz-Lutz et al. (1984) examined a patient with poor performance on auditory comprehension, repetition, and writing to dictation tasks. The patient's ability to recognize instrumental melodies was normal, but the patient suffered from impairments in the ability to reproduce rhythms. Despite the patient's aphasia, she remained able to discriminate intonation contours—she could classify twenty tape-recorded sentences as affirmative statements, questions, imperative statements, or negative statements, and was able to discriminate and identify foreign languages as well.

Yaqub et al. (1988) examined a patient who, following a stroke, became unable to understand speech, but who did not report any problems with music perception (the authors did not formally test his musical abilities). The patient remained, however, able to tell when the same sentence was delivered as a question, imperative, or exclamation, information contained within the intonation contour. Takahashi et al. (1992) described a patient almost totally incapable of understanding individual words and unable to recognize emotional prosody, but who remained able to recognize melodies. Despite this preserved melody recognition, however, the patient was impaired on a number

of other music abilities, including pitch discrimination, rhythm discrimination, and melody discrimination.

Godefroy et al. (1995) studied a patient who, following a stroke, became unable to understand speech. Tests revealed that she had trouble recognizing individual syllables, especially when the syllables differed by place of articulation; discrimination of the voicing feature was spared. The authors did not test the patient's prosody perception. Finally, Mendez (2001) examined a patient who had difficulty understanding both speech and environmental sounds whose appreciation for music increased following his stroke. The author reports, however, that his speech inflection and prosody were unaffected, and that he "responded to prosody". It is, unfortunately, unclear what the author means by this last remark, as no formal tests of prosodic processing were conducted, but it suggests that the patient may have had some degree of preserved prosodic perception.

As brain damage can occur in different places and in different degrees of concordance with boundaries between brain regions, the overall picture painted by the studies described above is rather complex. Overall, though, it's fair to say that presenting this body of work as evidence that music and speech do not overlap in the brain is an oversimplification. At the very least, it's clear that impaired music processing is often correlated with impaired prosodic processing, even when speech perception is otherwise normal. On the other hand, when speech perception is impaired and music perception is spared,

prosodic perception is often found to be normal. Whether it is possible for prosody perception to be impaired without impairment to music perception—or vice versa—is a question that can only be answered by further research.

**Evidence from congenital amusia**

Congenital amusia (popularly known as "tone deafness") has also been cited as evidence that music is processed in domain-specific modules. Ayotte et al. (2002) found a group of congenital amusics by advertising for subjects who had lifelong trouble understanding and performing music. These subjects were, in fact, impaired on a variety of musical tasks, including sensitivity to dissonance and tune identification, but were not impaired on identification of lyrics, voices, and environmental sounds. The subjects were able to distinguish statements from questions, but were unable to do so when the linguistic information was removed and intonational contours were replaced by a series of static tones.

Further research, however, has revealed that amusics suffer from a variety of low-level pitch processing deficits. Hyde et al. (2004), for example, demonstrated that amusics had trouble detecting pitch changes of less than two semitones, but were not impaired on an analogous rhythm task. Foxton et al. (2004), similarly, found that amusics had higher thresholds than control subjects for detection of continuous and segmented changes in pitch, discrimination of pitch sequences, and determination of the direction of pitch

changes. Patel et al. (2005) demonstrated that amusics are less able than control subjects to discriminate sentence-question pairs or detect shifts of focus when hearing only the intonational contour of a sentence—both when each syllable is replaced by a static tone and when each syllable is replaced by its original pitch contour. Finally, Patel et al. (2008) found that amusics, as a group, were not more able to discriminate statement/question pairs and sentences differing in location of focus when the original sentences were presented vs. intonational contours, and found that three amusics actually performed worse on the speech discrimination task than on the tone discrimination task.

To summarize, when patients with amusia are tested on their ability to perceive information contained within intonational contours, they more often than not show deficits. Also, those patients with pure word deafness or generalized auditory agnosia combined with preserved musical abilities whose prosodic abilities have been examined have shown preserved ability to process intonational contour. Finally, subjects who suffer from congenital amusia also seem to have some difficulty in perceiving intonational contours. This link between prosodic and musical processing is also supported by the finding that music lessons lead to an increased ability to decode speech prosody (Thompson et al. 2004).

## Evidence from syntactic processing

Recent studies have also cast doubt on the theory that linguistic syntax is processed by a completely domain-specific system. Patel et al. (1998) found that syntactic violations in language and harmonic violations in music both led to the P600 ERP component with an equivalent scalp distribution in the two conditions. Follow-up studies using MEG and fMRI found that perception of harmonic violations was linked to activation in Broca's area (Maess et al. 2001, Koelsche et al. 2002). Finally, patients suffering from Broca's aphasia, a disorder of speech and syntactical processing thought to be specific to language, have also been shown to have an impaired ability to process harmonic relations (Patel et al. 2008).

Taken as a whole, these studies suggest a great deal of overlap between the processes subserving music and language. However, the speech/song illusion discovered by Deutsch et al. (2008) demonstrates that there must be some difference in the resources upon which music and language processing draw. The dramatic perceptual shift experienced when a listener hears a spoken phrase suddenly turn into song implies that there are different processes at work in the two different modes of listening.

## Potential processing differences between music and language

To form a prediction about what those differences in processing between music and language listening might be, we can return to the question

of what the differences are between the acoustic characteristics of sung and spoken phrases. In speech, the majority of the relevant information is contained in higher-frequency information (rapid shifts in formant frequencies characterizing vowels and consonant-vowel transitions) and in very fast noise-bursts (consonants). While there is information contained within pitch, it can for the most part be described as secondary: some supplementary information about the syntax of the sentence, some emotional cues, information about what words in the sentence carry more focus, and information about the speaker's intent, such as whether a sentence was meant as a question or a statement. Relatively little of this information is preserved in written language, besides the occasional italicized word indicating emphasis and the use of a small range of punctuation marks, and while it can be somewhat harder to communicate subtle shades of meaning and intent via the written word, it is far from impossible.

In music, on the other hand, pitch forms the foundation of tonality, the major source of tension and resolution in many forms of music. It would be clearly impossible to notate music in such a way that pitch information were removed but the overall "meaning" of a piece was preserved. Pitch in both music and speech can be processed in terms of contour, but pitch in music can also be processed in terms of jumps between the different discrete segments of a given culture's scales, and in terms of tonality, that is, the degrees of stability of different pitches and chords according to their position on a particular scale.

Nearly every musical culture that has been studied makes use of scales of five to seven discrete pitches arranged within an octave range. For the most part these scales use unequal steps, and feature intervals (differences between adjacent note pairs) whose ratios are close to simple fractions. (McDermott and Hauser 2005).

Another important difference between song and speech lies in the regular pulse of music, a feature found to some extent in almost all of the world's music (Cross 2003). This feature may be linked to the impulse to move in response to music, both by singing along with what a listener is hearing, and by moving the body in ways unrelated to the production of the song (i.e. dancing). The combination of music's regular rhythm and tendency to elicit movement in listeners causes it to be very useful as a way of synchronizing the behavior of groups of people (Bispham 2009).

Given the large amount of information contained in the pitch of song that is missing in the pitch of speech and the finer-grained temporal and frequency representations of pitch information necessary for assigning pitches to scale degrees, listening to song likely involves directing more attention to the pitch of the signal. This could correspond to an increase in activation in pitch-sensitive cortex, which has been identified via imaging studies as lateral Heschl's gyrus and the planum polare just anterior to Heschl's gyrus (Griffiths et al. 2001, Patterson et al. 2002, Warren 2003, Schmithorst and Holland 2003, Penagos et al. 2004). Music perception may also involve a greater

degree of auditory-motor integration, given the stronger tendency to both sing back, either to oneself or out loud, what is being heard, and the impulse to move along to music in other ways (dancing, tapping fingers, clapping, etc.) Areas in the posterior Sylvian fissure and the supramarginal gyrus, where the temporal and parietal lobes meet, have been found to be activated during a wide range of tasks involving auditory-motor integration (Hashimoto and Sakai 2003, Hickok et al. 2003, Hickok and Poeppel 2007, Toyomura et al. 2007, Zarate and Zatorre 2008), and may therefore show more activation when subjects listen to song, as opposed to speech. In Chapter 4 I show that listening to spoken phrases that transform into song when repeated is, in fact, linked to an increase in activation in these two areas, as compared to listening to repeated spoken phrases that continue to sound like speech.

**Structure of the remainder of the dissertation**

In Chapters 1 and 2 I show that birdsong, human song, human speech, and human instrumental music share several statistical patterns, which likely stem from motor constraints. Chapters 3 and 4 examine two consequences of this finding. First, the auditory system may have adapted to process stimuli following these patterns. As a result such stimuli may be useful tools for probing the structure and function of the auditory system, as I demonstrate in a study of the representation of frequency information in the auditory system. Second, speech and song are acoustically overlapping categories, and as a

result it's possible to find ambiguous speech/song stimuli that can be used to examine the neural correlates of speech and song perception.

## References

Ayotte, J., Peretz, I., and Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. Brain *125*, 238.

Belin, P., Fillion-Bilodeau, S., and Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods *40*, 531.

Belin, P., and Zatorre, R. (2003). Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport *14*, 2105.

Belin, P., Zatorre, R., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. Cognitive Brain Research *13*, 17–26.

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

Bispham, J. (2009). Music's "design features": Musical motivation, musical pulse, and musical pitch. Musicae Scientiae *Special Issue 2009-2010*, 29-44.

Bolinger, D. (1978). Intonation across languages. In Universals of Human Language Vol. II: Phonology, J. Greenberg, ed. (Palo Alto, CA: Stanford University Press), pp. 471–524.

Connell, B., and Ladd, D. (1990). Aspects of pitch realization in Yoruba. Phonology *7*, 1–29.

Cross, I. (2003). Music, cognition, culture, and evolution. In The Origins of Music, N. Wallin, B. Merker, and S. Brown, eds. (Cambridge, MA: MIT press).

Deutsch, D., Lapidis, R., and Henthorn, T. (2008). The speech-to-song illusion. The Journal of the Acoustical Society of America *124*, 2471.

Fant, G., Kruckenberg, A., and Nord, L. (1991). Prosodic and segmental speaker variations. Speech Communication *10*, 521–531.

Foxton, J., Dean, J., Gee, R., Peretz, I., and Griffiths, T. (2004). Characterization of deficits in pitch perception underlying'tone deafness'. Brain *127*, 801-810.

Godefroy, O., Leys, D., Furby, A., and De Reuck, J. (1995). Psychoacoustic deficits related to bilateral subcortical hemorrhages: A case with apperceptive auditory agnosia. Cortex *31*, 149–159.

Goller, F., Mallinckrodt, M., and Torti, S. (2004). Beak gape dynamics during song in the zebra finch. Journal of Neurobiology *59*, 289–303.

Griffiths, T., Rees, A., Witton, C., Cross, P., Shakir, R., and Green, G. (1997). Spatial and temporal auditory processing deficits following right hemisphere infarction. A psychophysical study. Brain *120*, 785-794.

Griffiths, T., Uppenkamp, S., Johnsrude, I., Josephs, O., and Patterson, R. (2001). Encoding of the temporal regularity of sound in the human brainstem. Nature Neuroscience *4*, 633–637.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In Speech Prosody 2002, International Conference.

Hartley, R., and Suthers, R. (1989). Airflow and pressure during canary song: direct evidence for mini-breaths. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology *165*, 15–26.

Hashimoto, Y., and Sakai, K. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. Human Brain Mapping *20*, 22–28.

Herman, R. (1996). Final lowering in Kipare. Phonology *13*, 171–196.

Herman, R., Beckman, M., and Honda, K. (1996). Subglottal pressure and final lowering in English. In Proceedings, Fourth International Conference on Spoken Language 1996.

Hickok, G., Buchsbaum, B., Humphries, C., and Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. Journal of Cognitive Neuroscience *15*, 673–682.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience *8*, 393–402.

von Hippel, P. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. Music Perception *17*, 315–328.

von Hippel, P., and Huron, D. (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. Music Perception *18*, 59–86.

Huron, D. (1996). The melodic arch in Western folksongs. Computing in Musicology *10*, 3–23.

Huron, D., and Ollen, J. (2003). Agogic contrast in French and English themes: Further support for Patel and Daniele (2003). Music Perception *21*, 267–271.

Hyde, K., and Peretz, I. (2004). Brains that are out of tune but in time. Psychological Science *15*, 356-360.

Jusczyk, P., Hirsh-Pasek, K., Kemler Nelson, D., Kennedy, L., Woodward, A., and Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. Cognitive Psychology *24*, 252–293.

Jusczyk, P., and Krumhansl, C. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. Journal of Experimental Psychology: Human Perception and Performance *19*, 627–640.

Juslin, P., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin *129*, 770–814.

Klatt, D. (1975). Vowel Lengthening is Syntactically Determined in a Connected Discourse. Journal of Phonetics *3*, 129–140.

Koelsch, S., Gunter, T., Cramon, T., Zysset, S., Lohmann, G., and Friederici, A. (2002). Bach speaks: a cortical "language-network" serves the processing of music. Neuroimage *17*, 956–966.

Liberman, M., and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In Language, Sound, Structure: : Studies in Phonology Presented to Morris Halle by his Teacher and Students, M. Aronoff and R. Oehrle, eds. (Cambridge: MIT Press), pp. 157–233.

Liebenthal, E., Binder, J., Spitzer, S., Possing, E., and Medler, D. (2005). Neural substrates of phonemic perception. Cerebral Cortex *15*, 1621-1631.

Lindblom, B., and Sundberg, J. (2007). The human voice in speech and singing. In Springer Handbook of Acoustics (Springer Verlag), pp. 669-706.

Maess, B., Koelsch, S., Gunter, T., and Friederici, A. (2001). Musical syntax is processed in Broca's area: an MEG study. Nature Neuroscience *4*, 540–545.

Marin, O., Perry, D., and Deutsch, D. (1999). Neurological aspects of music perception and performance. In The Psychology of Music (Academic Press), pp. 653–724.

McDermott, J., and Hauser, M. (2005). The origins of music: Innateness, uniqueness, and evolution. Music Perception *23*, 29-60.

Mendez, M. (2001). Generalized auditory agnosia with spared music recognition in a left-hander. Analysis of a case with a right temporal stroke. Cortex *37*, 139–150.

Metz-Lutz, M., and Dahl, E. (1984). Analysis of word comprehension in a case of pure word deafness. Brain and Language *23*, 13–25.

Myers, S., and Hansen, B. (2007). The origin of vowel length neutralization in final position: Evidence from Finnish speakers. Natural Language & Linguistic Theory *25*, 157–193.

Narain, C., Scott, S., Wise, R., Rosen, S., Leff, A., Iversen, S., and Matthews, P. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. Cerebral Cortex *13*, 1362-1368.

Nowicki, S. (1987). Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere. Nature *325*, 53-55.

Patel, A. (2008). Music, Language, and the Brain (Oxford University Press, USA).

Patel, A., and Daniele, J. (2003). An empirical comparison of rhythm in language and music. Cognition *87*, B35–B45.

Patel, A., Foxton, J., and Griffiths, T. (2005). Musically tone-deaf individuals have difficulty discriminating intonation contours extracted from speech. Brain and Cognition *59*, 310-313.

Patel, A., Gibson, E., Ratner, J., Besson, M., and Holcomb, P. (1998). Processing syntactic relations in language and music: An event-related potential study. Journal of Cognitive Neuroscience *10*, 717–733.

Patel, A., Iversen, J., Wassenaar, M., and Hagoort, P. (2008a). Musical syntactic processing in agrammatic Broca's aphasia. Aphasiology *22*, 776–789.

Patel, A. D., Wong, M., Foxton, J., Lochy, A., and Peretz, I. (2008b). Speech intonation perception deficits in musical tone deafness (congenital amusia). Music Perception *25*, 357-368.

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

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

Penel, A., and Drake, C. (2004). Timing variations in music performance: Musical communication, perceptual compensation, and/or motor control? Perception and Psychophysics *66*, 545-562.

Peretz, I. (1993). Auditory atonalia for melodies. Cognitive Neuropsychology *10*, 21–56.

Peretz, I., and Coltheart, M. (2003). Modularity of music processing. Nature Neuroscience *6*, 688–691.

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., and Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. Brain *117*, 1283-1302.

Peretz, I., and Zatorre, R. (2005). Brain Organization for Music Processing. Annu. Rev. Psychol. *56*, 89-114.

Petkov, C., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N. (2008). A voice region in the monkey brain. Nature Neuroscience *11*, 367–374.

Piccirilli, M., Sciarma, T., and Luzzi, S. (2000). Modularity of music: evidence from a case of pure amusia. Journal of Neurology, Neurosurgery and Psychiatry *69*, 541-545.

Podos, J., Sherer, J., Peters, S., and Nowicki, S. (1995). Ontogeny of vocal tract movements during song production in song sparrows. Animal Behaviour *50*, 1287–1296.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. Journal of the Acoustical Society of America *90*, 2956–2970.

Repp, B. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. The Journal of the Acoustical Society of America *88*, 622-641.

Schafer, A., Speer, S., Warren, P., and White, S. (2000). Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research *29*, 169–182.

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

Scott, D. (1982). Duration as a cue to the perception of a phrase boundary. The Journal of the Acoustical Society of America *71*, 996-1007.

Scott, S., Blank, C., Rosen, S., and Wise, R. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain *123*, 2400-2406.

Scott, S., Rosen, S., Lang, H., and Wise, R. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech—a positron emission tomography study. The Journal of the Acoustical Society of America *120*, 1075-1083.

Slifka, J. (2006). Respiratory system pressures at the start of an utterance. In Dynamics of speech production and perception, P. Divenyi, ed. (IOS Press), pp. 45-57.

Smith, B. (2002). Effects of speaking rate on temporal patterns of English. Phonetica *59*, 232–244.

Smith, B. (2000). Variations in temporal patterns of speech production among speakers of English. The Journal of the Acoustical Society of America *108*, 2438-2442.

Steinke, W., Cuddy, L., and Jakobson, L. (2001). Dissociations among functional subsystems governing melody recognition after right-hemisphere damage. Cognitive Neuropsychology *18*, 411–437.

Sundberg, J. (2000). Emotive transforms. Phonetica *57*, 95–112.

Sundberg, J. (1977). The acoustics of the singing voice. Scientific American *236*, 82–91.

Takahashi, N., Kawamura, M., Shinotou, H., Hirayama, K., Kaga, K., and Shindo, M. (1992). Pure word deafness due to left hemisphere damage. Cortex *28*, 295-303.

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

Toyomura, A., Koyama, S., Miyamaoto, T., Terao, A., Omori, T., Murohashi, H., and Kuriki, S. (2007). Neural correlates of auditory feedback control in human. Neuroscience *146*, 499–503.

Truckenbrodt, H. (2004). Final lowering in non-final position. Journal of Phonetics *32*, 313–348.

Vaissière, J. (1995). Phonetic explanations for cross-linguistic prosodic similarities. Phonetica *52*, 123–130.

Vaissière, J. (1991). Rhythm, accentuation, and final lengthening in French. In Music, Language, and Brain, J. Sundberg, L. Nord, and R. Carson, eds. (Houndsmills: Macmillan), pp. 108-120.

Venditti, J., and Santen, J. (1998). Modeling vowel duration for Japanese text-to-speech synthesis. In Fifth International Conference on Spoken Language Processing.

Warren, J., Scott, S., Price, C., and Griffiths, T. (2006). Human brain mechanisms for the early analysis of voices. Neuroimage *31*, 1389–1397.

Warren, J., Uppenkamp, S., Patterson, R., and Griffiths, T. (2003). Separating pitch chroma and pitch height in the human brain. Proceedings of the National Academy of Sciences of the United States of America *100*, 10038-10042.

Westneat, M., Long, J., Hoese, W., and Nowicki, S. (1993). Kinematics of birdsong: functional correlation of cranial movements and acoustic features in sparrows. Journal of Experimental Biology *182*, 147-171.

Wild, J., Goller, F., and Suthers, R. (1998). Inspiratory muscle activity during bird song. Journal of Neurobiology *36*, 441-453.

Yaqub, B., Gascon, G., Nosha, M., and Whitaker, H. (1988). Pure word deafness (acquired verbal auditory agnosia) in an Arabic speaking patient. Brain *111*, 457-466.

Zarate, J., and Zatorre, R. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. Neuroimage *40*, 1871–1887.

# Chapter 1.

# Shared Pitch Patterns in Human Song and Birdsong.

## 1.1 Introduction

Diverse cultures differ widely in the rhythms and scales they use and the role music plays in everyday life. Nevertheless, recent research has begun to outline some patterns underlying this variation. This research has largely focused on western music but nonetheless reveals several commonalities within a rather diverse set of musical traditions. For example, large jumps in pitch, or "skips", are more often than not followed by changes in direction, or "reversals" (von Hippel and Huron 2000). Another pattern is found in the phrases of folk songs, which have been shown to follow descending contours more often than ascending contours and are more likely to follow arch contours, in which the pitch sharply rises at the beginning and falls at the end, than u-shaped contours (Huron 1996). Finally, classical pianists tend to lengthen the duration of notes just prior to major structural boundaries and do so to a greater extent if asked to perform expressively (Repp 1990), but this effect has yet to be demonstrated cross-culturally, and it is not yet known whether lengthening is also present separately from performance at the structural level. This final lengthening effect was present in song produced by a single professional opera singer as well (Sundberg 2000), but given that this study examined only a single opera singer performing a single musical excerpt, the finding is preliminary at best.

Although all three phenomena—the skip-reversal pattern, the melodic contour patterns, and phrase-final lengthening—have been discussed for

decades, the question of their origin has remained largely unaddressed. One possibility is that the tendency to produce these patterns is somehow directly specified genetically, perhaps because they help communicate with the listener—both the melodic arch and phrase-final lengthening, for example, could conceivably help listeners detect structural boundaries.

An alternative possibility is that these three patterns each stem from basic motor constraints. The vocal instrument consists of three parts: a power supply (the lungs), an oscillator (the vocal folds), and a resonator (the vocal and nasal cavities) (Sundberg 1977). Each of these three parts operates under separate constraints that could potentially give rise to the three patterns. The vocal folds, for example, are only able to produce pitches from a bounded distribution. Large jumps in pitch are likely to approach the edge of this range, at which point most of the pitches that could be produced would result in reversals. Supporting this interpretation of the pitch-skip pattern is the fact that only skips departing from or crossing the median are more likely to be followed by reversals; skips approaching the median are more likely to be followed by continuations (von Hippel and Huron 2000). The arch contour may be a consequence of the breath cycle, as subglottal pressure and the pitch of song are strongly correlated (Lindblom and Sundberg 2007) and subglottal pressure rises at the very beginning of a spoken utterance and falls at the very end. Finally, a need for the articulators, which are moving rapidly as they change the resonant frequencies of the vocal and nasal cavities, to slow down

before coming to a complete stop could account for final lengthening, as has been proposed for human speech (Myers and Hansen 2007).

One way to choose between the genetic specification and motor constraint hypotheses would be to examine the vocalizations of another species, such as birdsong. The bird vocal apparatus differs from that of the human in several important respects, including the much smaller size of its components and the fact that birds have two syrinxes rather than a single larynx. Nevertheless, like human song, bird song is produced by a system consisting of the lungs driving vibrating vocal folds, which produce a sound source that is then filtered through the vocal and nasal cavities. Birds, therefore, should operate under the same three constraints. Pitches in bird song are presumably drawn from a bounded distribution. Birds filter out the upper harmonics of each note by rapidly opening and closing their beaks as the fundamental rises and falls; the higher the fundamental frequency, the wider the bird needs to open its beak (Westneat et al. 1993, Goller et al. 2004, Podos et al. 2004). Since birds have two syrinxes rather than a single larynx, they are capable of very rapidly switching between two notes with rather different frequencies, which would require very rapid beak movement. Birds may, therefore, need to slow this movement as the end of a song approaches. Finally, birds are able to breathe in between individual notes of a song, a phenomenon that has been demonstrated in canaries (Hartley and Suthers 1989), zebra finches, and cowbirds (Wild et al. 1998). As a result, if the breath

cycle is tied to the arch contour, we would expect to find it at the level of the individual note. Evidence of similar patterns in birdsong would support the hypothesis that the patterns are due to shared motor constraints, as opposed to genetic specification, as the evolution of birds and humans diverged over 200 million years ago.

To test the motor constraint hypothesis, we conducted the first comparative study of human and avian song based on a taxonomically diverse sample of birdsong. Music and birdsong have been compared for centuries (Baptista and Keister 2005), but despite sustained interest in the topic, almost no empirical work has been done comparing the two. The little that has been done has largely focused on whether or not birdsong follows Western musical scales and has either used musical notation rather than examining the actual pitch contour (Wing 1951, Saunders 1959, Hartshorne 1973), or has shown negative results (Dobson and Lemon, 1977a, 1977b). The current study differs from these previous attempts in that it is hypothesis driven, that is, it postulates the existence of specific commonalities between humans and birds that stem from specific similarities in the structure of the song-production systems that the two animals share.

Even if a study were to find an empirical commonality between human song and a single birdsong, it would not imply any deeper link between the two. Given the thousands of birdsongs being sung in the wild, a few are likely to share features with human music solely by chance, and these songs will tend

to be given more attention by humans. Rigorous studies comparing birdsong to human music, therefore, should include songs from a variety of songbird families, to guard against such a sample bias.

### 1.2 Methods

In compiling the birdsong dataset, we gathered as taxonomically diverse a sample of birdsong as possible, based on the 84 songbird families listed in the *Howard and Moore Complete Checklist of the Birds of the World* (1984). All songs had at least five notes, a tonal quality strong enough for F0 analysis to be performed on them, low background noise, and significant tonal variation (that is, we excluded songs in which only a single note was repeated). All songs consisted of a sequence of notes preceded and followed by a long pause relative to the duration of the notes. Songs meeting these criteria were obtained from The Cornell Laboratory of Ornithology, the Borror Laboratory of Acoustics, the British Museum Library, compact discs accompanying *Music of the Birds* (Elliot 1999), *Nature's Music* (Marler and Slabbekorn 2004), and *The Singing Life of Birds* (Kroodsma 2004), and 12 internet sources. We were able to obtain songs meeting our criteria from 54 families; we then sampled one more species from 26 of those families, sampling when possible from the most speciose families, for a total of 80 species (see Table 1 for a full list of bird families, bird species, and bird common names used). This corpus is, to my knowledge, the most phylogenetically diverse collection of oscine songs ever analyzed. Note that, even with perfect knowledge of all existing oscine species,

not all 84 families would be represented, as at least a few (*Corvidae*, for example) do not produce songs. The F0 of each note was tracked using custom-written software in SIGNAL which traced the lowest harmonic in a narrow-band spectrogram with frequency resolution = 125 Hz, time resolution = 8 ms, and one FFT per 3 ms.

**Table 1.1.** Family, common name, and species name of each bird whose song we studied.

| **Family** | **Common Name** | **Species Name** |
|---|---|---|
| Acanthizidae | Striated fieldwren | Calamanthus fuliginosus |
| Aegithinidae | Common lora | Aegithina tiphia |
| Alaudidae | Crested lark | Galerida cristata |
| Alaudidae | Woodlark | Lullula arborea |
| Atrichornithidae | Noisy scrub-bird | Atrichornis clamosus |
| Bombycillidae | Cedar waxwing | Bombycilla cedrorum |
| Campephagidae | Madagascar cuckoo-shrike | Coracina cineria |
| Campephagidae | Scarlet minivet | Pericrocotus flammeus |
| Cardinalidae | Northern cardinal | Cardinalis cardinalis |
| Cardinalidae | Rose-breasted grosbeak | Pheucticus ludovicianus |
| Certhiidae | Eurasian treecreeper | Certhia familiaris |
| Certhiidae | Rusty-flanked treecreeper | Certhia nipalensis |
| Chloropseidae | Lesser green leafbird | Chloropsis cyanopogon |
| Cinclidae | American dipper | Cinclus mexicanus |
| Cisticolidae | Gray-capped warbler | Eminia lepida |
| Cisticolidae | Rattling cisticola | Cisticola chiniana |
| Colluricinclidae | Grey shrike-thrush | Colluricincla harmonica |
| Cracticidae | Pied butcherbird | Cracticus nigrogularis |
| Dicaeidae | Fire-breasted flowerpecker | Dicaeum ignipectus |
| Dicruridae | Lesser racket-tailed drongo | Dicrurus remifer |
| Dicruridae | Pale blue monarch | Hypothymis puella |

**Table 1.1.** Family, common name, and species name of each bird whose song we studied, continued.

| | | |
|---|---|---|
| Emberizidae | Field sparrow | Spizella pusilla |
| Emberizidae | White-throated sparrow | Zonotrichia albicollis |
| Estrildidae | Common waxbill | Estrilda astrild |
| Eupetidae | Chiming wedgebill | Psophodes occidentalis |
| Fringiliidae | European greenfinch | Carduelis chloris |
| Fringillidae | Oriental greenfinch | Carduelis sinica |
| Hirundinidae | Northern rough-winged swallow | Stelgidopteryx serripennis |
| Icteridae | Baltimore oriole | Icterus galbula |
| Icteridae | Eastern meadowlark | Sturnella magna |
| Malaconotidae | black-crowned tchagra | Tchagra senegala |
| Malaconotidae | Brubru shrike | Nilaus afer |
| Maluridae | Splendid fairywren | Malurus splendens |
| Maluridae | Superb fairy wren | Malurus cyaneus |
| Meliphagidae | Giant honeyeater | Gymnomyza viridis |
| Meliphagidae | Kauai oo | Moho braccatus |
| Mimidae | Northern mockingbird | Mimus polyglottos |
| Monarchidae | Black-faced monarch | Monarcha melanopsis |
| Motacillidae | Tree pipit | Anthus trivialis |
| Muscicapidae | Nightingale | Luscinia megarhynchos |
| Muscicapidae | White-browed bush robin | Tarsiger indicus |
| Nectariniidae | Bronze sunbird | Nectarinia kilimensis |
| Nectariniidae | Crimson sunbird | Aethopyga siparaja |
| Oriolidae | Indian golden oriole | Oriolus kundoo |
| Pachycephalidae | Golden whistler | Pachycephala pectoralis |
| Paridae | Black-capped chickadee | poecile atricapilla |
| Paridae | Turkestan tit | Parus bokharensis |
| Parulidae | Chestnut-sided warbler | Dendroica pensylvanica |
| Parulidae | Yellow warbler | Dendroica petechia |
| Passeridae | Mauritius fody | Foudia rubra |
| Petroicidae | New Zealand robin | Petroica australis |
| Petroicidae | South Island Tomtit | Petroica macrocephala |
| Peucedramidae | Olive warbler | Peucedramus taeniatus |
| Ploceidae | Spectacled weaver | Ploceus ocularis |

**Table 1.1.** Family, common name, and species name of each bird whose song we studied, continued.

| | | |
|---|---|---|
| Polioptilidae | Long-billed gnatwren | Ramphocaenus melanurus |
| Prunellidae | Brown accentor | Prunella fulvescens |
| Pycnonotidae | Orange-spotted bulbul | Pycnonotus bimaculatus |
| Pycnonotidae | Spectacled bulbul | Pycnonotus erythropthalmos |
| Regulidae | Firecrest | Regulus ignicapillus |
| Regulidae | Golden-crowned kinglet | Regulus satrapa |
| Remizidae | Eurasian penduline-tit | Remiz pendulinus |
| Rhipiduridae | Pied fantail | Rhipidura javanica |
| Rhipiduridae | White-browed fantail | Rhipidura aureola |
| Sittidae | Rock nuthatch | Sitta neumayer |
| Sturnidae | Common mina | Acridotheres tristis |
| Sylviidae | Japanese bush-warbler | Cettia diphone |
| Sylviidae | Willow warbler | Phylloscopus trochilus |
| Thraupidae | Summer tanager | Piranga rubra |
| Thraupidae | Variable seedeater | Sporophila corvina |
| Timaliidae | Hwamei | Garrulax canorus |
| Timaliidae | Spot-breasted laughingthrush | Garrulax merulinus |
| Trogloditidae | Carolina wren | Thryothorus ludovicianus |
| Troglodytidae | Canyon wren | Catherpes mexicanus |
| Turdidae | Slaty-backed nightingale thrush | Catharus fuscater |
| Turdidae | White-browed shortwing | Brachypteryx montana |
| Vangidae | Rufous vanga | Schetba rufa |
| Viduidae | Pin-tailed whydah | Vidua macroura |
| Vireonidae | Red-eyed vireo | Vireo olivaceus |
| Vireonidae | Green shrike-vireo | Vireolanius pulchellus |
| Zosteropidae | Mascarene white-eye | Zosterops borbonicus |

We also examined a database of over 18,000 folk songs taken from the Essen musical data package, which consists of songs from the modern states of Austria, Croatia, the Czech Republic, Germany, Luxembourg, the Netherlands,

Poland, Romania, Russia, Serbia, Slovakia, Slovenia, Switzerland, the Ukraine, China, Scotland, Ireland, and Nova Scotia (see Table 2 for a list of the number of phrases and songs taken from each region). The songs in this database are encoded using musical notation. Durations values are based on the time signature; for example, for a song with a time signature of 4/4, a quarter note would be given a value of one. Pitches were encoded as a distance in semitones from the root note of the scale; for example, if the song were in the key of A major, A440 would be given a value of 0. (See Figure 1.1 for examples of how we encoded the pitch of a sample birdsong and musical phrase.) This database is not diverse enough to make claims about true musical universals. Nevertheless, we can make a strong prediction, based on the motor constraint hypothesis, that patterns found in this database and in birdsong are likely to be found in music from other cultures as well.

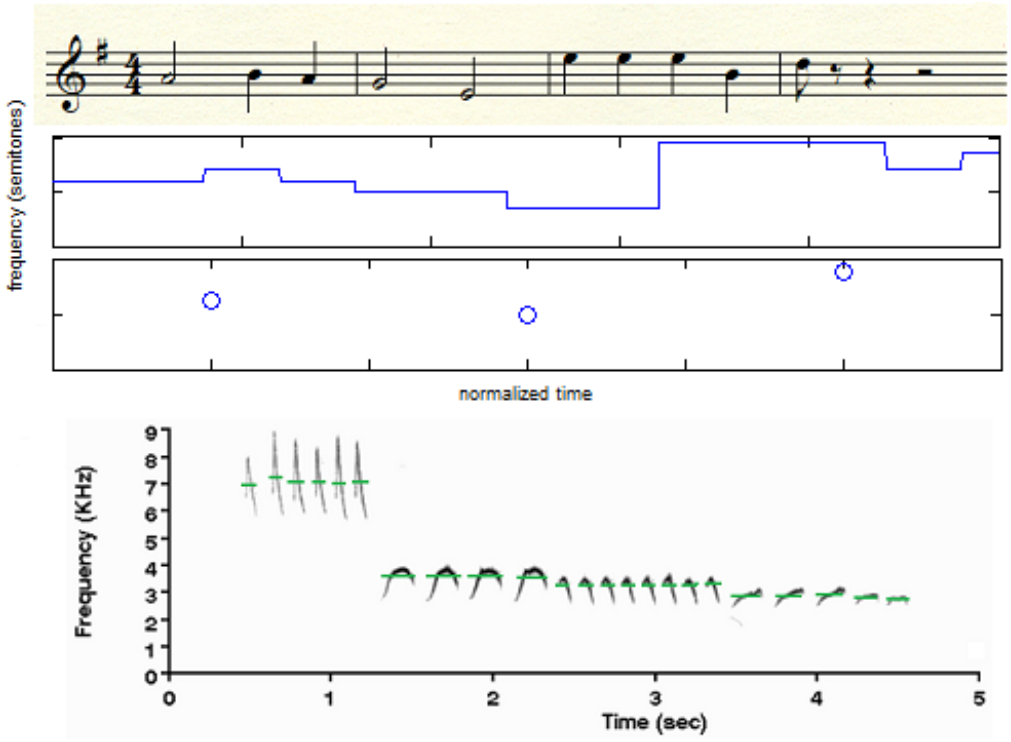


**Figure 1.1.** Pitch encoding method for human and bird song.

**Table 1.2.** Number of phrases and songs taken from each country of origin for the folksongs contained in the analyzed database.

| Country | Number of phrases | Number of songs |
|---|---|---|
| China | 10259 | 2041 |
| Germany | 30977 | 5252 |
| Ireland | 343 | 63 |
| Luxemburg | 3563 | 621 |
| Belgium | 130 | 26 |
| Netherlands | 313 | 62 |
| Switzerland | 550 | 99 |
| Yugoslavia | 303 | 119 |
| Denmark | 41 | 9 |
| Czech Republic | 165 | 43 |

**Table 1.2.** Number of phrases and songs taken from each country of origin for the folksongs contained in the analyzed database, continued.

| | | |
|---|---|---|
| Austria | 817 | 130 |
| Sweden | 52 | 12 |
| Hungary | 189 | 46 |
| Norway | 9 | 2 |
| Romania | 114 | 28 |
| France | 3113 | 589 |
| Poland | 117 | 25 |
| Greece | 3 | 1 |
| Russia | 213 | 51 |
| Finland | 4 | 1 |
| Brazil | 2 | 1 |
| England | 32 | 4 |
| USA | 61 | 10 |
| Italy | 18 | 4 |
| Canada | 4 | 1 |
| Iceland | 10 | 2 |
| Mexico | 77 | 11 |
| Japan | 7 | 1 |
| India | 6 | 1 |
| Indonesia | 14 | 1 |
| Scotland | 422 | 51 |
| Nova Scotia | 885 | 152 |

Bird song and human song each contain different hierarchical levels that are difficult to map onto one another. Human song can be divided into notes, phrases, and entire songs, and phrases are often separated by a breath and, therefore, a pause. Bird song can be divided into individual notes, motifs consisting of two or three notes, and entire songs. Unlike humans, birds are able to breathe in between notes, although they nonetheless take larger breaths between songs (Hartley and Suthers 1989, Wild et al. 1998). In this

paper we have chosen to compare entire bird songs to human song phrases. One argument for taking this approach is that human song phrases and bird songs are, on average, of comparable length, while human songs tend to be significantly longer (Figure 1.2).

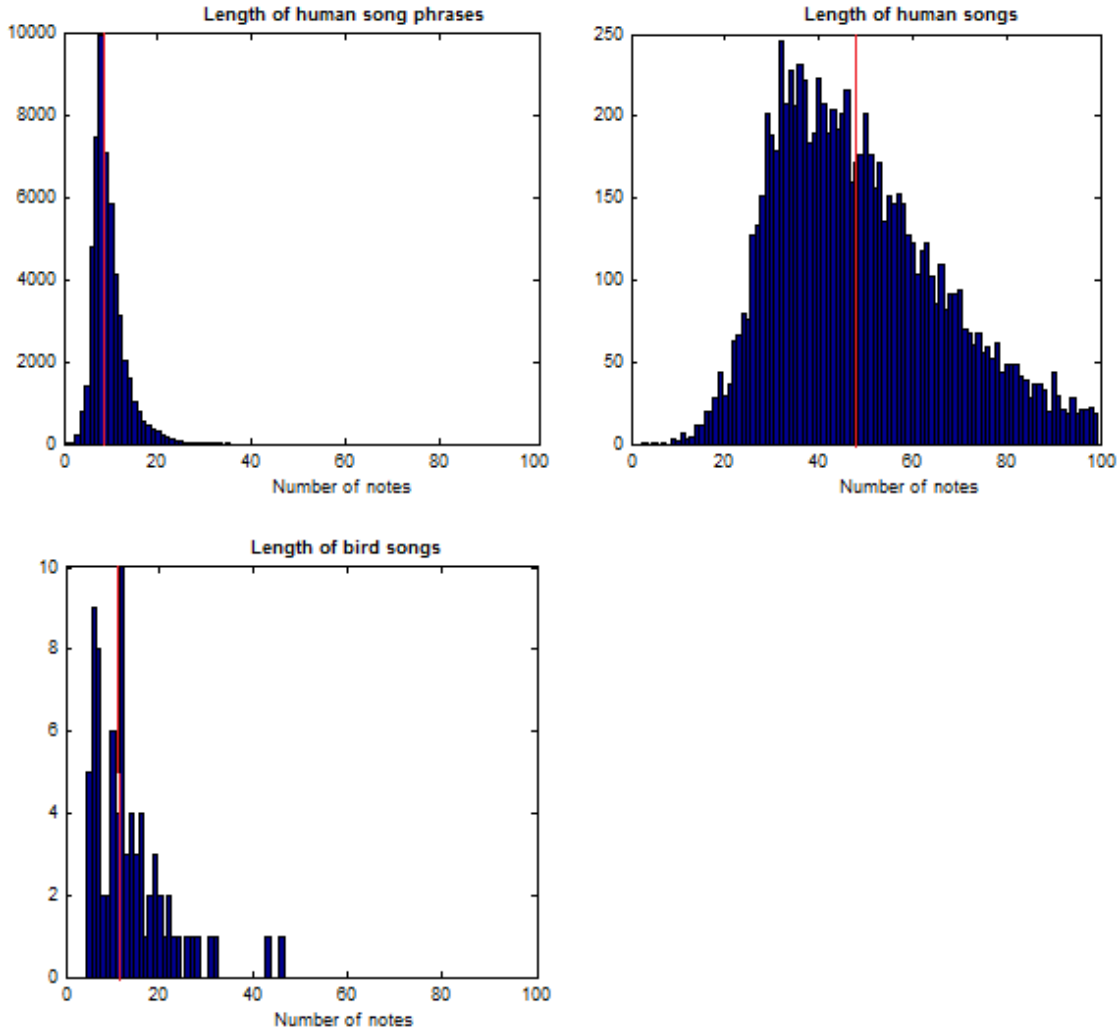


**Figure 1.2.** Histogram of length in number of notes of human songs, human song phrases, and birdsongs. Lengths of bird songs and human song phrases are quite similar, while human songs are, on average, much longer.

## 1.3 Results

To test for the existence of final lengthening in bird song we compared, for each note, the ratio of the duration of that note to the average duration of all the notes in a given song. For birdsong and human song, this ratio was significantly larger for the final note of a phrase than for all other notes. To determine whether this pattern holds true for a sample of non-western song and to examine the robustness of the phenomenon, we ran the same analysis for German and Chinese songs. The final notes of songs from both regions were significantly longer on average than all other notes (Figure 1.4). All significance testing was conducted using Mann-Whitney U tests.

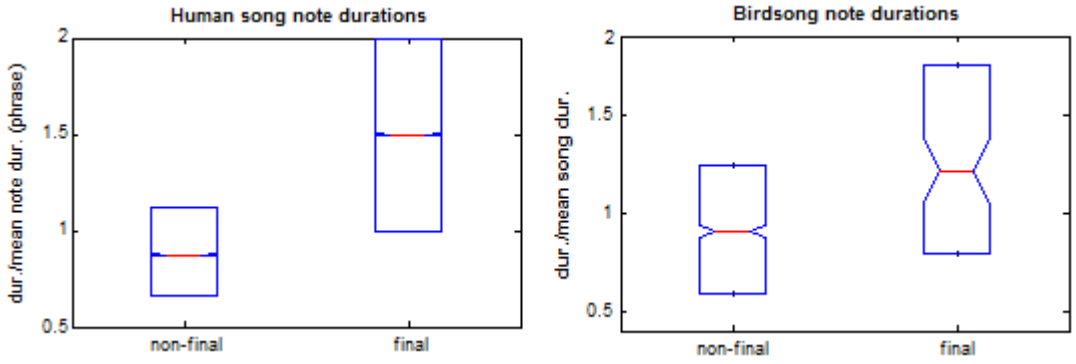


**Figure 1.3.** Final lengthening in human song and birdsong. For each note, we computed the ratio of the duration of that note to the mean duration of all notes for the entire song (for birds) or song phrase (for humans). For both birds and humans, this ratio was larger for final notes than for all other notes. The notches illustrate comparison intervals; two medians are significantly different at p = .05 if their intervals do not overlap.

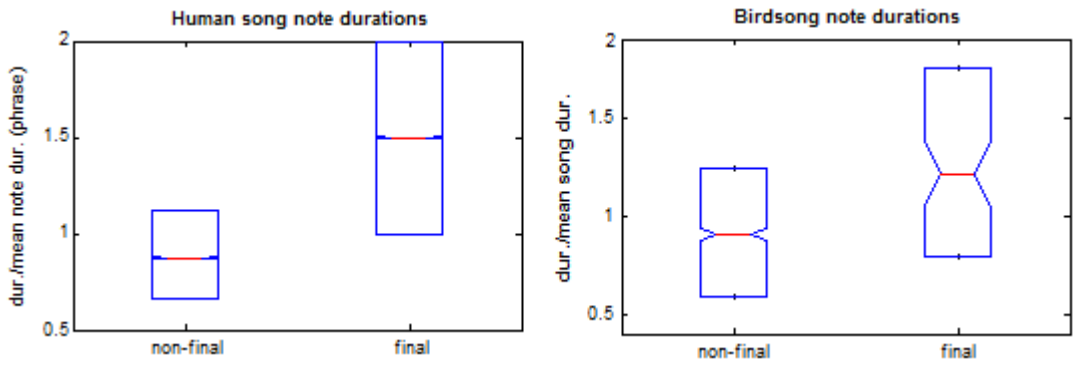


**Figure 1.4.** Final lengthening in German and Chinese song. The notches illustrate comparison intervals; two medians are significantly different at p = .05 if their intervals do not overlap.

Skips were defined in both human song and birdsong as pitch intervals greater than 50% of the tessitura. The tessitura was defined as the standard deviation of the distance in semitones of each pitch in a given sentence/theme from the mean pitch of that sentence/theme. In both domains skips were more likely to be followed by reversals; however, this only held true for skips departing from or crossing the median pitch. Skips approaching the median were instead more likely to be followed by continuations (Figure 1.5). For human song, skips landing on the median were more likely to be followed by continuations than reversals, suggesting that, all else being equal, there is a trend for human melodic lines to continue in the same direction. When examined separately, both Chinese and German songs exhibited the same pattern (Figure 1.6), suggesting that the effect is not limited to Western music. All of the differences described above were significant according to the binomial test.

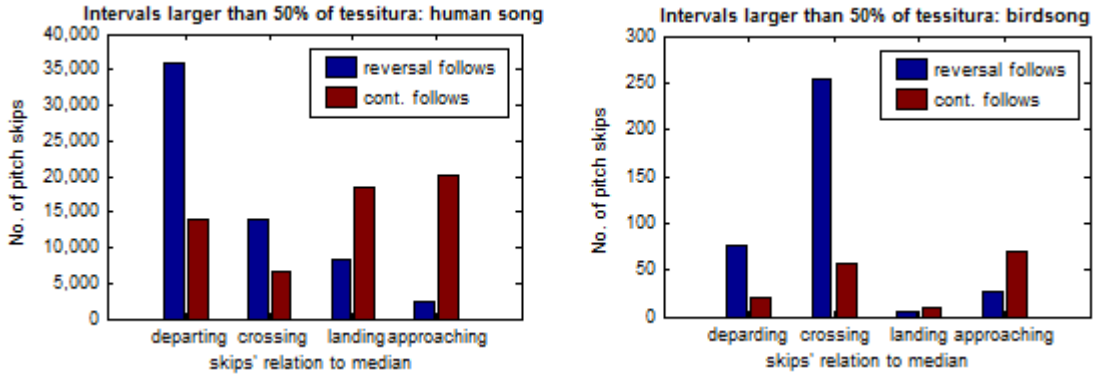


**Figure 1.5.** Skip-reversal patterns in human song and birdsong. In both domains, skips departing from or crossing the median are more likely to be followed by reversals, while skips approaching the median are more likely to be followed by continuations.

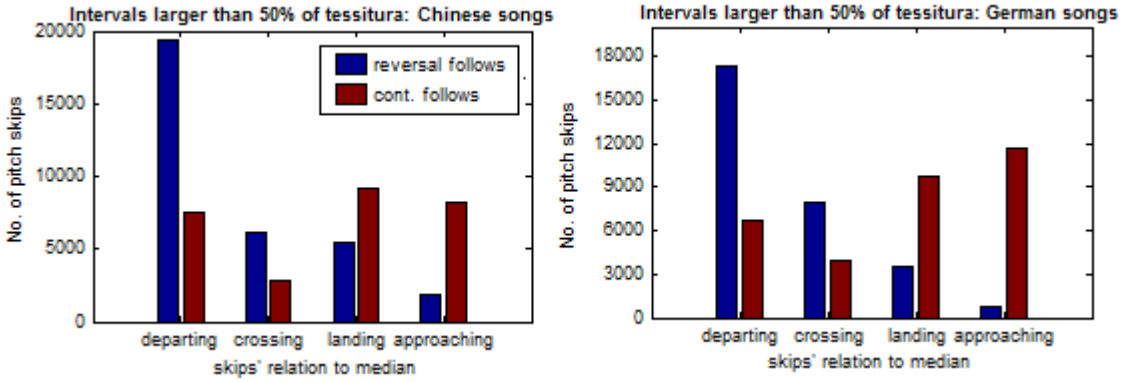


**Figure 1.6.** Skip-reversal patterns in Chinese and German song.

An alternative way to examine the skip-reversal pattern is to divide skips into those followed by a continuation and those followed by a reversal, then examine where each skip fell in relation to the mean pitch of the phrase in which it occurred. Figure 1.7 shows that, for human song and birdsong, skips followed by reversals tend to fall further away from the mean. For human

song, skips followed by reversals had an average absolute value z-score of 0.5547, while skips followed by continuations had an average z-score of 0.9787. For birdsong, skips followed by reversals had an average z-score of 0.5418, while skips followed by continuations had an average z-score of 0.9904. Figure 1.8 shows that the same trend holds true for German (average z-score of skips followed by reversals = 0.5266, average z-score of skips followed by continuations = .9682) and Chinese songs (average z-score of skips followed by reversals = 0.5867, average z-score of skips followed by continuations = .9920). All four of these differences were significant according to the Mann-Whitney U test.

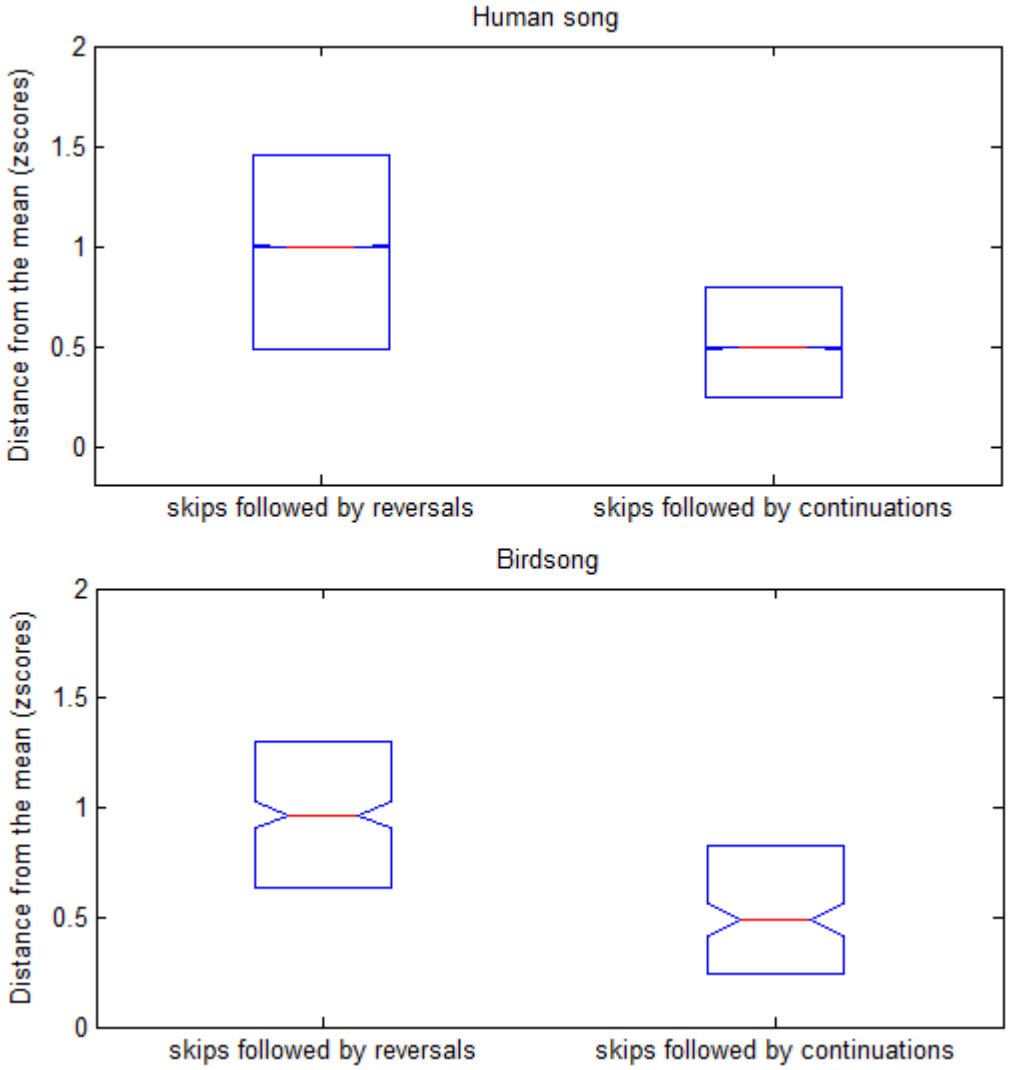


**Figure 1.7.** In human song and birdsong, skips that are followed by reversals fall further from the mean than skips that are followed by continuations. The notches illustrate comparison intervals; two medians are significantly different at p = .05 if their intervals do not overlap.

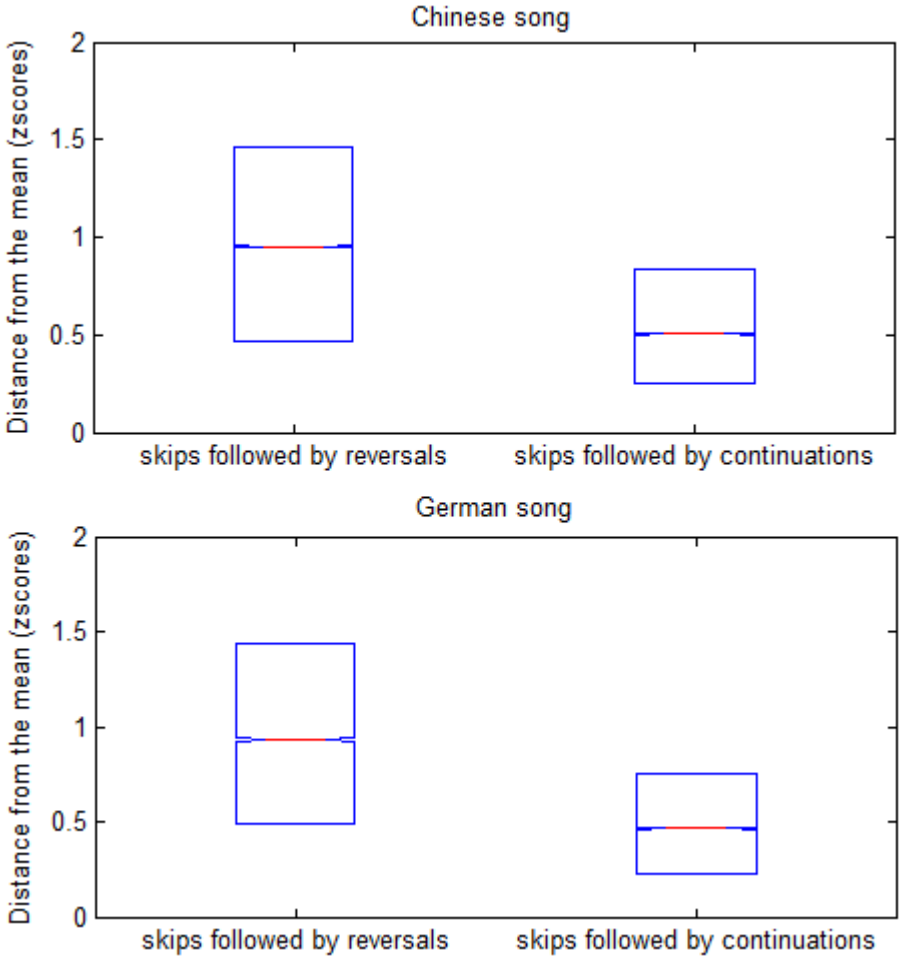


**Figure 1.8.** In Chinese song and German song, skips that are followed by reversals fall further from the mean than skips that are followed by continuations. The notches illustrate comparison intervals; two medians are significantly different at p = .05 if their intervals do not overlap.

To investigate contour classification patterns in human song phrases and birdsong individual notes, we divided each phrase or note into fifty equal time steps, computed the pitch of each time step as a distance in semitones from the average pitch of that contour, and then averaged across phrases/notes. In both birdsong notes and human song phrases the resulting

contour resembled a combination of a constant downward slope (declination) and an arch. The arch can be seen more clearly when the declination is removed via detrending (Figure 1.9). This same pattern was found to hold true when the Chinese and German songs were analyzed individually (Figure 1.10), suggesting that it is not limited to western folk songs.

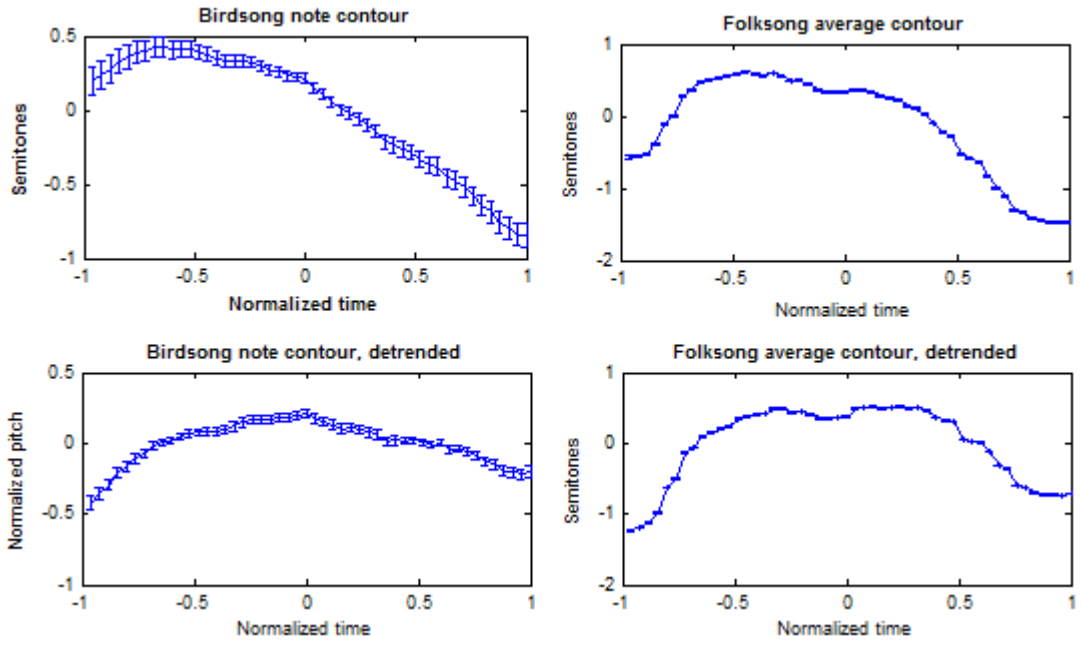


**Figure 1.9.** Average pitch contour of human song phrases and birdsong individual notes, both unaltered and detrended.

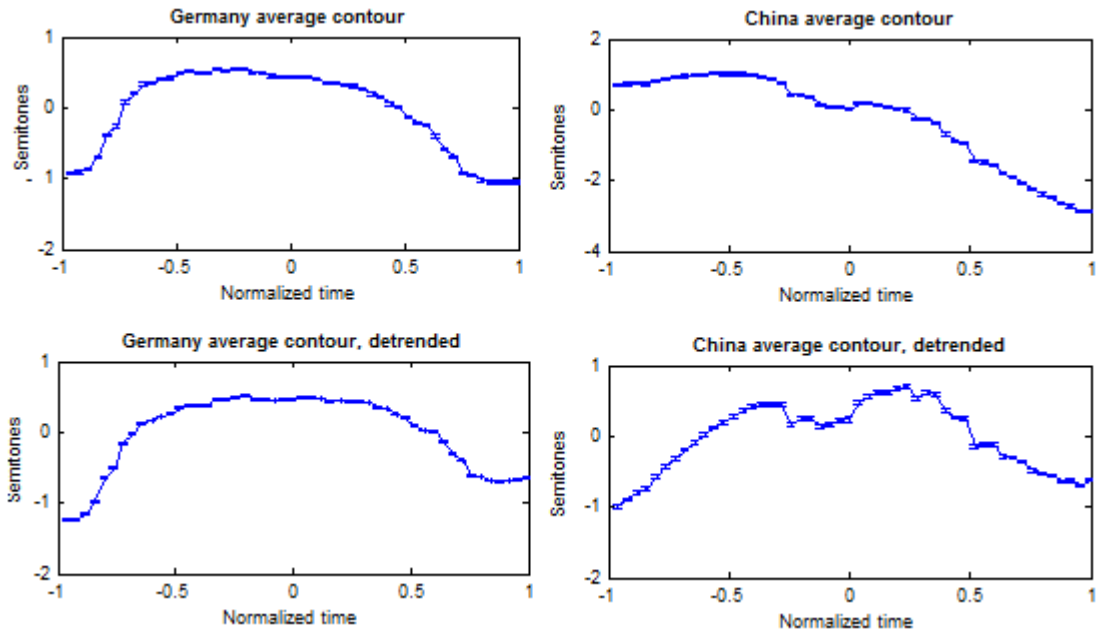


**Figure 1.10.** Average pitch contour of German and Chinese song phrases, both unaltered and detrended.

To further investigate pitch contour patterns in folksong and birdsong we divided each note or folk song into thirds, then took the average of all of the pitches in each third. Next, we assigned each pitch contour a classification based on the relationship between the middle third and the first and last third (pitches within 0.5 semitones of one another were marked as equivalent). Nine different classifications are possible: a constant descending slope (declination), a constant ascending slope, an convex arch shape, a concave u-shape, a horizontal line, and four combinations of horizontal and sloped contours—horizontal-ascending, horizontal-descending, ascending-horizontal, and descending-horizontal. (See Figure 1.11 for simple drawings of the shapes

of these nine different contours.) In both human song and birdsong individual notes, descending contours were far more common than ascending contours, and convex contour shapes were more common than concave contour shapes (Figure 1.11). This pattern was replicated for the German and Chinese song subsets as well (Figure 1.12). The differences in the incidence of descending versus ascending and convex versus concave contours were significant in all cases according to the binomial test.

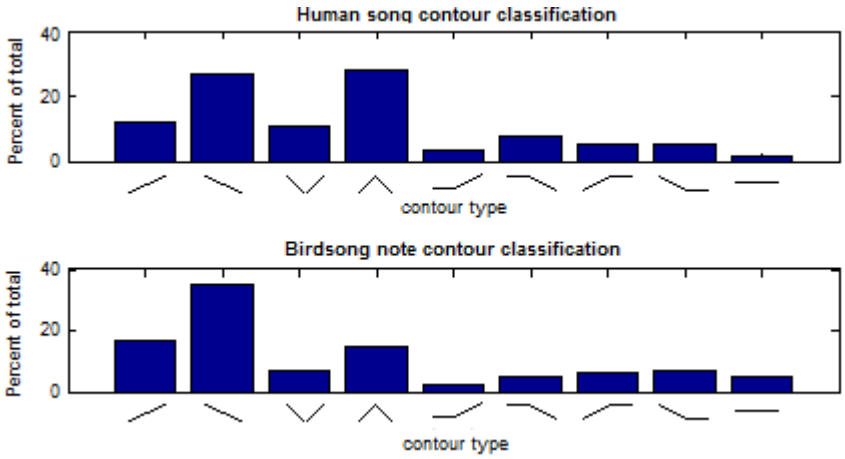


**Figure 1.11.** Classification of pitch contours in human song phrases and birdsong individual notes.

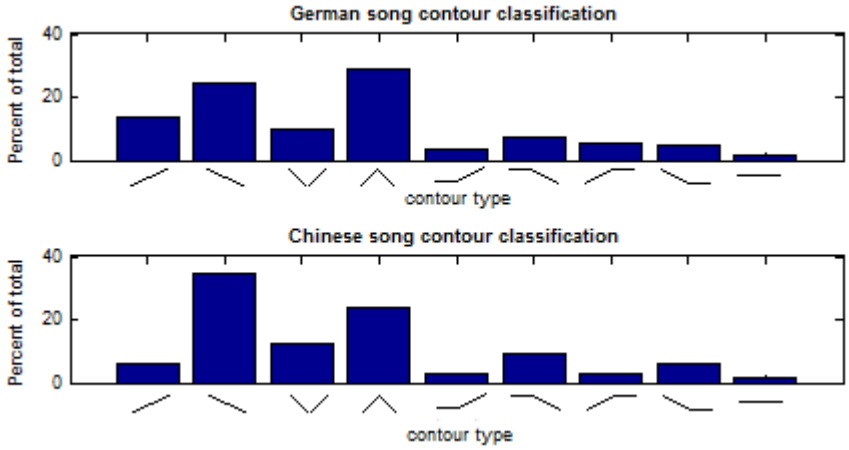


**Figure 1.12.** Classification of pitch contours in German and Chinese song phrases.

A corollary of our motor constraint hypothesis is that any salient differences in motor constraints on song in the two groups should result in statistical differences in song structure. Perhaps the most striking difference between the song-production systems of birds and humans is that while humans produce pitch using a single larynx, birds have instead two syrinxes, located just below the branching of the trachea. This could conceivably enable birds to make rapid pitch jumps which would be difficult or impossible for humans, as birds are able to modify the tension of the vocal folds within the two syrinxes independently. As a result, the tendency for small intervals to predominate seen in human song (Huron 2006) may be weaker or nonexistent in bird song.

To test this hypothesis, for each birdsong we computed the mean interval size. We then scrambled the notes of the song 100 times, computed the mean interval size for each scrambled song, and took the mean of all 100 means. Taking the ratio of the mean interval size to the average scrambled mean interval size gave us the "compression ratio", a measure of the degree to which small intervals dominate in a given song more than would be expected based on a random ordering of the pitches.

In order to directly compare the compression ratios of human and bird song, we used a Monte Carlo analysis. First, we excluded any song phrases from the Essen Musical Data Package that contained rests or did not contain any pitch variation. Then, for each birdsong, we randomly selected a human song phrase of the same length. (Five birdsongs were long enough that there were no human song phrases of the same length; these birdsongs were excluded from the analysis.) We computed the compression ratio for each song phrase in this set, and then finally took the median value of the resulting set of compression ratios. This entire process was iterated 1000 times, giving us a distribution of 1000 median compression ratios for human song.

The topmost panel of Figure 1.13 shows a histogram of the compression ratios for each birdsong example, along with a red line representing the median compression ratio for the entire set. The middle panel of Figure 1.13 shows a histogram of the compression ratios for one random sample of human song phrases matching the bird songs in length, along with a green line

representing the median compression ratio for the whole set. Finally, the lower panel of Figure 13 shows a histogram of all 100 median compression ratios for human song phrases generated during the Monte Carlo analysis in green, and a red line once again representing the median compression ratio for the birdsongs. The birdsong median compression ratio (1.235) was lower than the minimum of the entire distribution of median compression ratios for human song (1.254). Thus, the tendency for small intervals to predominate is in fact significantly weaker in bird song than it is in human song.

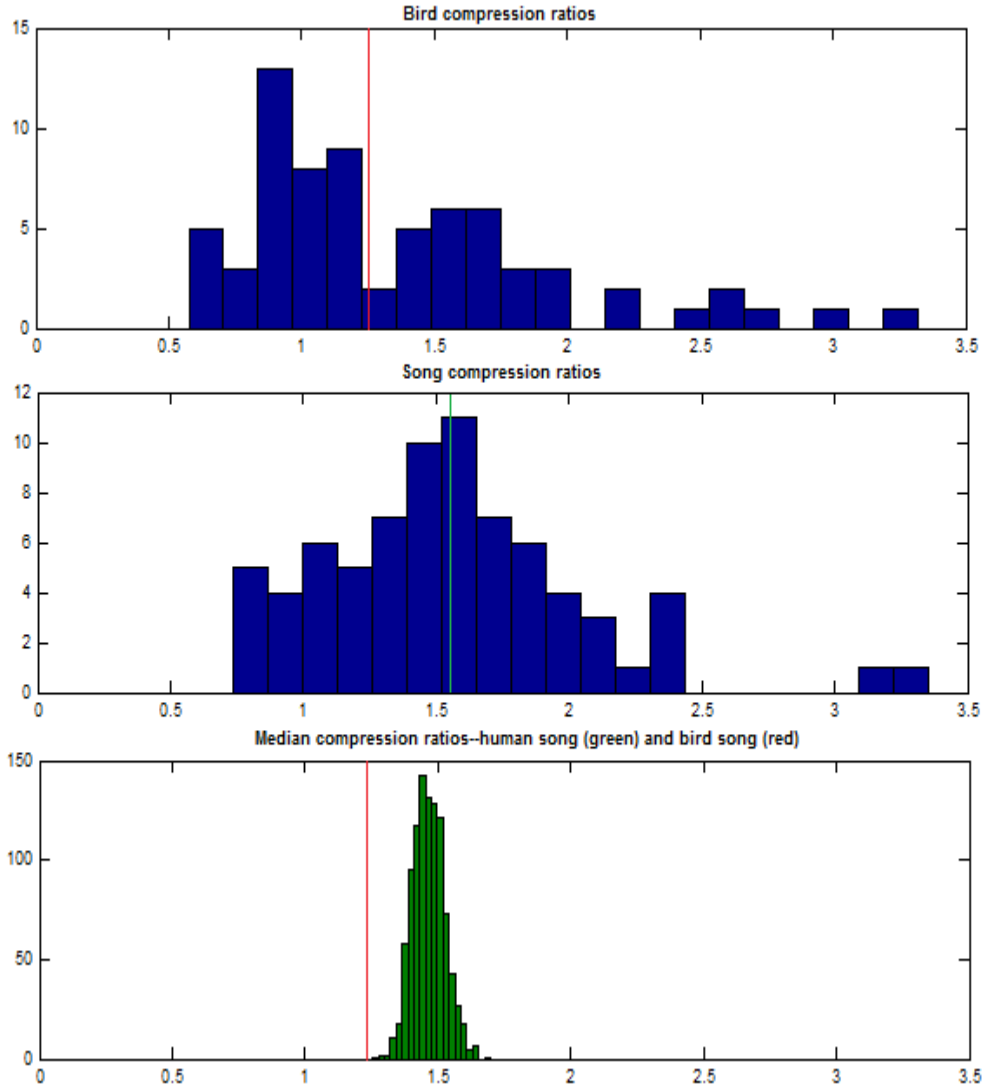


**Figure 1.13.** Tendency of bird songs and human songs to contain smaller pitch intervals than one would expect based solely on the pitch values used in a given song or song phrase. This tendency is weaker for bird songs than for human songs.

## 1.4 Discussion

Three patterns are found in a diverse sample of folk songs: final lengthening, a tendency for large jumps in pitch to be followed by reversals,

and a tendency for pitch contours to follow an arch shape. These patterns exist in a phylogenetically diverse sample of birdsong as well, suggesting that they stem from motor constraints rather than being specified in the genome.

One way to further test the motor constraint hypothesis would be to examine song structure in species that do not share one of the constraints that we argue lead to the three patterns described above. For example, we predict whale song will not show phrase-final lengthening, as whales sing very slowly compared to birds or humans, and they do not need to change the characteristics of a resonating cavity as they sing.

The regularities discussed above can have perceptual consequences for humans: for example, durational lengthening can be used as a cue for phrase boundaries in human speech (Krumhansl and Jusczyk 1990), while infants use both durational lengthening and falls in pitch to locate the ends of musical phrases (Jusczyk and Krumhansl 1993). It remains to be seen whether birds make use of these regularities when perceiving song—for example, whether the arch shape of individual notes helps set them apart from adjacent notes or notes of other birds in a noisy environment. Future work could also investigate the extent to which the auditory systems of birds and humans are adapted to process signals that share the characteristics of vocalizations. The auditory system may gain an advantage when processing simple, abstract sounds like pure tones if the sounds, for example, follow a pitch contour typical of a simple vocalization, and this advantage could range from lowered thresholds for

detection to increased temporal resolution. If such an advantage does exist, it might lead singers and musicians to exaggerate arch-shaped pitch contours, or produce them even when they are not necessitated by a physical constraint. In fact, in Chapter 2 I will demonstrate that composers of Western classical music do, on average, follow arch-shaped pitch contours, even when writing instrumental music. If advantages do exist for processing abstract stimuli with vocalization-like characteristics, one could then investigate the neural basis of these advantages, which could help shed light on the representation of basic characteristics of speech and song stimuli in auditory cortex. If the arch pattern, for example, does stem from the breath cycle, then it is rather ancient, even in evolutionary terms, and the nervous system would have had much more time to adapt to its presence than to, for example, the existence of speech formants.

In summary, I have shown evidence that there are several structural patterns shared by folk songs from a wide variety of geographical regions, and that these patterns are likely to have been shaped by physical production constraints rather than by genetic features governing universal structure. As a result, the action of these constraints on songs can be seen across species boundaries in a phylogenetically diverse corpus of oscine songs and human songs.

**Acknowledgement**

Chapter 1, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Frank Russo and Aniruddh Patel. Permission from all co-authors has been obtained.

## References

Baptista, L., and Keister, R. (2005). Why birdsong is sometimes like music. Perspectives in biology and medicine *48*, 426–443.

Dobson, C., and Lemon, R. (1977a). Bird song as music. The Journal of the Acoustical Society of America *61*, 888.

Dobson, C. W., and Lemon, R. E. (1977b). Markovian versus rhomboidal patterning in the song of Swainson's thrush. Behaviour *62*, 277–297.

Elliott, L. (1999). Music of the Birds: A Celebration of Bird Song (Houghton Mifflin Company).

Goller, F., Mallinckrodt, M., and Torti, S. (2004). Beak gape dynamics during song in the zebra finch. Journal of Neurobiology *59*, 289–303.

Hartshorne, C. (1973). Born to sing (Indiana University Press).

von Hippel, P., and Huron, D. (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. Music Perception *18*, 59–86.

Howard, R., and Moore, A. (1991). A Complete Checklist of the Birds of the World (Academic Press San Diego).

Huron, D. (2007). Sweet Anticipation: Music and the Psychology of Expectation (MIT press).

Huron, D. (1996). The melodic arch in Western folksongs. Computing in Musicology *10*, 3–23.

Jusczyk, P., and Krumhansl, C. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. Journal of Experimental Psychology: Human Perception and Performance *19*, 627–640.

Kroodsma, D. (2005). The Singing Life of Birds: the Art and Science of Listening to Birdsong (Houghton Mifflin Harcourt (HMH)).

Krumhansl, C., and Jusczyk, P. (1990). Infants' perception of phrase structure in music. Psychological Science *1*, 70-73.

Lindblom, B., and Sundberg, J. (2007). The human voice in speech and singing. In Springer Handbook of Acoustics (Springer Verlag), pp. 669-706.

Marler, P., and Slabbekoorn, H. (2004). Nature's Music: The Science of Birdsong (Academic Press).

Myers, S., and Hansen, B. (2007). The origin of vowel length neutralization in final position: Evidence from Finnish speakers. Natural Language & Linguistic Theory *25*, 157–193.

Podos, J., Sherer, J., Peters, S., and Nowicki, S. (1995). Ontogeny of vocal tract movements during song production in song sparrows. Animal Behaviour *50*, 1287–1296.

Repp, B. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. The Journal of the Acoustical Society of America *88*, 622-641.

Saunders, A. (1959). Octaves and kilocycles in bird songs. The Wilson Bulletin *71*, 280–282.

Sundberg, J. (2000). Emotive transforms. Phonetica *57*, 95–112.

Sundberg, J. (1977). The acoustics of the singing voice. Scientific American *236*, 82–91.

Wing, A. (1951). Notes on the song series of a hermit thrush in the Yukon. The Auk *68*, 189–193.

## Chapter 2.

## Pitch Patterns in Speech and Instrumental Music.

## 2.1 Introduction

Chapter 1 demonstrated that human song and birdsong share three statistical patterns, each of which may be the consequence of a motor constraint arising from one of the three parts of the song-production system. The oscillator can only produce pitches within a limited range, which leads large jumps in pitch to be more often followed by pitch movements in the opposite direction rather than continuations. The resonating cavities are rapidly moved during song production, and this movement needs to be slowed before it can be stopped entirely, which could lead to the phenomenon of final lengthening. Finally, the lungs power the system, and the increase and decrease in subglottal pressure at the beginning and end of breath groups may cause arch contours to be more common than u-shaped contours.

If these patterns are in fact due to motor constraints on pitch production, then we would expect to find these three patterns in any domain produced by a system that shares these three constraints. Speech, for example, uses essentially the same production mechanism as human song. While syllable production is more rapid in speech, and pitch production needs to be more finely controlled during song, the two signals are nonetheless produced in the same basic way: air is forced out of the lungs and through the larynx, where vocal chords vibrate, producing a periodic sound, which is filtered through a series of rapidly changing resonating cavities. As a result, we would expect to find the three patterns in any language we choose to examine.

Final lengthening has been demonstrated in many languages, including English (Klatt 1975), Japanese (Venditti and van Santen 1998), Swedish (Fant et al. 1991), and French (Vaissiere 1991). Moreover, it has been shown that English speakers will increase this effect when attempting to make clear a syntactic ambiguity (Scott 1982, Schafer et al. 2000). Also, a brief lowering in pitch at the end of intonational phrases has been demonstrated in a wide variety of languages, including English (Arvaniti 2007), Spanish (Prieto and Shih 1996), Kipare (Herman 1996), and Yoruba (Connell and Ladd 1990). Boundary effects at the beginnings of intonational phrases have not yet been empirically investigated, though the existence of arch-shaped contours has been postulated despite an absence of real data (Vaissiere 1995).

In the current study, we investigated a corpus of spoken declarative sentences in English, French, Japanese, and Yoruba. As this corpus is small and contains a minute fraction of the world's spoken languages, the results reported below should not be taken as evidence for universal pitch patterns in spoken language. Nevertheless, if these patterns are in fact due to constraints arising from the structure of the human body, they should be found in any language investigators choose to examine in future studies.

Instrumental music, on the other hand, is produced by a system that differs significantly from the human song/speech-production system. The pitch-production component, in many cases, does consist of a vibrating system that can produce only a limited range of pitches, and as a result one would

expect to find the skip-reversal pattern. For most instruments rapid movement is necessary for producing each note or for changing the frequency at which the oscillator vibrates, and therefore one would expect final lengthening as well. However, one would not necessarily expect to find the arch contour pattern, because while some instruments are powered by air pressure, it is not generally possible for a performer to begin playing before air pressure has reached a sustained level. Nevertheless, composers and improvisers may imitate the arch contour found in speech and song in order to mark the beginnings and ends of musical themes, to aid in their memorization and detection by listeners.

Of the three patterns discussed here, only final lengthening has been investigated in instrumental music. Research on final lengthening, however, has been limited to music performance rather than musical structure, and has almost exclusively examined players of keyboard instruments. Repp (1990), for example, found that pianists increased the duration of notes just prior to major structural boundaries, and did so to a greater extent if asked to perform expressively, as opposed to mechanically. Similarly, Penel and Drake (2004) found that pianists increased the duration of notes at the end of rhythmic groups, and that this effect was larger when the musicians were asked to perform musically. Analysis of performance of the jazz ballad “My Funny Valentine” revealed similar phrase-final lengthening (Ashley 2002): performers anticipated the last note of a phrase and delayed the first note of

the following phrase, such that the time between note onsets was greatest between motifs. Finally, Sundberg (2000) found lengthening of notes just prior to phrase boundaries in a single classical baritone singer, regardless of whether or not the singer was asked to perform expressively or mechanically.

In the current study, we investigated musical themes written by English, French, German, Italian, and Russian composers who lived at the turn of the 19th and 20th centuries. This is clearly a narrow sample of the extreme musical variety available to the researcher who wishes to investigate the structure of instrumental music. Our results should, therefore, be taken as preliminary, and will need to be confirmed by future work on non-western music and western music from genres other than classical music. Studies of improvised music could be particularly enlightening.

**2.2 Methods**

The French, English, and Japanese sentences were taken from the database of Nazzi et al. (1998). Four female speakers per language read five unique sentences each, for a total of twenty sentences per language. The data set also included three speakers of Yoruba reading seven unique sentences each; these were taken from the database of Marina Nespor and Jacques Mehler. These languages were selected because they possess a wide variety of rhythmic and prosodic features—Yoruba is a tone language, Japanese is a

pitch-accent language and is mora-timed, English is a stress-timed intonation language, and French is a syllable-timed intonation language.

Musical themes were selected from Barlow and Morgenstern's *Dictionary of Musical Themes* (1983). Following Patel and Daniele (2003), themes were selected for all English, French, German, Italian, and Russian composers in the dictionary who were born in the 1800s and died in the 1900s. In order to be included, themes were required to contain at least twelve notes and no internal rests, fermatas, or grace notes. Moreover, themes from pieces with titles suggestive of a particular rhythm (e.g. marches, waltzes) or an attempt to produce an exotic style (children's music, music evocative of another composer or country) were also excluded. These criteria yielded 136 English themes from 6 composers, 180 French themes from 10 composers, 112 German themes from 5 composers, 53 Italian themes from 4 composers, and 238 Russian themes from 6 composers.

Duration in musical themes was encoded relative to the time signature, such that the basic beat for each theme was assigned a duration of one. Thus, in the time signature 4/4, a quarter note would be assigned a value of 1, an eighth note would be given a 0.5, etc. Durational data was collected from the speech samples by marking vowel boundaries in each sentence using speech spectrograms generated with Praat running on a personal computer. Both the waveform and the spectrogram were available during this analysis, plus interactive playback.

Pitch sequences in music were encoded as a series of distances from A440. Thus, a note two half-steps above A440 would be encoded as 2, and a note three half-steps below would be encoded as -3. Pitch sequences in speech were encoded using the prosogram version 1.3.6 as instantiated in Praat. The prosogram is a representation of F0 contour based on human pitch perception. Vowels with a pitch change exceeding the glide threshold $0.32/T^2$ are marked as glides (where T = vowel duration in s). This threshold is based on meta-analysis of a number of studies of the threshold for human perception of pitch change in speech (t'Hart et al. 1990). Vowels with rates of pitch change below this threshold are treated as perceptually equivalent to level tones. Using the threshold $0.32/T^2$ a large majority of tones are represented as level tones, allowing melodic patterns in speech to be directly compared to music.

**2.3 Results**

Skips were defined in both speech and instrumental music as pitch intervals greater than 50% of the tessitura. The tessitura was defined as the standard deviation of the distance in semitones of each pitch in a given sentence/theme from the mean pitch of that sentence/theme.

If the skip-reversal pattern is due to the fact that pitches are being drawn from a limited range in both speech and instrumental music, one would expect that only skips moving away from the median would be more likely to be followed by reversals, while skips moving towards the median should be

more likely to be followed by continuations. We divided skips up into four different categories based on their relation to the median: departing from, crossing, landing on, or approaching the median. In both domains, skips departing from or crossing the median were more likely to be followed by reversals, while skips approaching the median were on average followed by continuations. In instrumental music, as in vocal music (see Chapter 1), skips that landed on the median were more likely to be followed by continuations, suggesting that, all else being equal, there is a tendency for melodic lines to continue in the same direction (Figure 2.1). No consistent pattern was found in speech for median-landing skips. All of these differences were significant according to the binomial test.

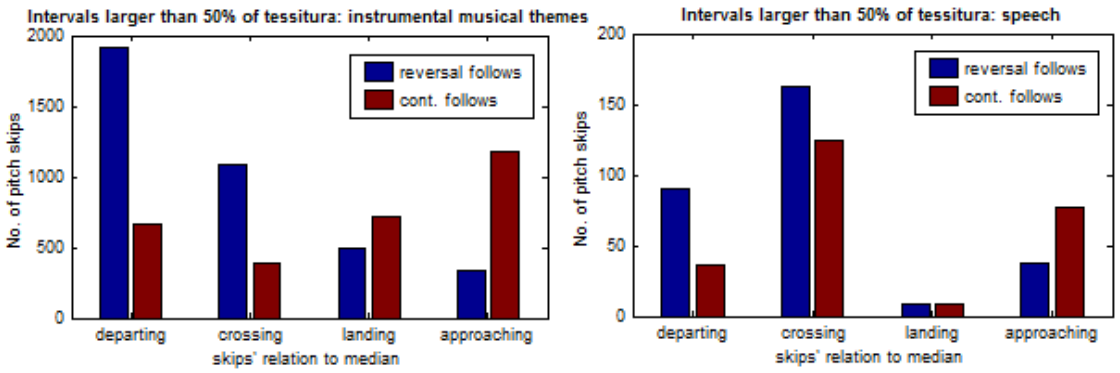


**Figure 2.1.** Skip-reversal patterns in instrumental music and speech. In both domains, skips departing from or crossing the median are more likely to be followed by reversals, while skips approaching the median are more likely to be followed by continuations.

An alternative way to examine the skip-reversal pattern is to divide skips into those followed by a continuation and those followed by a reversal, then examine where each skip fell in relation to the mean of the phrase in which it was contained. Figures 2.2 and 2.3 show that skips followed by reversals tend to fall further away from the mean. For instrumental music, skips followed by reversals had an average absolute value z-score of 0.5547, while skips followed by continuations had an average z-score of 0.9787. For speech, skips followed by reversals had an average z-score of 0.5418, while skips followed by continuations had an average z-score of 0.9904. Both of these differences were significant according to the Mann-Whitney U test.

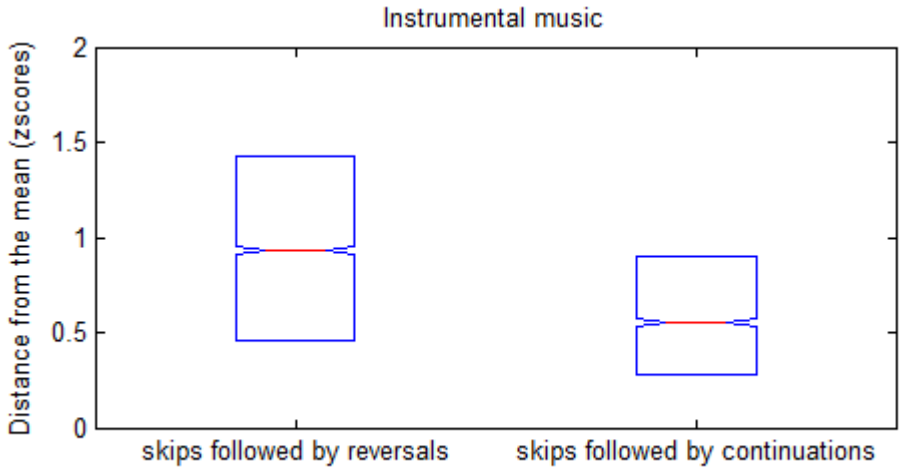


**Figure 2.2.** In instrumental music, skips that are followed by reversals fall further from the mean than skips that are followed by continuations.

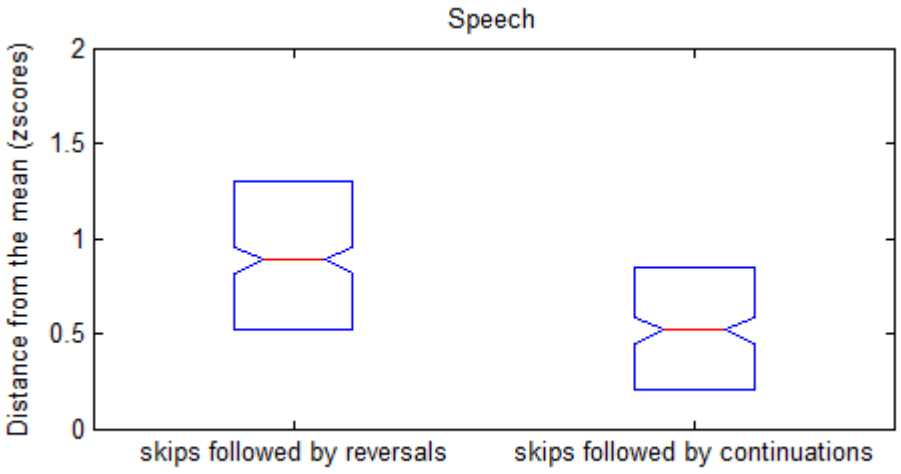


**Figure 2.3.** In speech, skips that are followed by reversals fall further from the mean than skips that are followed by continuations.

To test for the presence of final lengthening we compared, for each element (syllable or note), the ratio of the duration of that element to the average duration of all the elements in a given sentence or theme. For both speech and instrumental music, this ratio was significantly larger for the final element than for all other elements combined according to the Mann-Whitney U test (Figure 2.4).

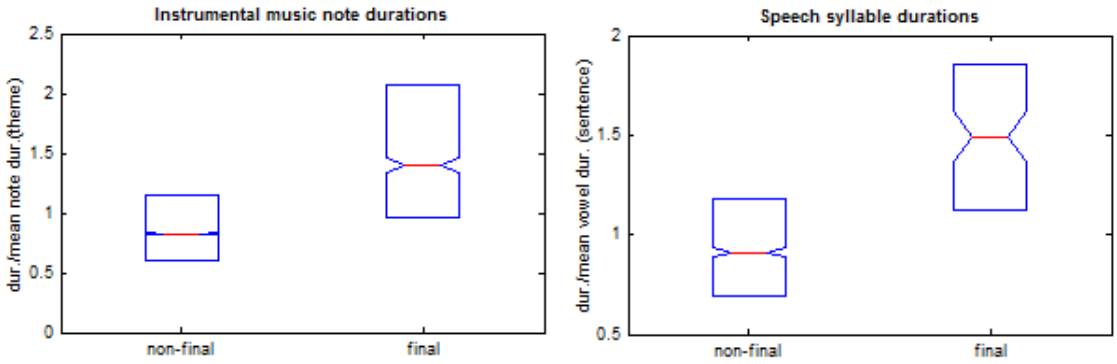


**Figure 2.4.** Final lengthening in instrumental music and speech. For each note, we computed the ratio of the duration of that note to the mean duration of all notes for the entire song or sentence. For both domains, this ratio was larger for final notes than for all other notes.

We also examined final lengthening within each language. Final lengthening was present in all languages (Figure 2.5), and the effect was significant according to the Mann-Whitney U test ($p < 0.005$). However, the ratio of the final syllable in each sentence to all other syllables was significantly smaller ($p < 0.05$ according to the Mann-Whitney U test) for Japanese (1.3023) than for Yoruba (1.61), English (1.8445), or French(1.5221).

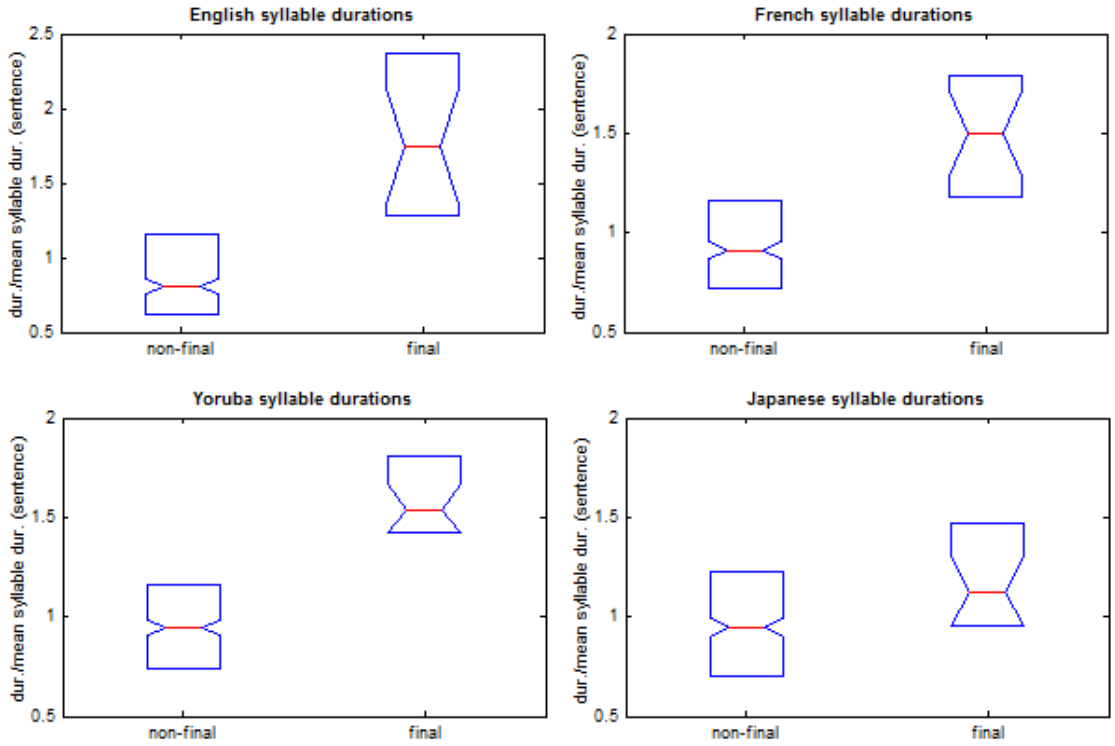


**Figure 2.5.** Final lengthening in English, French, Yoruba, and Japanese.

To test for the presence of an arch contour in speech and instrumental music, we generated an average pitch contour for each domain. First, we eliminated the pauses in the speech data and spliced the remaining segments together. We then normalized the time of each theme/sentence from 0 to 1 and divided each pitch contour into 50 time steps. Next, we normalized pitch by dividing the pitch of each time step by the mean pitch of each sentence or theme. Finally, we averaged across samples to generate an average pitch contour. The pitch contour for the instrumental music displays a clear arch shape, while the average pitch contour for the spoken sentences resembles an

arch superimposed onto a steady decrease in pitch. When this decrease is removed by detrending the contour of each speech sample before averaging them together, the result is a clear arch (Figure 2.6).

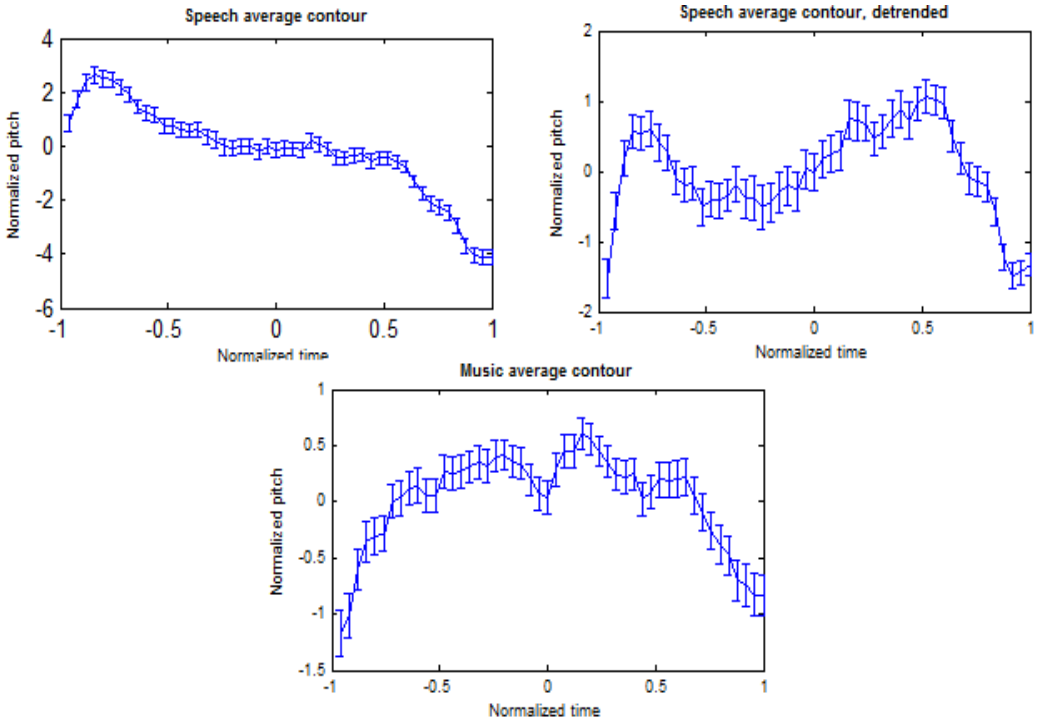


**Figure 2.6.** Average note contours for speech and instrumental music. The speech contour is shown both unaltered and detrended.

## 2.4 Discussion

The current study revealed three statistical patterns in corpuses of spoken declarative sentences in English, French, Japanese, and Yoruba and instrumental music written by composers born in Germany, England, France,

Russia, and Italy. In both domains, skips were more likely to be followed by reversals than by continuations, but only when the skips departed from or crossed the median; the last element of sentences and themes was longer on average than all other elements; and pitch contours tended to follow an arch shape, increasing in pitch at the beginning of a segment and decreasing at the end. These findings, combined with the results reported in Chapter 1, in which the same patterns were found in corpuses of human song and birdsong, strongly suggest that these patterns are due to motor constraints, rather than a cause unique to language processing or music cognition.

The arch contour found in instrumental music, however, is unlikely to be tied to the breath cycle, as I have claimed for speech, birdsong, and human song, as not all instruments use air pressure as their power source, and it is unlikely that performers would begin playing before air pressure had reached its sustained value. Moreover, while the average contours found in human song structure, birdsong notes, and speech all consisted of declination and arch shapes combined, only instrumental music followed on average an arch contour, but displayed no declination. It is possible that composers are imitating the rise in pitch at the beginning of a song phrase or utterance and drop at the end in order to emphasize the beginnings and ends of themes. Declination may, therefore, not have been imitated because it would not help convey this kind of structural information.

Future studies could investigate whether different instruments, which operate under slightly different motor constraints, produce somewhat different patterns. It is possible, for example, that instruments which require less movement when switching between notes, such as the flute, would exhibit less final lengthening than an instrument requiring greater movement, such as the piano. It also remains to be seen whether or not these patterns are present during improvisation, or, for that matter, in any form of instrumental music besides notated Western art music. Finally, although I have suggested that the arch contour is present because composers are using it to signal structural boundaries, its presence could instead somehow be tied to the breath cycle in players of wind instruments. Future work could distinguish between these two possibilities by comparing the prevalence of arch contours in the melodies written for wind and stringed instruments, which should be unaffected by the breath cycle.

**Acknowledgement**

Chapter 2, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Aniruddh Patel. Permission from all co-authors has been obtained.

## References

Arvaniti, A. (2007). On the presence of final lowering in British and American English. In Tones and Tunes, Vol. 2: Experimental Studies in Word and Sentence Prosody, C. Gussenhoven and T. Riad, eds. (Berlin and New York: Mouton de Gruyter), p. 317.

Ashley, R. (2002). Do [n't] change a hair for me: The art of jazz rubato. Music Perception *19*, 311–332.

Barlow, H., and Morgenstern, S. (1967). A dictionary of musical themes (Benn).

Connell, B., and Ladd, D. (1990). Aspects of pitch realization in Yoruba. Phonology *7*, 1–29.

Fant, G., Kruckenberg, A., and Nord, L. (1991). Prosodic and segmental speaker variations. Speech Communication *10*, 521–531.

Herman, R. (1996). Final lowering in Kipare. Phonology *13*, 171–196.

Klatt, D. (1975). Vowel Lengthening is Syntactically Determined in a Connected Discourse. Journal of Phonetics *3*, 129–140.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. Journal of Experimental Psychology: Human Perception and Performance *24*, 756–766.

Penel, A., and Drake, C. (2004). Timing variations in music performance: Musical communication, perceptual compensation, and/or motor control? Perception and Psychophysics *66*, 545-562.

Prieto, P., Shih, C., and Nibert, H. (1996). Pitch downtrend in Spanish. Journal of Phonetics *24*, 445–474.

Repp, B. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. The Journal of the Acoustical Society of America *88*, 622-641.

Schafer, A., Speer, S., Warren, P., and White, S. (2000). Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research *29*, 169–182.

Scott, D. (1982). Duration as a cue to the perception of a phrase boundary. The Journal of the Acoustical Society of America *71*, 996-1007.

Sundberg, J. (2000). Emotive transforms. Phonetica *57*, 95–112.

t'Hart, J., Collier, R., and Cohen, A. (1990). Perceptual Study of Intonation (Cambridge).

Vaissière, J. (1995). Phonetic explanations for cross-linguistic prosodic similarities. Phonetica *52*, 123–130.

Vaissière, J. (1991). Rhythm, accentuation, and final lengthening in French. In Music, Language, and Brain, J. Sundberg, L. Nord, and R. Carson, eds. (Houndsmills: Macmillan), pp. 108-120.

Venditti, J., and Santen, J. (1998). Modeling vowel duration for Japanese text-to-speech synthesis. In Fifth International Conference on Spoken Language Processing.

# Chapter 3.

## Tonotopic Mapping of Human Auditory Cortex Using Meaningful Stimuli.

## 3.1 Introduction

Chapter 1 demonstrated that the pitch and rhythm of birdsong and human song share a number of simple statistical patterns, while Chapter 2 showed that these same patterns exist in human speech. In fact, since these patterns seem to stem from motor constraints, one would expect to find them in any vocalization produced by a system similar to that of birds and humans. These patterns have likely been characteristic of vocalizations for hundreds of millions of years—ample time for the auditory system to adapt to their presence. As a result, probing the brain using simple stimuli following these patterns could be an ideal way to investigate the structure and function of the auditory system, and in particular, how the auditory system has adapted to process speech and song.

Auditory cortex in non-human primates consists of up to three core areas, surrounded by a belt of as many as seven secondary auditory areas. An indeterminate number of parabelt areas (at least two, possibly more) are located further lateral to the lateral belt. The cochlea is known to be represented topographically in the core and belt areas, but the core areas respond best to pure tones, while the belt areas respond better to band-passed noise and, in some cases, prefer certain monkey calls to others (Rauschecker et al., 1995; Kosaki et al. 1997; Petkov et al. 2006; Kaas and Hackett 2000). Core areas receive connections from the ventral portion of the medial geniculate

nucleus of the thalamus, while belt areas rely on the core areas for input (Rauschecker et al. 1997).

Early research on the organization of human auditory cortex was primarily dedicated to demonstrating that these basic characteristics of non-human primate primary auditory cortex applied to humans as well. Cytoarchitectonic mapping confirmed the existence of as many as three core areas centered on Heschl's gyrus, surrounded laterally and medially by several belt areas (Galaburda and Sanides 1980; Rivier and Clarke 1997; Morosan et al. 2001). Imaging studies using a variety of methodologies demonstrated the existence of tonotopy in human auditory cortex (Romani et al. 1982; Lauter et al. 1985; Verkindt et al. 1995; Howard et al. 1996; Wessinger et al. 1997), but due to the limited spatial resolution of electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET), and even functional magnetic resonance imaging (fMRI) using a standard 12-channel receive coil, the exact number of areas and the direction of the tonotopy gradients could not be determined.

Subsequent research, using fMRI with small surface coils and high spatial resolutions, revealed the existence of multiple topographic representations of frequency in human auditory cortex. Formisano and colleagues, using pure tones, found evidence that the human auditory core consists of three separate mirror-image tonotopic maps, while Talavage and colleagues, using noise bands with slowly changing frequencies, outlined six

potentially tonotopic regions including the core and part of the belt (Formisano et al. 2003; Talavage et al. 2000; Talavage et al. 2004).

Using surface coils for high-resolution fMRI is ideal for topographic mapping of the auditory system due to the relatively small size of the areas. Unfortunately, this method also restricts the area of cortex from which signal can be recovered. As a result, previous studies have been unable to simultaneously map those areas known to respond tonotopically in other primates and search for other areas—parabelt areas, for example—which may also contain tonotopic representations. We used a 32-channel receive coil to acquire high-resolution MR images from the whole of the left temporal lobe.

Several studies have now replicated the basic finding that the core areas in humans are tonotopically organized in a similar manner to the core areas of monkeys—that is, there are three core areas organized in mirror-image to one another. In addition, there is some limited evidence that human belt areas are tonotopically organized, as appears to be the case in other primates. However, no studies have previously reported successfully tonotopically mapping any areas outside of the core and belt areas, and no tonotopic maps yet reported extend below the superior temporal gyrus (STG) into the superior temporal sulcus (STS). It is possible, of course, that this failure is due to these areas not being tonotopically organized. However, up to this point all tonotopy studies have used ramped pure tones or noise bands as stimuli. Therefore, an alternative possibility is that auditory cortex inside the STS is not strongly

activated by these stimuli, and instead prefers stimuli of either greater complexity or greater ecological validity.

To explore this possibility, we used the Montreal Affective voices, a set of emotional non-verbal vocalizations, to tonotopically map human auditory cortex. These more ecologically valid stimuli enabled us to search for voice-selective tonotopically-organized areas in the STS. Moreover, the emotional nature of the vocalizations naturally attracts the attention of subjects, helping them maintain attention more consistently and for a longer time period, possibly leading to greater activation of belt areas, which are strongly modulated by attention (Woods et al. 2009).

Nearly all of these stimuli follow the arch contour that was revealed in Chapters 1 and 2 to be extremely prevalent in human speech and song. As suggested in Chapter 1, part of the auditory cortex may have been adapted to process signals with simple speech-like characteristics, which could include the arch contour and the existence of speech formants. Unfortunately, when doing tonotopic mapping one cannot rely on the presence of formants to activate voice-sensitive areas, as the center frequency of the pass band needs to be ramped from frequencies well below the levels of the formants to frequencies well above them. However, the arch-like pitch contour of the emotional vocalizations remains obvious, regardless of the filtering of the signal, enabling the stimuli to be clearly heard as vocalizations throughout the frequency ramp. This may enable us to topographically map areas which

previous studies, using noise bands or pure tones, did not find to be tonotopically organized.

### 3.2 Methods

Six subjects (one female) participated in the study. All subjects gave informed written consent. Each subject participated in at least two scanning sessions; two subjects participated in three scanning sessions. Each session consisted of at least four tonotopy scans, two using ascending frequency ramps and two using descending ramps.

Auditory stimuli were adapted from the Montreal Affective Voices (Belin et al. 2008), a series of recordings of emotional non-verbal vocalizations elicited by asking actors to produce sounds corresponding to a set of eight emotions (a ninth "neutral" emotion was not used in the current study, and only male voices were included). Tokens from this set were randomly selected such that no two recordings were repeated until the entire set had been presented. These tokens were then spliced together with no intervening pause to form 8-minute, 32-second passages. Next, these stimuli were amplitude-compressed in Adobe Audition 1.5 with additional manual editing; a bandpass filter was then cycled with a period of 64 seconds over the entire passage with center frequency logarithmically ascending from 150 Hz to 9600 Hz. Descending stimuli were created by reversing the entire waveform before filtering, then re-reversing at the final step. This initial filtering was cleaned

using similarly ascending low- and high-pass filters positioned an octave above and below the center frequency. Finally, a dynamic amplitude envelope was imposed over the frequency sweeps to equilibrate perceived loudness in the scanner environment. See Figure 3.1 for a spectrogram of a sample 8-minute, 32-second, downw-ramped run.

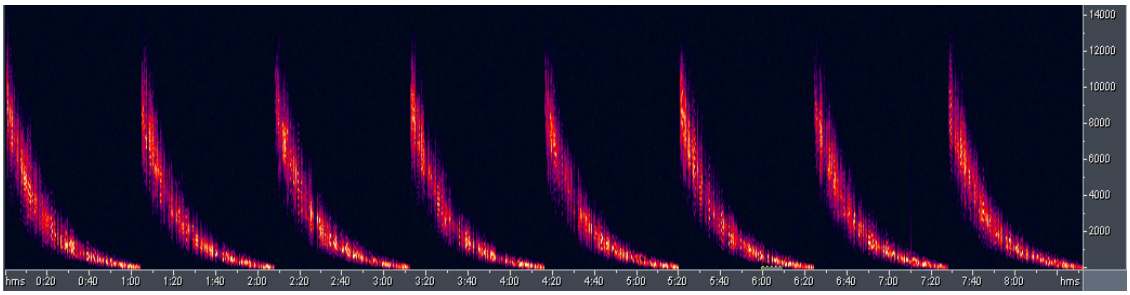


**Figure 3.1.** Spectrogram of the stimulus used for a single 8-minute, 32-second down-ramped tonotopy scan.

To investigate the pitch contour of the individual emotional vocalizations, I cut out any silences or unvoiced segments, spliced the remaining voiced vocalizations, and ran an analysis identical to the analysis that I ran on the individual birdsong notes in Chapter 1. Each sound was divided into fifty time steps, the pitch of each sound was computed as a distance in semitones from the average pitch of the sound, and an average contour was calculated across sounds. The resulting contour is strongly arch-shaped (Figure 3.2). Then, following the same procedure used in Chapters 1 and 2, I divided up each normalized pitch contour into thirds, and then classified each third into one of nine different pitch contour shapes based on

the relationship between the middle and the other two thirds. Arch-shaped contours were more common by far than u-shape contours, and declining contours were much more common than ascending contours (Figure 3.3). Figure 3.4 displays the contour of a single affective vocalization from the set used in this study, representative of how strongly arch-like the pitch contours of the majority of the stimuli are.

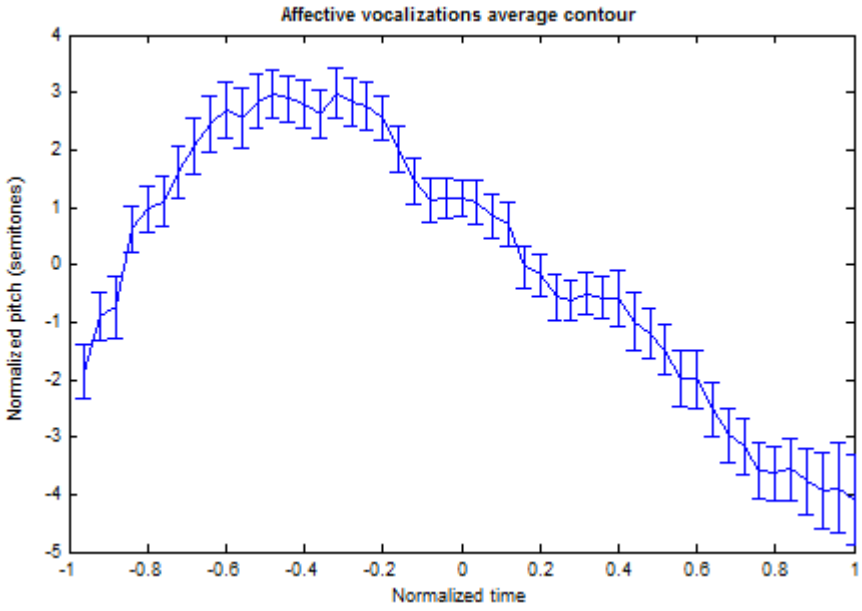


**Figure 3.2.** Average pitch contour of affective voice stimuli.

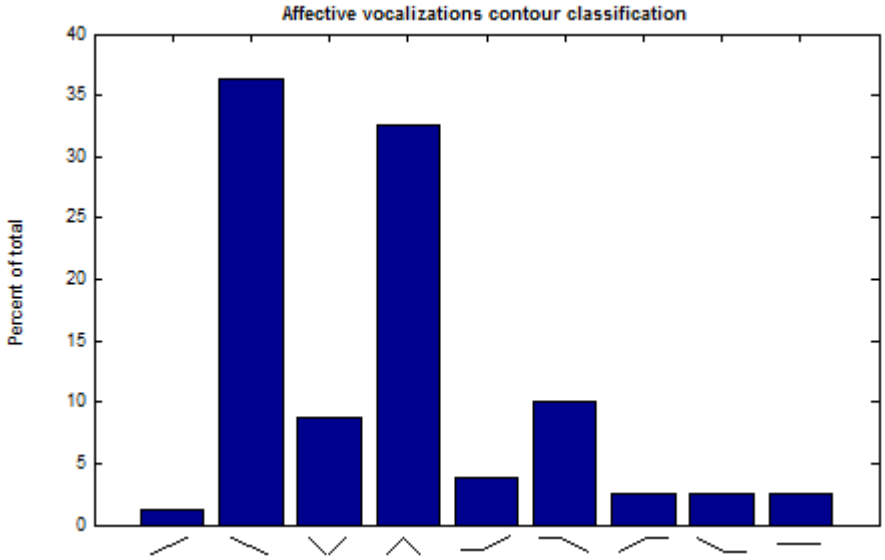


**Figure 3.3.** Pitch contour classification of affective voice stimuli.

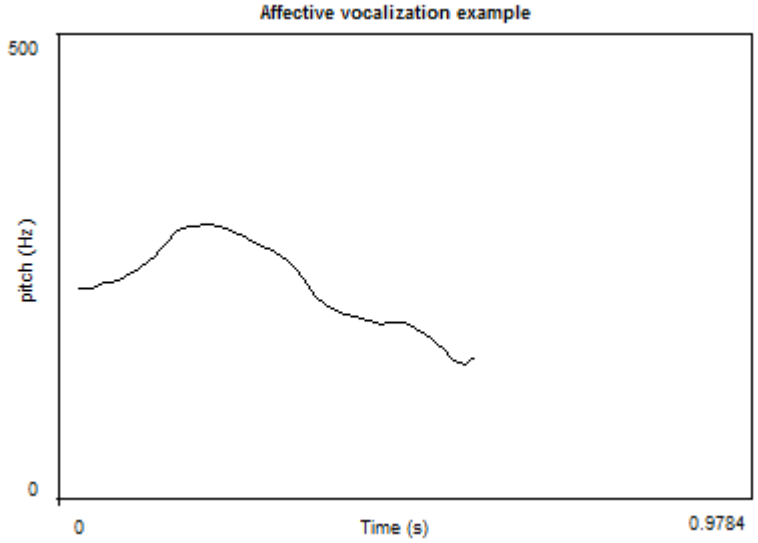


**Figure 3.4.** Representative example of the pitch contour of a single affective voice stimulus.

Stimuli were delivered by an electrodynamic, magnetless headphone using an isodynamically driven lightweight membrane coupled to an ear insert that also acted as a passive attenuator of acoustical scanner noise. The headphone gives a high amplitude, smooth, and quite undistorted response over a wide bandwidth (approximately 14 KHz). Stimuli were delivered to only the left ear. Subjects were asked to monitor the stimuli and press a button whenever the vocalization they heard resembled laughter.

Data were acquired with a 1.5 T Avanto magnetic resonance imaging scanner (Siemens) using a 32-channel head coil. MR slices were 3.2 mm thick, with an in-plane resolution of 3.2 x 3.2 mm. A single scan took 512 seconds, with 256 single-shot EPI images per slice (TR = 2 seconds, 22 slices, TE = 27.4 ms, flip angle = 90 degrees). A total of sixty-one functional scans were carried out on the 6 subjects: fifteen for subjects 1 and 2, eight for subjects 3, 4, and 5, and seven for subject 6. Each subject's cortical surface was reconstructed from at least one separate structural scan (T1-weighted MPRAGE, 1 x 1 x 1 mm, 224 x 256 x 160 voxels, TR = 16 ms, TE = 6.6 ms, flip angle = 18 degrees). Each subject also underwent at least one alignment scan in order to register the functional data to the reconstructed surface.

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 two images were manually registered and averaged prior to surface reconstruction.

Functional scans were motion-corrected using the volreg function included in Analysis of Functional Neuroimages (AFNI). Images were registered to the middle of the first functional run. Phase-encoded analysis of the functional images was performed using FreeSurfer. This method calculates the Fourier transform of the time course of activation at each voxel. Any voxel preferentially responding to a particular point in the stimulus cycle should show a higher amplitude at the frequency of stimulus cycling than at any other frequency. Therefore, the significance of a voxel's preference for particular frequencies can be estimated by dividing the squared amplitude at the stimulus "signal" frequency by the sum of the squared amplitudes of all other "noise" frequencies. This procedure results in a ratio that follows the F distribution, from which a P value can be calculated. The phase of the signal, which corresponds to a particular point of the stimulus ramp, can then be mapped to the color wheel, while the amplitude of the signal is mapped to the voxel's color saturation. Finally, the resulting maps are painted onto the reconstructed cortical surface. "Down" scans were reversed and averaged with "up" scans before analysis began.

Group averaging was performed by inflating each subject's reconstructed surface to a sphere, performing 10 steps of surface-based smoothing, and aligning it to the FreeSurfer spherical atlas. Results were then painted back onto a single subject's surface for viewing. Averaging of phase-encoded mapping data was performed using the methodology developed by

Hagler and Sereno (2006) in which the real and imaginary components of the signal are averaged across subjects, preserving any phase information consistent between subjects.

### 3.3 Results

Figures 3.5–3.16 show single-subject mapping data from both hemispheres of all six subjects, with green corresponding to low frequencies, blue to mid frequencies, and red to high. The upper-left panel shows higher-threshold data, zoomed out, and the upper-right panel displays the cortical surface, so that the relationship between the mapping data and the structure of the temporal lobe can be easily seen. In the lower panel the threshold has been lowered and the superior temporal lobe, including the STS, STG, and temporal plane, fills the entire panel so that the maps can be examined in greater detail.

Each voxel that is periodically activated at the frequency of stimulus presentation responds at a particular phase. This phase is mapped onto the color wheel and corresponds to the frequency of the signal to which that voxel responds. As a result, to investigate the arrangement of tonotopic maps within auditory cortex, we can gradually change the phase from that of the beginning of the 64-second ramp (low frequencies) to that of the end of the ramp (high frequencies). Any area containing a tonotopic map should show a gradual progression of activation across the cortical surface as the phase being investigated is changed. By mapping these spatial gradients we can distinguish

between areas in which there is a gradual change in the preferred phase from areas in which there is a sharp discontinuity. Moreover, as auditory cortical maps appear to be organized in mirror-image fashion (as are those of the visual system), mapping the spatial gradients can help delineate the boundaries between cortical areas, which should correspond to areas where the change in preferred phase reverses direction as one moves across the cortical surface. By examining movies of the progression of activation across the cortical surface as the phase changed, I manually drew spatial gradients for both hemispheres of all six subjects, progressing in the low frequency to high frequency direction. I also marked discontinuities, high-frequency reversals, low-frequency reversals, and reversals at frequencies other than the extreme end of the frequency range presented (which are to be expected if tonotopic maps representing a limited range of human-audible frequencies exist).

The most obvious feature shared by the maps of all six subjects is a frequency progression gradient not quite perpendicular to Heschl's gyrus that runs across a large section of the temporal plane. In all subjects this gradient is not exactly parallel to the STG, extending instead both anterior/posterior and medial/lateral (moving towards the STG as it moves more posterior, and away from the STG as it moves more anterior). The gradient is present and oriented in roughly the same position in both hemispheres.

In both hemispheres of every subject at least two reversals are present—corresponding to three different frequency progressions and at least three

different tonotopic maps. In each case, as the gradient extends from anterior to posterior, the frequencies to which cortex best responds change from high, to low, to high, and back to low. In some cases, there is an additional posterior high-frequency area, such that there are three mirror-image reversals; this is the case in the right hemispheres of subjects 2, 3, 4 and 6, and the left hemispheres of subjects 3 and 5. In each case, if this posterior high-frequency area exists, it lies on the STG, in roughly the place one would expect if the main frequency-progression gradient found in each subject were to be further extended back and towards the STG. In some cases there is also an additional anterior low-frequency area located on the STG, rather than on the medial temporal plane. This is clearly present in the left hemisphere of subject 1, the left hemisphere of subject 3, the left hemisphere of subject 4, the left and right hemispheres of subject 5, and the right hemisphere of subject 6.

In several cases there is an additional anterior high-frequency area inferior to the temporal plane, lying either on the lower lip of the STG or on the upper bank of the STS, interior to the high-frequency area on the temporal plane. This feature is present in the right hemisphere of subject 2, the right and left hemispheres of subject 3, the left hemisphere of subject 4, and the left hemisphere of subject 6. Even in subjects where this feature is not clearly present, there is often some suggestion of an additional superior-to-inferior frequency progression in the same direction. In subject one, for example, there is a mid-frequency area on the anterior STG. Finally, in a few cases there is a

second inferior high-frequency area on either the lower STG or upper bank of the STS. This area is evident in the right and left hemispheres of subjects 2, 3, and 4 and the left hemisphere of subject 5.

To summarize the single-subject data, in all subjects there is a frequency gradient not quite perpendicular to Heschl's gyrus which, as it runs from anterior to posterior, responds to high, then low, then high, then low frequencies. In some subjects there is an additional high-frequency area in the very posterior section of this gradient. Some subjects also have an additional low-frequency anterior area on the STG. Finally, some subjects have two additional inferior high-frequency areas, one anterior and one posterior, both more or less parallel to the high-frequency areas on the temporal plane.

Figures 3.17 and 3.18 show the total average for the scans of all six subjects, painted onto a single subject's cortical surface. In the left hemisphere (Figure 3.17) there is a frequency gradient perpendicular to Heschl's gyrus containing two reversals, with responses running from high to low to high to low frequency. In the right hemisphere (Figure 3.18) this same gradient exists, but with an extra low-frequency area on the anterior STG, such that the progression from anterior to posterior runs from low to high to low to high to low frequency, and the progression bends slightly. (Were the progression to run in a straight line, the anterior low-frequency area would be on the planum polare rather than on the STG.) In the right hemisphere only, there is an additional set of frequency progressions located along the STS, with areas of

frequency response roughly parallel to those of the gradient perpendicular to Heschl's gyrus. However, these areas respond only to phases corresponding to the lowest two-thirds of the frequencies presented. As a result, from anterior to posterior, this progression runs from low to medium to low to medium to low frequencies.

Overall, there is direct evidence for three frequency progressions in the left hemisphere on the temporal plane, oriented perpendicular to Heschl's gyrus. The right hemisphere data is more complex—there is clear evidence for four frequency progressions on the temporal plane, once again perpendicular to Heschl's gyrus, and at least three in the STS. Additionally, there are high-low-high progressions running superior to inferior between the posterior high-frequency areas.

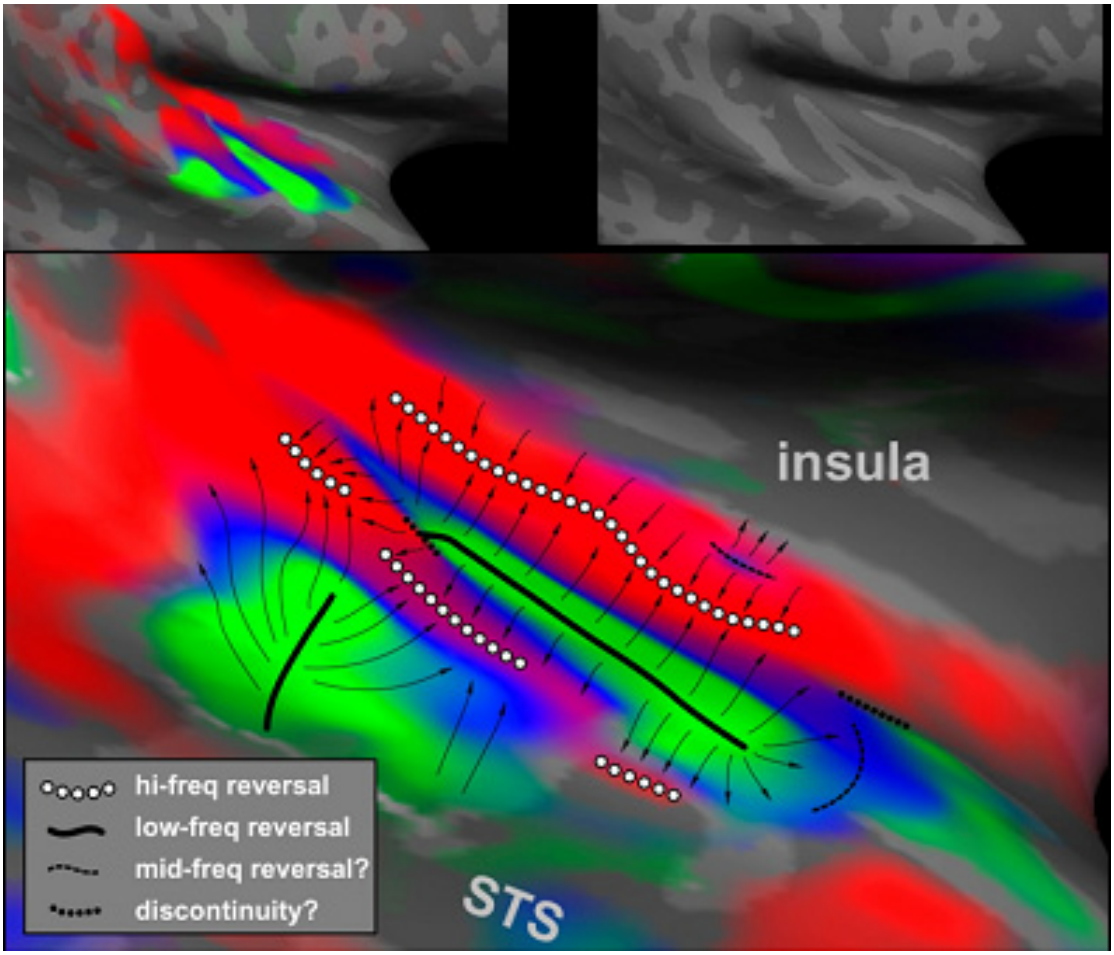


**Figure 3.5.** Tonotopic activation in the right hemisphere of subject 1. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

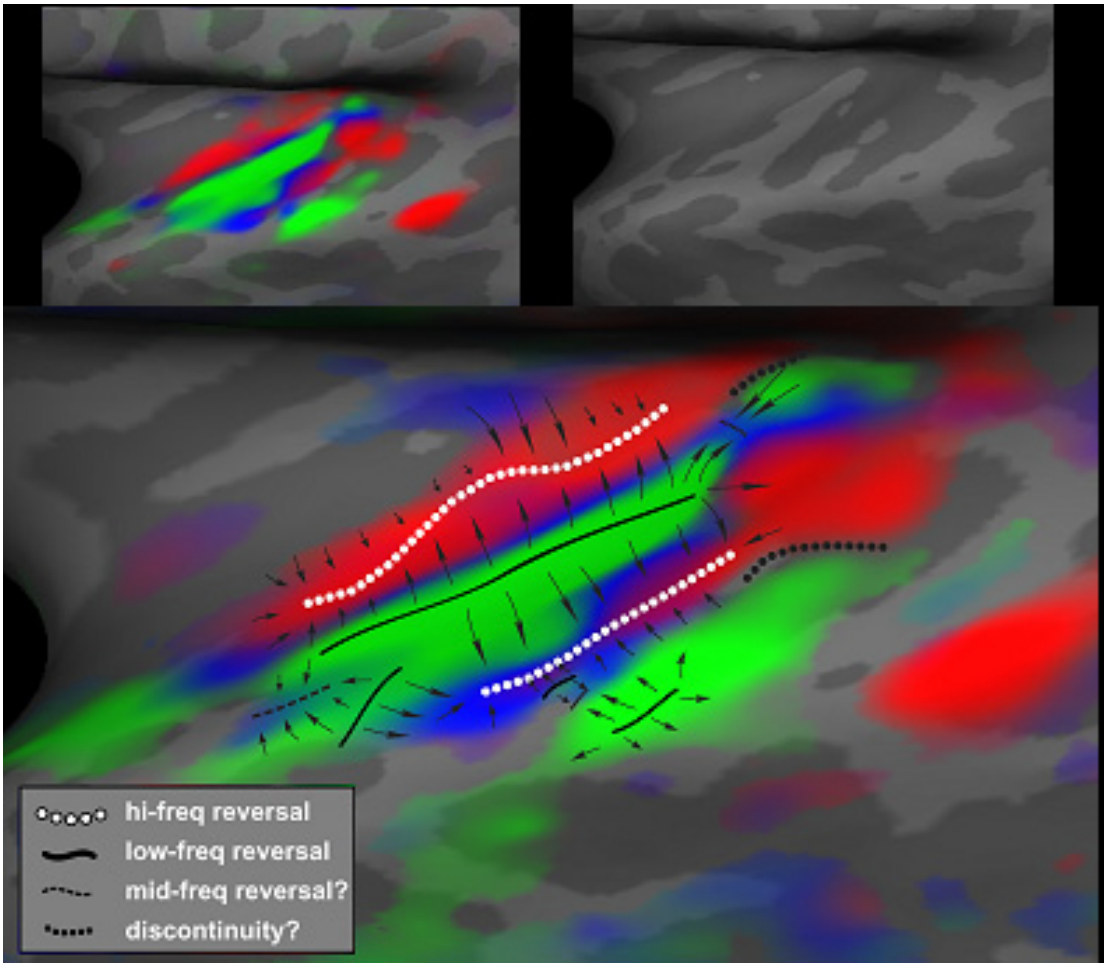


**Figure 3.6.** Tonotopic activation in the left hemisphere of subject 1. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

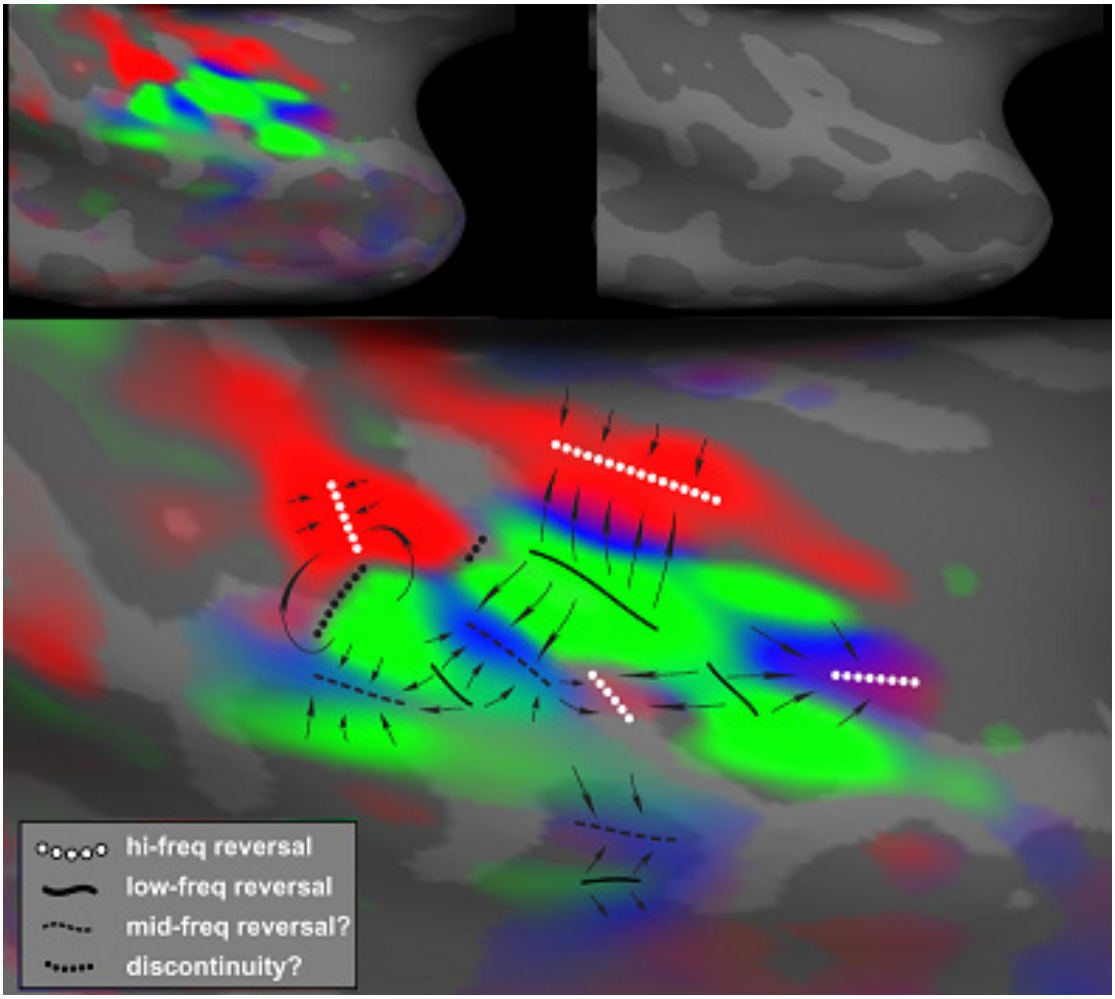


**Figure 3.7.** Tonotopic activation in the right hemisphere of subject 2. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

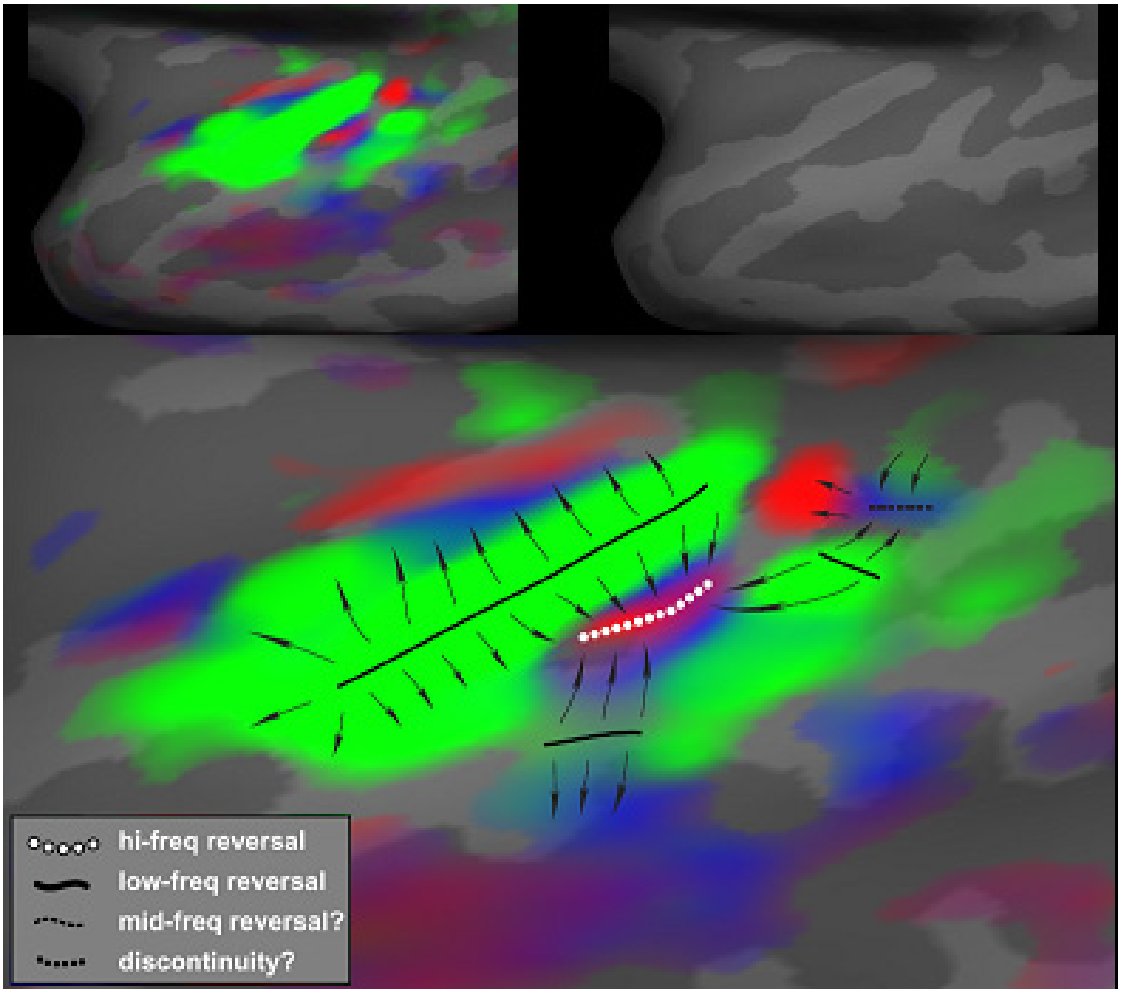


**Figure 3.8.** Tonotopic activation in the left hemisphere of subject 2. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

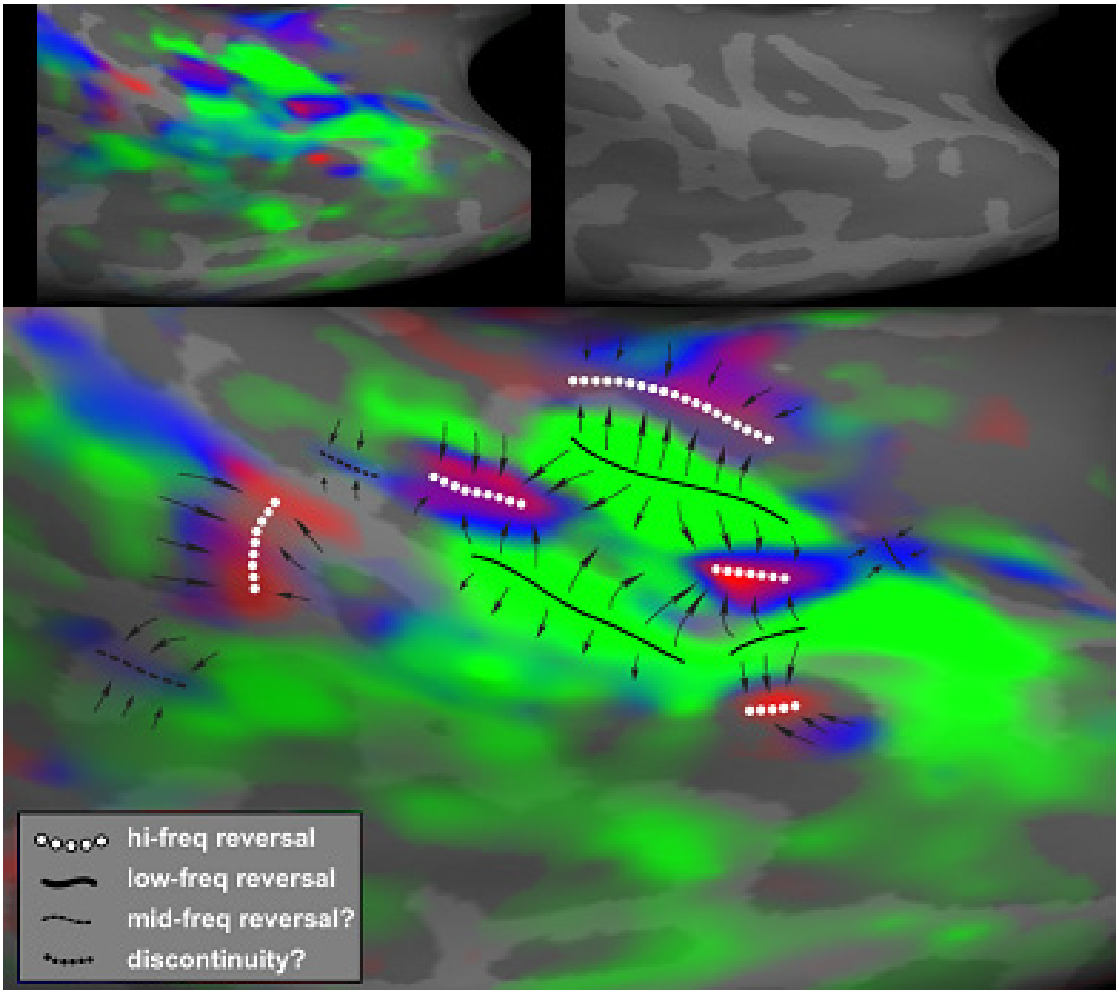


**Figure 3.9.** Tonotopic activation in the right hemisphere of subject 3. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

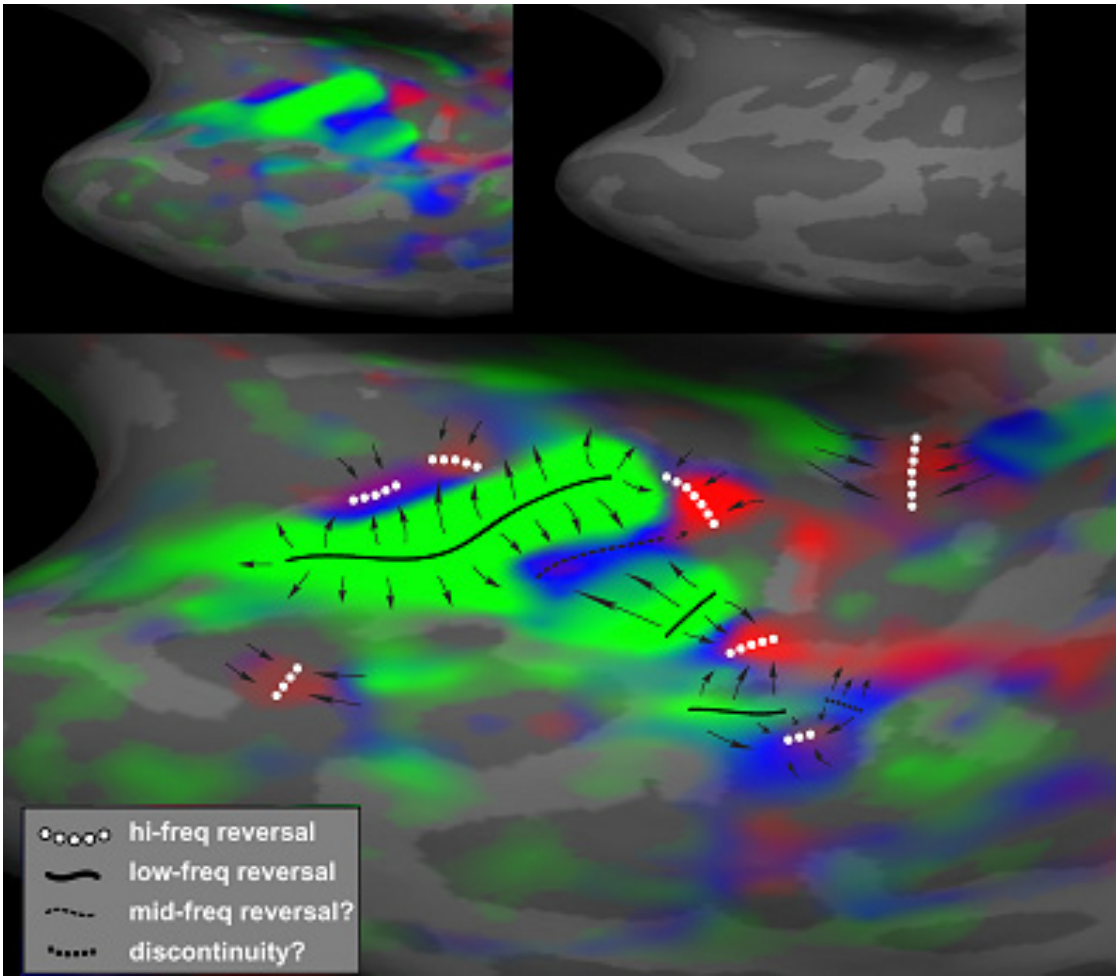


**Figure 3.10.** Tonotopic activation in the left hemisphere of subject 3. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

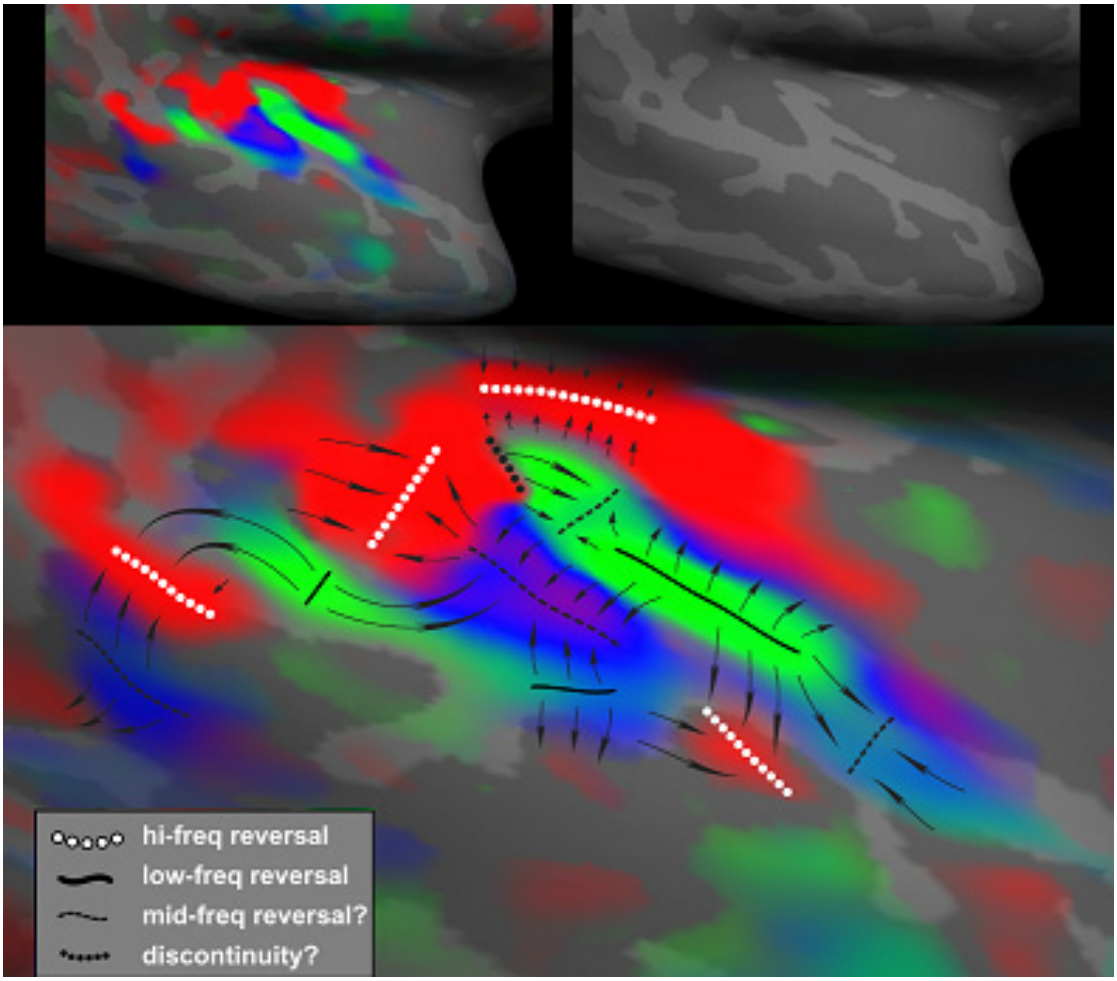


**Figure 3.11.** Tonotopic activation in the right hemisphere of subject 4. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

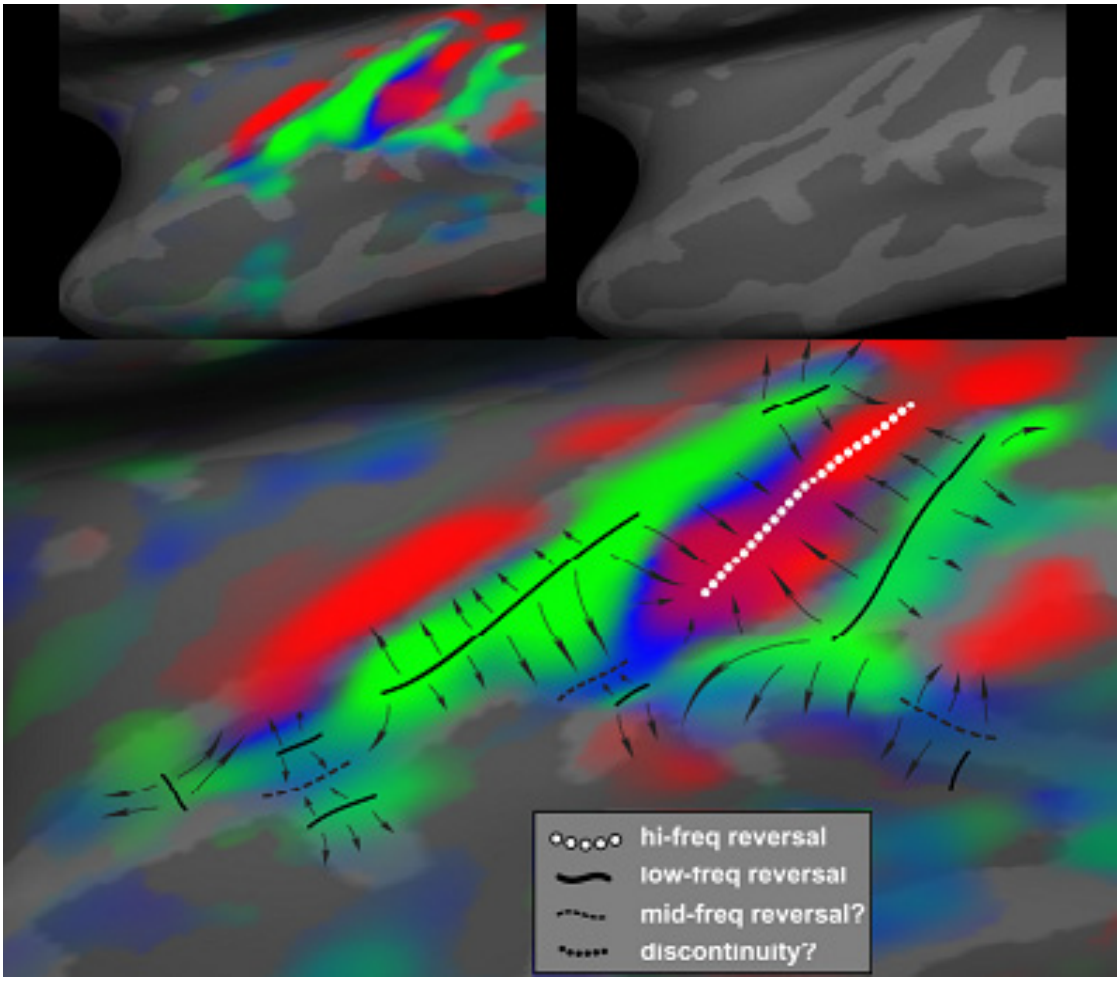


**Figure 3.12.** Tonotopic activation in the left hemisphere of subject 4. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

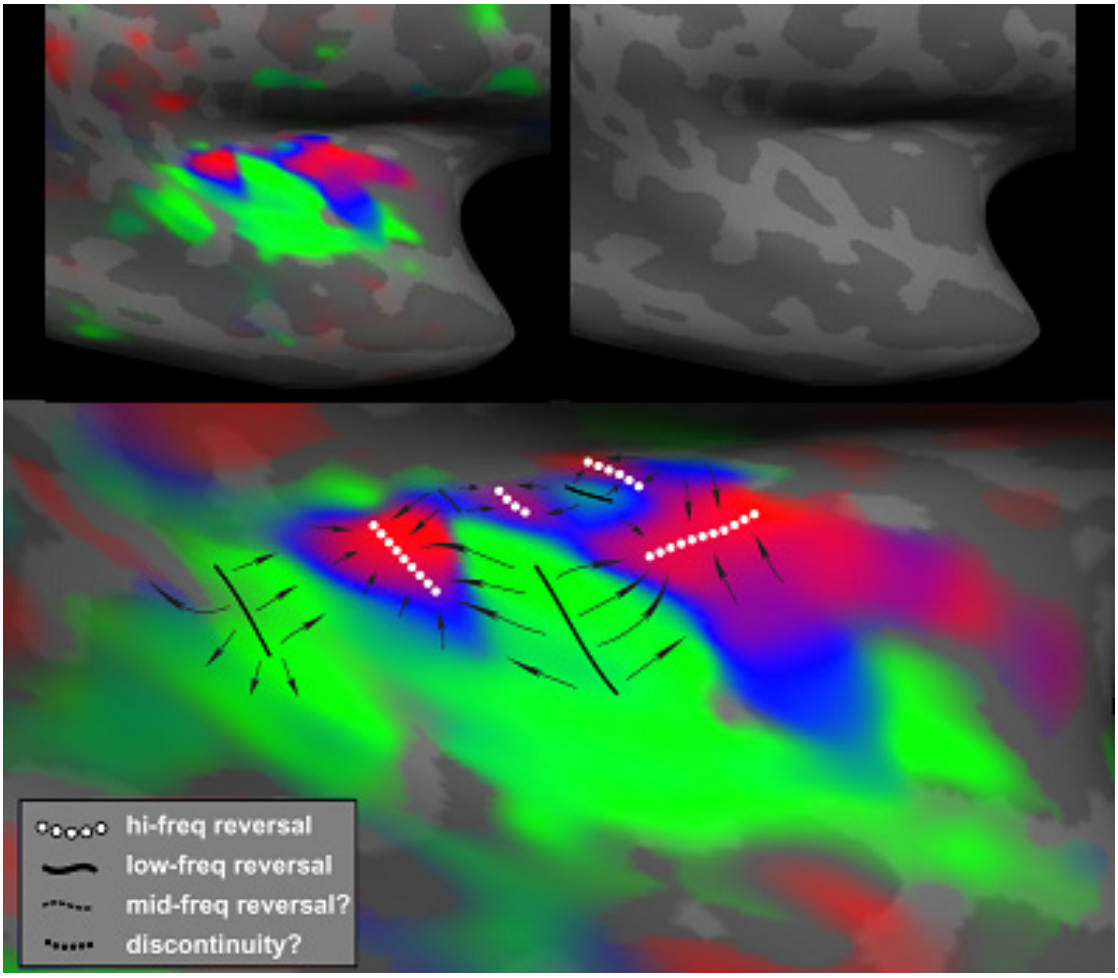


**Figure 3.13.** Tonotopic activation in the right hemisphere of subject 5. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

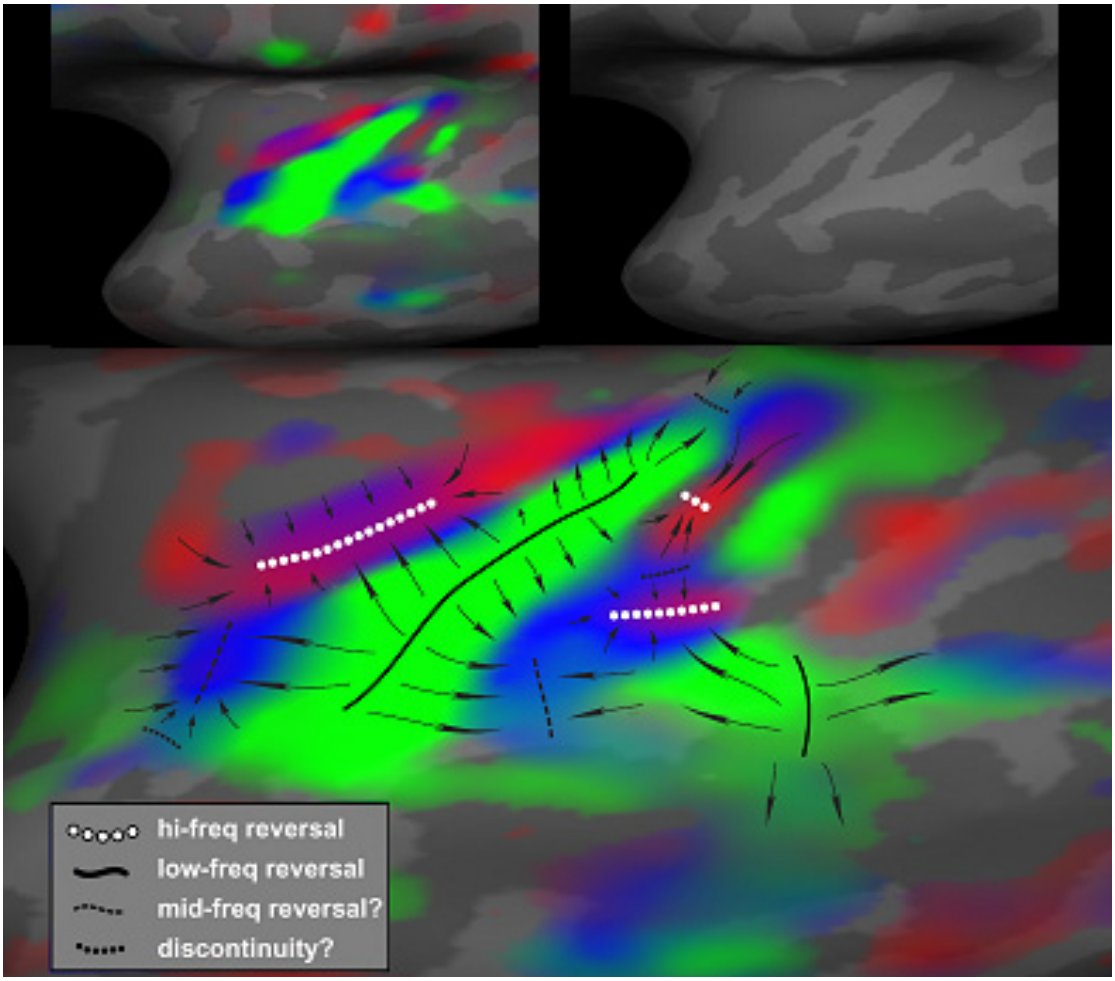


**Figure 3.14.** Tonotopic activation in the left hemisphere of subject 5. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

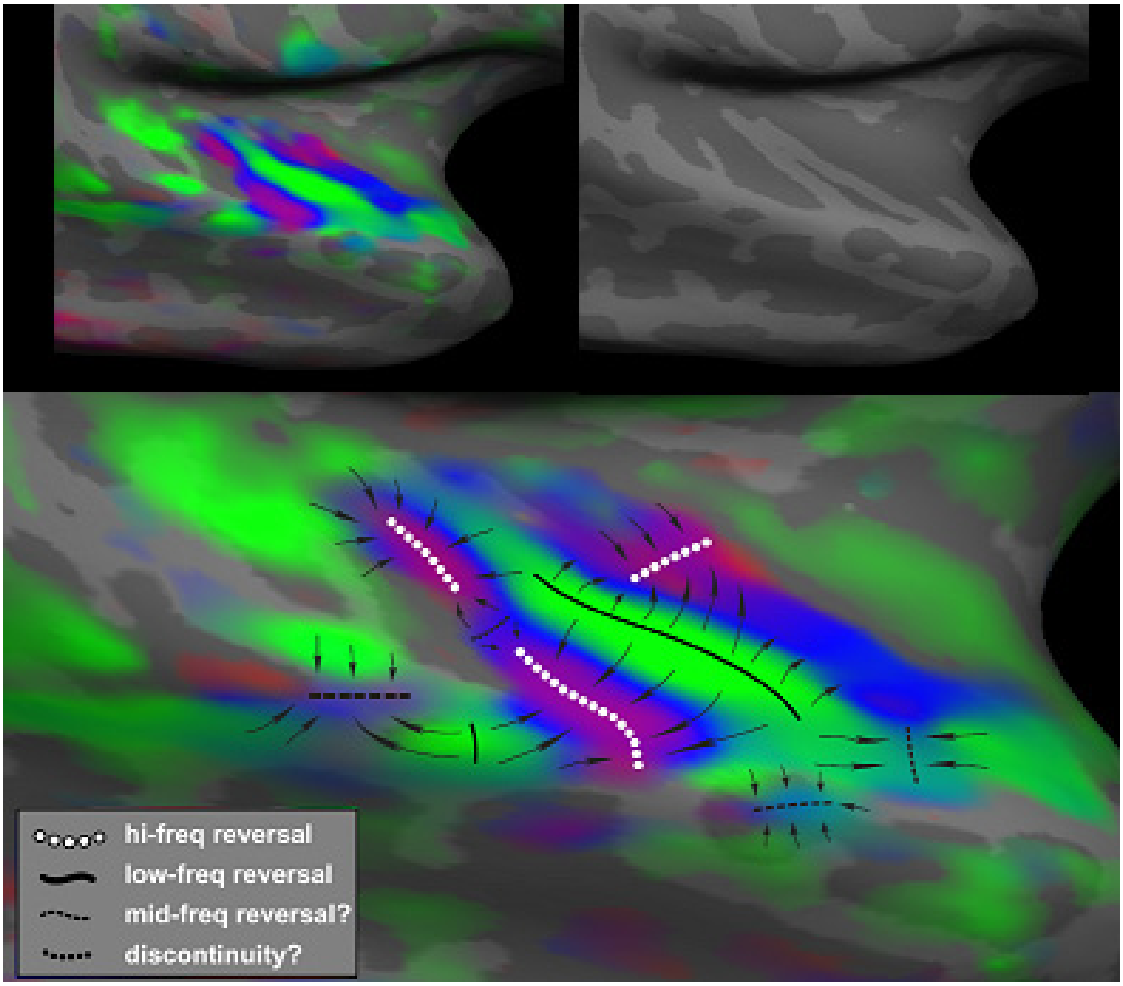


**Figure 3.15.** Tonotopic activation in the right hemisphere of subject 6. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

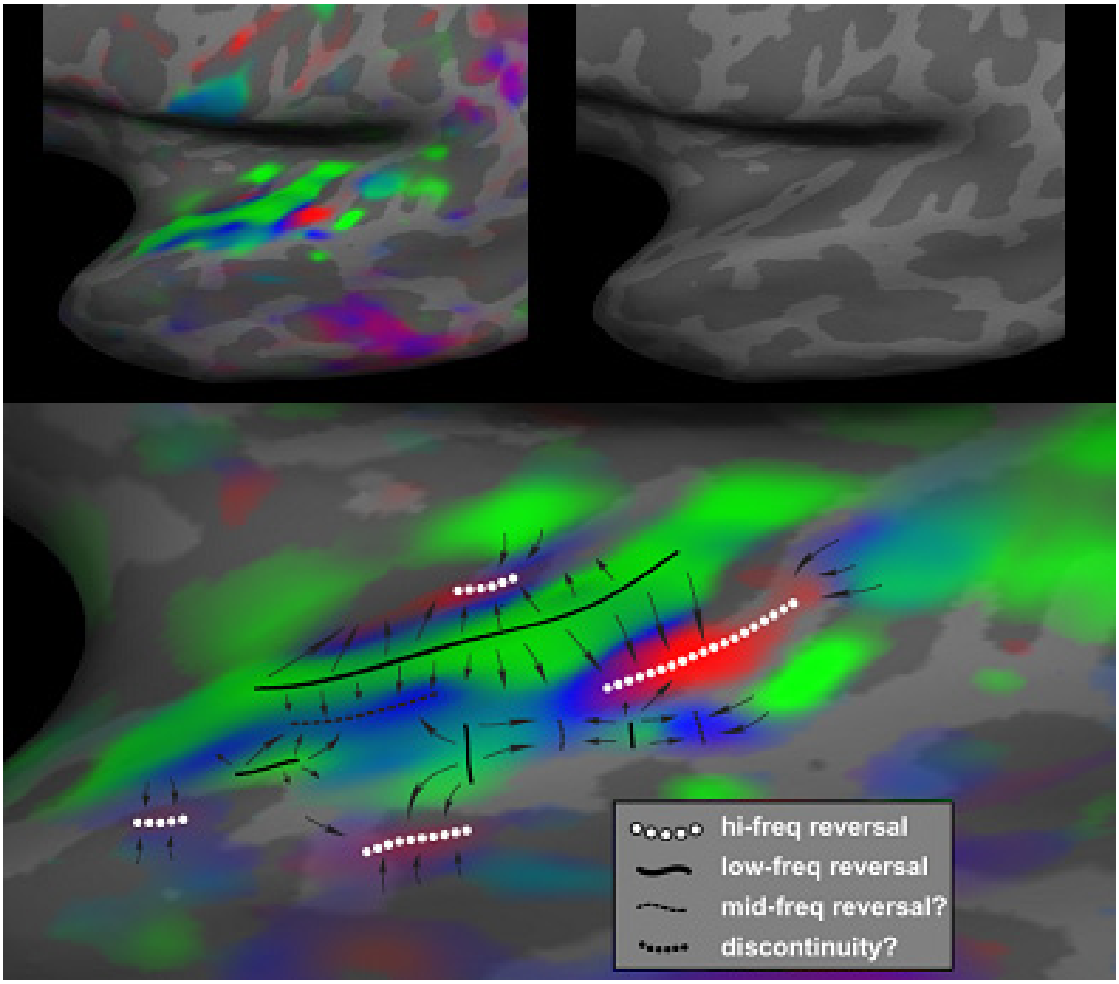


**Figure 3.16.** Tonotopic activation in the left hemisphere of subject 6. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

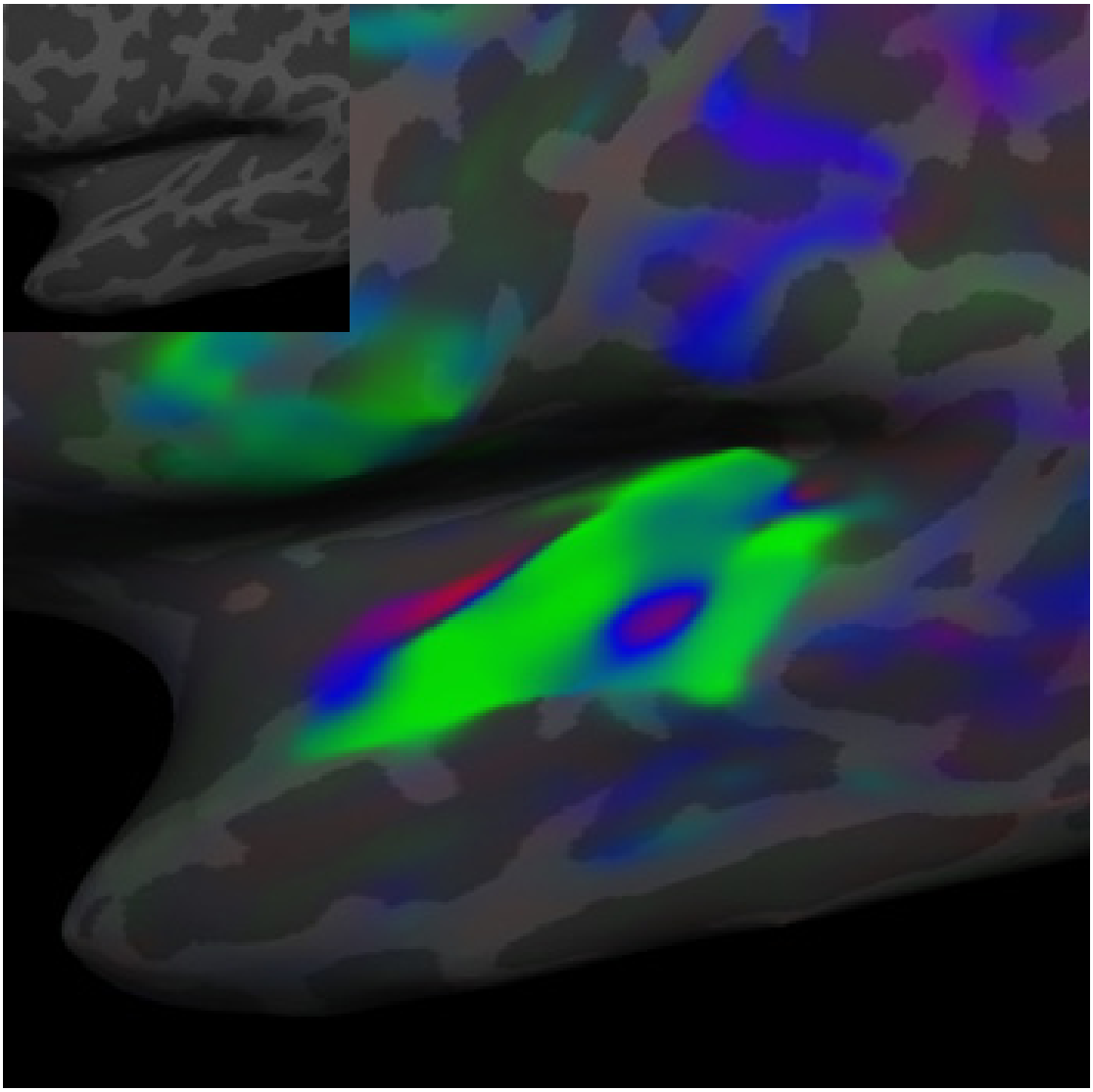

**Figure 3.17.** Tonotopy mapping results of left hemisphere between-subjects average. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

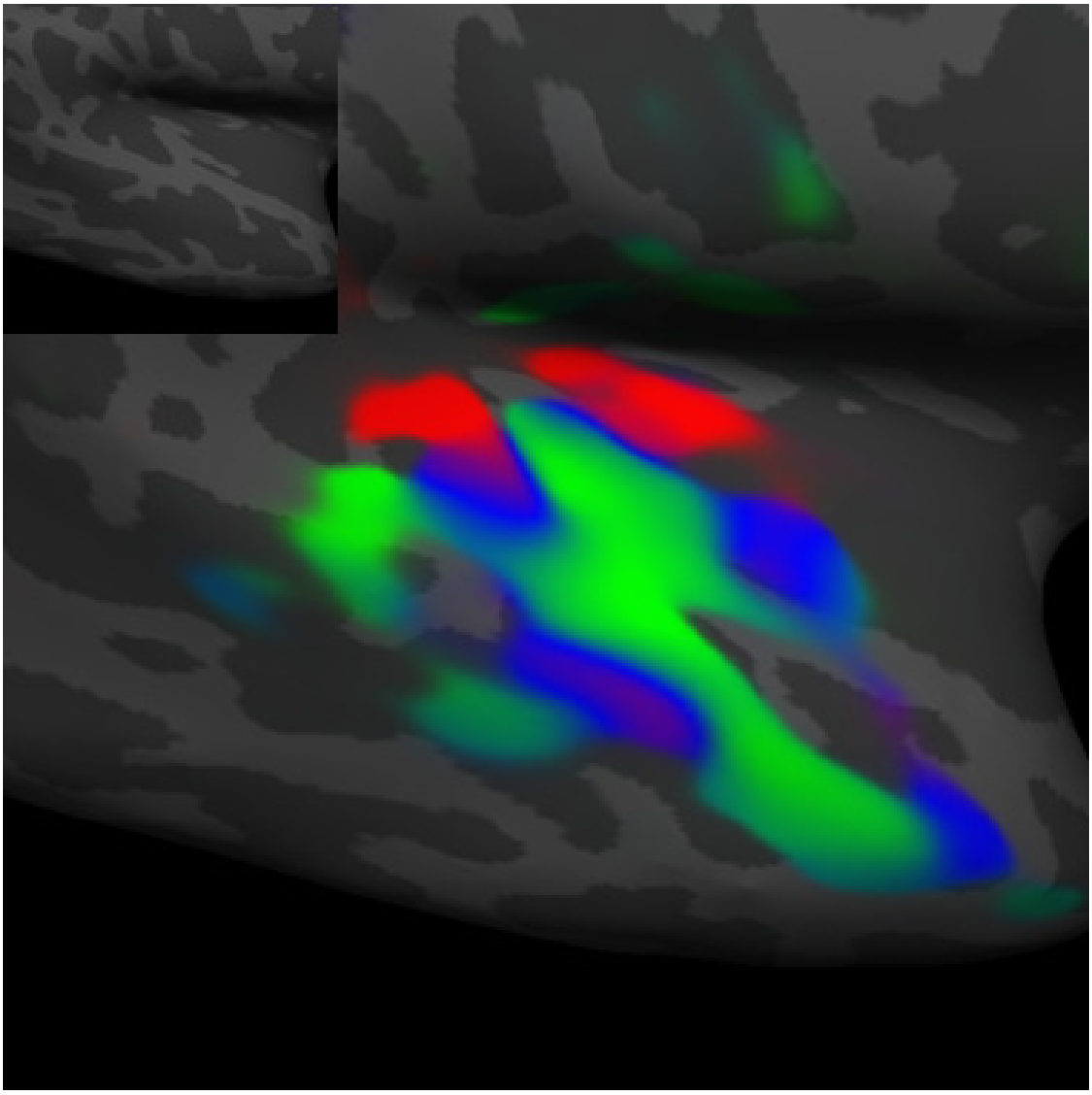

**Figure 3.18.** Tonotopy mapping results of right-hemisphere between-subjects average. Areas highlighted in red respond best to high frequencies, in blue respond best to mid frequencies, and in green respond best to low frequencies.

## 3.5 Discussion

We investigated the tonotopic organization of human auditory cortex using band-passed affective human vocalizations. These stimuli proved to be both attention-grabbing and meaningful even when the center frequency of the pass band was very high or very low, allowing us to search for tonotopically organized areas that are difficult to activate with less complex and less meaningful sounds.

Our results from the STG and in the superior temporal plane closely mirror the organization of auditory areas found in monkeys by Petkov et al. (2006) and in humans by Talavage et al. (2004), Woods et al. (2009), and Humphries et al. (2010): a series of mirror-symmetric tonotopic gradients running roughly parallel to the STG. In the right hemisphere, our results follow those of Petkov et al. (2006) more closely than any previous study; while Humphries et al. (2010) found only two gradients, Talavage et al. (2004) found three, and Woods et al. (2009) found three, we found four. (Tonotopic activation was more robust in the right hemisphere, which is likely due to the fact that auditory stimulation was delivered only to the left ear; as a result, the remainder of this discussion will focus on data from the right hemisphere.) Moving from front to back along the superior temporal plane, areas responded best to low, then to high, then to low, then to high, then to low frequencies—the exact same pattern found by Petkov et al. (2006).

The tonotopic gradients that have been mapped in the medial and lateral belt areas in monkeys (Rauschecker et al. 1995, Kosaki et al. 1997, Kusmierek and Rauschecker 2009) have been found to move parallel to the gradients in adjacent core areas. As a result, the boundaries between the core and the medial and lateral belt should not be accompanied by gradient reversals, and therefore cannot be detected from tonotopic activation alone. It is likely, therefore, that though our data from the superior temporal plane contains only four gradients and therefore is only conclusive evidence for at least four areas, this activation is produced by many more areas, including the core, lateral belt, and possibly some or all of the medial belt.

Several options exist for distinguishing between core and belt areas. Petkov et al. (2006) used the strength of the average tone response, which should be higher in primary auditory cortex, to outline the core by setting the activation threshold such that the outlined area roughly corresponded to the size of the core areas, as reported by structural studies. Another possibility would be to use very high-resolution structural scans to detect R1 (the reciprocal of T1), which is affected by tissue myelin content. The auditory core has been shown to be characterized by high myelin content (Hackett et al. 2001), which should lead to an increase in R1. This methodology has been shown to delineate regions that closely correspond to expected locations of koniocortex (Sigalovsky et al. 2006).

Following Fullerton and Pandya (2007), we suggest that the edge of the lateral belt roughly corresponds to the inferior lip of the STG. Here we differ from Woods et al. (2009), who place the parabelt areas along the STG, while we place the parabelt areas in the STS. We find three frequency gradients in the STS, the first published evidence for tonotopy in parabelt areas in either monkeys or humans. Belt areas in monkeys are topographically connected to adjacent parabelt areas (Hackett et al. 1998). As a result, one would expect any frequency gradients in the parabelt to be parallel to those in belt and core areas, and in fact, we found that the three gradients in the STS corresponded closely to those on the STG and in the superior temporal plane. The anterior high-frequency areas form a band reaching from the superior temporal plane into the STS, while the posterior high-frequency areas in the superior temporal plane and STS are separated by areas that respond best to low frequencies. As a result, there are, in total, three reversals and four frequency progressions in the STS, implying the possible existence of at least four different auditory areas in the STS.

Previous work in macaques has only tentatively parcellated parabelt cortex into an anterior and a posterior region based on connection patterns (Hackett et al. 1998). It cannot be determined at the present time whether human parabelt cortex has split into more areas since macaques and humans diverged from a common ancestor, or whether there exist finer subdivisions in macaque parabelt cortex that have not yet been discovered. It could be

enlightening to use fMRI to examine frequency representations in chimpanzees using either the same stimuli used in the present study or a set of filtered primate vocalizations to determine whether there are clear homologues for these STS divisions, or whether they are potentially unique to humans.

The orientation of the frequency gradients with respect to Heschl's gyrus has proved controversial. Several studies in humans, mostly using volume-based fMRI, have found tonotopy gradients running along Heschl's gyrus, finding either that higher frequencies are represented more medially and lower frequencies more laterally (Lauter et al. 1985, Howard et al. 1996, Wessinger et al. 1997, Bilecen et al. 1998, Lockwood et al. 1999, Engelien et al. 2002, Langers et al. 2007), or that there is a reversal and two tonotopy gradients (Formisano et al. 2003). Studies of the structure of human auditory cortex have also found several auditory areas located along Heschl's gyrus (Morosan et al. 2001, Hackett et al. 2001). Several other imaging studies, on the other hand, using mostly surface-based analyses, have found tonotopy gradients running in an anterior-to-posterior direction, perpendicular to Heschl's gyrus (Talavage et al. 2000, Schonwiesner et al. 2002, Talavage et al. 2004, Upadhyay et al. 2006, Woods et al. 2009, Humphries et al. 2010). This picture would appear to be more consistent with the organization of the auditory areas of the macaque, in which the frequency gradients largely run anterior-to-posterior across the STG and superior temporal plane

(Rauschecker et al. 1995, Rauschecker and Tian 2004, Petkov et al. 2006, Rauschecker et al. 2009).

Humphries et al. (2010) attempted to reconcile these contradictory findings by suggesting that using volume-based fMRI may cause activation from the high-frequency areas in the sulci posterior and anterior to HG—which are very close together in three-dimensional space—to blend together. This would explain why volume-based studies report fewer gradients, as well as why the gradients tend to be oriented parallel to HG, but leaves open the question of why the results of electrophysiological and structural studies of human auditory cortex differ so consistently from results collected via surface-based fMRI. Nonetheless, our data lend further support to the idea that the frequency gradients are oriented orthogonal to Heschl's gyrus; not only did we find this to be the case in both hemispheres in the between-subjects average, but the pattern holds true in both hemispheres for all six subjects.

We have found evidence of tonotopy that extends into the STS, while other studies have previously reported that tonotopically organized cortex did not extend past the STG. There are a number of factors that may have contributed to our ability to detect tonotopic activity in these areas. We used a 32-channel head coil, affording both an increased signal-to-noise ratio and coverage of the entire brain. All prior studies used either 12-channel or 8-channel send-receive coils, resulting in decreased signal-to-noise, or surface coils, limiting the area of cortex from which signal could be received. As

activity in response to sound tends to be strongest in core areas and drop off in the belt and parabelt areas, increased signal-to-noise could have helped us detect topographic activity in the STS. Increased attention may have also played a role; the affective voices are rather arresting stimuli, capturing and sustaining subjects' attention easily, and the lateral auditory areas are more strongly modulated by attention (Woods et al. 2009).

A third possibility, however, is that the STS responds particularly strongly to human vocalizations. Belin et al. (2002) found the right STS to be preferentially activated by non-speech vocalizations, as compared to a variety of other complex sounds. Moreover, Belin and Zatorre (2003) found that the right anterior STS showed adaptation to constant speaker identity, as opposed to constant syllable identity. The fact that the band-passed affective voices remained speech-like even when only very low or very high frequencies were let through the pass band, therefore, may have been crucial for mapping the tonotopic organization of the STS. The arch-like pitch contour of the vocalizations may be a particularly important factor leading to activation in some of these STS regions, as it is likely to be one of the only speech-like characteristics remaining once formant frequencies are filtered out.

This idea could be tested using fMRI by filtering the affective voices stimuli such that the formant frequencies are removed, then inverting the pitch contours such that the predominant arches become predominant u-shapes (and thus less ecologically valid). Activation resulting from the filtered

and inverted stimuli could be compared to activation resulting from the stimuli that were merely filtered. If the arch-like pitch contours are the main characteristic of the stimuli which causes them to sound like vocalizations, therefore activating the STS, then inverting the pitch contour should lead to a decrease in activation in the STS.

In summary, we mapped the tonotopic response of human auditory cortex using a 32-channel set of RF coils, a custom set of headphones with excellent frequency response over a large range, and a set of stimuli derived from recordings of human emotional vocalizations. Using this methodology, we replicated the basic tonotopic organization of the temporal plane described by previous fMRI studies on both chimpanzees and humans. In addition, however, we found evidence for at least three previously undiscovered tonotopic gradients within the STS, possibly corresponding to several areas within human parabelt cortex.

**Acknowledgement**

Chapter 3, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Fred Dick and Martin Sereno. Permission from all co-authors has been obtained.

## References

Belin, P., Fillion-Bilodeau, S., and Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods *40*, 531.

Bilecen, D., Scheffler, K., Schmid, N., Tschopp, K., and Seelig, J. (1998). Tonotopic organization of the human auditory cortex as detected by BOLD-FMRI. Hearing Research *126*, 19–27.

Dale, A., Fischl, B., and Sereno, M. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. Neuroimage *9*, 179-194.

Engelien, A., Yang, Y., Engelien, W., Zonana, J., Stern, E., and Silbersweig, D. (2002). Physiological mapping of human auditory cortices with a silent event-related fMRI technique. Neuroimage *16*, 944–953.

Fischl, B., Sereno, M., and Dale, A. (1999). Cortical surface-based analysis II: inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195–207.

Formisano, E., Kim, D., Di Salle, F., van de Moortele, P., Ugurbil, K., and Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron *40*, 859–869.

Fullerton, B., and Pandya, D. (2007). Architectonic analysis of the auditory-related areas of the superior temporal region in human brain. The Journal of Comparative Neurology *504*, 470–498.

Galaburda, A., and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. The Journal of Comparative Neurology *190*, 597–610.

Hackett, T., Preuss, T., and Kaas, J. (2001). Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. Evolution *197-222*, 222.

Hackett, T., Stepniewska, I., and Kaas, J. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. Journal of Comparative Neurology *394*, 475–495.

Hagler Jr, D., and Sereno, M. (2006). Spatial maps in frontal and prefrontal cortex. Neuroimage *29*, 567–577.

Howard III, M., Volkov, I., Abbas, P., Damasio, H., Ollendieck, M., and Granner, M. (1996). A chronic microelectrode investigation of the tonotopic organization of human auditory cortex. Brain Research *724*, 260–264.

Kaas, J., and Hackett, T. (2000). Subdivisions of auditory cortex and processing streams in primates. Proceedings of the National Academy of Sciences of the United States of America *97*, 11793-11799.

Kosaki, H., Hashikawa, T., He, J., and Jones, E. (1998). Tonotopic organization of auditory cortical fields delineated by parvalbumin immunoreactivity in macaque monkeys. The Journal of Comparative Neurology *386*, 304–316.

Kusmierek, P., and Rauschecker, J. (2009). Functional specialization of medial auditory belt cortex in the alert rhesus monkey. Journal of neurophysiology *102*, 1606-1622.

Langers, D., Backes, W., and van Dijk, P. (2007). Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. Neuroimage *34*, 264–273.

Lauter, J., Herscovitch, P., Formby, C., and Raichle, M. (1985). Tonotopic organization in human auditory cortex revealed by positron emission tomography. Hearing Research *20*, 199–205.

Lockwood, A., Salvi, R., Coad, M., Arnold, S., Wack, D., Murphy, B., and Burkard, R. (1999). The functional anatomy of the normal human auditory system: responses to 0.5 and 4.0 kHz tones at varied intensities. Cerebral Cortex *9*, 65-76.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., and Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage *13*, 684–701.

Petkov, C., Kayser, C., Augath, M., and Logothetis, N. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. PLoS Biology *4*, e215.

Rauschecker, J., and Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. Journal of Neurophysiology *91*, 2578-2589.

Rauschecker, J., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. Science, 111–114.

Rauschecker, J., Tian, B., Pons, T., and Mishkin, M. (1997). Serial and parallel processing in rhesus monkey auditory cortex. The Journal of Comparative Neurology *382*, 89–103.

Rivier, F., and Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. Neuroimage *6*, 288–304.

Romani, G., Williamson, S., and Kaufman, L. (1975). Tonotopic organization of the human auditory cortex. Science *216*, 1139-1140.

Sigalovsky, I., Fischl, B., and Melcher, J. (2006). Mapping an intrinsic MR property of gray matter in auditory cortex of living humans: a possible marker for primary cortex and hemispheric differences. Neuroimage *32*, 1524–1537.

Talavage, T., Ledden, P., Benson, R., Rosen, B., and Melcher, J. (2000). Frequency-dependent responses exhibited by multiple regions in human auditory cortex. Hearing Research *150*, 225–244.

Talavage, T., Sereno, M., Melcher, J., Ledden, P., Rosen, B., and Dale, A. (2004). Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. Journal of Neurophysiology *91*, 1282-1296.

Upadhyay, J., Ducros, M., Knaus, T., Lindgren, K., Silver, A., Tager-Flusberg, H., and Kim, D. (2006). Function and connectivity in human primary auditory cortex: a combined fMRI and DTI study at 3 Tesla. Cerebral Cortex, 2420-2432.

Verkindt, C., Bertrand, O., Perrin, F., Echallier, J., and Pernier, J. (1995). Tonotopic organization of the human auditory cortex: N100 topography and multiple dipole model analysis. Electroencephalography and Clinical Neurophysiology *96*, 143–156.

Wessinger, C., Buonocore, M., Kussmaul, C., and Mangun, G. (1997). Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. Human Brain Mapping *5*, 18–25.

Woods, D., Stecker, G., Rinne, T., Herron, T., Cate, A., Yund, E., Liao, I., and Kang, X. (2009). Functional maps of human auditory cortex: effects of acoustic features and attention. PLoS One *4*.

# Chapter 4.

# Neural Correlates of Perceiving Speech and Song.

## 4.1 Introduction

As demonstrated in Chapters 1 and 2, the pitch contours and rhythms of speech and song share several statistical patterns. Moreover, song contains essentially every acoustic element found in speech—the vowels and consonants produced in speech are produced in song as well. In addition, the music of a given country also appears to mirror a number of patterns specific to the language native to that country. For example, both English music and English speech display a high contrast between the duration of neighboring elements (notes and vowels, respectively), while French music and speech show a low contrast (Patel and Daniele 2003). Patel et al. (2006) showed that the degree of pitch interval variability of the two languages is also reflected in the music written by composers from each country: both English speech and English music tend to have higher interval variability than French speech and music.

These facts suggest a somewhat surprising question: why is speech not normally heard as song? There are, to be sure, several obvious differences between speech and song. For example, analysis of the instrumental music and speech corpuses used in Chapter 2 reveals that the distribution of music intervals contains a peak at two semitones not present in the distribution of speech intervals, while the distribution of intervals in speech is simply dominated by small intervals. In addition, music often uses isochronous rhythm, in which events occur at more or less strict intervals. This does not

occur in speech, despite many years of attempts to search for such a phenomenon in so-called stress-timed languages.

These differences do not, however, seem sufficient to account for the large, striking perceptual difference between speech and music. Isochrony is not a required characteristic of music, and indeed unaccompanied song is unlikely to be strongly isochronous. Speech may not contain intervals of two semitones more often than would be expected given a normal distribution, but the predomination of smaller intervals does ensure that two-semitone intervals will not be rare—and besides, such an effect takes place over the course of a large number of themes, while the difference between music and speech can be audible after three or four notes/syllables.

The most obvious remaining difference between speech and song, then, is that speech consists of variable pitch whereas song consists of level tones. Research on the perception of the pitch of speech and other sounds, however, have revealed that 1) when listening to sentences, subjects mainly perceive the pitch during vowels, and 2) when listening to short-duration glissandos, subjects can tolerate a surprisingly large amount of pitch movement before actually hearing the glissando, rather than a steady tone. These two facts combined suggest that what subjects are actually hearing when listening to speech is often closer to a rapid sequence of tones than the graded pitch variation displayed on a speech spectrogram.

House (1990) presented listeners with artificial stimuli consisting of a single type of vowel and a single type of consonant. The pitch of the stimuli varied, either during the vowels or during the consonants, and subjects were asked to categorize the stimuli (specifically, they were presented with three stimuli, and were asked if the third was more similar to the first or the second). The subjects' responses were largely influenced by only the portions of the pitch movements that occurred during vowels, and they ignored the pitch variation occurring during spectral transitions, i.e., consonants. In summary, when discussing listeners' perception of the pitch of spoken words, pitch variation during most consonants can probably be safely discounted.

In order to determine exactly what listeners are hearing when exposed to speech, then, we need to turn to literature on the perception of the pitch of short-duration tones and speech sounds. Schouten (1985), for example, asked subjects to classify pure tone glissandos, which he called "sweep tones", as going up or going down. A variety of sweep rates, durations, and center frequencies were used. Shorter durations and lower sweep rates caused the tones to be interpreted as steady-state stimuli; the shorter the duration, the greater the sweep rate needed to be in order for the pitch to be heard as changing. In a meta-analysis of this and similar studies, t'Hart (1990) determined that the threshold for pitch change, called the "glissando threshold", is consistently $g = 0.16T^2$, in semitones per second. Of the thirty-three data points that t'Hart plotted, twenty-five were within a factor of two

from the line formed by that equation plotted on double-logarithmic paper. t'Hart also calculated the "differential glissando threshold", that is, the threshold at which listeners are able to hear a change in the rate of pitch movement. Once again combining results from multiple studies, he reports a consistent threshold of around a factor of two.

These thresholds were used by d'Alessandro et al. (1995) to create a method of intonation stylization meant to transform an F0 curve for a given section of speech into a representation of the pitch information contained in the speech that is perceptually relevant. This was done through a recursive process, starting with the selection of a window of recorded speech. At the beginning of the procedure, this window is the entire section of speech. Then, a straight line is drawn from the pitch value at the start of the window to the pitch value at the end of the window. Next, the point with the largest distance from that line is selected as a "turning point", and the window is divided into two segments, separated by that point. Those segments are then subjected to that same analysis, and so on, until one of two boundary conditions is reached: the glissando rate is below threshold, or the pitch difference at the turning point is below a critical value (one semitone). After this segmentation is complete, each segment is marked as a glissando or a static tone based, once again, on the glissando threshold.

In order to test the validity of the model, subjects were presented with pairs of sentences from a set composed of an unaltered sentence and one of

three stylizations with differing glissando thresholds, and were asked to determine if the two stimuli were the same or different. (The first stimulus was always the unaltered sentence). Overall, the results showed that the higher the glissando threshold, the more likely it was that subjects would be able to distinguish between the stylized sentence and the unaltered sentence. Nevertheless, a threshold of 0.32 only led subjects to mark 35.09% as different, and a threshold of 0.64 led subjects to mark 69.84% as different, despite the fact that stylizations made with a threshold of 0.32 consisted of more static tones than glissandos, and stylizations made with a threshold of 0.64 were almost entirely composed of static tones.

This model was updated by Mertens (2004) to include the finding that the pitch underlying vowels is more important perceptually than the pitch underlying consonants. In this stylization, called the "prosogram", the analyst first marks the onset and offset of each vowel. The system then uses a differential glissando threshold of 20 ST/s to break each vowel up into a series of pitch slopes. If these slopes are below a glissando threshold of $0.32/T^2$, they are marked as static tones. The result is a transcription that represents speech, more often than not, as a series of static tones. Thus, the "obvious" difference between speech and music mentioned above—that sung words consist of static tones and spoken words consist of constant gradient pitch variation—breaks down.

In fact, even when a sound is perceived as a glissando, researchers have found that subjects are capable of assigning specific perceived tones to it. d'Alessandro et al. (1993), for example, presented subjects with pairs of tones, a "vibrato" tone undergoing sinusoidal variation, and a static tone, and asked subjects to adjust the static tone so that it matched the pitch of the vibrato tone. The researchers found that these pitch judgments were influenced by the position of the vibrato at the end of the tone—higher pitch judgments were associated with vibrato tones ending with phase π, while lower pitch judgments were associated with tones ending in phase 0. Rather than simply taking the mean between the higher and lower frequencies, therefore, subjects were treating the end of the tone as a sort of pitch target towards which the rest of the variation was heading.

d'Alessandro et al. (1998) later asked subjects to adjust a flat tone to either the end or the beginning of a glissando. The stimuli were pairs of synthetic vowels. One of the two vowels was a glissando, either a rise or a fall, while the other was a static tone. Overall, both falls and rises were more likely to be matched closer to the end.

To summarize the reviewed research on human perception of the pitch of speech, pitch plays a much more important perceptual role during vowels as compared to during consonants, and thus the pitch stream that people actually hear can be described as a series of short-duration tones during the vowels. If the change in pitch of these vowels is over a particular threshold, the listener

hears a glissando, but this threshold is high enough that most vowels are probably heard as having a constant pitch. Even if a vowel is heard as a glissando, subjects can still consistently categorize that vowel as having a certain pitch, and when they do so, they tend to overemphasize the end of the glissando.

So, to return to the question at hand, if the objects of speech perception really are short-duration static tones, why isn't speech heard as music? As it turns out, certain spoken phrases *can* be heard as music, if they're removed from context and repeated. This illusion was first reported by Deutsch (2003) and further investigated by Deutsch et al. (2008). A spoken phrase ("sometimes behave so strangely") from her CD "Musical Illusions and Paradoxes" (1995) sounds normal when heard in context, but when repeated by itself eventually begins to sound as if it were being sung. Once this transformation occurs, the phrase will sound like song in context, and is in fact difficult to not hear as song when encountered later on. This phenomenon suggests that different perceptual processes are being engaged when listeners hear the phrase as speech or as song. Therefore, a corpus of similar ambiguous stimuli could be used to investigate both the nature of the different perceptual processes underlying speech and song perception, and the minimal acoustic cues that must be present if a stimulus is to be perceived as music.

By closely examining audio book recordings, I have found twenty-four spoken phrases that, when taken out of context and repeated, on average

sound more like song than like speech to subjects, and twenty-four phrases spoken by the same readers that sound more like speech than like song when repeated. The song and speech phrases are matched on a wide variety of basic acoustical measures; therefore, something in the pitch patterns underlying the phrases must be driving the illusion. In this chapter I suggest a few potential characteristics of the stimuli that may be driving the effect, and report the results of an fMRI investigation of the neural correlates of hearing the "song" versus hearing the "speech" stimuli.

In 1965 Alexander Luria published an account of his patient V.G. Shebalin, a composer who suffered three strokes that decimated his language abilities but left his ability to compose intact. Ever since, the extent to which music and language draw upon distinct or overlapping neural resources has remained a topic of interest for neuroscientists. Over the next thirty years, a consensus emerged that damage to the right temporal lobe is associated with a wide range of music-related deficits, including tonal memory (Zatorre 1991), melodic contour processing (Peretz 1990), melody recognition and discrimination of familiar melodies (Ayotte et al. 2000), and perception of missing fundamental tones (Zatorre 1988). Moreover, it has been claimed that these music deficits can occur in the absence of any obvious language deficits (Marin and Perry 1999).

More recent work, however, has revealed this association of music processing with the right hemisphere and speech processing with the left to be

overly simplistic. For example, unlike non-musicians, musicians show a left-hemisphere bias when processing music (Messerli 1995, Harris 1999). Moreover, recent brain imaging work has suggested a far greater degree of overlap between the two systems than was previously assumed. The P600, an ERP component correlated with syntactic processing, has been shown to be generated by violation of harmonic expectations in music as well (Patel et al. 1998). Moreover, processing of these harmonic violations is linked to an increase of activation in Broca's area (Koelsche et al. 2002, Maess et al. 2001).

Understanding the neural relationship between language and music could illuminate the effects of brain damage on these two abilities, and could help shed light on the evolution of one or both of these capabilities. Nevertheless, there have been very few imaging studies directly comparing the neural correlates of listening to speech and listening to music. Almost all of the few studies that have been conducted have compared spoken words, syllables, or sentences to tones or sequences of tones played on musical instruments (Hugdahl et al. 1999, Tervaniemi et al. 2000, Hickok et al. 2003, Tervaniemi et al. 2006). The majority of these studies found that the STG/STS of the right hemisphere was more activated by the musical stimuli than the speech stimuli, and vice versa for the left hemisphere. Unfortunately, because none of these studies controlled for the basic acoustical properties of the stimuli, the comparisons made do not necessarily capture the characteristic processing involved in the hearing of speech and music.

The only fMRI study to date to attempt to control for many of the basic acoustical differences in speech and music was Callan et al. (2006), who used as stimuli well-known Japanese songs that were both sung and spoken by trained singers. The speakers were asked to recite the lyrics at the same tempo at which they were sung. Unfortunately, this process was unlikely to generate natural speech, given that the speakers would have strong associations between the lyrics and the original song, and given that the lyrics were spoken at an unnaturally slow tempo, both factors that may have led the speech stimuli to be more rhythmic and musical than natural speech. Activation linked to hearing song as opposed to speech was found in the right anterior STG.

We compared speech and music processing using an fMRI block design while controlling as much as possible for any basic acoustical influences by choosing as stimuli the recorded spoken phrases mentioned above that, when taken out of context and repeated, sound as if they are being sung. These phrases were matched with an equal number of phrases spoken by the same speakers that do not undergo this transformation when repeated.

One possible finding is that hearing the illusion stimuli will be linked to an increase in activation within pitch-sensitive areas in auditory cortex, due to an increase in attention to this dimension of the stimuli. Activation in lateral Heschl's gyrus bilaterally has been demonstrated to correlate with the pitch strength of auditory stimuli (Griffiths et al. 2001, Patterson et al. 2002,

Penagos et al. 2004), and this same area has been shown to contain a pitch map in marmoset monkeys (Bendor and Wang 2005). Pitch variation/melody processing has been linked to activation just anterior to HG (Patterson et al. 2002, Warren 2003, Schmithorst 2003). This area was also found to be more strongly activated in the right hemisphere when subjects were asked to sing the words to a favorite song, compared to a control condition in which they were asked to recite the same lyrics (Jeffries et al. 2003).

### 4.2 Methods

Fourteen subjects (mean age 32.57 years, nine female) participated in the study. All subjects gave informed written consent. Each subject underwent four 512-second 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 discarded due to head motion.

All stimuli were collected with permission from audio books (via audiobooksforfree.com and librivox.org). 24 spoken phrases that do, on average, sound like song when removed from context and repeated and 24 that do not were found. Both illusion and non-illusion stimuli were taken from the same three talkers in the same proportions. All phrases were taken from passages that were intended to be heard as speech by the readers, rather than song (a fact which could be inferred from the context in which each sample was embedded). Figure 4.1 shows examples of the pitch traces (made using

Praat) of one “speech” and one “song” stimulus, illustrating that there is no obvious difference distinguishing the two.

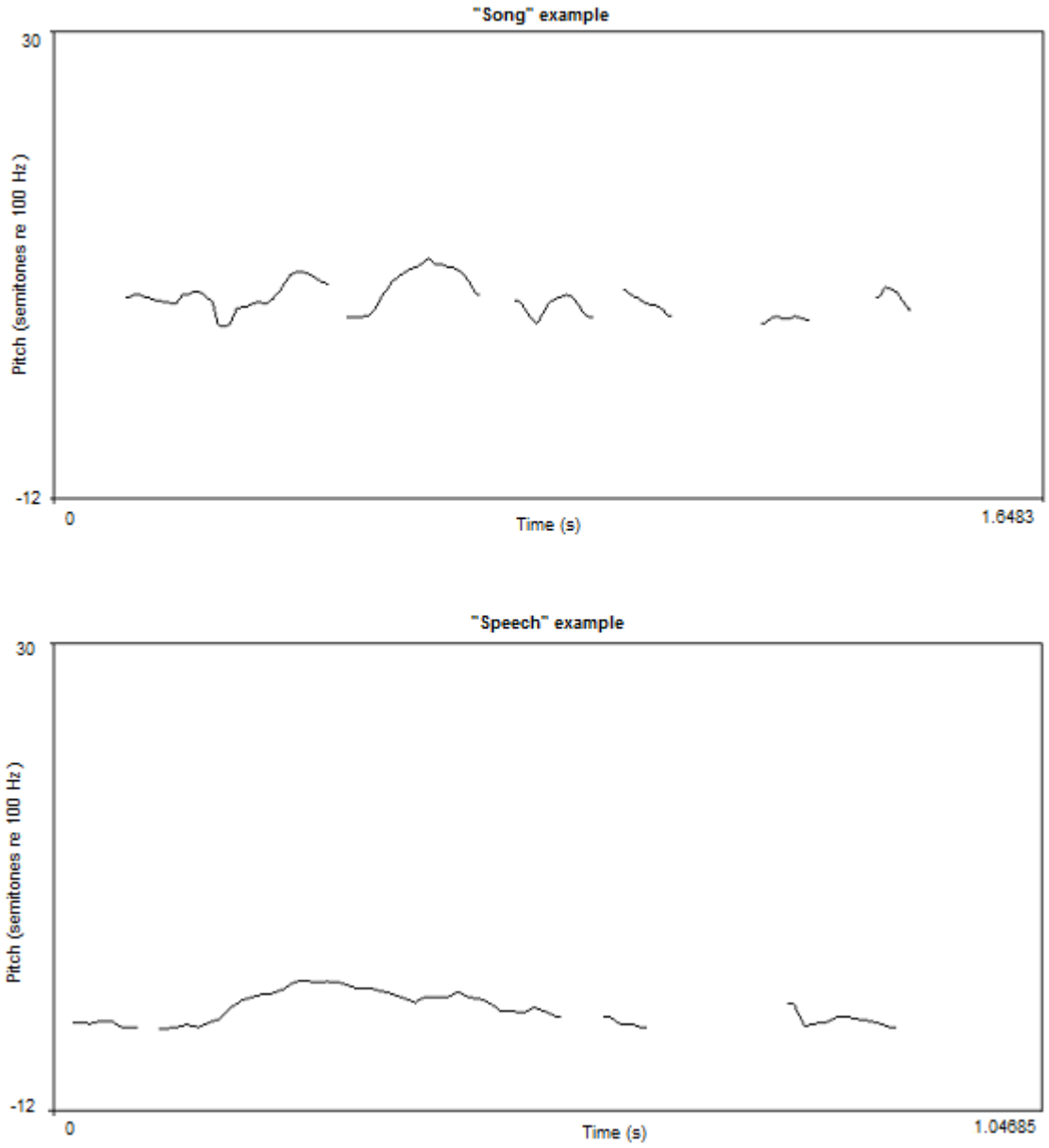


**Figure 4.1:** Pitch contour of single examples of “speech” and “song” illusion stimuli.

Illusion phrases had an average duration of 1305 milliseconds, while non-illusion phrases had an average duration of 1431 milliseconds. Illusion and non-illusion stimuli were matched for syllable length. The average syllable rates, in syllables per second, of the illusion and non-illusion stimuli were 5.13 and 5.00 (Figure 4.2). Illusion phrases had an average median pitch of 141.75 Hz, while non-illusion phrases had an average median pitch of 134.83 Hz (Figure 4.3). Illusion and non-illusion phrases did not significantly differ along any of these dimensions except that of average median pitch according to the Mann-Whitney U test ($p < 0.05$). This difference is, however, extremely small, and is therefore unlikely to cause the difference in the way in which the two sets of stimuli are perceptually categorized or any differences in brain activation found.

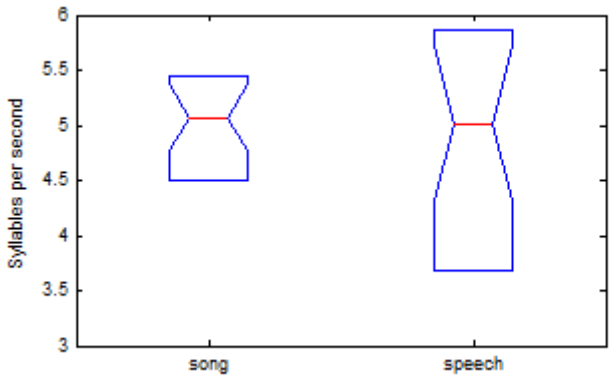


**Figure 4.2:** Speech rate, in syllables per second, of "speech" and "song" stimuli.

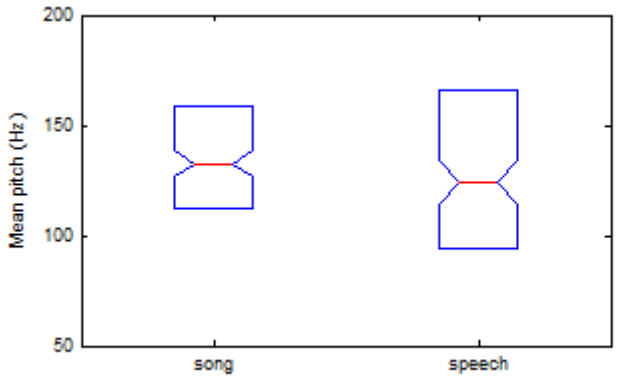


**Figure 4.3.** Mean pitch of "speech" and "song" stimuli.

On average, 21.2% of vowels in non-illusion phrases were classified as glides by prosogram analysis (see Chapter 2), while only 6.6% of vowels in illusion phrases were classified as glides (Figure 4.4), a significant difference according to the Mann-Whitney U test. It would appear that one factor driving the illusion is that the pitch contour within a syllable must be sufficiently flat for listeners to be able to assign a single pitch to that syllable. This may be a necessary condition for a spoken phrase to be heard as song, but it cannot be sufficient, as several non-illusion phrases follow quite flat pitch contours (see, for example, the "speech" example in Figure 1.1).

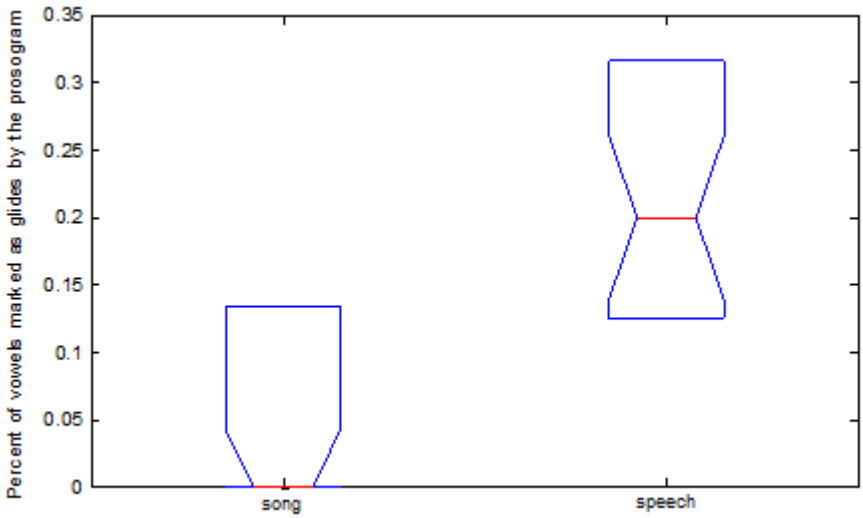


**Figure 4.4.** Percent of vowels marked as glides by the prosogram for "speech" and "song" stimuli.

To ensure that people do, on average, hear the illusion stimuli as song when repeated, and the non-illusion stimuli as speech when repeated, fifteen subjects in a pilot study were asked to listen to eight repetitions of each stimulus and indicate whether what they heard more closely resembled speech or song. On average, illusion stimuli were heard more as song than as speech by 12.67 subjects, while non-illusion stimuli were heard more as song than as speech by only 2.75 subjects (Figure 4.5). fMRI subjects were given the same test prior to being scanned; to ensure that only subjects who actually heard the illusion wwere tested, only subjects whose classifications agreed with the predetermined song/speech classifications in at least 80% of the cases were scanned.

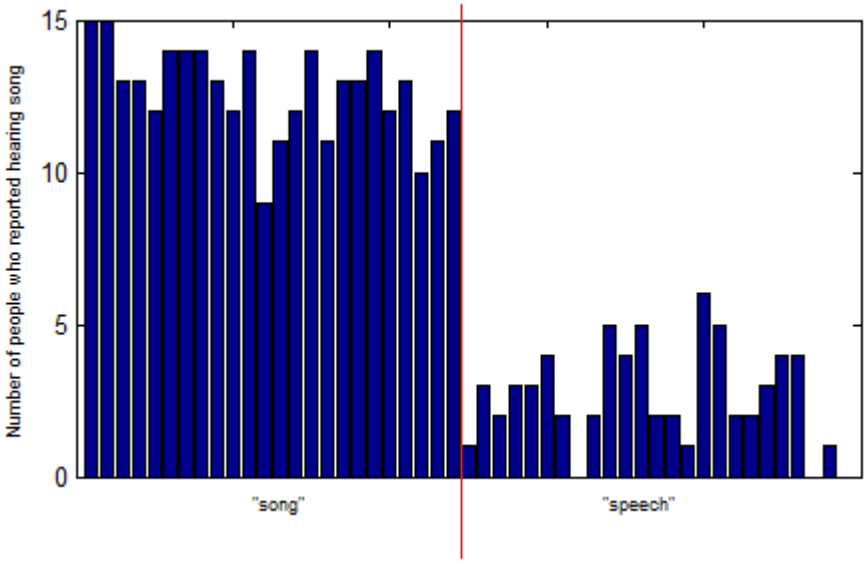


**Figure 4.5.** Number of people who reported hearing song for the "song" and "speech" stimuli.

During functional imaging, stimuli were presented in 16-second blocks. During each block a single phrase was repeated with a 0.5-second inter-stimulus interval as many times as possible within sixteen seconds. Speech, Song, and Silence blocks were presented in pseudo-random order. Subjects were asked to listen carefully to each phrase and mentally note whether the phrase sounded like speech or like song. 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. MR slices were 3.2 mm thick, with an in-plane resolution of 3.2 x 3.2 mm. A single scan took 516 seconds, with 258 single-shot EPI images per slice (TR = 2 seconds, 22 slices, TE = 27.4

ms, flip angle = 90 degrees). Stimulus presentation began after the first two TRs, which were discarded, as the magnetization needs about four seconds to reach a steady state. Each of the 14 subjects underwent four functional scans, for a total of 56 scans.

Each subject's cortical surface was reconstructed from at least one separate structural scan (T1-weighted MPRAGE, 1 x 1 x 1 mm, 224 x 256 x 160 voxels, TR = 16 ms, TE = 6.6 ms, flip angle = 18 degrees). Each subject also underwent at least one alignment scan in order to register the functional data to the reconstructed surface. 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 two images were manually registered and averaged prior to surface reconstruction.

The functional scans were superimposed on the alignment scan, then registered to the higher-resolution structural scan. Statistical analysis was carried out using AFNI (Cox 1996); the resulting activation data was then painted onto the reconstructed cortical surface. After being concatenated, the functional scans were motion-corrected using AFNI's volreg function. Images were registered to the middle of the first functional run. This registration results in six estimated motion parameters.

BOLD activations were analyzed using AFNI's 3dDeconvolve function, which uses multiple linear regression to estimate the extent to which each voxel's time series data fits estimated system impulse response functions. Four

different lag times (1, 2, 3, and 4 TRs) were used when modeling hemodynamic responses. A quadratic polynomial was used to model the baseline. The motion parameters acquired via volreg were included as regressors. We carried out three general linear model tests: Song and Speech versus baseline, and Song versus Speech.

Group averaging was performed by inflating each subject's reconstructed surface to a sphere, performing 10 steps of surface-based smoothing, and aligning it to the FreeSurfer spherical atlas. Results were then painted back onto a single subject's surface for viewing.

### 4.3 Results

To demonstrate the degree of inter-subject variability in activation patterns, Figure 4.6 displays individual subject data from six different hemispheres, one from each of six different subjects. Although there is a consistent pattern of results, there exists noticeable variability as well. For example, there is an activated area just anterior to Heschl's gyrus on the STG in panels A, C, and E, but not in B, D, and F. Activation in the supramarginal gyrus or in the adjacent posterior sylvian fissure, on the other hand, is seen in all six subjects. Activation in the inferior pre-central and post-central gyrii (primary somatosensory and motor cortex) exists in two of the six subjects. In panel E, this activation exists in two distinct areas within both motor and somatosensory cortex, one more lateral than the other, while in panel C only

the more lateral area appears to be activated. Subjects A, C, and E also show activation in left inferior parietal cortex.

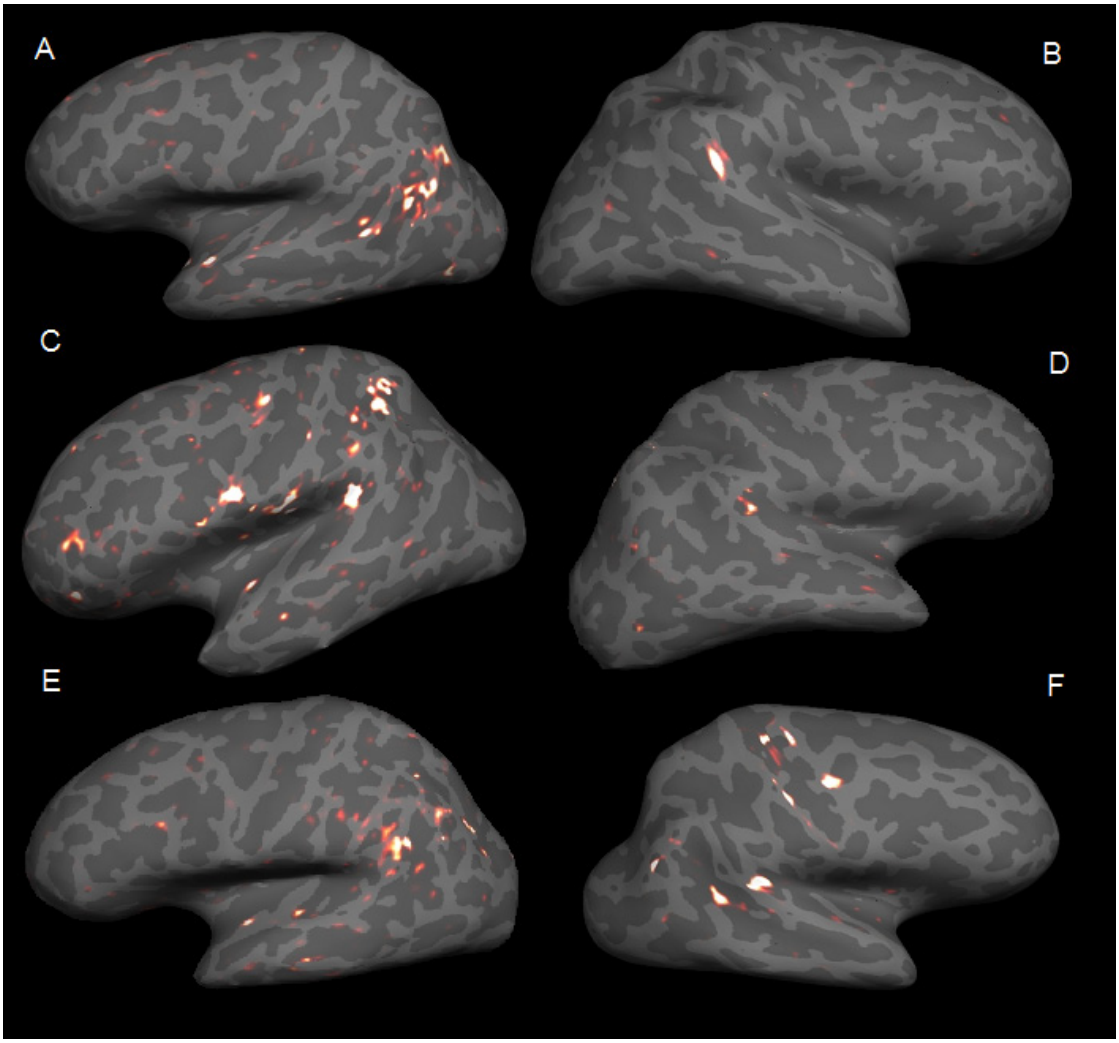


**Figure 4.6.** Examples of single-subject fMRI block design data, activation elicited by "song" stimuli subtracted from activation elicited by "speech" stimuli. Each panel shows data from a different subject.

The inter-subject average is shown in Figures 4.7 and 4.8, uncorrected for multiple comparisons, at a threshold of p = 0.0005. In both hemispheres two distinct areas on the superior temporal lobe are activated. One lies just in

front of Heschl's gyrus, in an area roughly corresponding to that activated by studies of pitch salience which used stimuli that changed rapidly in pitch (Patterson et al. 2002, Warren 2003, Schmithorst 2003). In addition, there is activation in the left supramarginal gyrus and in the posterior sylvian fissure adjacent to the supramarginal gyrus. There is noticeable activation in the inferior precentral gyrus in the right hemisphere but not the left, while the middle temporal gyrus is activated on the left but not the right.

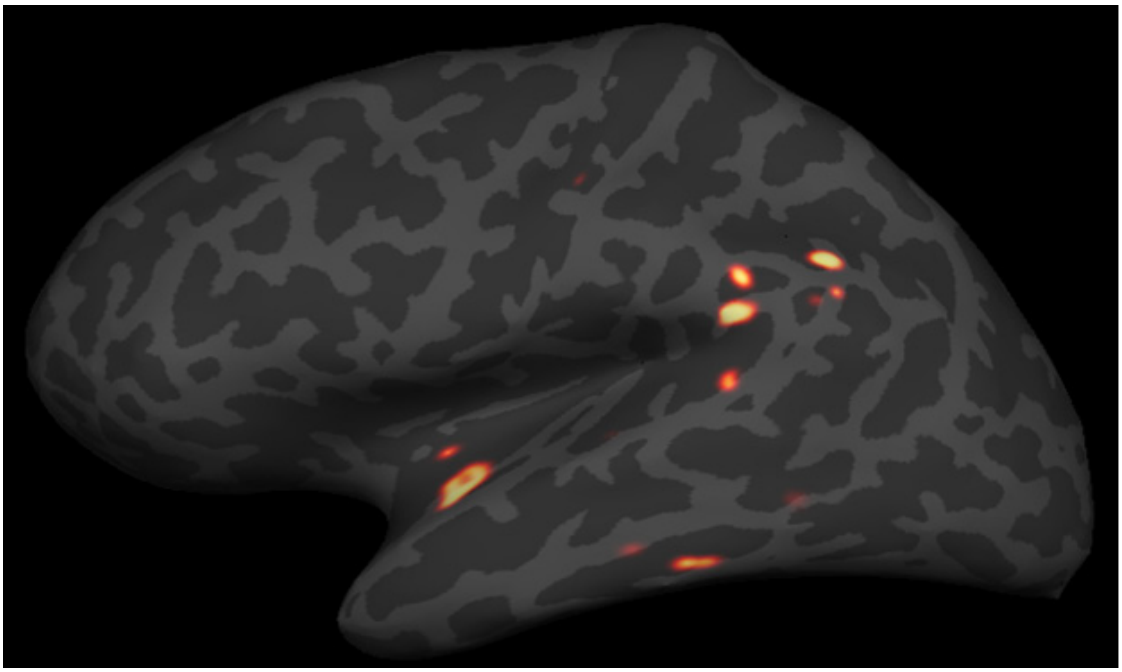

**Figure 4.7.** Between-subjects average, activation in left hemisphere elicited by "song" stimuli subtracted from activation elicited by "speech" stimuli.

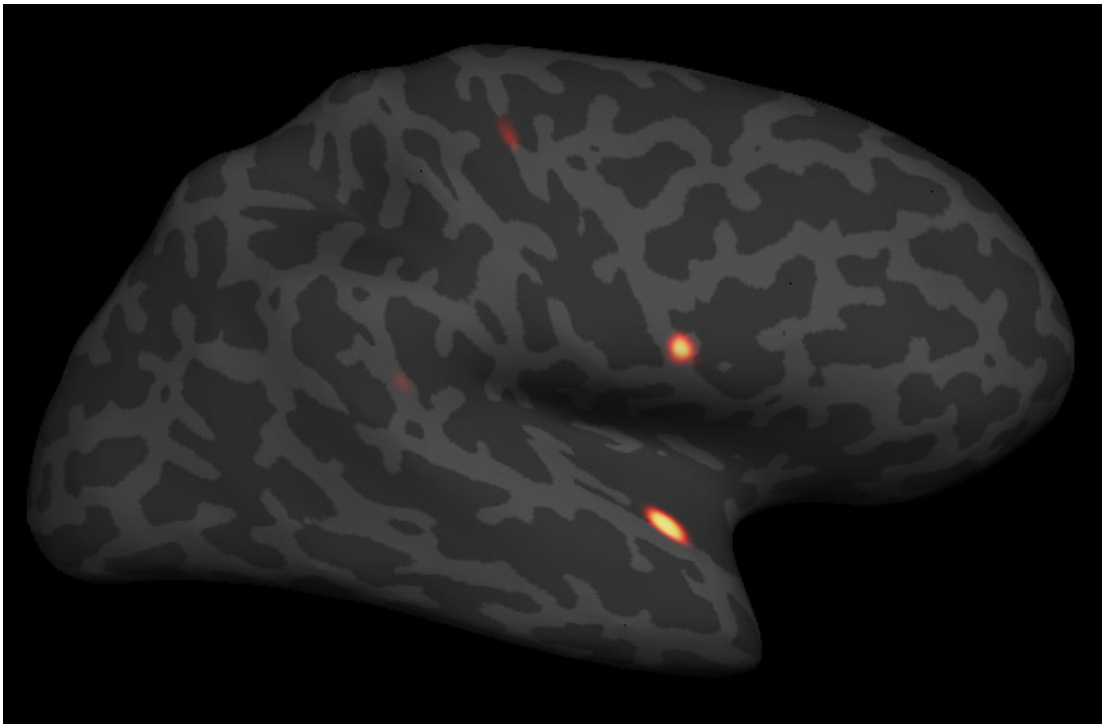

**Figure 4.8.** Between-subjects average, activation in right hemisphere elicited by "song" stimuli subtracted from activation elicited by "speech" stimuli.

Chapter 3 describes a study of tonotopic activation conducted on a subject pool that overlaps partially with the subject pool used in this study. Therefore, for subjects used in both studies, we can ask whether the areas of activation found in the temporal lobes overlap with the parts of auditory cortex also found to be tonotopically activated by band-passed affective voices. As Figure 4.9 shows, in these two subjects, the area just anterior to Heschl's gyrus was inside the area shown to be tonotopically organized in Chapter 3, most likely corresponding to a single anterior belt area. The activation within the supramarginal gyrus and posterior sylvian fissure, however, lies outside the area shown to be tonotopically organized in both hemispheres.

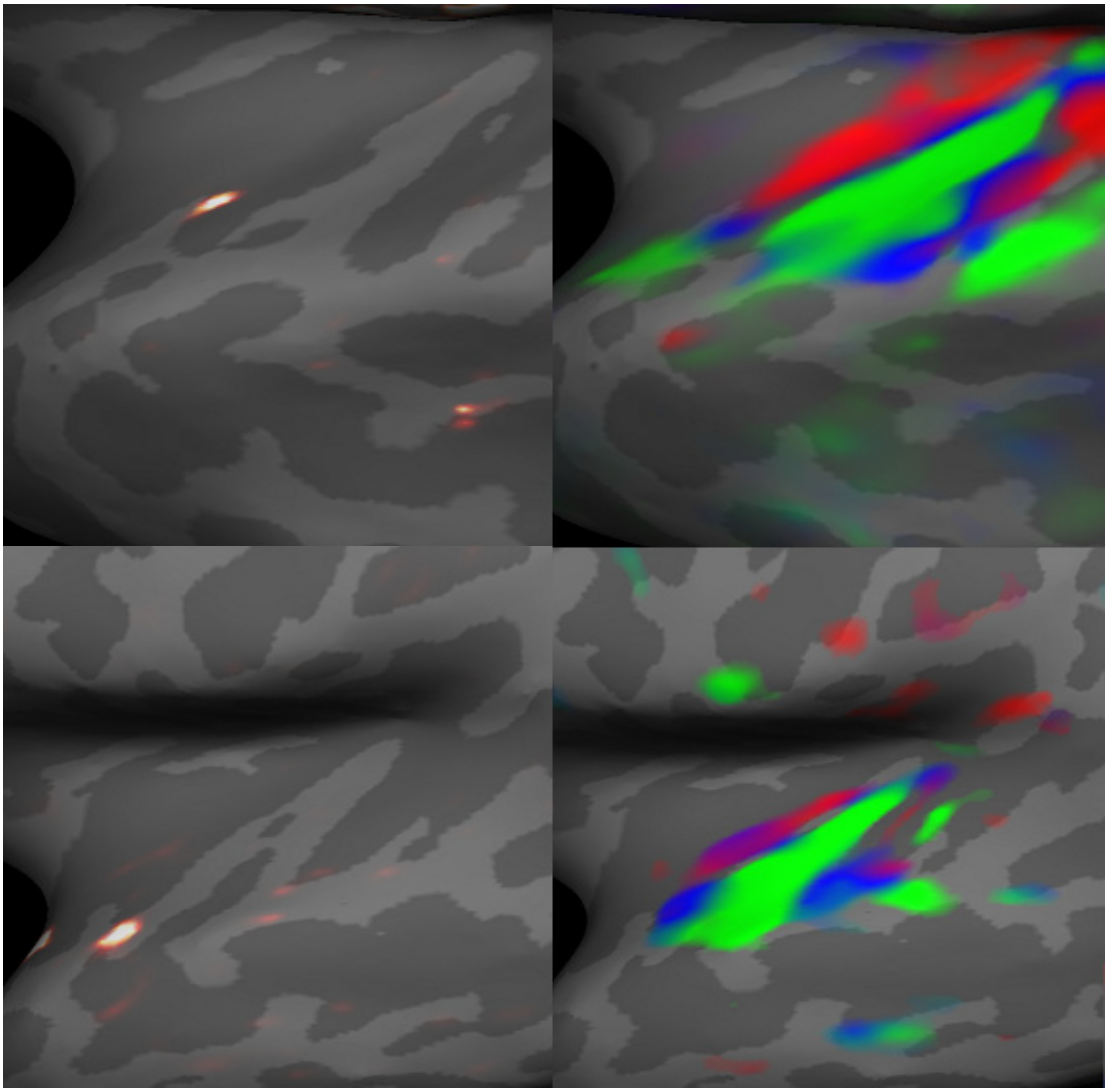

**Figure 4.9.** Comparison of single-subject data from Chapters 3 and 4. Area activated by listening to "song" stimuli anterior to HG falls within region successfully tonotopically mapped in Chapter 3.

## 4.4 Discussion

This study used fMRI to compare the neural correlates of perceiving speech and song. We used as stimuli a set of spoken phrases taken from audio

books; some sound as if they were being sung when they are removed from context and repeated, and others (recorded by the same speakers) do not undergo such a transformation. The two categories of stimuli should, therefore, be evenly matched for any low-level acoustic difference, and in fact we were unable to find any simple acoustic factor differentiating them.

When subjects heard the song stimuli, as compared to the speech stimuli, activation increased bilaterally in the temporal plane just anterior to Heschl's gyrus and in far posterior superior temporal cortex. No areas were significantly more activated when subjects listened to the speech stimuli, as compared to the song stimuli. This implies that perceiving a stimulus as song rather than speech does not involve switching between two different modes of listening; if this were the case, one would expect there to be some areas within auditory cortex that would be more active for speech listening as compared to song listening. Instead, it appears that hearing stimuli as song rather than speech involves the recruitment of extra processing resources in addition to those involved in normal speech processing. Therefore, hearing stimuli as song rather than speech should not hinder listeners from extracting linguistic information, be it phonemic or syntactic, and in fact may make some prosodic information easier to process.

The area activated anterior to Heschl's gyrus is located partially on the superior temporal gyrus and partially on the planum polare. It closely corresponds to an area found to be activated in studies that investigated the

neural correlates of the pitch salience of dynamic, moving pitch sequences (Patterson et al. 2002, Warren 2003, Schmithorst 2003). In the right hemisphere, this area has also been found to be more active when subjects were asked to sing the words to their favorite song, as compared to when they were asked to speak the same words (Jeffries et al. 2003). It was also found to be active bilaterally when subjects listened to six well-known Japanese songs, as opposed to a spoken recitation of the lyrics (Callan et al. 2006). Studies using static stimuli which give rise to a pitch percept have found activation slightly more posterior on lateral Heschl's gyrus (Griffiths et al. 2001, Patterson et al. 2002, Penagos et al. 2004). In those subjects who were studied in both Chapters 3 and 4 and who show evidence of activation in this anterior superior temporal area, the area falls within the bounds of tonotopically activated cortex, and is most likely an anterior belt area.

Taken as a whole, this evidence suggests that hearing the stimuli as song rather than speech involves an increase in attention to pitch information. Although there is linguistic information contained in the pitch of speech—changes in pitch can signal that a particular word is emphasized, for example—information can be conveyed via much smaller pitch movements in music than in speech. An increase in pitch of a single semitone, for example, can turn an in-key note into an out-of-key note, and jazz and blues singing often involve pitch movements of less than a semitone that are stylistically crucial. It's

possible, therefore, that pitch-sensitive cortex is recruited during song listening to facilitate discrimination of fine differences in pitch.

A number of functions have been ascribed to the posterior superior temporal/inferior parietal region that was more strongly activated by the song stimuli than the speech stimuli. This region extends from the posterior end of the Sylvian fissure onto the supramarginal gyrus. The posterior end of the Sylvian fissure, which has been dubbed area SPT (Hickok et al. 2003), is commonly activated by both perceiving and producing speech and song, which has led to suggestions that this area performs auditory-motor integration, closing some sort of auditory-motor feedback loop (Hickok and Poeppel 2007). Hickok et al. (2003), for example, found that the posterior Sylvian fissure was activated both when subjects were asked to listen to hummed melodic tonal sequences or spoken nonsense sentences, and when they were asked to produce these sounds. Moreover, this area appears to be specifically devoted to vocal auditory-motor integration—when skilled pianists were asked to listen to novel melodies and covertly reproduce them by imagining either humming or playing them on a keyboard, the area was significantly more active for the humming than for the playing condition (Pa and Hickok 2008).

The adjacent supramarginal gyrus has also been implicated in auditory feedback during speech production. Hashimoto and Kuniyoshi (2003), for example, presented subjects with either delayed or unaltered auditory feedback during speech production, and found that delayed feedback was

linked to increased activation in the supramarginal gyrus bilaterally and the middle temporal gyrus. Toyomura et al. (2007) asked subjects to vocalize and hold the pitch of their voice constant while they were presented with either unaltered feedback or feedback that was pitch-shifted (which the subjects were asked to compensate for). The altered feedback was correlated with an increase in activation in the supramarginal gyrus bilaterally. Finally, Zarate and Zatorre (2008) asked musically experienced subjects to sing while they were presented with either unaltered feedback or pitch-shifted feedback. Altered feedback led to increased activation in the left supramarginal gyrus, regardless of whether the subjects were asked to compensate for the shifted pitch or ignore it.

Activation within the supramarginal gyrus has also been linked to the performance of auditory working memory tasks. Paulesu et al. (1993) found that the left supramarginal gyrus was activated by a short-term letter memory task, as compared to a task in which subjects were asked to make rhyming judgments on the same (visually presented) letters. Gaab et al. (2003) found that the supramarginal gyrus was also activated when subjects performed a short-term pitch memory task, and that the level of activity in this area was positively correlated with performance on the task. Gaab et al. (2006) trained subjects on the same pitch memory task, divided subjects into strong and weak learners based on their degree of improvement in performance, and found that strong learners, as compared to weak learners, showed an increase in

activation in the left supramarginal gyrus while performing the task. Finally, Vines et al. (2006) found that stimulating the left supramarginal gyrus with transcranial magnetic stimulation (TMS) led to a decrease in performance on a pitch memory task.

In summary, the posterior end of the Sylvian fissure and the supramarginal gyrus have been linked to two main functions: integration of auditory feedback with speech and song production, and short-term memory for both speech information and pitch patterns. These functions are not mutually exclusive, as keeping an item in auditory short-term memory may require subjects to convert an auditory stimulus to a motor program, then covertly run that motor program while updating it with new auditory information. Listening to song, especially repetitive song, often gives rise to an impulse to sing along. Therefore, the increase in activity in the supramarginal gyrus and posterior Sylvian fissure found in this study when subjects listened to song as opposed to speech stimuli may indicate that the subjects had a greater tendency to covertly repeat the phrases that sounded like song; that is, they were internally singing along with the phrases. Alternatively, it's possible that subjects are covertly repeating the stimuli in both cases, but hearing the phrases as song rather than speech causes them to switch attention to the pitch of the signal, leading to activation in areas involved with integration of larynx motor representations and pitch information in auditory signals.

In addition to the bilateral activation described above, hearing the song stimuli (as compared to hearing the speech stimuli) led to increased activation in the most lateral section of the pre-central gyrus, but only in the right hemisphere. This activation clearly lies within the cortical region that has been shown to contain mouth 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 obvious structural landmarks, as only in recent years has this issue begun to be investigated using brain imaging techniques. Brown et al. (2007), for example, found that lip movements were linked to activation in the most inferior extent of the precentral gyrus, that an area activated by tongue movements was located dorsal to the lip area, and that an area activated by larynx movements (both glottal stops and song production) was located dorsal to the tongue area. Loucks et al. (2007) found that breathing also led to activation in this potential larynx motor area. Olthoff et al. (2008), on the other hand, found that phonation (humming a single pitch) and intonation (humming a series of different pitches) led to activation in two different areas: one lateral to an area activated by hand movement corresponding to the area indicated by Brown et al. 2007, and a second in the most inferior part of the precentral gyrus and central sulcus. Finally, Takai et al. (2010) conducted a meta-analysis plotting the points of peak activation of 54 imaging studies of respiration, lip movement, tongue movement, and swallowing. The researchers found that while some separation existed, in a pattern more or less in line with the

findings of Brown et al. (2007), there was a large degree of overlap. Studies of swallowing, which primarily involves pharyngeal muscles, found activation closer to the Sylvian fissure than the other three movements studied.

It's safe to say, therefore, that the area along the precentral gyrus just above the Sylvian fissure, which is activated in the current study when subjects listened to "song" stimuli, is a motor area containing a representation of some part of the mouth/vocal apparatus. Exactly which part is represented, however remains somewhat unclear (with possibilities including the lips, the pharyngeal muscles, or the larynx). If this area does contain a representation of the larynx, its increase in activation may indicate that subjects are covertly repeating the stimuli as well as devoting more attention to the pitch of the song stimuli than that of the speech stimuli.

This area was found to be activated only in the right hemisphere. Previous studies of singing and speech have found conflicting results regarding whether the resulting activation in primary motor cortex is bilateral or biased towards one of the hemispheres. Ricker et al. (2000) asked subjects to either covertly recite the months of the year or repeatedly sing a single tune, and found that covert singing was linked to right primary motor cortex activation, while covert speech was linked to left primary motor cortex activation. Brown et al. (2004) asked subjects to either rest, repeat a melody, harmonize with it, or vocalize monotonically; all three vocal production conditions led to bilateral activation of the mouth region of primary motor cortex. Finally, Ozdemir et al.

(2006) asked subjects to either sing or speak two-syllable words; both conditions led to bilateral activation in the lateral pre-central and post-central gyri. Ozdemir 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 been found to lead to bilateral activation; our findings fit this pattern.

In summary, we have shown that perceiving stimuli as song, as opposed to as speech, is linked to an increase in activation in several areas. Three of these areas are potentially involved in pitch processing: an area anterior to Heschl's gyrus activated in many studies of the neural correlates of pitch salience, a region in the posterior sylvian fissure and on the supramarginal gyrus that has been linked to pitch memory and auditory-motor integration, and a portion of the mouth area on the precentral gyrus that may be one of the areas controlling the larynx. These findings suggest that song processing differs from speech processing in large part simply by an increase in attention to pitch.

**Acknowledgement**

Chapter 4, in part, is being prepared for publication. The dissertation author was the primary investigator and author of this paper. The dissertation author acknowledges the following co-authors as collaborators on this manuscript: Diana Deutsch and Martin Sereno. Permission from all co-authors has been obtained.

## References

Ayotte, J., Peretz, I., Rousseau, I., Bard, C., and Bojanowski, M. (2000). Patterns of music agnosia associated with middle cerebral artery infarcts. Brain *123*, 1926-1938.

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. Nature *436*, 1161–1165.

Brown, S., Martinez, M., Hodges, D., Fox, P., and Parsons, L. (2004). The song system of the human brain. Cognitive Brain Research *20*, 363–375.

Brown, S., Ngan, E., and Liotti, M. (2007). A larynx area in the human motor cortex. Cerebral Cortex *18*, 837-845.

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

Cox, R. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research *29*, 162–173.

Dale, A., Fischl, B., and Sereno, M. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. Neuroimage *9*, 179-194.

d'Alessandro, C. (1994). The pitch of short-duration vibrato tones. J. Acoust. Soc. Am. *95*, 1617.

d'Alessandro, C., and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. Computer Speech & Language *9*, 257–288.

d'Alessandro, C., Rosset, S., and Rossi, J. (1998). The pitch of short-duration fundamental frequency glissandos. J. Acoust. Soc. Am *104*, 2339-2348.

Deutsch, D. (1995). Musical Illusions and Paradoxes. Philomel Records.

Deutsch, D. (2003). Phantom Words, and other Curiosities. Philomel Records.

Deutsch, D. (2008). The speech-to-song illusion. The Journal of the Acoustical Society of America *124*, 2471.

Fischl, B., Sereno, M., and Dale, A. (1999). Cortical surface-based analysis II: inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195–207.

Gaab, N., Gaser, C., and Schlaug, G. (2006). Improvement-related functional plasticity following pitch memory training. Neuroimage *31*, 255–263.

Gaab, N., Gaser, C., Zaehle, T., Jancke, L., and Schlaug, G. (2003). Functional anatomy of pitch memory–an fMRI study with sparse temporal sampling. Neuroimage *19*, 1417–1426.

Griffiths, T., Uppenkamp, S., Johnsrude, I., Josephs, O., and Patterson, R. (2001). Encoding of the temporal regularity of sound in the human brainstem. Nature Neuroscience *4*, 633–637.

Hagler Jr, D., Saygin, A., and Sereno, M. (2006). Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. Neuroimage *33*, 1093–1103.

Harris, P., and Silberstein, R. (1999). Steady-state visually evoked potential (SSVEP) responses correlate with musically trained participants' encoding and retention phases of musical working memory task performance. Australian Journal of Psychology *51*, 140–146.

Hashimoto, Y., and Sakai, K. (2003). Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study. Human Brain Mapping *20*, 22–28.

Hickok, G., Buchsbaum, B., Humphries, C., and Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. Journal of Cognitive Neuroscience *15*, 673–682.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience *8*, 393–402.

House, D. (1990). Tonal Perception in Speech (Lund University Press).

Hugdahl, K., Br\onnick, K., Kyllingsbrk, S., Law, I., Gade, A., and Paulson, O. (1999). Brain activation during dichotic presentations of consonant-vowel and musical instrument stimuli: a 15O-PET study. Neuropsychologia *37*, 431–440.

Jeffries, K., Fritz, J., and Braun, A. (2003). Words in melody: an H215O PET study of brain activation during singing and speaking. Neuroreport *14*, 745-749.

Koelsch, S., Gunter, T., Cramon, T., Zysset, S., Lohmann, G., and Friederici, A. (2002). Bach speaks: a cortical "language-network" serves the processing of music. Neuroimage *17*, 956–966.

Luria, A., Tsvetkova, L., and Futer, D. (1965). Aphasia in a composer. Journal of the Neurological Sciences *2*, 288–292.

Maess, B., Koelsch, S., Gunter, T., and Friederici, A. (2001). Musical syntax is processed in Broca's area: an MEG study. Nature Neuroscience *4*, 540–545.

Marin, O., Perry, D., and Deutsch, D. (1999). Neurological aspects of music perception and performance. In The Psychology of Music (Academic Press), pp. 653–724.

Mertens, P. (2004). The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In Speech Prosody 2004, International Conference.

Messerli, P., Pegna, A., and Sordet, N. (1995). Hemispheric dominance for melody recognition in musicians and non-musicians. Neuropsychologia *33*, 395–405.

Olthoff, A., Baudewig, J., Kruse, E., and Dechent, P. (2009). Cortical sensorimotor control in vocalization: a functional magnetic resonance imaging study. The Laryngoscope *118*, 2091–2096.

Özdemir, E., Norton, A., and Schlaug, G. (2006). Shared and distinct neural correlates of singing and speaking. Neuroimage *33*, 628–635.

Pa, J., and 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.

Patel, A., and Daniele, J. (2003). An empirical comparison of rhythm in language and music. Cognition *87*, B35–B45.

Patel, A., Gibson, E., Ratner, J., Besson, M., and Holcomb, P. (1998). Processing syntactic relations in language and music: An event-related potential study. Journal of Cognitive Neuroscience *10*, 717–733.

Patel, A., Iversen, J., and Rosenberg, J. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. The Journal of the Acoustical Society of America *119*, 3034-3047.

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

Paulesu, E., Frith, C., and Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. Nature *362*, 342-345.

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

Peretz, I. (1990). Processing of local and global musical information by unilateral brain-damaged patients. Brain *113*, 1185-1205.

Riecker, A., Ackermann, H., Wildgruber, D., Dogil, G., and Grodd, W. (2000). Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. Neuroreport *11*, 1997-2001.

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

Schouten, M. (1985). Identification and discrimination of sweep tones. Perception and Psychophysics *37*, 369-376.

Takai, O., Brown, S., and Liotti, M. (2010). Representation of the speech effectors in the human motor cortex: Somatotopy or overlap? Brain and Language *113*, 39-44.

Tervaniemi, M., Medvedev, S., Alho, K., Pakhomov, S., Roudas, M., Van Zuijen, T., and Naatanen, R. (2000). Lateralized automatic auditory processing of phonetic versus musical information: a PET study. Human Brain Mapping *10*, 74–79.

Tervaniemi, M., Szameitat, A., Kruck, S., Schroger, E., Alter, K., De Baene, W., and Friederici, A. (2006). From air oscillations to music and speech: functional magnetic resonance imaging evidence for fine-tuned neural networks in audition. Journal of Neuroscience *26*, 8647-8652.

t'Hart, J., Collier, R., and Cohen, A. (1990). Perceptual Study of Intonation (Cambridge).

Toyomura, A., Koyama, S., Miyamaoto, T., Terao, A., Omori, T., Murohashi, H., and Kuriki, S. (2007). Neural correlates of auditory feedback control in human. Neuroscience *146*, 499–503.

Vines, B., Schnider, N., and Schlaug, G. (2006). Testing for causality with transcranial direct current stimulation: pitch memory and the left supramarginal gyrus. Neuroreport *17*, 1047-1050.

Ward, B., and Revision, L. (2000). Deconvolution analysis of fMRI time series data. AFNI 3dDeconvolve Documentation, Medical College of Wisconsin.

Warren, J., Uppenkamp, S., Patterson, R., and Griffiths, T. (2003). Separating pitch chroma and pitch height in the human brain. Proceedings of the National Academy of Sciences of the United States of America *100*, 10038-10042.

Zarate, J., and Zatorre, R. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. Neuroimage *40*, 1871–1887.

Zatorre, R. (1988). Pitch perception of complex tones and human temporal-lobe function. The Journal of the Acoustical Society of America *84*, 566–572.

Zatorre, R., and Samson, S. (1991). Role of the right temporal neocortex in retention of pitch in auditory short-term memory. Brain *114*, 2403-2417.

# Chapter 5.

# Conclusions

The work presented in this dissertation revolves around a single simple finding: that several statistical patterns found in speech and music—arch-shaped pitch patterns, final lengthening, and the tendency for large jumps in pitch to be followed by reversals—have a common source in the structure of the human body. These same constraints are shared with birds and, most likely, with almost all vocalizing mammals, as they would be found in any animal producing vocalizations powered by the lungs, produced by a vibrating larynx, and filtered through the vocal cavity. The motor constraint hypothesis, therefore, motivates a wide range of studies on patterns in the calls and songs of birds and mammals. If a particular singing animal lacks a given constraint, that animal's songs or calls may not necessarily follow the pattern that arises from that constraint.

Portions of chimpanzee pant-hoots, for example, are produced during both inspiration and expiration. One might, therefore, expect only the portions of the call produced during expiration to feature the arch-shaped pitch contour, as portions produced during inspiration would not be driven by sub-glottal pressure. The hypothesis linking rapid movement to final lengthening could be tested by examining either whales, which sing very slowly and thus may not need to slow down at the ends of song phrases, or frogs, which do not incorporate rapid supraglottal articulation into their vocalizations. Future work could also replicate our birdsong findings with other animals sharing these same constraints—mice, for example, produce songs, and do so using a

system that includes the lungs, the larynx, and (potentially) rapid alteration of the resonating cavity.

The results presented here demonstrating final lengthening, arch contours, and skip-reversal patterns in speech and music are at best preliminary. Given the staggering diversity in musical systems and languages found throughout the world, our corpuses of four spoken languages, themes taken from Western classical music, and songs from Europe and China are at best a tiny sample. In particular, work demonstrating that these patterns hold true in corpuses primarily consisting of non-Western music would go a long way towards indicating that these patterns stem from bodily features shared by all humans. Further research on final lengthening and lowering could also investigate whether composers and performers exaggerate these effects at major structural boundaries. Also, while final lowering has been repeatedly documented and discussed in human speech, the existence of an initial rise in pitch has been almost completely ignored, and never before empirically demonstrated, perhaps because the pitch of speech is often notated entirely in terms of the pitch of stressed syllables. While final lowering is widely assumed to convey a feeling of finality and therefore facilitates conversational turn-taking, it remains to be seen whether the initial rise in pitch plays any such communicative role.

While it is likely that the auditory system takes advantage of the regularities found in human vocalizations, either through learning or through

evolutionary adaptation, the exact nature of this specialization remains, for now, unclear. In Chapter 3 I report success using simple arch-like emotional vocalizations to tonotopically map areas in the STS previously found to be voice-sensitive. This suggests that only relatively simple characteristics of vocalizations may be necessary to activate these areas. One way to test this hypothesis would be to use fMRI to compare the resulting brain activation from hearing pure-tone pitch contours that are either arch-shaped or u-shaped. U-shaped contours are rather unnatural to produce using the human vocal system, and through personal observation and informal experimentation I have found that arch-shaped pure-tone contours sound more like vocalizations than u-shaped contours. This suggests that arch-shaped pure-tones may activate the STS. A more general prediction would simply be that stimuli following final lengthening and arch contours would generate more auditory cortex activation than closely matched stimuli not following these patterns.

Another line of research could examine the processing advantages that exist when the auditory system perceives stimuli following these patterns. It's possible, for example, that subjects would be better able to distinguish between pairs of arch-shaped contours in which the exact location of the peak differs slightly, as opposed to between pairs of u-shaped contours in which the exact location of the trough differs. It is also possible that subjects would be better able to make fine temporal judgments between stimuli that feature final

lengthening, as opposed to stimuli that do not. If either of these advantages is found, one could then investigate whether they are also present in birds. If so, one could then attempt to determine whether or not the processing advantages are innate or learned by depriving birds of auditory input or raising them normally before testing them in adulthood.

The finding that speech and music share several statistical patterns, along with the more obvious acoustic characteristics they share by virtue of being produced by the same vocal system, suggests that it should be possible to find ambiguous stimuli that can be perceived as speech or song. In addition to Diana Deutsch's original example, I have now found twenty-four additional examples, and given enough time it should be possible to find many more. Using these stimuli, I showed in Chapter 4 that perceiving a stimulus as song, as opposed to speech, is linked to an increase in activation in just a few areas: an auditory belt area which has been shown to be strongly activated by pitch salience, an area in primary motor cortex that may correspond to a representation of the larynx, and an auditory-motor integration area located approximately at the junction of the temporal and parietal lobes. Future work could further determine what exactly the differences are between the speech and song listening modes. It's possible, for example, that hearing stimuli as song rather than speech is linked to better pitch-tracking at the level of the brainstem, which could be investigated using the EEG frequency-following response. Future work could also attempt to determine in more detail the

exact characteristics that must be present for isolated and repeated spoken stimuli to sound as if they were sung, with the end goal of being able to either manufacture such stimuli or find them in large speech corpuses.