

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Cortical representation of vocal pitch during speech perception in human superior temporal gyrus

Permalink

<https://escholarship.org/uc/item/3pc386t8>

Author

Tang, Claire

Publication Date

2019

Peer reviewed|Thesis/dissertation

Cortical representation of vocal pitch during speech perception in human superior temporal gyrus

by
Claire Tang

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

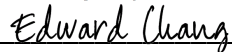


99D647545A30488...

Christoph Schreiner, MD, PhD

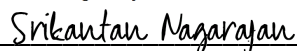
Chair

DocuSigned by:



Edward Chang

DocuSigned by:



4C21BC1B9EB145E...

Srikantan Nagarajan

Committee Members

Copyright 2019

by

Claire Tang

Acknowledgments

Completing my PhD was both challenging and rewarding, and it would not have been possible without the support of many people. First, I would like to thank my PhD advisor, Dr. Eddie Chang, who inspired me to think big, gave me freedom to explore and create, guided my thinking about my findings and results, and helped me see the larger context of my work. He was also there for me in times of both personal and professional distress, and it is in thinking back to those hard times that I can fully appreciate the amount of support I received.

I would like to thank Dr. Christoph Schreiner, who was the chair of both my qualifying exam committee and thesis committee, for encouraging me to be the best scientist I could be. He provided wise guidance as I navigated graduate school, beginning to end. I would also like to thank Dr. Srikantan Nagarajan, for serving on my thesis committee and for giving me critical feedback on my thesis research.

I am also extremely grateful for the home that I found in the UCSF Neuroscience program and in the Chang lab. In particular, I would like to thank Jiggy Athilingam, Jermyn See, Yelena Kulik, and the rest of my 2013 incoming Neuroscience class for making the transition to grad school so easy and fun. I would like to thank Matt Leonard for mentoring me through my rotation and showing me how interesting and exciting speech science could be. When I joined the Chang lab, I could not have felt a greater sense of belonging and that was because of the welcome I received from Liberty Hamilton, David Conant, Leah Muller, Ben Dichter, and David Moses. I would also like to thank Neal Fox for being a great friend and someone I could talk to about anything. It's hard to express just how much I enjoyed all of the conversations, lunches, conferences, dances, movies, dinners, and nights out.

I would also like to thank Patrick Johnston and Drew Friedmann, for making our home a place that I was always happy to head back to after work, and everyone in my two running families and in my dance communities.

Finally, I would like to thank my sister, who has always been there for me, and my parents who raised me and supported me through all of the years of my life.

Parts of this dissertation are reprints of material already published in peer-reviewed journals.

Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. *Science* 357:797–801.

Abstract

Cortical representation of vocal pitch during speech perception
in human superior temporal gyrus

by

Claire Tang

Pitch plays a crucial role in all spoken languages. In tone languages, pitch is used to distinguish between different words, such that the same syllable can have multiple lexical meanings depending on its pitch contour. In all other spoken languages, pitch conveys linguistic meaning at the sentence level through speech intonation. For example, in English, raising the pitch at the end of an utterance can change a statement into a question. Despite the importance of pitch for spoken language, we have limited understanding of how the human brain processes speech to represent pitch that is linguistically relevant. One difficulty that arises for the encoding of linguistic meaning in pitch is that the vocal pitch range varies vastly across different people. Thus, languages cannot use absolute values of pitch to convey meaning since some values of high absolute pitch may be out of the range of a low-pitched speaker, and vice versa. Instead, linguistic meaning must be transferred through a speaker-normalized representation of pitch. This dissertation seeks to understand how the human auditory cortex represents pitch information during speech perception. Using electrocorticography to record neural activity directly from the cortical surface of participants as they listen to both natural speech and controlled speech stimuli, I discovered populations of neurons in the human superior temporal gyrus that have activity patterns that differentiate lexical tones and intonation contours. These neural populations are separate from the neural populations that encode the phonetic features that make up different

consonants and vowels and from the neural populations that encode information about speaker identity. Furthermore, I show that the activity of these tone and intonation neural populations can be explained by the encoding of speaker-normalized relative pitch and pitch change. These results suggest that human auditory cortex processes speech to extract vocal pitch and abstracts absolute pitch values to encode linguistically relevant, speaker-normalized pitch information at the level of human non-primary auditory cortex.

Table of Contents

1	Introduction	1
2	Intonational speech prosody encoding in human auditory cortex	7
3	Speaker-normalized pitch encoding in human superior temporal gyrus in response to Mandarin tones.....	41
4	Conclusion.....	67
	References.....	69

List of Figures

Figure 2.1. Neural activity in STG differentiates intonational pitch contours.	9
Figure 2.2. Independent neural encoding of intonation, sentence, and speaker information at single electrodes.	12
Figure 2.3. Similar neural responses to intonation in speech and non-speech contexts.	14
Figure 2.4. Cortical representation of intonation relies on relative, not absolute, pitch encoding.	18
Figure 2.5. Individual and group cortical maps of intonation, sentence, and speaker condition encoding.	31
Figure 2.6. Activity that differentiates sentence conditions is driven by phonetic feature selectivity.	34
Figure 2.7. Absolute and relative pitch contours and variability in TIMIT speech corpus.	35
Figure 2.8. Example relative pitch encoding electrodes tuned for low relative pitch and high-to-low relative pitch.	37
Figure 2.9. Example absolute pitch encoding electrodes tuned for high and low absolute pitch.	39
Figure 3.1. Speaker-normalized relative pitch and pitch change are acoustic features that define lexical tones in continuous Mandarin speech.	44
Figure 3.2. Lexical tones in continuous Mandarin speech evoke differential neural responses in discrete populations in STG.	46
Figure 3.3. Encoding of speaker-normalized relative pitch height and pitch change underlies lexical tone discriminability.	48
Figure 3.4. STG responses to Mandarin tones largely reflects language-independent encoding of speaker-normalized pitch.	51
Figure 3.5. Speech responsive and tone discriminating electrodes for all participants.	62
Figure 3.6. Examples of different types of speaker-normalized pitch feature encoding.	64
Figure 3.7. Models fit on Mandarin data perform better than models fit on English data for both Mandarin and English speakers.	65

Chapter 1

1 Introduction

One of the most amazing feats of human cognition is our ability to communicate with each other through speech. At its core, speech is just a sound, which is processed like all other sounds by the human auditory system. However, speech is a highly complex acoustic signal that changes quickly over time, containing structure related to the phonology of spoken language. And the auditory processing required for speech comprehension goes far beyond hair cells in the cochlea responding to mechanical vibrations in the air. In speech, acoustic features such as sound frequency over time signal differences between individual consonant and vowel sounds. Other acoustic features that typically occur over longer timescales, such as pitch, loudness, and duration also convey linguistic meaning. Understanding how the human brain processes speech to ultimately derive meaning is one of the great undertakings of cognitive neuroscience. This dissertation adds to this body of knowledge by focusing on how human auditory cortex represents information about vocal pitch during speech perception.

Linguistic uses of pitch in speech

All spoken languages use pitch to convey linguistic meaning. In tonal languages, such as Mandarin, pitch differentiates words, such that changing the pitch contour of an otherwise identical sequence of consonants and vowels changes which word is being spoken (Chao, 1965). In all other spoken languages, pitch conveys sentence-level meaning through speech intonation (Cutler et al., 1997; Ladd, 2008). The same string of words can have different meanings when the pitch contour is changed. For example, in English, rising intonation can turn statements into

questions (“Anna likes oranges” vs. “Anna likes oranges?”). Local changes in pitch, called pitch accents, can also change sentence semantics through emphasis. For example, “*Anna* likes oranges” answers the question “Who likes oranges?”, while “Anna likes *oranges*” answers the question “What does Anna like?”.

Pitch is an auditory sensation that can be characterized on the dimension from low to high and is associated with musical melodies (Acoustical Terminology, 1960). In speech, it is produced by the vibration of the vocal folds and is correlated with the acoustic feature of fundamental frequency (f_0). Because vocal pitch ranges are determined by the physical length of the vocal folds and size of the larynx (Titze, 1989), they vary tremendously between people such that two people can have pitch ranges that do not overlap. In fact, the absolute pitch of a person’s voice is a primary cue to their speaker identity (van Dommelen, 1990). Because of this, absolute pitch per se cannot be used to encode linguistic meaning. In order for pitch to be useful as a linguistic feature, it must be normalized to each speaker and considered in a relative way. Despite the fact that vocal pitch is an important feature in spoken language, we have only a limited understanding of how the human auditory cortex encodes pitch information relevant to language during speech perception.

Neural basis of speech intonation and tone perception

Research on the neural basis of speech intonation and tone perception has primarily focused on identifying the gross brain regions involved. For speech intonation processing, early lesion studies implicated cortical regions in both hemispheres (Pell and Baum, 1997; Witteman et al., 2011), with some specifically implicating temporoparietal regions (Weintraub et al., 1981). Other studies used neuroimaging techniques (e.g. fMRI and PET) to investigate which cortical

areas modulate their activity during an intonation manipulation, either by comparing activation to different stimuli (Meyer et al., 2002, 2004) or by comparing activation to the same stimuli under different task conditions (Kreitewolf et al., 2014). These studies generally point to a fronto-temporal network, including inferior frontal gyrus, superior temporal gyrus (STG), and superior temporal sulcus (Plante et al., 2002; Friederici and Alter, 2004; Gandour et al., 2004; Meyer et al., 2004). However, these studies have not provided a detailed description of what the neural activity in each brain area is actually doing, whether that is encoding a stimulus or task feature or implementing an auditory or cognitive process. Additionally, studies that use task manipulations with the same set of stimuli assume that tasks that direct attention to certain acoustic features of a stimulus will cause greater activation of regions involved in processing those acoustic features (Zatorre et al., 1992; Kreitewolf et al., 2014). However, it is unclear whether this activation is a result of processing specific features or related to attention and decision strategies (Tong et al., 2005).

For tone perception, previous neuroscience research also has largely focused on anatomic localization of processing. A recent meta-analysis that aimed to clarify the neuroanatomy of lexical tone perception included dozens of neuroimaging studies that compared activations during tone discrimination to activations during passive listening or silence (Gandour et al., 2000; Klein et al., 2001; Liang and Du, 2018). These studies have primarily focused on questions of hemispherical asymmetry and how language experience may bias processing towards one hemisphere over the other (Gu et al., 2013; Ge et al., 2015). Like with speech intonation, there has not been much research on the mapping of stimulus features to neural activity in different brain regions.

Auditory processing for pitch perception

Another area of research that is relevant to understanding the neurophysiology of vocal pitch perception is the neuroscience of pitch perception in general. There have been many studies on the types of information carried by auditory nerve fibers in the auditory periphery about the pitch of simple and complex tones (Licklider, 1954; Cariani and Delgutte, 1996a; Plack et al., 2006). The processing of vocal pitch in speech must also start at the auditory periphery, however the auditory processing that underlies speech perception may be better understood at the cortical level. Electrophysiological studies in non-human primates suggest that anterior regions of non-primary auditory cortex may play a role in processing pitch. These studies have shown that neurons in the anterolateral belt region are selective for frequency sweeps at the slow rates found in vocalizations (Tian and Rauschecker, 2004) and can respond similarly to pure tones and missing fundamental complex sounds that are matched in pitch (Bendor and Wang, 2005). Human neuroimaging studies have also localized cortical pitch processing to non-primary auditory cortex in lateral Heschl's gyrus and STG (Griffiths et al., 1998; Patterson et al., 2002; Penagos et al., 2004). However, it is still an open question how neural activity in these regions relate to pitch values during speech perception.

Electrocorticography

In order to study the physiology of human auditory cortex during speech perception for this dissertation, I employed electrocorticography (ECoG) to record electrical activity directly from the brain surface. In this method, multielectrode grids are subdurally implanted during a neurosurgical procedure. Because this is a highly invasive procedure, it is only performed in specific circumstances on patients undergoing clinical treatment for intractable epilepsy or

certain types of brain cancer. For patients with epilepsy, the ECoG grids allow clinicians to monitor seizure activity and localize seizure foci in preparation for surgical resection. For patients with brain tumors, grids can be placed intra-operatively for language mapping so that neurosurgeons can minimize the risk of postoperative aphasia by avoiding areas found to be involved in language (Ojemann et al., 1989). During this time, many patients volunteer to participate in research studies.

Electrocorticographic signals are an excellent way to study the neurophysiology of speech perception. Their temporal resolution is high enough to track the rapidly changing acoustic features in speech. The spatial resolution and signal-to-noise ratio of ECoG is better compared to methods such as scalp electroencephalography because the signals are not spatially blurred and low-pass filtered passing through the skull (Pfurtscheller and Cooper, 1975). Early studies on event-related changes in the ECoG power spectrum focused on lower frequencies in the alpha (8-13 Hz) and beta (15-25 Hz) ranges (Crone, 1994; Ohara et al., 2004), but it was found that activity in higher frequencies (75-100 Hz) may more closely reflect cortical activation for sensorimotor (Crone et al., 1998), visual (Lachaux et al., 2005; Vidal et al., 2010), and auditory or speech-related tasks (Crone et al., 2001; Edwards et al., 2005; Mesgarani et al., 2014). Recent studies have also shown that high-gamma frequencies in the ECoG signal are correlated with the number of neuronal action potentials and can be used as a measure of neural activity (Ray and Maunsell, 2011).

Overview of dissertation

In Chapter 2, I investigate how intonational speech prosody is represented in human auditory cortex. I show that local neural populations in the superior temporal gyrus respond

differentially to intonation contours regardless of phonetic content or speaker. These populations were spatially distinct from regions that encoded phonetic content or speaker identity. I further show that the neural responses could be explained by the encoding of relative pitch, which is normalized by speaker, but not by the encoding of absolute pitch.

In Chapter 3, I investigate how lexical tones are represented in human auditory cortex. I show that activity patterns from certain neural populations in STG differentiate tones. Similar to findings in Chapter 2, neural activity in response to tones could not be explained by absolute pitch encoding. Instead, I found that two speaker-normalized pitch features, relative pitch height and pitch change, were encoded in tone discriminating populations. Additionally, I show that the responses to tones in STG do not depend on language comprehension and instead reflect general auditory processing of pitch in speech.

Overall, I have discovered neural populations in human non-primary auditory cortex that encode speaker-normalized representations of pitch during speech perception. These speaker- and phonetic-content- invariant representations provide a neurobiological basis for a listener's ability to recognize phonologically identical pitch contours despite variation between speakers and underlying phonetic content and provide insight into STG's role in representing abstract, linguistically-relevant information during speech perception.

Chapter 2

2 Intonational speech prosody encoding in human auditory cortex

Abstract

Speakers of all human languages regularly use intonational pitch to convey linguistic meaning, such as to emphasize a particular word. Listeners extract pitch movements from speech and evaluate the shape of intonation contours independent of each speaker's pitch range. We recorded neural population activity directly from the brain surface using high-density electrocorticography while participants listened to sentences that varied in intonational pitch contour, phonetic content, and speaker. Cortical activity at single electrodes over the human superior temporal gyrus selectively represented intonation contours. These electrodes were intermixed with, yet functionally distinct from sites that encoded different information about phonetic features or speaker identity. Furthermore, the representation of intonation contours directly reflected the encoding of speaker-normalized relative pitch, but not absolute pitch.

Introduction

Humans precisely control the pitch of their voices to encode linguistic meaning (Cutler et al., 1997; Ladd, 2008). All spoken languages use suprasegmental pitch modulations at the sentence-level, or speech intonation, to convey meaning not explicit in word choice or syntax (Shattuck-Hufnagel and Turk, 1996). Raising the pitch on a particular word can change the meaning of a sentence. While “*Anna* likes oranges” communicates that it is Anna, not Lisa, who likes oranges, “Anna likes *oranges*” communicates that Anna likes oranges, not apples. Similarly, rising pitch at the end of an utterance can signal a question (“Anna likes oranges?”). Confounding

the listener's task, pitch also varies with the length of a speaker's vocal folds (Titze, 1989), such that the highest pitch values reached by some low voices are still lower than the lowest of a higher-pitched voice.

Lesion and neuroimaging studies have implicated bilateral frontal and temporal regions in the perception of speech intonation (Ross ED, 1981; Heilman et al., 1984; Pell and Baum, 1997; Meyer et al., 2002, 2004; Plante et al., 2002; Doherty et al., 2004; Friederici and Alter, 2004; Gandour et al., 2004; Tong et al., 2005; Witteman et al., 2011; Kreitewolf et al., 2014). Human neuroimaging and primate electrophysiology have also suggested the existence of a putative general pitch center in lateral Heschl's gyrus (HG) and adjacent superior temporal gyrus (STG) (Griffiths et al., 1998; Patterson et al., 2002; Penagos et al., 2004; Bendor and Wang, 2005; Schönwiesner and Zatorre, 2008). However, a more fundamental question than anatomical localization is what the neural activity in those regions encodes—that is, the precise mapping between specific stimulus features and neural responses. How is intonational pitch in speech encoded and does its representation contain concurrent information about what is being said and by whom? Furthermore, because the same auditory feature of pitch is the primary cue to both intonation and speaker identity (van Dommelen, 1990), how does the auditory cortex represent both kinds of information?

Results

We designed and synthesized a controlled set of spoken sentences that independently varied intonation contour, phonetic content, and speaker identity (Fig. 2.1A). The four intonation conditions, Neutral, Emphasis 1, Emphasis 3, and Question, are linguistically distinct (Ladd, 2008) (Fig. 2.1B). By systematically manipulating the pitch (fundamental frequency: f_0) contour of each

token, we ensured that only pitch, not intensity or duration of segments, differed between intonation conditions. The three speakers consisted of one synthesized male (f_0 : 83 Hz $\pm$ 10 Hz) and two synthesized female speakers (f_0 : 187 Hz $\pm$ 23 Hz). The two female speakers had the same f_0 , but differing formant frequencies, one of which matched the male speaker's formant frequencies.

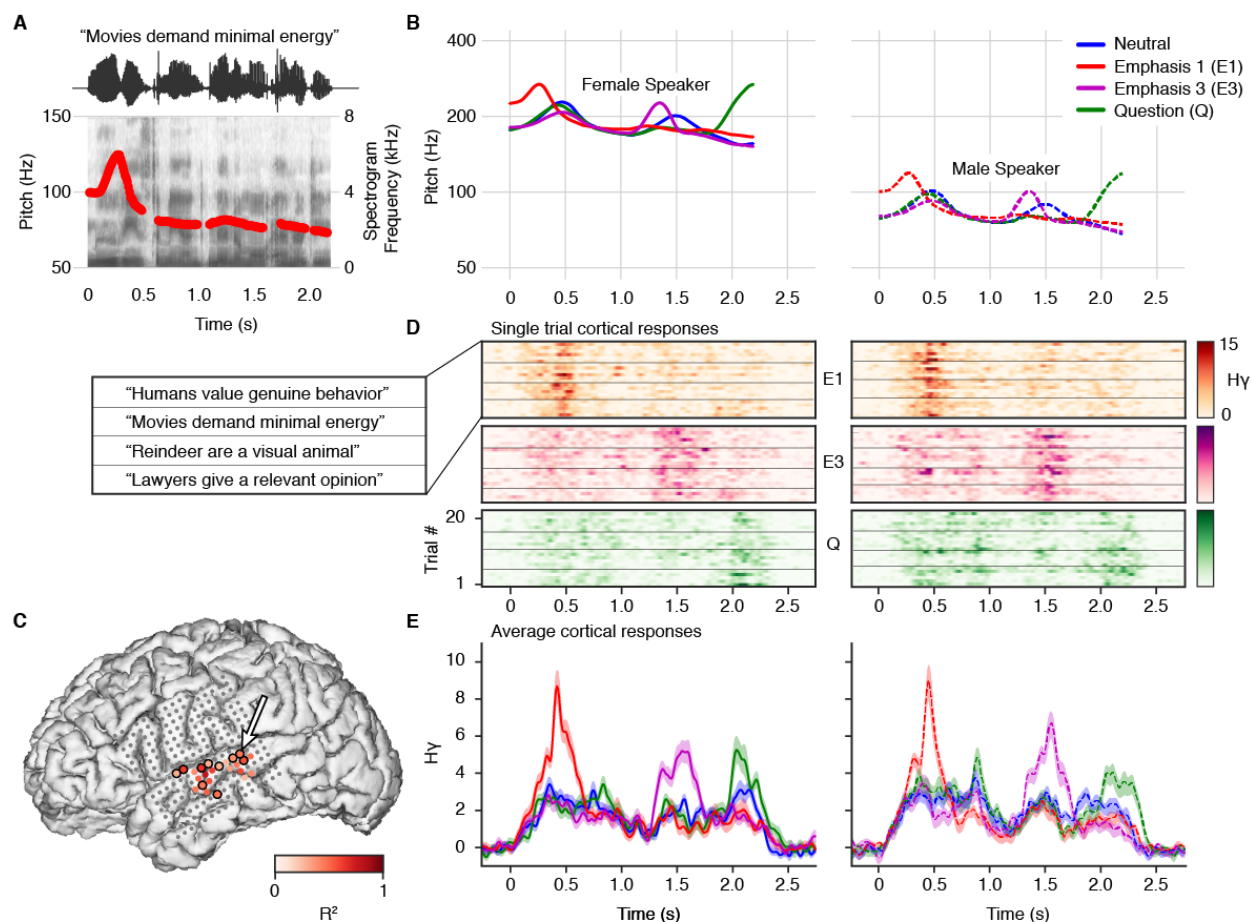


Figure 2.1. Neural activity in STG differentiates intonational pitch contours. (A) Stimuli consisted of spoken sentences synthesized to have different intonation contours. Example token with pitch accent on the first word (Emphasis 1) with amplitude signal, spectrogram, and pitch (f_0) contour shown. (B) Pitch contours for four intonation conditions, shown for a female speaker (left) and male speaker (right). (C) Electrode locations on a participant's brain. Maximum variance in neural activity explained by intonation, sentence, and speaker on electrodes where the full model was significant at more than two time points (omnibus F -test; $p < 0.05$, Bonferroni corrected). Non-significant electrodes are shown in gray. Circled electrodes had a significant (F -test, $p < 0.05$, Bonferroni corrected) main effect of intonation. Activity from the indicated electrode is shown in D and E (arrow). (D) Single trial responses from the indicated

electrode in C, divided by intonation condition (top, middle, bottom) and speaker (left, right). Horizontal lines within each intonation and speaker pair further divide trials by sentence (legend, left). (E) Average neural activity within each intonation condition. Average responses (± 1 sem) to a female (left) and male speaker (right) with non-overlapping absolute pitch values (B).

Participants ($N = 10$) passively listened to these stimuli while we recorded cortical activity from subdurally implanted, high-density grids (placed for clinical localization of refractory seizures). We examined the analytic amplitude of the high-gamma (70-150 Hz) band of the local field potential, which correlates with local neuronal spiking (Crone et al., 2001; Steinschneider et al., 2008; Edwards et al., 2009; Ray and Maunsell, 2011). We aligned the high-gamma responses to sentence onset and used time-dependent general linear models to determine whether and how neural responses on each electrode systematically depended on stimulus conditions. The fully specified encoding model included categorical variables for intonation, sentence, and speaker condition as well as terms for all pairwise interactions and the three-way interaction. Figure 2.1C shows the maximum variance explained in the neural activity for significant electrodes (defined as electrodes where the full model reached significance at >2 time points; omnibus F -test, $p < 0.05$, Bonferroni corrected) in one subject. We next found the electrodes whose activity differentiated intonation contours (F -test, $p < 0.05$, Bonferroni corrected) (circled electrodes in Fig. 2.1C). For one electrode on STG, single trial activity increased after the pitch accents in the Emphasis 1 and Emphasis 3 conditions and after the pitch rise in the Question condition (Fig. 2.1D). The same pattern of activity by intonation contour was seen for each sentence in the stimulus set (Fig. 2.1D) and for the two formant-matched speakers whose absolute vocal pitch values did not overlap (Fig. 2.1E).

We next calculated the unique contribution of the main effect of each stimulus dimension as well as their interactions to variance explained in the neural activity at each significant electrode. Some electrodes showed differences between intonation conditions, but not sentence or speaker

conditions (Fig. 2.2A, D, G) (max unique $R^2_{\text{intonation}} = 0.69$, $p = 1 \times 10^{-49}$, Fig. 2.2J). These electrodes would respond similarly to the sentence “Movies demand minimal energy” and “Humans value genuine behavior” if they were presented with the same intonation contour (e.g. emphasis on the first word), despite very different phonetic content. Other electrodes showed differences between sentence conditions, but not intonation or speaker conditions (Fig. 2.2B, E, H) (max unique $R^2_{\text{sentence}} = 0.85$, $p = 1 \times 10^{-73}$, Fig. 2.2K). In these electrodes, the response to a sentence was the same regardless of whether it was said neutrally, with emphasis, or as a question, but the responses strongly differed for a sentence with different phonetic content. Other electrodes differentiated between speaker conditions, but not intonation or sentence conditions (Fig. 2.2C, F, I) (max unique $R^2_{\text{speaker}} = 0.67$, $p = 1 \times 10^{-47}$, Fig. 2.2L). These electrodes mainly distinguished between the male speaker versus the two female speakers (15/16 electrodes; 1/16 differentiated between high and low formants). The anatomical distribution of encoding effects is shown as pie charts on the cortical surface indicating the proportion of variance explained (Fig. 2.2M, 2.5). Some intonation encoding sites were adjacent to lateral HG on STG, but others were found throughout the STG, interspersed with other electrodes that encoded phoneme- and speaker-related information.

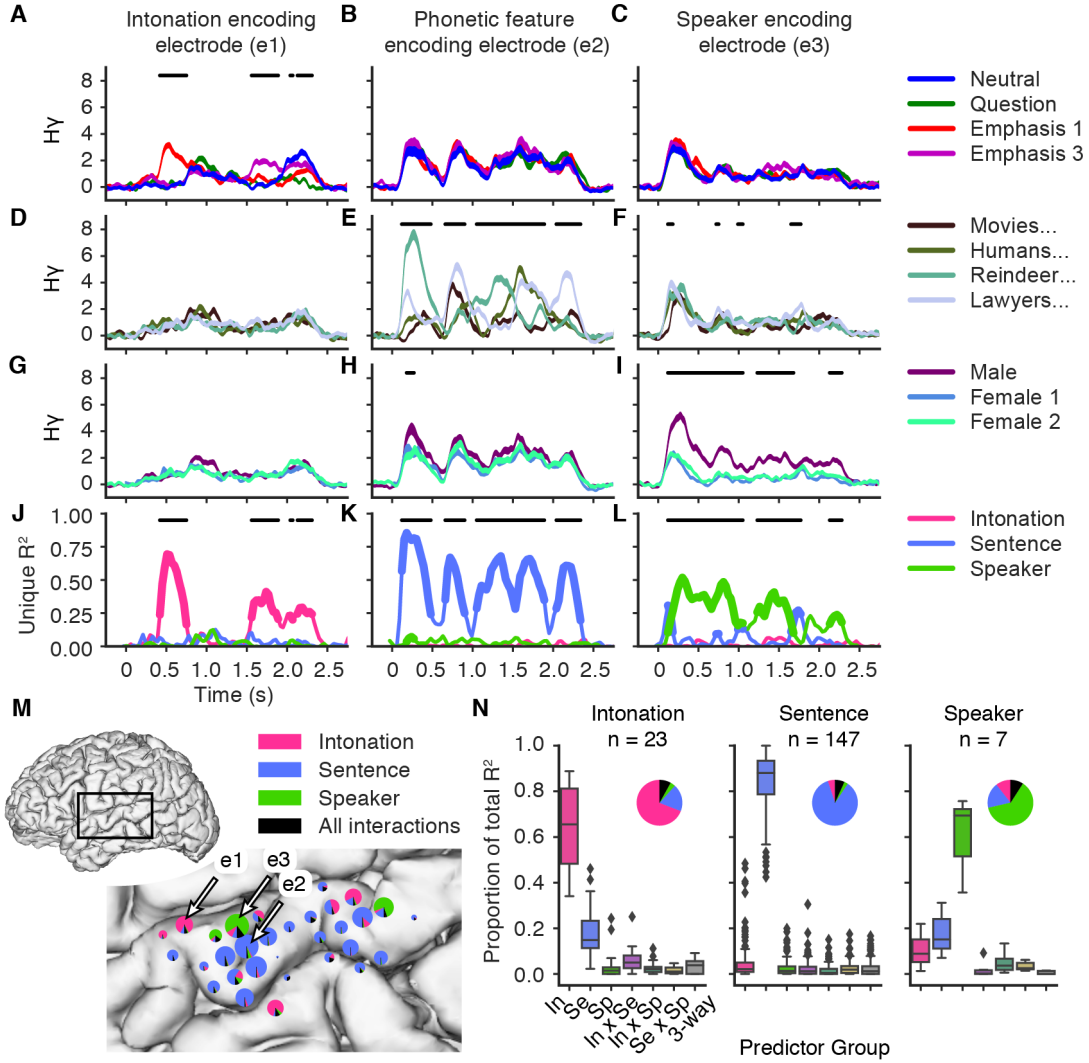


Figure 2.2. Independent neural encoding of intonation, sentence, and speaker information at single electrodes. (A-C) Neural response averaged over intonation contour for three example electrodes (mean $\pm$ 1 sem). Neural activity on electrode 1 (A) differentiates intonation contours, while activity on electrode 2 (B) and 3 (C) do not. Black lines indicate time points where means were significantly different between intonation conditions (F -test, $p < 0.05$, Bonferroni corrected). (D-F) Average neural response to each sentence condition for same electrodes as in A-C. Black lines indicate significant differences between sentence conditions. (G-I) Average neural response to each speaker for the same electrodes as in A-C and D-F. Black lines indicate significant differences between speaker conditions. (J-L) Unique variance explained by main effects for each example electrode. Bold lines indicate time points of significance for each main effect. Black lines indicate time points where the full model was significant (omnibus F -test; $p < 0.05$, Bonferroni corrected). (M) Map of intonation, sentence, and speaker encoding for one subject. Locations of electrodes 1, 2, and 3 are indicated. The area of the pie chart is proportional to the total variance explained. Wedges show the relative variance explained by each stimulus dimension (color) or for pairwise and three-way interactions (black) for each significant electrode. (N) Proportion of variance explained by main effects and interactions across time points where the full model was significant for all significant electrodes

across all ten participants with each electrode classified as either Intonation (In), Sentence (Se), or Speaker (Sp) based on which stimulus dimension was maximally encoded. Pie charts show the average proportions of the total variance explained.

We assigned each electrode into one of three categories, Intonation, Sentence, or Speaker, and then examined the proportion of variance explained by each group of predictors (Fig. 2.2N). The contributions of interactions were minimal (median total proportion of variance explained: 6.4%) (Fig. 2.2N), indicating orthogonal encoding of each stimulus dimension.

Based on previous work (Mesgarani et al., 2014), we hypothesized that electrodes whose activity differentiated between sentence conditions responded to particular classes of phonetic features. We therefore calculated the average phoneme selectivity index (PSI) (Mesgarani et al., 2014) for each significant electrode from its responses to a separate, phonetically-transcribed speech corpus (TIMIT) (Garofalo et al., 1993) (Fig. 2.6A, B), and correlated it with the maximum unique variance explained by each main effect (Fig. 2.6C, D). Average PSI and R^2_{Sentence} were positively correlated ($r = 0.64$, $p < 1 \times 10^{-20}$, Fig. 2.6C), while average PSI was negatively correlated with both $R^2_{\text{Intonation}}$ and R^2_{Speaker} ($r = -0.18$, $p < 0.05$; $r = -0.15$, $p > 0.05$, respectively, Fig. 2.6D). Therefore, Sentence electrode activity could be explained by the specific phonetic features in each stimulus token (Fig. 2.6E).

The phonetically-invariant representation of intonation suggests that intonation is encoded as an isolated pitch contour, irrespective of any lexical information or phonetic content. We thus created a set of non-speech stimuli that preserved intonational pitch contours but did not contain spectral information related to phonetic features (Sonntag and Portele, 1998) (Fig. 2.3A, B). To test that responses are due to the psychoacoustic attribute of pitch, rather than acoustic energy at the fundamental frequency, we also created a set of missing fundamental stimuli (Licklider, 1954; Plack et al., 2006) (Fig. 2.3C). Neural responses to intonation contours were similar between

speech and non-speech contexts, including the missing f_0 context (Fig. 2.3D). To quantify this similarity, we used linear discriminant analysis to fit a model to predict intonation condition from neural responses to speech and then tested this model on responses to the two types of non-speech stimuli (Fig. 2.3E). Model performance on the non-speech responses was as good as model performance on speech responses in almost all cases (with f_0 : 97%, 117/121 electrodes; missing f_0 : 96%, 47/49 electrodes) (Fig. 3F).

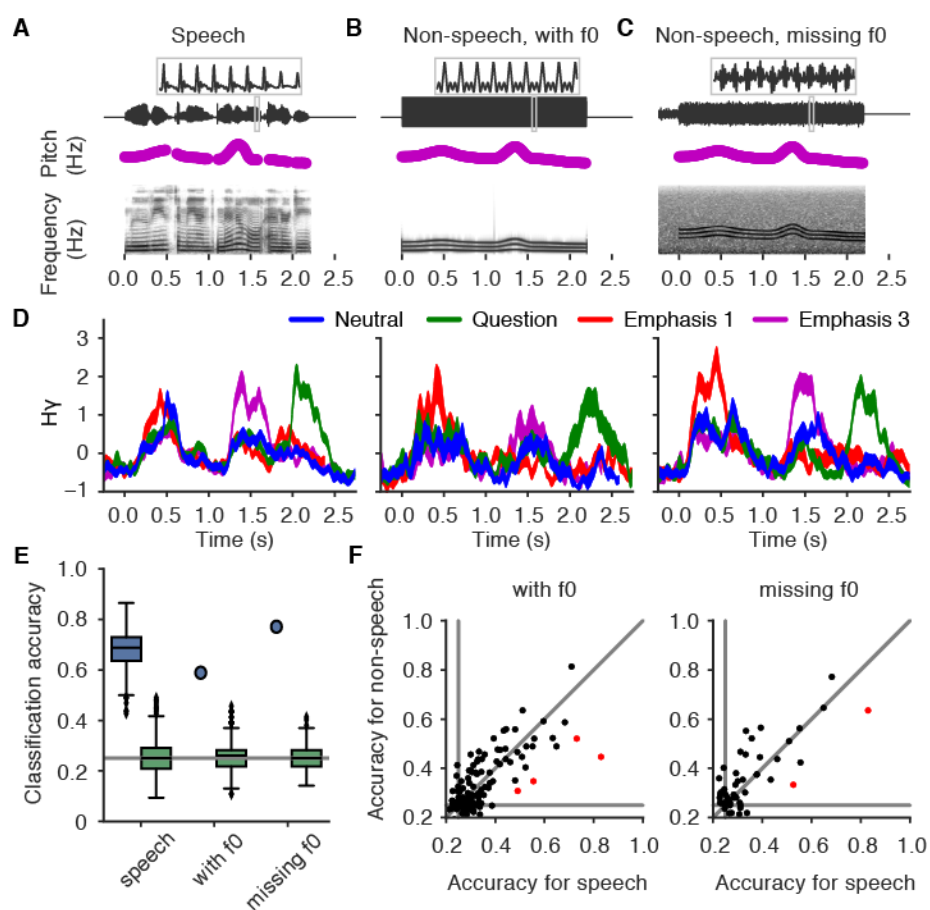


Figure 2.3. Similar neural responses to intonation in speech and non-speech contexts. (A) Acoustic signal, pitch contour, and spectrogram of an example speech token. A portion of the acoustic signal is expanded to show quasi-periodic amplitude variation characteristic of speech. (B) Non-speech token containing energy at the fundamental frequency (f_0) with pitch contour matching A. Three bands of spectral power can be seen at the fundamental, second harmonic, and third harmonic. (C) Non-speech token with same pitch contour as A and B that does not contain f_0 . Pink noise was added from 0.25 s before the onset of the pitch contour to pitch

contour offset. (D) Average neural response by intonation contour to speech (left), non-speech, with f0 (middle), and non-speech, missing f0 (right) stimuli at an example electrode (mean $\pm$ 1 sem). (E) Classification accuracy of an LDA model fit on neural responses to speech stimuli to predict intonation condition for the electrode in D (blue, shuffled: green). The accuracy of the speech trained model on the non-speech data, both with f0 and missing f0, was within the middle 95% of accuracies for speech stimuli. (F) Mean accuracy for speech stimuli vs. accuracy for non-speech stimuli (left: with f0, right: missing f0). Each marker represents a significant electrode from participants who listened to each type of non-speech stimuli (with f0: $N = 8$, missing f0: $N = 3$). Red markers indicate electrodes whose model performance on non-speech stimuli was below the middle 95% of accuracy values from speech stimuli. Gray lines indicate chance performance at 25% and the unity line.

The pitch contour of an utterance can be described in either absolute or relative terms (Fig. 2.4A-C). While absolute pitch is the primary acoustic feature for speaker identification (van Dommelen, 1990), behavioral evidence that listeners perceptually normalize pitch by speaker (Gussenhoven et al., 1997; Wong and Diehl, 2003) suggests the existence of a relative pitch representation in the brain. Indeed, for electrodes discriminating intonation contours, responses to a low pitched male voice and a high pitched female voice were statistically identical (see Fig. 2.1E), so it is unlikely that the amount of neural activity was directly related to the absolute pitch value.

To test the hypothesis that neural activity differentiating intonation contours can be explained by relative pitch, but not absolute pitch, we presented tokens from the TIMIT speech corpus containing sentences spoken by hundreds of male and female speakers (Fig. 2.4A, 2.7A-C) (Garofalo et al., 1993) to participants. We then compared encoding models (Theunissen et al., 2001) containing absolute pitch (Fig. 2.4B), relative pitch (Fig. 2.4C), or both to compute the unique variance (R^2) explained by absolute and relative pitch features at each electrode. We also used these models to predict neural responses to the original set of synthesized intonation stimuli to compare the prediction performance between absolute and relative pitch models.

Figure 2.4D shows the absolute and relative pitch receptive fields for one example electrode that had a significant increase in R^2 when relative, but not absolute pitch features were included in the model (permutation test, see Methods). This electrode was tuned to high relative pitch, but did not respond differentially to different absolute pitch levels. Other relative pitch encoding electrodes were tuned for low relative pitch (Fig. 2.8A-C) or for both high and low relative pitch at different delays (Fig. 2.8D, E), indicating an increased response to pitch movement. Across absolute pitch encoding electrodes, some were tuned to high absolute pitch, while others were tuned to low absolute pitch (Fig. 2.9).

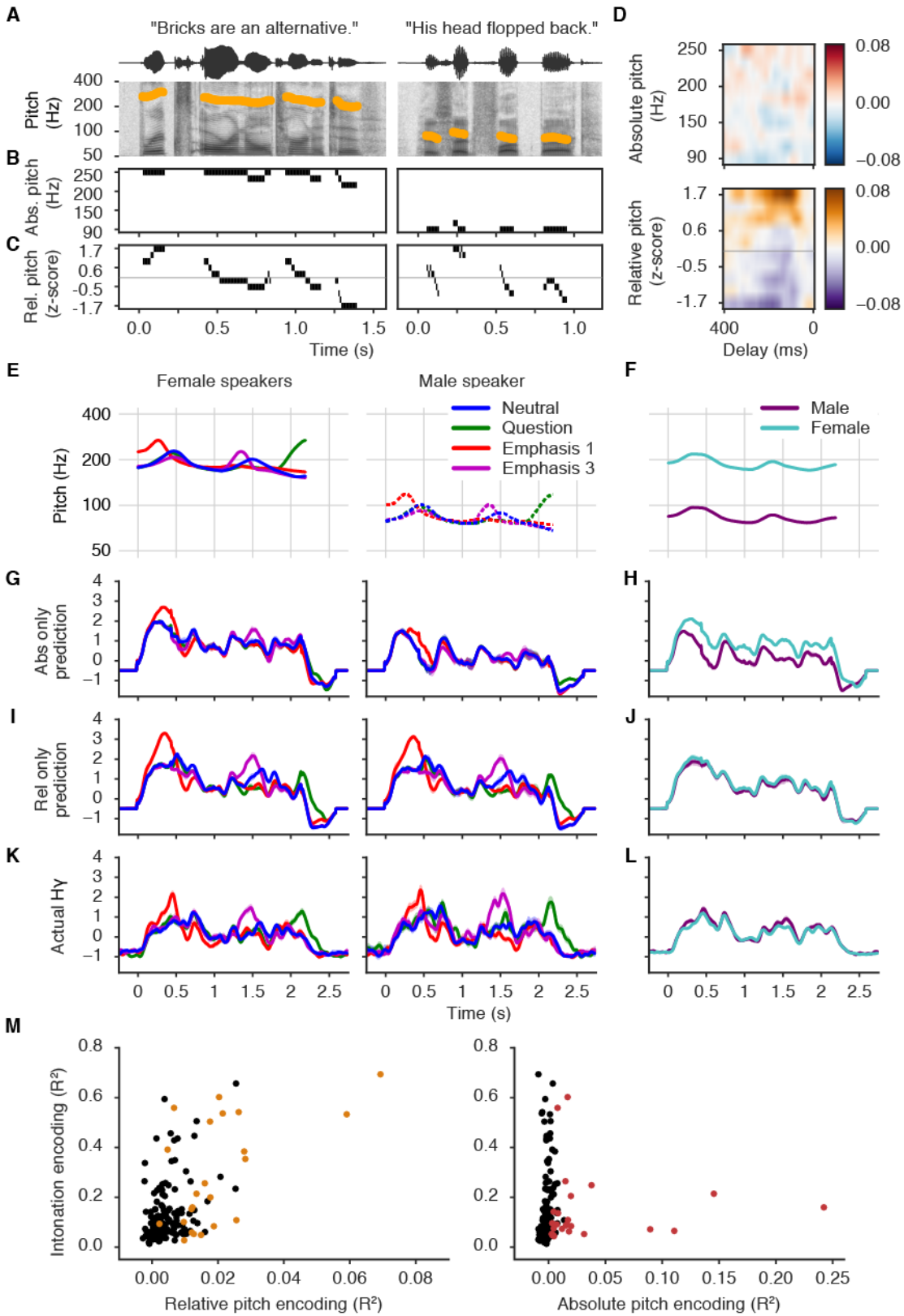


Figure 2.4. Cortical representation of intonation relies on relative, not absolute, pitch encoding. (A) Example tokens from the TIMIT speech corpus. (B) Absolute pitch (ln Hz) feature representation. Bins represent different values of absolute pitch. (C) Relative pitch (z-score of ln Hz within speaker) feature representation. Gray line indicates relative pitch value of 0. (D) Pitch temporal receptive field from one example electrode that encoded relative, but not absolute pitch ($R^2_{\text{relative}} = 0.03$, significant by permutation test; $R^2_{\text{absolute}} = 0.00$, not significant). The receptive field shows which stimulus features drive an increase in the neural response, in this case, high values of relative pitch. Color indicates regression weight (a.u.) (E) Pitch contours of the original stimulus set. (F) Average pitch contours for male and female speakers in the original stimulus set across intonation conditions. (G) Prediction of the model fit with only absolute pitch features. (H) Average predicted response across all male and female tokens from the absolute-pitch-only model. (I) Prediction of the model fit with only relative pitch features. (J) Average predicted response across all male and female tokens from the relative-pitch-only model. (K) Actual neural responses to original stimulus set (mean $\pm$ 1 sem). The actual response of this electrode was better predicted by the relative pitch only model ($r_{\text{rel_pred}} = 0.85$; $r_{\text{abs_pred}} = 0.66$). (L) Actual neural responses averaged over intonation conditions. (M) Scatterplot between relative and absolute pitch encoding with neural discriminability of intonation contours, showing that intonation contour discriminability is correlated with relative, but not absolute, pitch encoding ($r_{\text{relative_intonation}} = 0.57$, $p < 1 \times 10^{-16}$; $r_{\text{absolute_intonation}} = 0.03$, $p > 0.05$). Colored markers show electrodes with significant (permutation test; $R^2 > 95^{\text{th}}$ percentile of null distribution) relative pitch and absolute pitch encoding for the left and right panels, respectively.

We next determined which pitch features (absolute or relative) better predicted the neural responses to the original stimulus set (Fig. 2.4E, F). For the electrode shown in Fig. 2.4D, the absolute pitch only model predicted that the pattern of responses to different intonation contours differed for the female and male speaker (Fig. 2.4G), with a greater response to the female speakers over the male speaker (Fig. 2.4H). Conversely, the relative pitch only model predicted similar responses for the female and male speaker (Fig. 2.4I, J). The actual neural response to these stimuli is shown in Fig. 2.4K (Figure 2.4L shows an additional view of the actual responses averaged over intonation conditions), and was more similar to the prediction from the relative pitch only model than the prediction from the absolute pitch only model ($r_{\text{rel_pred}} = 0.85$; $r_{\text{abs_pred}} = 0.66$). Responses of 84% of intonation electrodes (38/45 electrodes) were better predicted by relative pitch. In addition, relative pitch encoding predicted neural discriminability of intonation contours ($r = 0.57$,

$p < 1 \times 10^{-16}$), while absolute pitch encoding did not ($r = 0.03$, $p = 0.67$) (Fig. 4M).

We demonstrated direct evidence for the concurrent extraction of multiple, socially and linguistically-relevant dimensions of speech information at the level of human non-primary, high-order auditory cortex in the STG. Our results are consistent with the idea that the main types of voice information, including speech and speaker identity, are processed in dissociable pathways (Belin et al., 2004, 2011).

Discussion

The importance of relative pitch to linguistic prosody is well established because vocal pitch ranges differ across individual speakers (Jakobson et al., 1951; Ladd, 2008). Additionally, in music, melodies can be recognized even when the notes are transposed. The representation of relative auditory features may, in fact, be a general property of the auditory system because contours can be recognized in multiple auditory dimensions, such as loudness and timbre (Kluender et al., 2003; McDermott et al., 2008). fMRI BOLD activity increases in non-primary areas of human lateral HG, planum polare, and anterolateral planum temporale for harmonic tone complexes that change in pitch height and pitch chroma (Warren et al., 2003), where activity depends on the variability of pitch height changes (Zatorre et al., 2012; Allen et al., 2016). However, in addition to relative pitch encoding, we also found coexisting absolute pitch encoding in STG consistent with reports that differences in cortical activation for different speakers are correlated with differences in fundamental frequency (Formisano et al., 2008).

In animal model studies, spectral and temporal features important for pitch are encoded at many levels of the auditory system from the auditory nerve (Cariani and Delgutte, 1996a, 1996b) to primary auditory cortex (Fishman et al., 2013). Single neurons can encode information about

pitch by systematically varying their firing rate to sounds with different pitch (Bizley et al., 2009, 2010), and some respond similarly to pure tone and missing fundamental tones with matched pitch (Bendor and Wang, 2005). Multiunit activity containing information about whether a target sound was higher or lower in pitch than a previous sound may play a role in relative pitch processing (Bizley et al., 2013). However, a direct neural encoding of relative pitch or its dissociation from sites encoding absolute pitch has not been previously demonstrated.

Perceptual studies have demonstrated that speaker normalization for pitch occurs in the sentence context (Pierrehumbert, 1979; Wong and Diehl, 2003), and can also occur as rapidly as within the first six glottal periods (~20-50 ms) (Lee, 2009). We have demonstrated how intonational pitch undergoes specialized extraction from the speech signal, separate from other important elements, such as the phonemes themselves. An outstanding future question is how such components are then integrated to support a unified, meaningful percept for language comprehension.

Materials and Methods

The experimental protocol was approved by the Institutional Review Board at the University of California, San Francisco. Participants gave their written, informed consent before testing.

Participants

Ten individuals, with self-reported normal hearing, participated in this study. Each participant was a neurosurgical patient with intractable epilepsy who had a high-density electrode grid implanted subdurally for clinical monitoring of seizure activity. The placement of

the grid was determined solely by clinical needs, typically covering the lateral cortical surface. 6 subjects had left-hemisphere grids and 4 subjects had right-hemisphere grids. Electrodes were localized by aligning preimplantation MRI and post-implantation CT scans.

Data acquisition and neural signal processing

During experimental tasks, neural signals were recorded from the 256 channel (16 x 16, 4 mm spacing) ECoG grid (or from two 128 channel ECoG grids, 8 x 16, 4 mm spacing) using a multichannel amplifier optically connected to a digital signal processor. The local field potential at each electrode contact was amplified and sampled at 3052Hz. The raw waveform was visually examined, and channels containing continuous epileptiform activity or signal variation too low to be detectable from noise were removed. Time segments on remaining channels that contained electrical/movement-related artifacts or discrete epileptiform activity were manually marked and excluded. The signal was common average referenced and notch-filtered at 60 Hz, 120 Hz, and 180 Hz to remove line noise. Using the Hilbert transform, the analytic amplitude of eight Gaussian filters (center frequencies: 70-150 Hz) was computed. The high-gamma signal was taken as the average analytic amplitude across these eight bands. This signal was downsampled to 100 Hz and z-scored either to a silent baseline or across the entire recording block ($H\gamma$). For each token, we analyzed the neural data in the window from 150 ms before stimulus onset to 650 ms after stimulus offset.

Stimulus design

Stimuli consisted of spoken sentences synthesized to have four linguistically distinct intonational contours. These contours, depicted in Fig. 1b, were Neutral, Emphasis 1, Emphasis

3, and Question. In the Neutral condition, the less variable pitch contour and falling pitch at the end of the sentence do not impart additional meaning beyond the declarative meaning of the words. In the Emphasis 1 condition, a pitch accent (rising followed by falling pitch) on the first word and low pitch throughout the rest of the sentence indicates the first word as the focus of the sentence, whereas the pitch accent on the third word in Emphasis 3 indicates that the third word was emphasized. Finally, rising pitch through the last word in the Question condition signals that the utterance was interrogative.

We applied these intonation contours to four declarative sentences (“**Humans** value **genuine** behavior”, “**Movies** demand **minimal** energy”, “**Lawyers** give a **relevant** opinion”, and “**Reindeer** are a **visual** animal”) constructed such that rising intonation at the end of the utterance would signal a question and a pitch accent could be added to either of the bolded words (referred to as the first word and third word even though in the latter two sentences the second bolded word is the fourth word). In order to precisely align the pitch contours across sentences, the sentences were designed to contain the same number of syllables.

Each sentence and intonation combination was recorded from a native female speaker of standard American English. To create base tokens, we duration-matched the syllables of the Neutral sentence recordings and equalized the average root-mean-square intensity across sentences. The total duration of each sentence was 2.2 seconds. To create the four intonation contours, the average pitch trace of recorded Neutral, Question, Emphasis 1, and Emphasis 3 sentences was taken and smoothed. We then applied each intonation contour to each base token using the pitch-synchronous-overlap-add (PSOLA) method (Moulines and Charpentier, 1990).

Finally, we manipulated baseline pitch and formant values to create three speakers (two female, one male). In general, voices from different speakers are perceived on two main acoustic

dimensions, fundamental frequency (f_0) and formant frequencies, based on the length of the vocal folds and shape of the vocal tract, respectively (Fant, 1971; Baumann and Belin, 2010). To model these two dimensions, the three speakers consisted of a low-pitch, low-formant male speaker (median pitch = 80 Hz, formants lowered by 15% of original recording), a high-pitch, low-formant female speaker (median pitch = 180 Hz, formants lowered by 15% of original recording), and a high-pitch, high-formant female speaker (median pitch = 180 Hz, formants lowered by 5% of original recording). The two female speakers had the same f_0 , but differing baseline formant frequencies, one of which matched the male speaker's baseline formant frequencies. Baseline pitch values were manipulated using PSOLA and baseline formants were manipulated by shifting the entire sound spectrum while maintaining duration and fundamental frequency. All speech synthesis was done using the linguistics software, PRAAT (Boersma, 2002).

The resulting stimulus set consisted of 4 intonation contours x 4 sentences x 3 speakers = 48 tokens (Audio S1-48). The tokens were each played twice in one recording block in random order. The total experimental time for each block was about 5 minutes. We collected 2-4 blocks per participant (mean 2.7 blocks).

Data Analysis

All analyses were carried out using custom software written in MATLAB and Python. Open-source scientific Python packages used included numpy, scipy, pandas, scikit-learn, and statsmodels. Figures were created using matplotlib and seaborn. The code used to analyze the data and produce the figures is publicly available on Github at

<https://github.com/ChangLabUcsf/intonatang>. The accompanying documentation can be found at <https://changlabucsf.github.io/intonatang>.

Raw data, experimental stimuli, and analysis code is accessible at <https://doi.org/10.5281/zenodo.826950>.

Single-electrode encoding analysis

Single-electrode encoding analyses were not restricted anatomically *a priori*. Using ordinary least-squares regression, we fit encoding models that predicted neural activity ($H\gamma$) from stimulus conditions (e.g. Neutral, Question, Emphasis 1, and Emphasis 3 formed the set of intonation conditions). To determine how variance in the neural activity was explained by the intonation, sentence, and speaker conditions, we represented stimulus conditions and interactions with sets of dichotomous predictor variables. This regression analysis is mathematically equivalent to three-way, crossed ANOVA. To reduce timepoint by timepoint variability, we took neural activity as the average of $H\gamma$ in 60 ms windows, moving in 30 ms steps. For each model, the coefficient of determination, R^2 , provides a measure of the proportion of variance in neural activity that is explained by stimulus conditions and interactions. The p -value associated with the omnibus F -statistic provides a measure of significance. We set the significance threshold at $\alpha = 0.05$ and corrected for multiple comparisons using the Bonferroni method, taking individual time points and electrodes as independent samples. The average number of significant electrodes for each participant was 17.7 (min: 5, max: 32).

Variance partitioning and evaluation of interaction terms

The predictor variables were grouped into seven mutually exclusive sets. Three of the seven groups represented the main effects of intonation (In), sentence (Se), and speaker (Sp) condition. An additional three of the seven represented pairwise interactions, intonation $\times$ sentence (InSe), intonation $\times$ speaker (InSp), and sentence $\times$ speaker (SeSp). The last group of predictor variables represented the three-way interaction (InSeSp). For each token with intonation condition i , sentence condition j , and speaker condition k , the high-gamma was modeled as:

$$\begin{aligned} H\gamma_{ijk}(t) = & \beta_0(t) + \beta_{In}(t) \cdot In_i + \beta_{Se}(t) \cdot Se_j + \beta_{Sp}(t) \cdot Sp_k + \beta_{InSe}(t) \cdot InSe_{ij} + \\ & \beta_{InSp}(t) \cdot InSp_{ik} + \beta_{SeSp}(t) \cdot SeSp_{jk} + \beta_{InSeSp}(t) \cdot InSeSp_{ijk} \end{aligned}$$

The contribution of each group of predictor variables, including the groups for interaction terms, was evaluated by comparing the variance explained by the fully specified model with one that excluded the group. The proportion of variance uniquely explained by each group, R^2_G , was calculated as the difference in R^2 between those two models:

$$R^2_G = R^2_{full} - R^2_{wo_G}$$

The significance of each group of predictors was evaluated using the F -test($m, N-k-1$) with the following F -statistic:

$$F_G = \frac{R^2_G}{m} \bigg/ \frac{1 - R^2_{full}}{N - k - 1}$$

where m is the number of predictor variables coding for the group G , k is the number of predictor variables in the fully specified model, and N is the number of trials. By applying a Bonferroni correction on the p -value, this method takes a conservative stance on finding significant values.

De-lexicalized non-speech control stimuli

To determine whether the cortical representation of intonational pitch contours is independent from the processing of phonetic information, we created a set of non-speech control stimuli that completely removed spectral information related to phonetic features and consisted of only a pitch contour (for 5 of 8 subjects who listened to non-speech control stimuli, these stimuli also had amplitude contours corresponding to the amplitude contours of each sentence condition (Audio S49 – S80); remaining 3 of 8 subjects heard (Audio S81 – S88)). These stimuli were created by summing a sinusoid and its second and third harmonics (Sonntag and Portele, 1998). The varying frequency of the sinusoid was matched to the pitch contour of each intonation condition.

To test whether neural responses to non-speech stimuli were similar in pattern to responses to the original speech stimuli, we used linear discriminant analysis (LDA) to ask whether the pattern of neural activity that differentiates intonation contours in the speech context generalizes to a non-speech context. To do this, we fit the model using neural responses to speech to predict the intonation condition from the neural activity time series from a single electrode (average Hy in 60 ms windows centered at -0.15 s to 2.85 s in 30 ms steps) and then tested the model on non-speech data. We then determined whether model performance on the

non-speech data, measured as classification accuracy, was as good as performance for the speech data.

Specifically, we first computed a distribution of classification accuracies for speech. We trained an LDA classifier on a random 80% of the speech trials and then tested the model on a set with $N_{\text{non-speech}}$ trials bootstrapped from the remaining 20%. To prevent overfitting, we used a form of regularized LDA, diagonal LDA, which uses a diagonal covariance matrix that is shared between classes (i.e. intonation conditions). We performed this procedure 1000 times to arrive at the distribution of accuracies for speech data. We then trained an LDA classifier on 100% of the speech trials and used this model to test the non-speech trials and determined whether this accuracy fell within the 95% of values from the 2.5 to the 97.5 percentile for the speech trials.

Missing fundamental, non-speech control stimuli

To determine whether the neural activity we observed was a response to the psychoacoustic, perceptual attribute of pitch, rather than the physical, acoustic energy at the fundamental frequency (f_0), we created a second type of non-speech control stimuli that did not contain energy at f_0 (30, 31) (Audio S89 – S96). These missing f_0 stimuli contained the fourth, fifth, and sixth harmonics. To mask distortion products that may be introduced at the level of the cochlea (Oxenham and Plack, 1997), we also added pink noise from 0.25s before pitch contour onset through the duration of the stimulus. We presented these stimuli in random order to three participants, while we recorded their cortical activity. We then used the same analysis, described above, to test whether the pattern of neural activity to intonation contours was similar between the speech context and the missing f_0 context. Briefly, we trained an LDA classifier to predict intonation condition from neural responses to speech stimuli. We then tested this classifier on

neural responses to missing f_0 and determined whether classification accuracy was as good for missing f_0 data as it was for speech data (see section above for more details).

Absolute and relative pitch temporal receptive field analysis

To investigate how the cortical representation of intonation reflected the encoding of pitch values, we recorded neural activity as participants listened to a subset of the TIMIT continuous speech corpus (Garofalo et al., 1993). This stimulus set contained sentences spoken by hundreds of male and female speakers allowing for a statistical separation of absolute and relative pitch values (Fig. 2.7). Vocal pitch values were extracted using an automated autocorrelation method and corrected for halving and doubling errors. We defined absolute pitch as the natural logarithm of pitch values in Hz. There are two main methods used to compute relative pitch. One method normalizes values by interpolating each value between a speaker's minimum pitch and maximum pitch (Earle, 1975), while the other uses z-scoring (Rose, 1987). Here, we used the z-score method to compute relative pitch values, and normalized absolute pitch values in $\ln$ Hz by the mean and standard deviation of each sentence as a proxy for speaker. We then discretized absolute and relative pitch values into 10 bins, equally spaced in pitch space from the 2.5 percentile to the 97.5 percentile value (Fig. 2.7H, I). The bottom and top 2.5% of the pitch values were placed into the first and last bins, respectively. By defining these percentile bounds, we prevent unstable estimates that can occur when the top and bottom bins contain too few data points.

To determine how absolute and relative pitch values in speech drive neural activity, we fit temporal receptive field models (Theunissen et al., 2001) that predicted neural activity from pitch values in the immediately preceding 400 ms window (sampled at 100 Hz) using L_2

regularized multiple linear regression. By including both absolute and relative pitch values as features, we could assess the unique contribution of absolute and relative pitch in predicting neural activity. We also included two additional temporal features using a continuous variable with intensity information and a binary variable when pitch values were present. This binary feature allows us to statistically control for the contribution of the presence of pitch (or voicing in the speech signal) when evaluating the contribution of absolute and relative pitch levels. To calculate the unique contribution of absolute and relative pitch, we calculated the R^2 gained when absolute or relative pitch features, respectively, were included in the model.

We used L_2 regularization (ridge regression) and cross-validation to prevent overfitting since the number of features in temporal receptive field models is typically large (>500). We evaluated the models using the correlation between actual and predicted values of neural activity on held on data. Specifically, we divided the data into three mutually exclusive sets containing 80%, 10%, and 10% of the total number of sentences. The first set of 80% was used as the training set. The second set was used to fit the L_2 regularization hyperparameter, and the final tenth was used as the test set. We performed this procedure 25 times and the performance of the model was taken as the mean of performance across all testing sets.

To calculate the significance of the R^2_{absolute} and R^2_{relative} values computed for each electrode, we used a permutation test. We shuffled the pitch and intensity contours between TIMIT sentences before using the same analysis pipeline to compute null values of R^2_{absolute} and R^2_{relative} . We ran this analysis 200 times to arrive at the null distribution. R^2_{absolute} and R^2_{relative} values above the 95th percentile were considered significant.

Predicted neural activity from pitch temporal receptive fields

To determine whether the absolute or relative pitch temporal receptive field models captured stimulus features that were relevant for intonation-encoding electrodes, we used the pitch temporal receptive field (ptrf) models fit using TIMIT speech data to predict neural responses to the original intonation stimuli containing the four intonation conditions. To determine whether absolute or relative pitch better explained the neural responses, we compared the performance of the different ptrf models. We parameterized the original intonation stimuli using bins for absolute and relative pitch determined by the TIMIT data. We then used the absolute ptrf model and the relative ptrf model to predict the neural responses to the intonation stimuli. Since there are no features in the ptrf models associated with the spectral information that determine phonetic information, we averaged predictions and real neural responses over sentence conditions. We additionally averaged predictions and real neural responses over the two female speakers who did not differ in pitch. We then took the correlation (Pearson's r) between the predicted and actual neural responses for the absolute pitch only model ($r_{\text{abs_pred}}$) and relative pitch only model ($r_{\text{rel_pred}}$).

Supplementary Figures

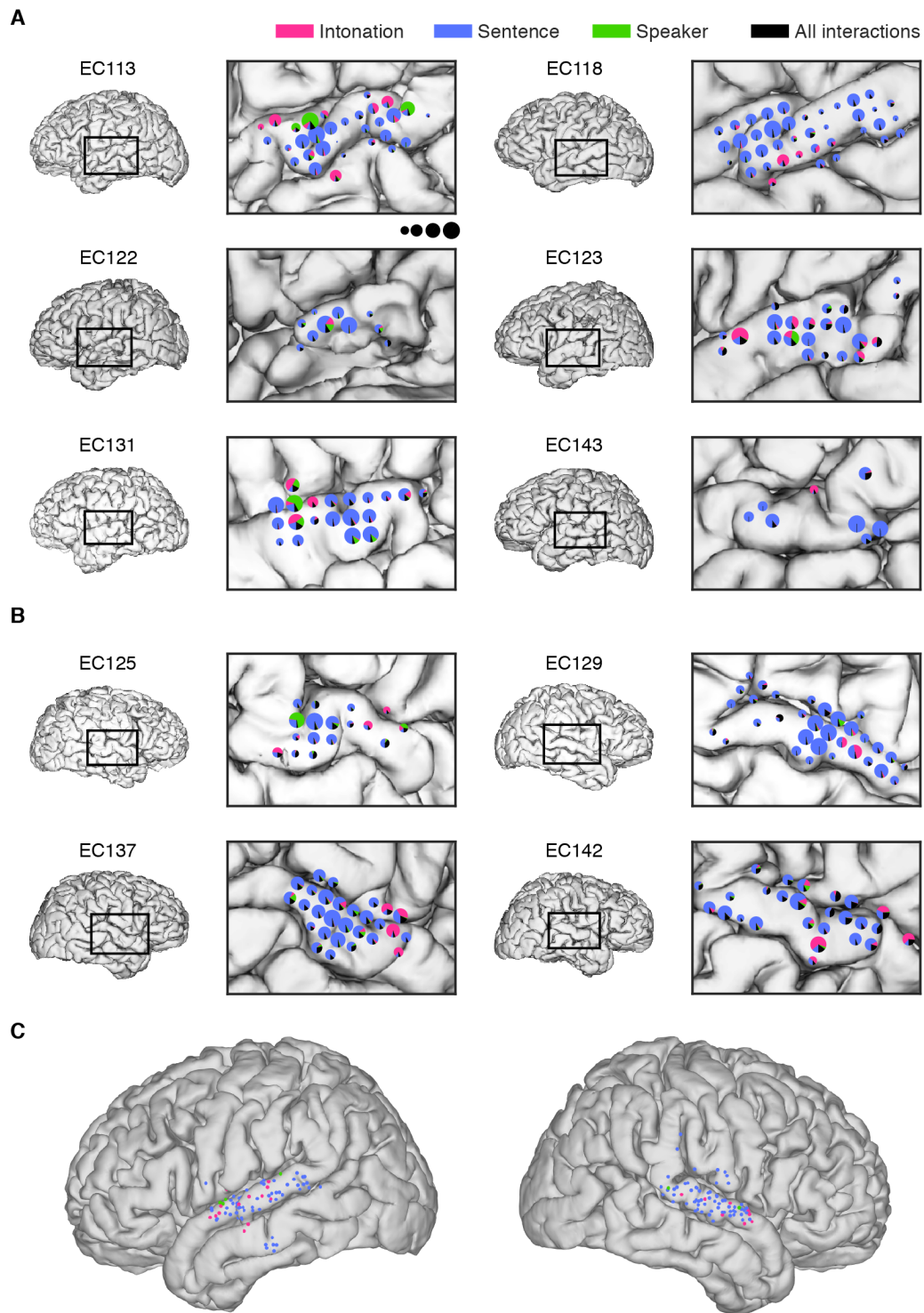


Figure 2.5. Individual and group cortical maps of intonation, sentence, and speaker condition encoding. Individual cortical maps in A and B show the proportion of variance

explained by the main effects of intonation, sentence, and speaker condition, as well as the proportion explained by the sum of all pairwise and three-way interactions as pie charts. All significant electrodes (defined as those where the full encoding model was significant for at least 2 time points at $\alpha = 0.05$, Bonferroni corrected) are shown with the area of the pie chart proportional to the total R^2 of the fully specified model. The black circles represent the areas for 25%, 50%, 75%, and 100% of the maximum total R^2 across all significant electrodes for one subject. (A) Maps for participants with left hemisphere grids. (B) Maps for participants with right hemisphere grids. (C) Group cortical map showing Intonation, Sentence, and Speaker electrodes from all ten subjects warped to a common MNI brain.

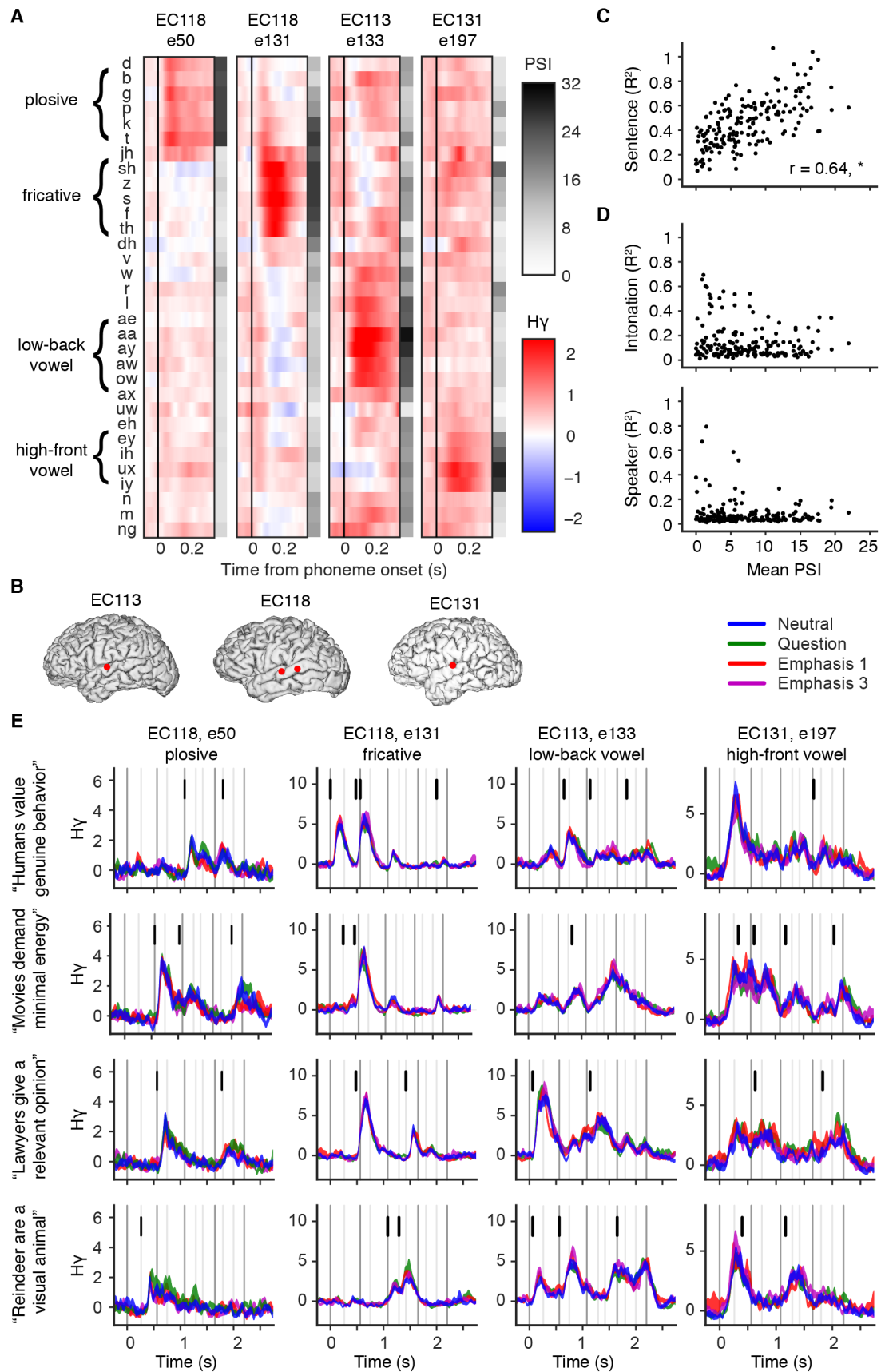


Figure 2.6. Activity that differentiates sentence conditions is driven by phonetic feature selectivity. (A) Average neural responses time-locked to the onsets of individual phonemes from sentences in the TIMIT speech corpus. Each column shows the average response of an individual electrode. The phonetic selectivity index measures whether a response to a given phoneme can be discriminated from the response to all other phonemes. Grouping of phonemes into four phonetic categories is shown to the left. (B) Anatomical location of electrodes shown in A. Each electrode is located in the STG. (C) Scatter plot showing each significant electrode's sentence condition encoding and average phonetic selectivity index ($r = 0.64$, $p\text{-value} < 1 \times 10^{-20}$). Data from 177 significant electrodes across 10 participants are shown. (D) Scatter plots of intonation encoding and average PSI on the top and of speaker encoding and average PSI on the bottom ($r = -0.18$, $p\text{-value} < 0.05$; $r = -0.15$, $p\text{-value} > 0.05$, respectively). (E) Average neural response of each example electrode in A to original stimulus set. Each row shows responses to a different sentence. For each column, tick marks indicate the onsets of phonemes which fall into the class written at the top of the column. The responses are colored by intonation condition. These phonetically-selective, sentence electrodes are not sensitive to intonation and have a similar response regardless of what the intonation condition was.

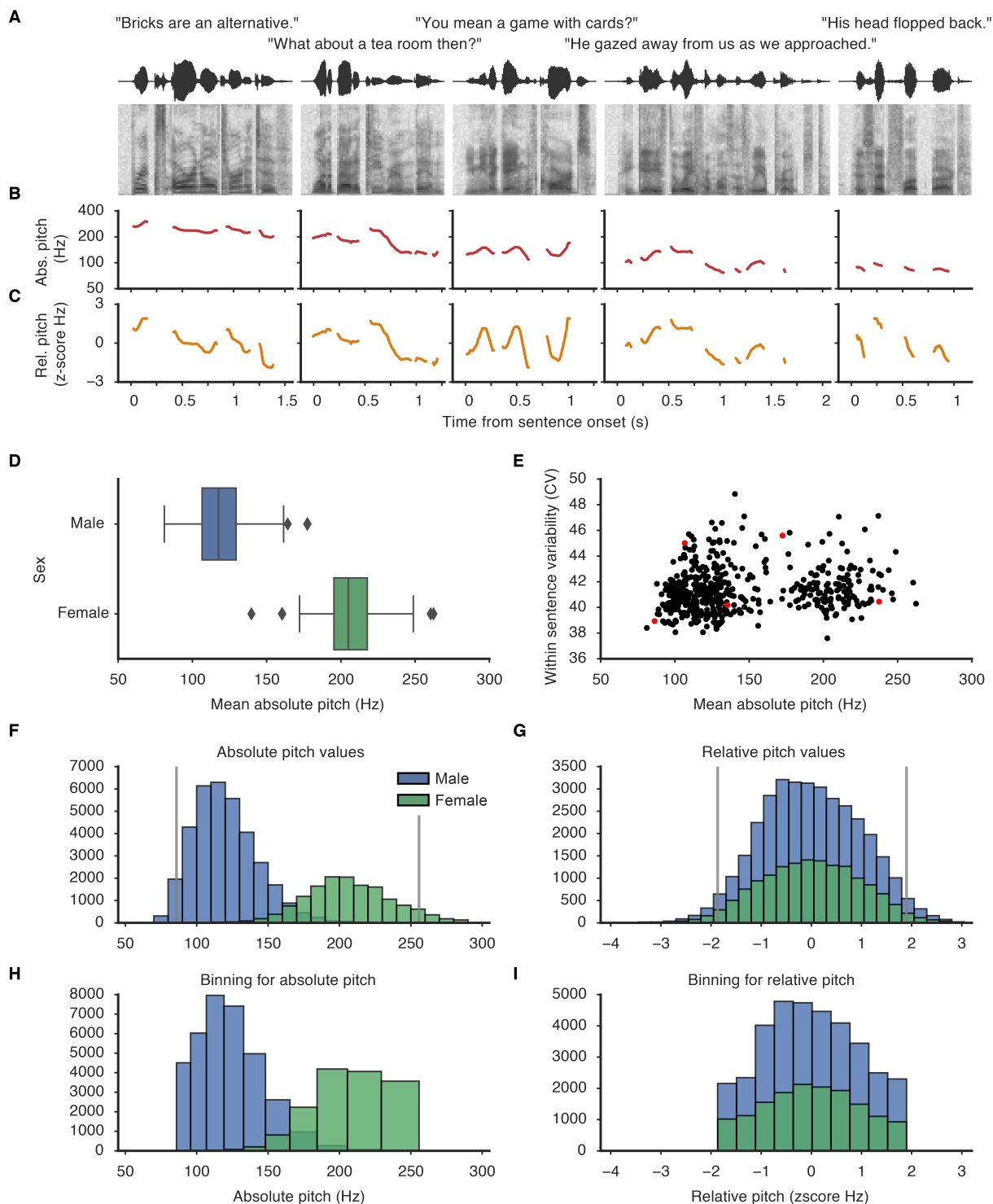


Figure 2.7. Absolute and relative pitch contours and variability in TIMIT speech corpus.

The TIMIT corpus contains continuous speech recorded from hundreds of male and female speakers. Having many speakers who differ in their baseline absolute pitch allows for the statistical separation of absolute and relative pitch. (A) Five example tokens from the TIMIT dataset with their acoustic amplitude signal and spectrograms. The first two were spoken by

female speakers while the last three were spoken by male speakers. (B) Absolute pitch contours for the five tokens in (A). The fundamental frequency was extracted from the speech signal using an autocorrelation method. (C) Relative pitch contours for the five tokens in (A). The relative pitch was calculated as the z-score of absolute pitch values (in ln Hz) for each token. (D) Distribution of mean pitch values for the speakers in the subset of TIMIT used in this study. (E) Scatterplot of mean pitch values and pitch variability (coefficient of variation expressed as percentage of the mean). Each dot represents one token and red dots indicate the five tokens from (A). (F, G) Histogram of all the absolute pitch (F) and relative pitch (G) values calculated from TIMIT tokens, with values from male and female speakers shown separately. Gray lines indicate the 2.5 and 97.5 percentile. (H, I) Histograms showing the binning used to parameterize absolute (H) and relative (I) pitch values for the pitch temporal receptive field models. Ten equally spaced bins (for absolute pitch, bins are equally spaced on a logarithmic scale) were created between the 2.5 and 97.5 percentile of all pitch values.

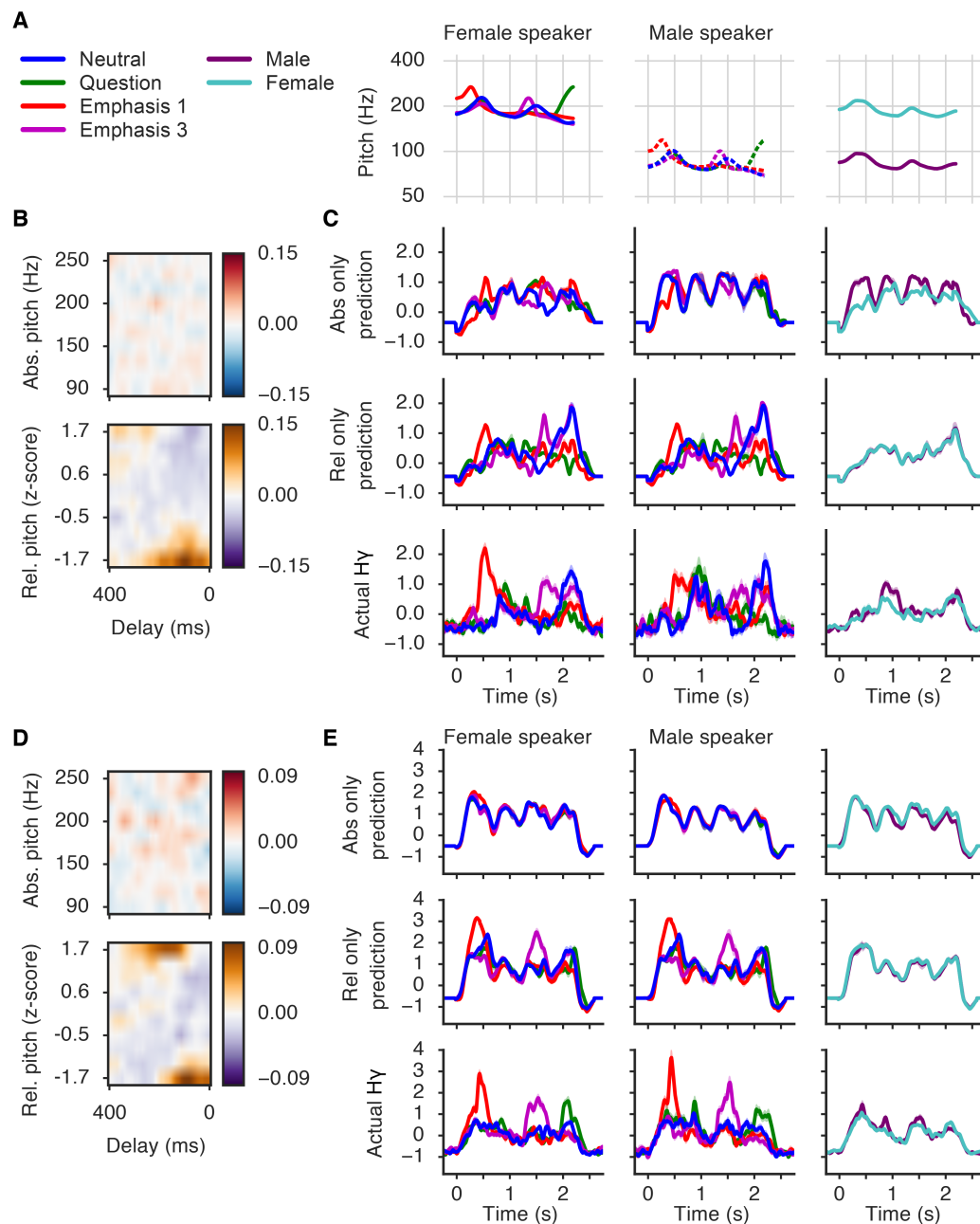


Figure 2.8. Example relative pitch encoding electrodes tuned for low relative pitch and high-to-low relative pitch. The pitch temporal receptive field (ptrf) fit using neural responses to the TIMIT speech corpus and the prediction of this encoding model to the original set of stimuli are shown for two example relative pitch encoding electrodes. The ptrf indicates how different values of absolute and relative pitch at different time delays affects neural activity. (A) Pitch contours of original set of stimuli. The left panel shows the pitch contours for each intonation condition for the female speakers. The middle panel shows the pitch contours for each intonation condition for the male speaker. The right panel shows the average pitch contour for the male versus the female speakers. (B) Electrode that encoded relative pitch ($R^2_{\text{relative}} = 0.07$, significant by permutation test; $R^2_{\text{absolute}} = -0.01$, not significant) and was tuned to low relative pitch. (C)

The top two rows show the predicted neural responses from the absolute pitch only and relative pitch only models. The bottom row shows the actual neural responses. The actual response of this electrode to the original stimulus set was better predicted by the relative pitch only model ($r_{\text{rel_pred}} = 0.76$; $r_{\text{abs_pred}} = 0.55$). (D) Electrode that encoded relative pitch ($R^2_{\text{relative}} = 0.02$, significant by permutation test; $R^2_{\text{absolute}} = -0.01$, not significant) and was tuned to high relative pitch at a delay of ~ 180 ms and low relative pitch at a delay of ~ 100 ms. (E) The activity on this electrode was better predicted by the relative pitch only model than the absolute pitch only one ($r_{\text{rel_pred}} = 0.85$; $r_{\text{abs_pred}} = 0.74$).

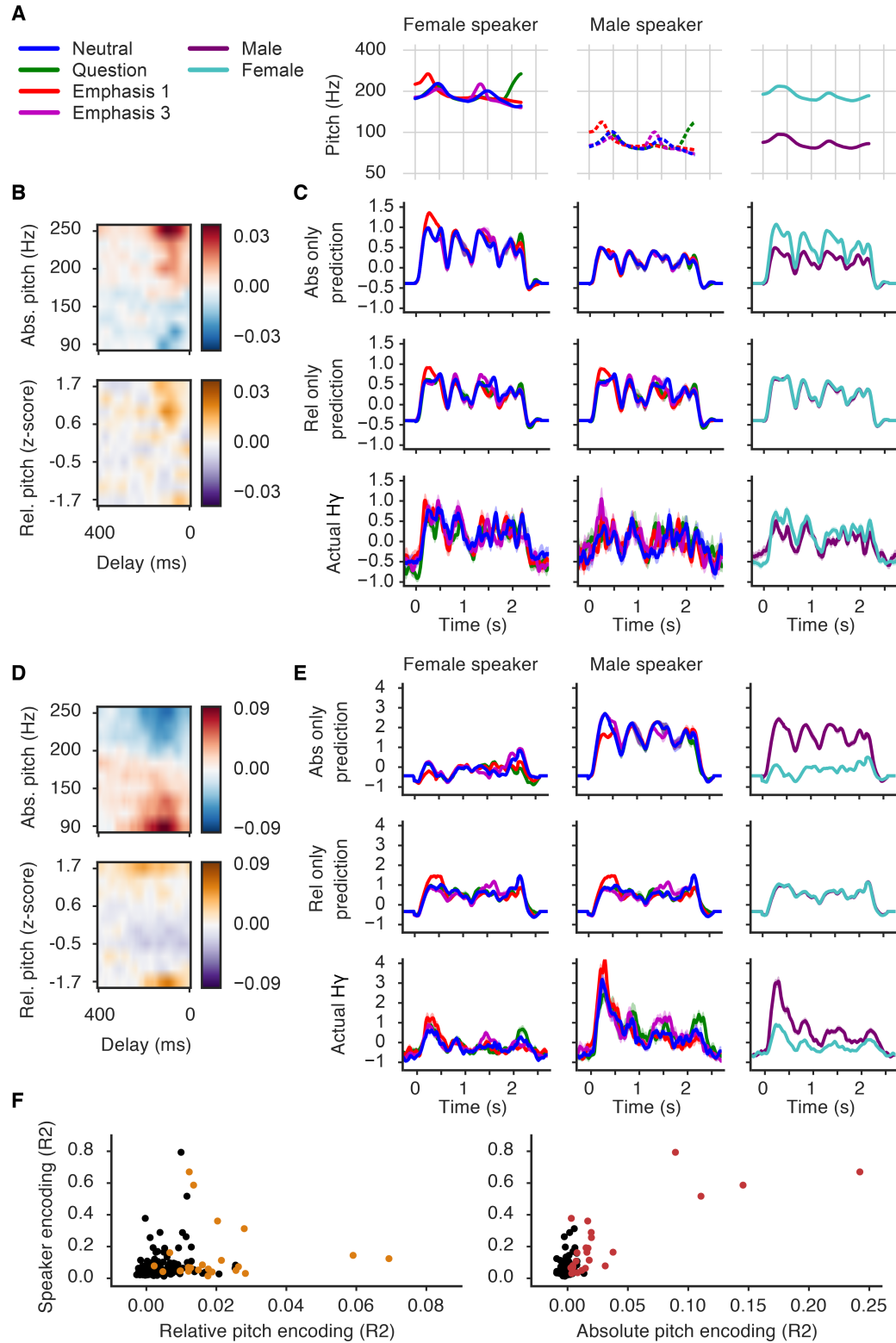


Figure 2.9. Example absolute pitch encoding electrodes tuned for high and low absolute pitch. The pitch temporal receptive field, predicted responses of the ptrf model to the original set

of stimuli, and the actual responses are shown for two example absolute pitch encoding electrodes. The ptrf indicates how different values of absolute and relative pitch at different time delays affects neural activity. (A) Pitch contours of original set of stimuli. (B) Electrode that encoded absolute pitch ($R^2_{\text{relative}} = 0.00$, not significant by permutation test; $R^2_{\text{absolute}} = 0.02$, significant) and was tuned to absolute pitch values greater than 180 Hz. (C) The top two rows show the predicted neural responses from the absolute pitch only and relative pitch only models. The bottom row shows the actual neural responses. This electrode had a greater response to the female speakers than the male speaker. The actual response of this electrode was better predicted by the absolute pitch only model ($r_{\text{rel_pred}} = 0.75$; $r_{\text{abs_pred}} = 0.78$). (D, E) Electrode that encoded absolute pitch and was tuned to low absolute pitch ($R^2_{\text{relative}} = 0.01$, significant; $R^2_{\text{absolute}} = 0.15$, significant; $r_{\text{rel_pred}} = 0.59$; $r_{\text{abs_pred}} = 0.68$). (F) Scatterplot between relative and absolute pitch encoding with neural discriminability of speaker conditions ($r_{\text{relative_speaker}} = 0.21$, $p\text{-value} < 0.05$; $r_{\text{absolute_speaker}} = 0.79$, $p\text{-value} < 1 \times 10^{-38}$). Colored markers indicate electrodes with significant (permutation test; $R^2 > 95^{\text{th}}$ percentile of null distribution) relative pitch and absolute pitch encoding for the left and right panels, respectively.

Chapter 3

3 Speaker-normalized pitch encoding in human superior temporal gyrus in response to Mandarin tones

Abstract

In tone languages such as Mandarin Chinese, the pitch trajectory of a syllable distinguishes word meanings. Previous studies into the neural basis of lexical tone perception have primarily attempted to determine which areas in the brain are involved in tone processing. However, there is limited knowledge about how the specific acoustic features that underlie tone differences are represented in human non-primary auditory cortex. To understand the auditory cortical representation of lexical tones, we used electrocorticography to record neural activity from Mandarin-speaking participants while they passively listened to natural, continuous Mandarin speech. We found that select neural populations in the superior temporal gyrus (STG) have activity patterns that differentiate tones. Using encoding models, we show that these activity patterns can be explained by tuning for the speaker-normalized pitch features of relative pitch height and pitch change. Furthermore, we show that these responses are largely language-independent. By having Mandarin-speaking participants listen to English speech, we could fit highly predictive pitch encoding models using neural responses to English, which revealed similar tuning for pitch features as the Mandarin-speech-fit models and were also able to predict responses to Mandarin tones. These results show that the representation of lexical tones in STG is based in a language-general, shared encoding of speech related pitch.

Introduction

In tone languages like Mandarin Chinese, changing the pitch contour of a syllable can change word meaning. For example, “ma” spoken with a high, level pitch contour means “mother”, while “ma” spoken with a low, dipping pitch contour means “horse”. Many studies into the neural basis of lexical tone perception have attempted to pin down the neural substrates involved in tone processing (Gandour et al., 2000; Klein et al., 2001; Liang and Du, 2018). These studies have primarily focused on questions of hemispherical asymmetry and how language experience may bias processing towards one hemisphere over the other (Gu et al., 2013; Ge et al., 2015). However, a separate question is what features related to lexical tones (in the following often referred to simply as ‘tones’) are being encoded in the neural responses at different levels of the auditory hierarchy, including non-primary auditory cortex in human superior temporal gyrus (STG), a region that encodes phonetic features and prosodic features important for speech perception (Mesgarani et al., 2014; Tang et al., 2017). One hypothesis is that there are representations of tone category in STG that are invariant to acoustic variation within tones. This type of representation could be the neural basis for categorical perception of lexical tones (Wang, 1976; Peng et al., 2010). Another hypothesis is that neural responses to tones in STG are determined by the acoustic features of each individual tone. Neural responses to the lexical tones would be distinct because each tone has different underlying acoustics.

The primary acoustic cue for lexical tone is vocal pitch, related to the fundamental frequency of voiced portions of speech (Howie, 1976). However, because different speakers have varying pitch ranges, determined by the physical length of their vocal folds and size of their larynxes (Titze, 1989), pitch must be normalized by speaker in order to provide speaker-invariant information about which tone was spoken (Rose, 1987). Given the linguistic relevance of

speaker-normalized pitch, studies that have shown perceptual normalization of pitch values in tone perception (Wong and Diehl, 2003; Huang and Holt, 2009), and a previous study which showed that neural responses that differentiated English intonation encoded relative pitch (Tang et al., 2017), we hypothesized that human auditory cortex processes tonal languages to represent information about speaker-normalized pitch features during speech perception.

Results

To understand how lexical tones in Mandarin are represented in human speech cortex, we recorded neural activity from Mandarin-speaking participants while they passively listened to natural, continuous, Mandarin speech. Figure 3.1A and 3.1B shows the acoustic waveform and spectrogram of an example Mandarin sentence spoken by a female and male speaker, respectively. Each syllable in Mandarin has a lexical tone that is primarily cued by the auditory feature of pitch, but the absolute pitch values of a Mandarin tone vary from speaker to speaker due to the speaker's baseline pitch (Fig. 3.1C, D). Figure 3.1I shows the average absolute pitch contour for each lexical tone for each of the ten speakers in the stimulus corpus. The contours occupy different regions of the absolute pitch space, such that no absolute pitch contour determines tone identity. In order for pitch to be linguistically useful, it must be invariant to speaker. Two pitch features that have been shown to be important in tone perception are relative pitch height and pitch change (Gandour, 1983; Chandrasekaran et al., 2007; Khouw and Ciocca, 2007; Tsang et al., 2011). We examined these two features in our stimulus corpus and found that the contours for both relative pitch and pitch change for every tone are consistent between speakers and distinct from each other (Fig. 3.1E-H, J, K), so we considered relative pitch and pitch change to be speaker-normalized pitch features.

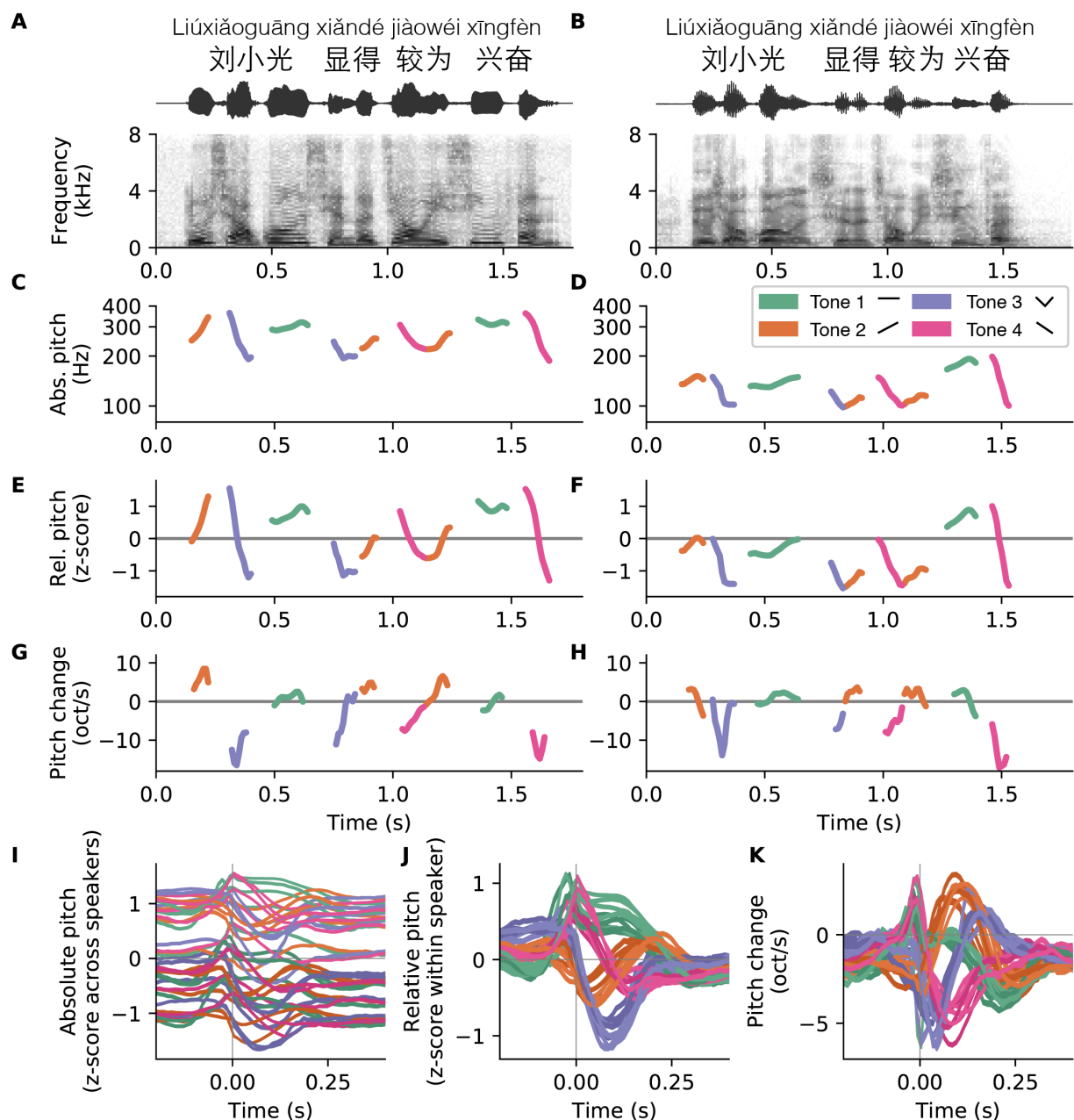


Figure 3.1. Speaker-normalized relative pitch and pitch change are acoustic features that define lexical tones in continuous Mandarin speech. (A, B) Two example sentences spoken by a female and male speaker, respectively, from the Annotated Speech Corpus of Chinese Discourse which consists of spoken passages of Mandarin Chinese. The pinyin, Chinese characters, acoustic amplitude signal, and spectrogram of the sentence is shown. Each syllable is associated with a lexical tone indicated in the pinyin above the vowel of the syllable. (C, D) The absolute pitch in Hz of the two sentences in A, B. Each syllable is colored according to its lexical tone. (E, F) The relative pitch contours of the sentences in A, B, colored by the lexical tone of each syllable. Relative pitch is calculated by normalizing pitch values by speaker (z-score of log Hz). (G, H) Pitch change of the example sentences in A, B, colored by the lexical tone of each

syllable. (I) Average absolute pitch contours for each lexical tone aligned to vowel onset for the ten speakers in ASCCD (five female, five male). The two speakers of the example sentences in A, B are shown in solid colors. The lighter colored traces are from the female speaker and the darker colored traces are from the male speaker. The absolute pitch contours from the other speakers are shown more transparently. (J) Average relative pitch for each tone aligned to vowel onset for all ten speakers. (E) Average pitch change for each tone aligned to vowel onset for all ten speakers.

To study the neurophysiology of tone perception, we recorded voltage potentials from subdurally implanted, multielectrode grids with 128 contacts placed onto the lateral brain surface (see Fig. 3.2A for example grid placement). We then computed for each electrode contact the amplitude of signals in the high-gamma band (70 to 150 Hz), a measure correlated with local neuronal activity (Ray and Maunsell, 2011). After aligning high-gamma activity to speech onset, we confirmed widespread activation of neural populations in the superior temporal gyrus in response to Mandarin (Fig. 3.2A), a region known to be crucial for speech perception (Geschwind, 1965; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Leonard and Chang, 2014).

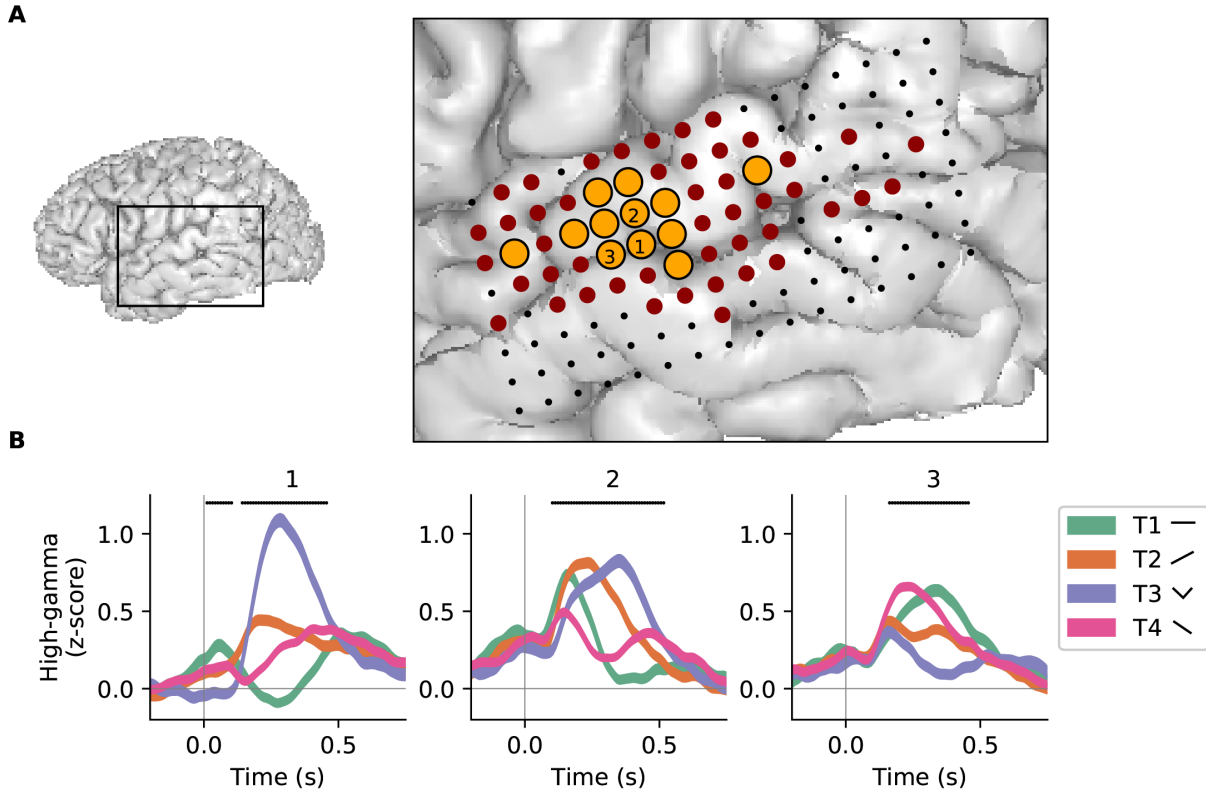


Figure 3.2. Lexical tones in continuous Mandarin speech evoke differential neural responses in discrete populations in STG. (A) Electrode locations for one participant. Dark red indicates electrodes that were responsive to speech. Circled orange electrodes had significantly different response patterns to tones. (B) Average neural activity (shaded area represents mean $\pm$ 1 s.e.m.) in response to each tone on three individual electrodes from participant shown in A, aligned to vowel onset. Black lines indicate time points where means were significantly different between tones (F -test, $p < 0.05$, Bonferroni corrected).

We next aligned high-gamma responses to vowel onsets in each syllable to determine whether differentiable patterns of neural activity are evoked in response in lexical tones. Figure 3.2A shows electrodes which have significantly dissimilar responses to different lexical tones (F -test, $p < 0.05$, Bonferroni corrected) for one participant. Across 8 Mandarin-speaking participants, tone discriminating electrodes made up an average of 14% ($\pm$ 9% std) of speech responsive electrodes (Fig. 3.5). There were a variety of neural response patterns that differentiated tones. Figure 3.2B shows the average neural response to each tone for three STG

electrodes. The first electrode had the largest response to tone 3. The second electrode had an initial response to tone 1 that subsided, an early response to tone 2, and a delayed response to tone 3. The third electrode responded most to tones 4 and 1. One common aspect of the neural responses on these electrodes, however, is that every tone evoked a response pattern. We did not observe instances where an electrode responded only to one particular tone.

Because the response patterns from tone-discriminating electrodes always revealed neural activity evoked by more than one tone category, we next wanted to determine which acoustic or auditory features were driving those responses. An obvious candidate would be voice pitch since pitch is the primary cue to Mandarin tones (Howie, 1976). However, absolute pitch does not determine tone category in a speaker-invariant way. Instead, the pitch values must be normalized by each speaker. Additionally, there is perceptual and neurophysiological evidence that Mandarin tones vary on two different speaker-normalized pitch dimensions, height and change (also known as pitch direction) (Gandour, 1983; Chandrasekaran et al., 2007; Khouw and Ciocca, 2007; Tsang et al., 2011).

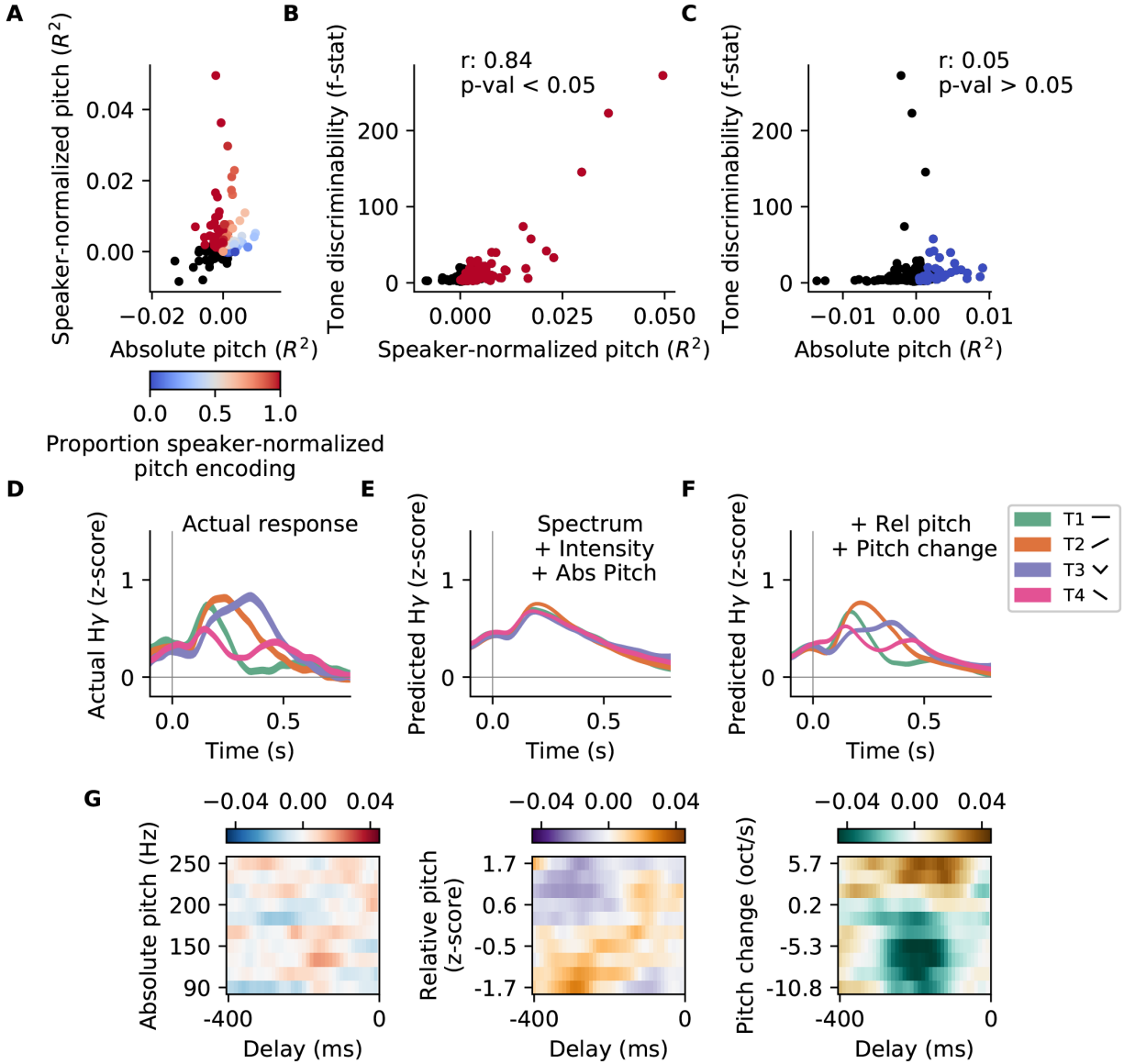


Figure 3.3. Encoding of speaker-normalized relative pitch height and pitch change underlies lexical tone discriminability. (A) Scatterplot of the unique variance explained by absolute pitch and the unique variance explained by speaker-normalized pitch features for every speech responsive electrode from all Mandarin-speaking participants. Each circle represents a single electrode. Colored electrodes had significant encoding of either absolute pitch or speaker-normalized pitch features, with the color indicating the proportion of the variance explained by speaker-normalized pitch features of the total variance explained by pitch. (B) Scatterplot showing the relationship between the unique variance explained by speaker-normalized pitch features and tone discriminability. Colored dots indicate electrodes that had significant encoding of speaker-normalized pitch features. (C) Scatterplot showing the relationship between the unique variance explained by absolute pitch and tone discriminability. Colored dots indicate electrodes that had significant encoding of absolute pitch. (D) Actual high-gamma responses from a single example electrode that differentiates lexical tones. (E) Predicted responses from an

encoding model where speaker-normalized pitch features are excluded. (F) Predicted responses from an encoding model where speaker-normalized pitch features are included. (G) Regression weights from the model whose predictions are shown in F. This electrode is not tuned for absolute pitch, is tuned for low to high relative pitch height over time, and is tuned to pitch change increases.

To test the hypothesis that neural responses differentiating lexical tones can be explained by speaker-normalized pitch features, namely relative pitch height and pitch change, we fit encoding models to test whether speaker-normalized pitch features contributed to the variance explained in the neural activity by stimulus features. Out of 429 speech responsive electrodes across all participants, 100 (23.3%) had significant encoding of speaker-normalized pitch features, explaining up to 5% of the variance of neural activity (Fig. 3.3A). Consistent with previous work on absolute versus relative pitch encoding (Tang et al., 2017), there were fewer sites in STG that encoded absolute pitch (49/429, 11.4%), and absolute pitch encoding explained less variance in the neural activity (up to 1%, see Fig. 3.3A). Furthermore, the variance explained by speaker-normalized pitch features is correlated with tone discriminability whereas the variance explained by absolute pitch is not (Fig. 3.3B, C). This suggests that tone discriminability reflects encoding of speaker-normalized pitch features.

We then looked at model predictions to understand how encoding of speaker-normalized pitch features leads to tone discriminability. The actual responses and model predictions of an example STG electrode is shown in Figure 3.3D and Figure 3.3E, F. An encoding model that has the spectrum, intensity, and absolute pitch as predictor variables does not predict differences between lexical tones (Fig. 3.3E). However, adding speaker-normalized features to this model separates the predictions to each of the lexical tones to match the actual responses (Fig. 3.3F). The regression weights for this example electrode indicate selectivity for increasing pitch change (Fig. 3.3G). Across all tone discriminating electrodes, we saw a variety of tuning patterns,

including selectivity for no pitch change (Fig. 3.6A), low relative pitch height (Fig. 3.6B), and high relative pitch height (Fig. 3.6C). The defining similarity was the encoding of speaker-normalized pitch features, rather than absolute pitch.

Finally, we wanted to understand whether the processing of lexical tones relies on language-specific mechanisms or whether it can be explained by general auditory mechanisms for pitch processing. To do this, Mandarin-speaking patients listened to English speech while we recorded their high-gamma activity. We then fit encoding models to predict neural responses from acoustic features based solely on the responses to English speech (English data) and applied these models to predict high-gamma responses to Mandarin speech. Figures 3.4A and 3.4B show model-predicted neural responses from a model fit on Mandarin data and English data, respectively, for one STG electrode. Figure 3.4C shows the actual high-gamma responses to different tones. The pattern of the English data predicted responses is qualitatively similar to both the pattern of the Mandarin data predicted responses and the pattern of the actual responses (Fig. 3.4A-C).

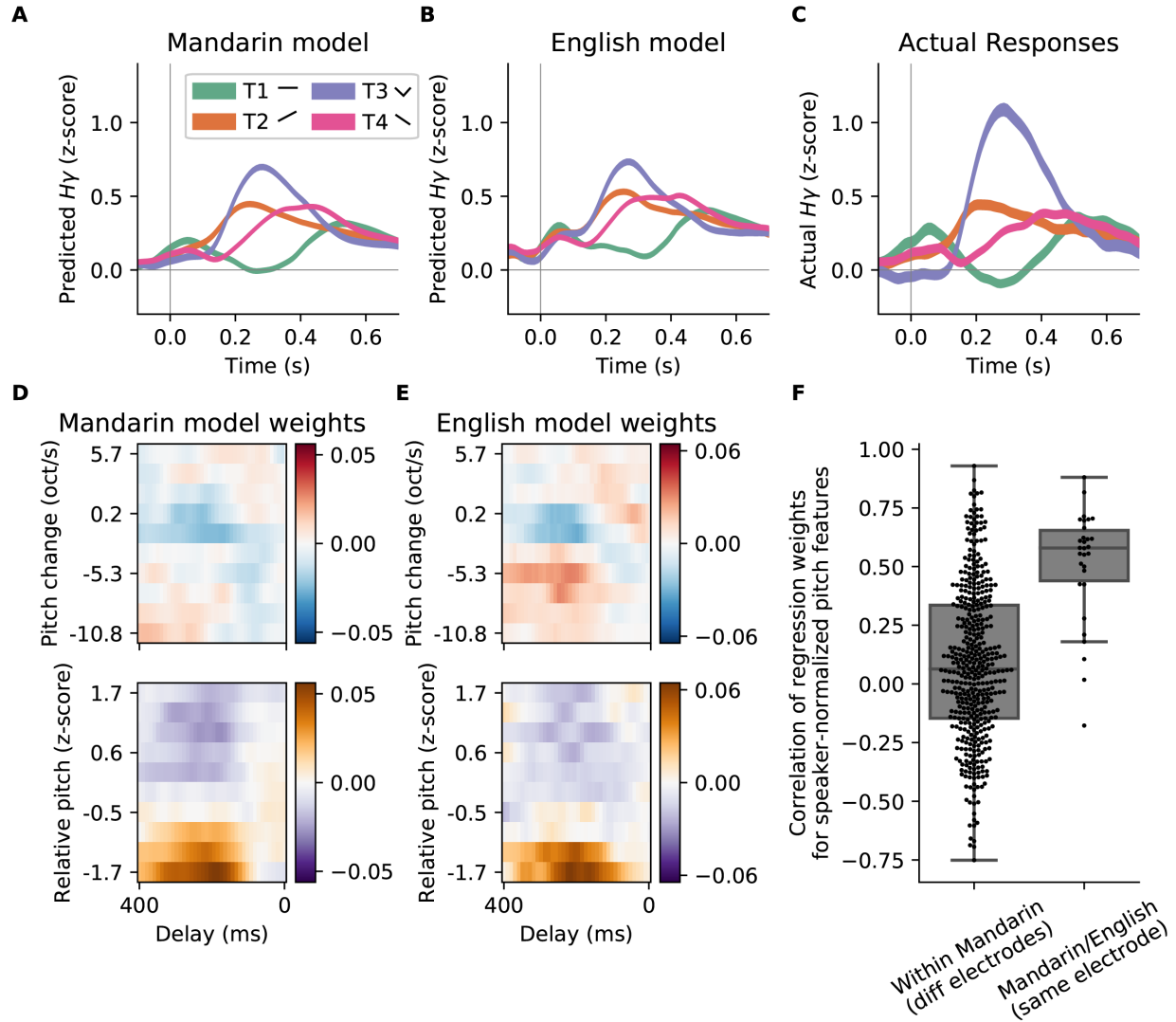


Figure 3.4. STG responses to Mandarin tones largely reflects language-independent encoding of speaker-normalized pitch. (A) Average of predicted neural responses by tone on a single electrode, predicted from the Mandarin encoding model, which was fit on neural responses to Mandarin speech and contained the features of spectrum, intensity, absolute pitch, relative pitch, and pitch change. (B) Average of predicted neural responses by tone from the English encoding model, which was fit on neural responses to English speech. The English model of B contained the same acoustic features as the Mandarin model from A. (C) Average of the actual neural responses, divided by lexical tone. (D) Regression weights for speaker-normalized pitch features over time from the Mandarin model whose predictions are shown in A. (E) Regression weights for speaker-normalized pitch features over time from the English model whose predictions are shown in B. (F) Distribution of the correlation values between the regression weights for speaker-normalized pitch features. The within Mandarin group shows correlation values between the regression weights for each pair of tone discriminating electrodes, both from the Mandarin model. The Mandarin/English group shows correlations between the weights from the Mandarin model and English model for each tone discriminating electrode.

To quantify the similarity between the predictions, we calculated the correlations between the actual neural responses and the two sets of predicted responses, one predicted from English data and one predicted from Mandarin data. By bootstrapping the individual trials, we could derive distributions of correlation values for the English model and Mandarin model. In the majority of cases, we found that the distributions were significantly different (Kolmogorov-Smirnov test, $p < 0.05$), with the Mandarin model performing better than the English model (single electrode example shown in Fig. 3.7A, B, all tone discriminating electrodes shown in Fig. 3.7G and Fig. 3.7I). However, all of the English models are predictive and have correlation values that are on average about 80% of the Mandarin models. What explains the marginal better performance of the Mandarin models? There are two potential explanations. First, it is possible that there are some features missing from the encoding model that explain some variance in the neural response. These unaccounted-for features could be acoustic in nature and language-independent or they could be higher-level features that are dependent on language. To establish that these unaccounted-for features could be language-independent, we ran the same experiment on English-speaking participants. We found that neural responses to Mandarin tones were better predicted by Mandarin data models than by English data models for non-Mandarin-speaking participants as well (single electrode example shown in Fig. 3.7C, D, all tone discriminating electrodes shown in Fig. 3.7H and 3.7I). This result shows that the better performance of the Mandarin model cannot always be due to a speaker's understanding of Mandarin, since the performance increase is seen in English-speaking participants (Fig. 3.7G-I). It suggests that the better performance of the Mandarin model in Mandarin speakers is not based on language-dependent features, however, it does not rule it out conclusively. The second explanation of the models' performance gap is that the stimulus space of the two spoken languages is different.

Because there is more pitch variability and pitch change in Mandarin, it is possible that the encoding models fit on English data are not well trained at the extremes of the relative pitch or pitch change distributions, leading to lower performance of the English models. In this case, if an electrode encodes values of relative pitch and pitch change that are well covered by both Mandarin and English, the performance between the two models should be more similar. Indeed, in some electrodes from Mandarin-speaking participants, the English data model performed just as well as the Mandarin data model (Fig. 3.7E, F).

Although there is a performance gap between the Mandarin and English models, the English data model still is largely able to capture the response patterns to tones. To determine whether speaker-normalized pitch tuning underlies the ability of the English model to predict responses to tones, we looked at the regression weights from the Mandarin and English models. Figures 3.4D and 3.4E show the temporal receptive field for speaker-normalized pitch features for the Mandarin model and English model, respectively, corresponding to the models whose predictions are shown in Figures 3.4A and 3.4B. Both show tuning for low relative pitch height, and a change from decreasing pitch to increasing pitch. To quantify this similarity, we calculated the correlation between the regression weights for the Mandarin and English models (Fig. 3.4F). For almost all electrodes, the correlation was positive, indicating that the tuning for speaker-normalized pitch features was similar during listening to both Mandarin and English.

Discussion

We showed that lexical tones in Mandarin evoke differential responses in neural populations in STG. Furthermore, these responses could be largely explained by the encoding of speaker-normalized pitch features in a language-general manner. The neural populations we

identify here as encoding speaker-normalized pitch features are likely the same ones identified as encoding relative pitch over absolute pitch in a previous study on pitch encoding during speech intonation perception (Tang et al., 2017). That pitch change, another speaker-normalized pitch feature, is also encoded in these neural populations adds to our understanding of the mapping between stimulus features and neural activity.

The result that STG responses encode speaker-normalized pitch features is consistent with psychophysical studies which have shown speaker normalization for tone perception (Moore and Jongman, 1997; Wong and Diehl, 2003). In these studies, tones were evaluated at the end of a speech precursor in which the speaker's pitch range was different between conditions. The effect was contrastive, such that an acoustically identical syllable is perceived as low following a high-pitched precursor and high following a low-pitched precursor. Another study showed the same effect when the precursor was a non-speech sound with an average pitch percept, providing evidence that speaker normalization for tone may be based on general auditory mechanisms (Huang and Holt, 2009).

Behavioral and neurophysiological studies have shown that both relative pitch height and pitch change are important for tone perception (Gandour, 1983; Chandrasekaran et al., 2007; Khouw and Ciocca, 2007; Tsang et al., 2011). In one study, pitch height and contour were independently manipulated, creating a continuum of pitch patterns between Mandarin Tone 1 and Tone 2. It was found that both features influenced tone identification, with the contribution of one cue increased when the other was more ambiguous (Massaro et al., 1985). These results are consistent with our results that STG responses may encode relative pitch or pitch change.

It is perhaps surprising that we did not uncover a prominent language-specific response to lexical tones. Previous studies have shown that the perception of tones varies between people

who speak a tonal language versus people who do not (Chan et al., 1975; Burnham et al., 1996; Bent et al., 2006). There is also some evidence that Mandarin speakers may perceive tones in a categorical way (Wang, 1976; Peng et al., 2010), however some studies did not find evidence for categorical perception (Abramson, 1979; Francis et al., 2003). Our results do not preclude an effect of language on STG activity. Instead, they suggest that the largest differences in neural activity are driven by speaker-normalized pitch features. Indeed, we also found that neural responses to Mandarin tones were similar between Mandarin speakers and English speakers. Although neuroimaging studies on tone perception have focused on differences between tone language and non-tone language speakers, our results are still consistent with their findings that activation in bilateral STG is similar between the two groups (Gandour et al., 1998, 2000; Klein et al., 2001).

Because the features of relative pitch and pitch change are abstracted from absolute pitch and are speaker invariant, they can thus be linguistically relevant. Overall, our results suggest that Mandarin and English and, by extension, perhaps all spoken languages share an underlying auditory code for speech-related pitch.

Materials and Methods

The experimental protocol was approved by the Institutional Review Board at the University of California, San Francisco (UCSF) and by the Huashan Hospital Institutional Review Board associated with Fudan University. All participants gave their written, informed consent prior to testing.

Participants

This study included participants who were neurosurgical patients at either Huashan Hospital or UCSF. Participants from Huashan Hospital were cancer patients undergoing surgery to remove tumors. High-density electrode grids were placed onto the lateral cortical surface intraoperatively for clinical language mapping. The experiments in this study were run after grid placement before language mapping. Electrodes were localized by recording the three-dimensional positions of the corners of the grid using Brainlab and aligning them to pre-surgery MRI (Hamilton et al., 2017). The remaining electrodes were localized using interpolation and extrapolation from those points. The 8 participants from Huashan Hospital had left-hemisphere grids.

Participants from UCSF were patients with intractable epilepsy who had high-density electrode grids implanted for clinical monitoring of seizure activity. The placements of the grids were determined solely by clinical needs. All 3 participants had left-hemisphere grids. Electrodes were localized by aligning preimplantation MRI and post-implantation CT scans.

Data acquisition and neural signal processing

During experimental tasks, neural signals were recorded from one or two 128 channel ECoG grids (8 x 16, 4mm spacing) using a multichannel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies). The local field potential at each electrode contact was amplified and sampled at 3052Hz. The raw voltage waveform was visually examined, and channels containing signal variation too low to be detectable from noise or continuous epileptiform activity were removed. Time segments on remaining channels that contained electrical or movement-related artifacts were manually marked and excluded. The signal was then notch-filtered to remove line noise (at 50Hz, 100Hz, and 150Hz for Huashan participants and at 60Hz, 120Hz, and 180Hz for UCSF participants).

Using the Hilbert transform, the analytic amplitude of eight Gaussian filters (center frequencies: 70-150Hz) was computed. The high-gamma signal was taken as the average analytic amplitude across these eight bands. The signal was down-sampled to 100Hz and z-scored across the recording block ($H\gamma$).

Experimental stimuli

The acoustic stimuli used in this study consisted of natural, continuous speech in both Mandarin and American English. The Mandarin speech was a subset of the Annotated Speech Corpus of Chinese Discourse (ASCCD), which included read texts of a variety of discourse structures, such as narrative and prose from five male and five female speakers (Aijun et al., 2000). Mandarin Chinese has four lexical tones, called Tone 1-4. The prototypical pitch pattern for each tone can be described as high-level, mid-rising, low-dipping, and high-falling,

respectively (Chao, 1965). The primary acoustic cue of Mandarin tones is voice pitch (Howie, 1976).

The English speech stimuli consisted of materials from the TIMIT corpus (Garofalo et al., 1993) and the Boston University Radio Speech Corpus (BURSC) (Ostendorf et al., 1995).

Data analysis software

All analyses were carried out using custom software written in MATLAB and Python. Open-source scientific Python packages used included numpy, scipy, pandas, scikit-learn, and statsmodels. Figures were created with matplotlib.

Speech responsive electrodes

To find speech responsive electrodes, we first aligned high-gamma responses to the onsets of speech. Onsets were defined as times in the stimulus where sound was preceded by at least 400ms of silence. We then used a paired sample t-test to test whether the average response after speech onset (window from 100ms after onset to 400ms after onset, accounting for a neural delay of ~100ms) was significantly different from the response before speech onset (-250ms to 50ms). Because the temporal dynamics of the response differ between electrodes, we also tested neural responses at speech offsets, comparing average high-gamma before speech offset to average high-gamma after speech offset. This comparison captured electrodes that did not have a strong onset response, but whose response was elevated before speech offset and dropped off after speech offset. Offsets were defined as times in the stimulus where sound was followed by at least 400ms of silence. All t-tests were corrected for multiple comparisons using the Bonferroni method.

Single-electrode tone discrimination analysis

To determine whether neural responses to different tones were distinguishable, we aligned high-gamma responses to the onsets of the vowels of syllables and divided trials by the lexical tone of the syllable. For each 10ms timepoint between 0.2s before the vowel onset and 1s after the vowel onset, we used a one-way ANOVA to test whether differences in high-gamma between tones were significant. We used the Bonferroni method to correct for multiple comparisons across both electrodes and time, taking the alpha level to be $0.05/(128 \text{ electrodes} * 120 \text{ timepoints})$. To determine whether a single electrode discriminated tones, we checked whether it had three consecutive significant time points.

Pitch temporal receptive field analysis

To determine whether speaker-normalized pitch features drive neural responses in lexical tone discriminating neural populations, we used time-delayed linear encoding models known as temporal receptive field models (Theunissen et al., 2001; Crosse et al., 2016). Temporal receptive field (TRF) models allow us to predict neural activity based on stimulus features in a window of time preceding neural activity.

In the full TRF model, we included features for the sound spectrum, sound intensity, absolute pitch, relative pitch, and pitch change. To calculate the unique contribution of speaker-normalized pitch features (relative pitch and pitch change), we fit TRF models that excluded these features and calculated the difference in R2 between the full model and exclusionary ones.

To prevent model overfitting, we used L2 regularization and cross-validation. Specifically, we divided the data into three mutually exclusive sets of 80%, 10% and 10% of

samples. The first set of 80% was used as the training set. The second set was used to the L2 regularization hyperparameter, and the final set was used as the test set. We evaluated the models using the correlation between actual and predicted values of neural activity on held out data. We performed this procedure 5 times and the performance of the model was taken as the mean of performance across all testing sets.

To calculate the significance of unique portions of variance explained, we employed permutation testing. We shuffled the acoustic features between all phrases in the stimuli before computing null values of the unique variance explained by either absolute pitch or by speaker-normalized pitch features by running the same analysis pipeline. We ran this procedure 200 times to get a null distribution of values. Unique R^2 values above the 95th percentile were considered significant.

Comparison of predicted neural activity from Mandarin-fit and English-fit models

To determine whether the neural response to Mandarin tones could be predicted from neural responses to English speech, we recorded cortical activity while participants listened to English speech (English data), in addition to the cortical activity recorded while participants listened to Mandarin (Mandarin data). We constructed the stimulus representations for both the English and Mandarin speech stimuli using the same feature descriptions. These features included the acoustic spectrogram, intensity, absolute pitch, relative pitch, and pitch change. We then fit TRF models on the English data and used the regression weights to predict neural responses to Mandarin speech.

To quantify the performance of the Mandarin and English models, we calculated the correlation between the predicted and actual neural responses. To compare the Mandarin and

English model performances, we used a bootstrapping procedure to derive a distribution of correlation values by sampling from the trials of individual tones. Specifically, we randomly chose 500 trials with replacement, calculated the correlation between the actual and predicted response to those trials, and then performed that procedure 200 times to arrive at two distributions, one for the Mandarin model and one for the English model. We tested whether those distributions were significantly different using the Kolmogorov-Smirnov test.

To determine whether speaker-normalized pitch encoding was similar for individual electrodes between Mandarin and English, we calculated the correlation of the regression weights of the two models. To reduce noise in the weights, we first spatially blurred the weights and then calculated the correlation. Positive correlations indicated that an electrode had similar tuning for speaker-normalized pitch features across the two languages.

Supplementary Figures

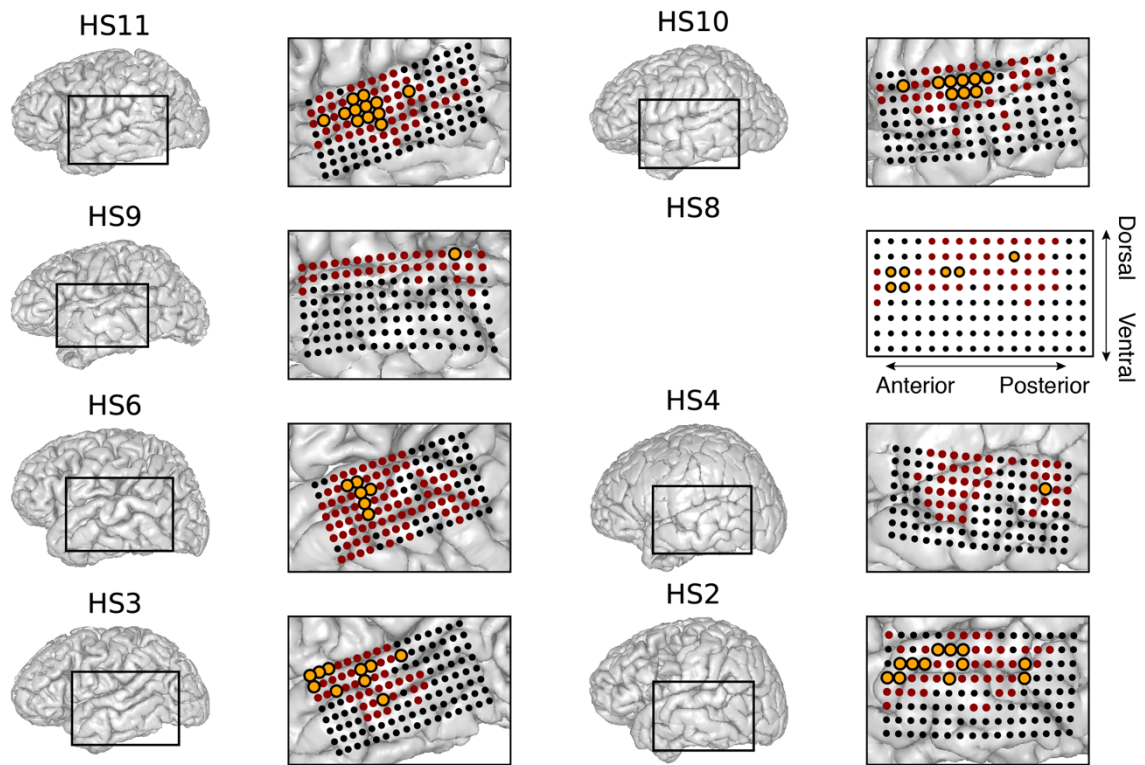


Figure 3.5. Speech responsive and tone discriminating electrodes for all participants. ECoG grids covered the lateral temporal lobe of all participants. The length between adjacent electrodes is 4mm. Dark red indicates electrodes that were responsive to speech. Circled orange electrodes indicate tone discriminating electrodes. We were unable to reconstruct the brain anatomy for HS8.

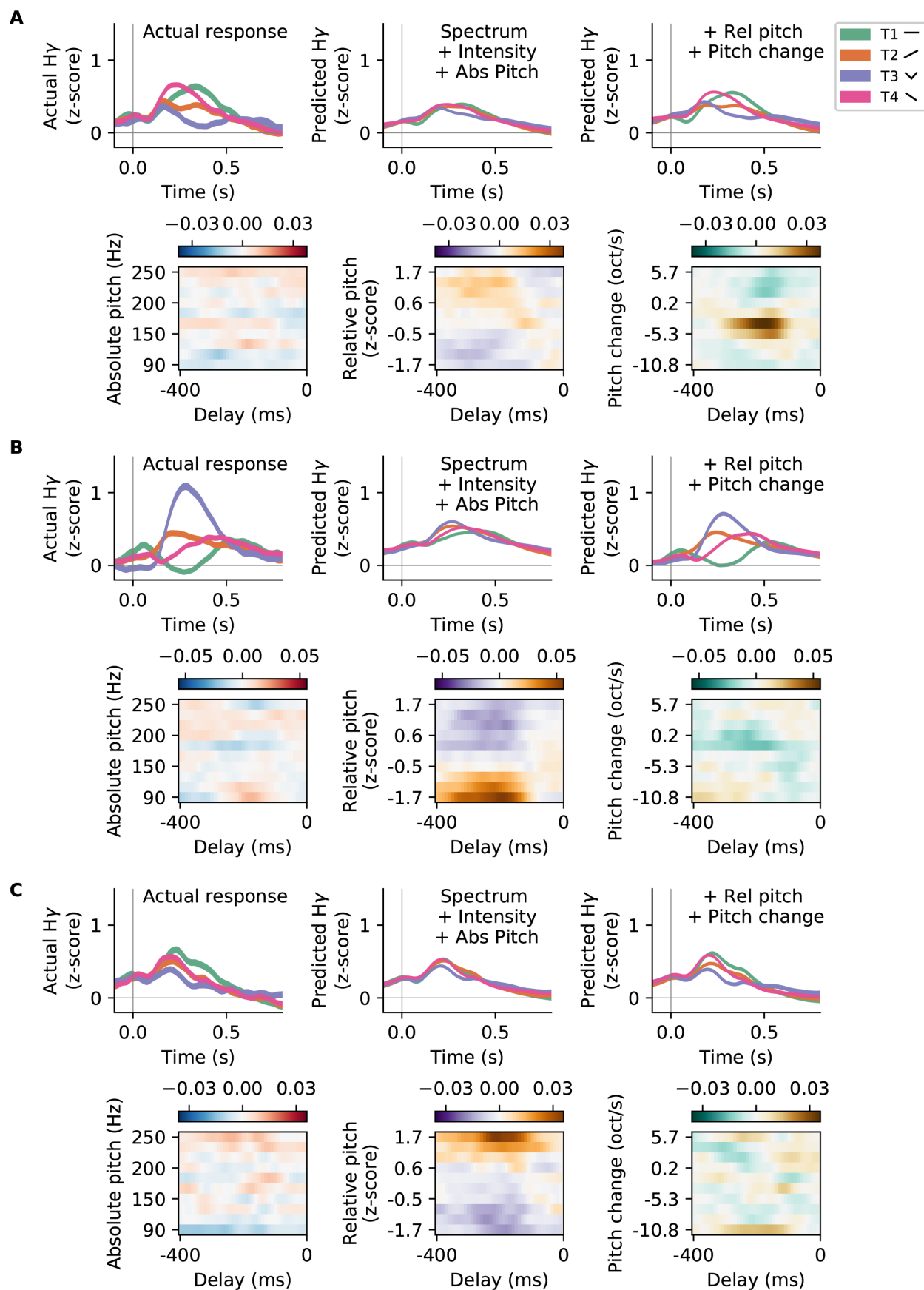


Figure 3.6. Examples of different types of speaker-normalized pitch feature encoding. (A)

Top row shows actual high-gamma responses and predicted responses from a single example electrode that differentiates lexical tones. The middle and right panel of the top row show predicted responses from an encoding model where speaker-normalized pitch features are left out and predicted responses from an encoding model where speaker-normalized pitch features are included, respectively. Bottom row shows regression weights from the model, indicating increased activity to no pitch change. (B) Same as A, but for an electrode tuned to low relative pitch. (C) Same as A, but for an electrode tuned to high relative pitch.

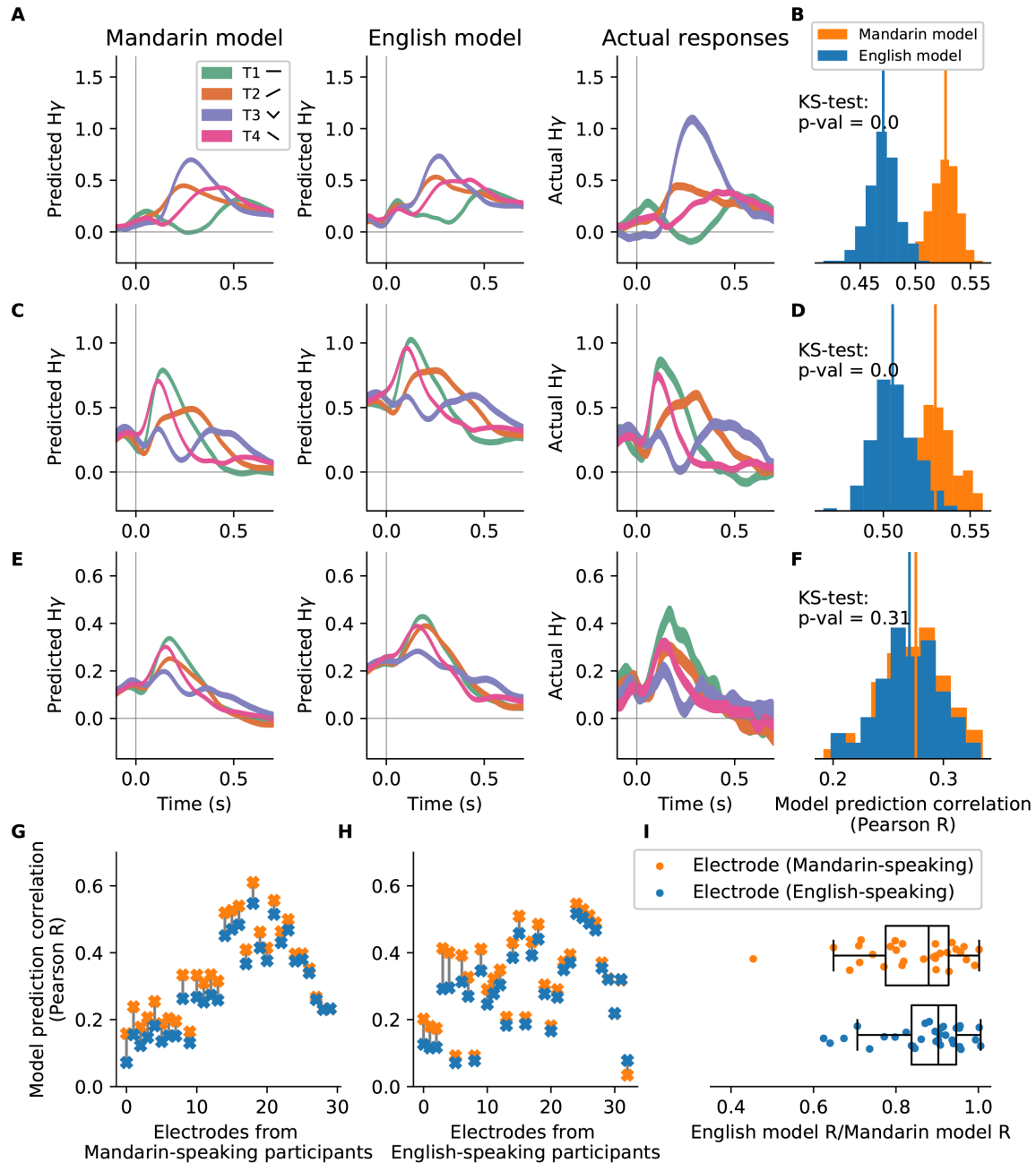


Figure 3.7. Models fit on Mandarin data perform better than models fit on English data for both Mandarin and English speakers. (A) Predicted responses and actual responses for an example electrode from a Mandarin-speaking participant. The left panel shows predicted responses from a model fit on Mandarin data while the right panel shows predicted responses from a model fit on English data. (B) Distribution of bootstrapped correlations between actual and predicted responses. The correlations from the Mandarin model are significantly better than the correlations from the English model in this example. (C, D) Same as A, B, but for an example electrode from an English-speaking participant who does not speak Mandarin. (E, F) Same panels as above, but showing an electrode from a Mandarin-speaking participant where the

English model performs as well as the Mandarin model. (G) Plot showing the correlation between actual and predicted responses for the Mandarin and English models for all tone-selective electrodes from all Mandarin-speaking patients. The electrodes are ordered from left to right by the proportion of the English model correlation over the Mandarin model correlation. (H) Same as G, but for tone-selective electrodes from English-speaking participants. (I) The fraction of English model correlation over Mandarin model correlation. For both Mandarin-speaking and English-speaking participants, the Mandarin model performs better than the English model.

Chapter 4

4 Conclusion

The results presented in this dissertation add to our knowledge of how the human brain is capable of understanding spoken language. Specifically, we have shown how neural activity in human non-primary auditory cortex encodes linguistically relevant, vocal pitch information during speech perception.

In Chapter 2, we analyzed auditory cortical responses to speech that varied in its intonation contour. We found a subset of neural populations in human superior temporal gyrus whose activity patterns differentiated intonational pitch contours. These neural populations were not sensitive to the phonetic content of speech or to who was speaking. We further showed that the activity of these neural populations could be explained by the encoding of speaker-normalized relative pitch values. These results established the encoding of the linguistically relevant, abstract feature of relative pitch in human non-primary auditory cortex during speech perception.

In Chapter 3, we examined auditory cortical activity in Mandarin-speaking participants in response to lexical tones in continuous Mandarin speech. We found that the representation of lexical tones in human superior temporal gyrus was based on the encoding of speaker-normalized relative pitch and pitch change, but not absolute pitch. Furthermore, we showed that neural responses to Mandarin tones could be predicted by looking at neural responses to an unfamiliar language. These results suggest that all spoken languages share a common underlying auditory code for speech related pitch.

Future Directions

The findings of dissertation lay a foundation for understanding how pitch information is extracted from speech and represented by neural activity in human non-primary auditory cortex in a speaker-normalized, relative way. However, there are still many unanswered questions. One future goal is to understand the auditory mechanism that results in the encoding of relative pitch at a computational or neural implementation level. This mechanism needs to be able to account for calculating statistics, such as the average pitch, and representing a value relative to those statistics. Understanding the mechanism of relative auditory encoding may also shed light on other auditory perceptual phenomenon, such as the effect whereby an ambiguous sound between ‘da’ and ‘ga’ is biased towards one or the other when it occurs after a ‘al’ or ‘ar’ sound (Lotto and Kluender, 1998). This context dependence can be explained by the relative height values of the third formant, which is one of the cues that distinguishes ‘da’ and ‘ga’. The same auditory mechanism could also underlie the ability to perceive contours in other auditory dimensions such as loudness and timbre (McDermott et al., 2008).

Finally, it is still not clear how the neural representation of vocal pitch information is used to arrive at linguistic meaning. Does the information about vocal pitch need to be integrated with the information about other linguistic features to form a percept of something that is linguistically meaningful? Are pitch-related representations transformed into tone category representations at some level in the auditory or language comprehension systems? Although we have established that auditory processing of speech results in a neural representation of relative pitch in certain neural populations in human non-primary auditory cortex, there is still much to be learned about how vocal pitch in speech is understood by the human brain.

References

- Abramson A (1979) The noncategorical perception of tone categories in Thai. *Frontiers of speech communication research*, edited by B. Lindblom & S. Ohman, 127-134.
- Acoustical Terminology S (1960) New York: American Standards Association Std.
- Aijun L, Maocan L, Yiqing Z, Guohua S, Wu H, Zhigang Y, Jingzhu Y (2000) Speech Corpus of Chinese Discourse and the Phonetic Research. In, pp 7.
- Allen EJ, Burton PC, Olman CA, Oxenham AJ (2016) Representations of Pitch and Timbre Variation in Human Auditory Cortex. *J Neurosci*:2336–16.
- Baumann O, Belin P (2010) Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychol Res* 74:110–120.
- Belin P, Bestelmeyer PEG, Latinus M, Watson R (2011) Understanding voice perception. *Br J Psychol Lond Engl* 102:711–725.
- Belin P, Fecteau S, Bédard C (2004) Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 8:129–135.
- Bendor D, Wang X (2005) The neuronal representation of pitch in primate auditory cortex. *Nature* 436:1161–1165.
- Bent T, Bradlow AR, Wright BA (2006) The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *J Exp Psychol Hum Percept Perform* 32:97.

- Bizley JK, Walker KMM, King AJ, Schnupp JWH (2010) Neural Ensemble Codes for Stimulus Periodicity in Auditory Cortex. *J Neurosci* 30:5078–5091.
- Bizley JK, Walker KMM, Nodal FR, King AJ, Schnupp JWH (2013) Auditory cortex represents both pitch judgments and the corresponding acoustic cues. *Curr Biol* CB 23:620–625.
- Bizley JK, Walker KMM, Silverman BW, King AJ, Schnupp JWH (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *J Neurosci Off J Soc Neurosci* 29:2064–2075.
- Boersma P (2002) Praat, a system for doing phonetics by computer. *Glot Int* 5:341–345.
- Burnham D, Francis E, Webster D, Luksaneeyanawin S, Attapaiboon C, Lacerda F, Keller P (1996) Perception of lexical tone across languages: evidence for a linguistic mode of processing. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, pp 2514–2517 vol.4.
- Cariani PA, Delgutte B (1996a) Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J Neurophysiol* 76:1698–1716.
- Cariani PA, Delgutte B (1996b) Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *J Neurophysiol* 76:1717–1734.
- Chan SW, Chuang C, Wang WS (1975) Cross- linguistic study of categorical perception for lexical tone. *J Acoust Soc Am* 58:S119–S119.

- Chandrasekaran B, Gandour JT, Krishnan A (2007) Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity. *Restor Neurol Neurosci* 25:195–210.
- Chao YR (1965) A grammar of spoken Chinese. Univ of California Press.
- Crone N (1994) Regional cortical activation during language and related tasks identified by direct cortical electrical recording. *Brain Lang* 47:466–468.
- Crone NE, Boatman D, Gordon B, Hao L (2001) Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol* 112:565–582.
- Crone NE, Miglioretti DL, Gordon B, Lesser RP (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain J Neurol* 121:2301–2315.
- Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Front Hum Neurosci* 10 Available at: <https://www.frontiersin.org/articles/10.3389/fnhum.2016.00604/full> [Accessed April 3, 2019].
- Cutler A, Dahan D, van Donselaar W (1997) Prosody in the comprehension of spoken language: a literature review. *Lang Speech* 40 (Pt 2):141–201.

- Doherty CP, West WC, Dilley LC, Shattuck-Hufnagel S, Caplan D (2004) Question/statement judgments: an fMRI study of intonation processing. *Hum Brain Mapp* 23:85–98.
- Earle MA (1975) An acoustic phonetic study of Northern Vietnamese tones. Speech Communications Research Laboratory.
- Edwards E, Soltani M, Deouell LY, Berger MS, Knight RT (2005) High Gamma Activity in Response to Deviant Auditory Stimuli Recorded Directly From Human Cortex. *J Neurophysiol* 94:4269–4280.
- Edwards E, Soltani M, Kim W, Dalal SS, Nagarajan SS, Berger MS, Knight RT (2009) Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J Neurophysiol* 102:377–386.
- Fant G (1971) Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Walter de Gruyter.
- Fishman YI, Michey C, Steinschneider M (2013) Neural representation of harmonic complex tones in primary auditory cortex of the awake monkey. *J Neurosci Off J Soc Neurosci* 33:10312–10323.
- Formisano E, De Martino F, Bonte M, Goebel R (2008) “Who” is saying “what”? Brain-based decoding of human voice and speech. *Science* 322:970–973.
- Francis AL, Ciocca V, Chit Ng BK (2003) On the (non)categorical perception of lexical tones. *Percept Psychophys* 65:1029–1044.

- Friederici AD, Alter K (2004) Lateralization of auditory language functions: a dynamic dual pathway model. *Brain Lang* 89:267–276.
- Gandour J (1983) Tone perception in Far Eastern languages. *J Phon* 11:149–175.
- Gandour J, Tong Y, Wong D, Talavage T, Dziedzic M, Xu Y, Li X, Lowe M (2004) Hemispheric roles in the perception of speech prosody. *NeuroImage* 23:344–357.
- Gandour J, Wong D, Hsieh L, Weinzaepfel B, Lancker DV, Hutchins GD (2000) A crosslinguistic PET study of tone perception. *J Cogn Neurosci* 12:207–222.
- Gandour J, Wong D, Hutchins G (1998) Pitch processing in the human brain is influenced by language experience. *Neuroreport* 9:2115–2119.
- Garofalo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL (1993) The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom. Linguist Data Consort.
- Ge J, Peng G, Lyu B, Wang Y, Zhuo Y, Niu Z, Tan LH, Leff AP, Gao J-H (2015) Cross-language differences in the brain network subserving intelligible speech. *Proc Natl Acad Sci* 112:2972–2977.
- Geschwind N (1965) Disconnexion syndromes in animals and man. *Brain* 88:585–585.
- Griffiths TD, Büchel C, Frackowiak RS, Patterson RD (1998) Analysis of temporal structure in sound by the human brain. *Nat Neurosci* 1:422–427.
- Gu F, Zhang C, Hu A, Zhao G (2013) Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. *Neuroimage* 83:637–645.

- Gussenhoven C, Repp BH, Rietveld A, Rump HH, Terken J (1997) The perceptual prominence of fundamental frequency peaks. *J Acoust Soc Am* 102:3009–3022.
- Hamilton LS, Chang DL, Lee MB, Chang EF (2017) Semi-automated Anatomical Labeling and Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography. *Front Neuroinformatics* 11 Available at: <https://www.frontiersin.org/articles/10.3389/fninf.2017.00062/full> [Accessed March 23, 2019].
- Heilman KM, Bowers D, Speedie L, Coslett HB (1984) Comprehension of affective and nonaffective prosody. *Neurology* 34:917–921.
- Hickok G, Poeppel D (2007) The cortical organization of speech processing. *Nat Rev Neurosci* 8:393.
- Howie JM (1976) *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press.
- Huang J, Holt LL (2009) General perceptual contributions to lexical tone normalization. *J Acoust Soc Am* 125:3983–3994.
- Jakobson R, Fant G, Halle M (1951) Preliminaries to speech analysis. The distinctive features and their correlates.
- Khouw E, Ciocca V (2007) Perceptual correlates of Cantonese tones. *J Phon* 35:104–117.
- Klein D, Zatorre RJ, Milner B, Zhao V (2001) A Cross-Linguistic PET Study of Tone Perception in Mandarin Chinese and English Speakers. *NeuroImage* 13:646–653.

- Kluender KR, Coady JA, Kieft M (2003) Sensitivity to change in perception of speech. *Speech Commun* 41:59–69.
- Kreitewolf J, Friederici AD, von Kriegstein K (2014) Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *NeuroImage* 102 Pt 2:332–344.
- Lachaux J-P, George N, Tallon-Baudry C, Martinerie J, Hugueville L, Minotti L, Kahane P, Renault B (2005) The many faces of the gamma band response to complex visual stimuli. *NeuroImage* 25:491–501.
- Ladd DR (2008) *Intonational phonology*. Cambridge University Press.
- Lee C-Y (2009) Identifying isolated, multispeaker Mandarin tones from brief acoustic input: a perceptual and acoustic study. *J Acoust Soc Am* 125:1125–1137.
- Leonard MK, Chang EF (2014) Dynamic speech representations in the human temporal lobe. *Trends Cogn Sci* 18:472–479.
- Liang B, Du Y (2018) The Functional Neuroanatomy of Lexical Tone Perception: An Activation Likelihood Estimation Meta-Analysis. *Front Neurosci* 12.
- Licklider JCR (1954) “Periodicity” Pitch and “Place” Pitch. *J Acoust Soc Am* 26:945–945.
- Lotto AJ, Kluender KR (1998) General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Percept Psychophys* 60:602–619.
- Massaro DW, Cohen MM, Tseng C (1985) The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *J Chin Linguist*:267–289.

- McDermott JH, Lehr AJ, Oxenham AJ (2008) Is relative pitch specific to pitch? *Psychol Sci* 19:1263–1271.
- Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343:1006–1010.
- Meyer M, Alter K, Friederici AD, Lohmann G, von Cramon DY (2002) FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Hum Brain Mapp* 17:73–88.
- Meyer M, Steinhauer K, Alter K, Friederici AD, von Cramon DY (2004) Brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain Lang* 89:277–289.
- Moore CB, Jongman A (1997) Speaker normalization in the perception of Mandarin Chinese tones. *J Acoust Soc Am* 102:1864–1877.
- Moulines E, Charpentier F (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun* 9:453–467.
- Ohara S, Crone N, Weiss N, Lenz F (2004) Attention to pain modulates electrocorticographic event-related desynchronization during cutaneous laser stimulation in humans. *Clin Neurophysiol* 115:1641–1652.
- Ojemann G, Ojemann J, Lettich E, Berger M (1989) Cortical language localization in left, dominant hemisphere: An electrical stimulation mapping investigation in 117 patients. *J Neurosurg* 71:316–326.

- Ostendorf M, Price PJ, Shattuck-Hufnagel S (1995) The Boston University radio news corpus. *Linguist Data Consort*:1–19.
- Oxenham AJ, Plack CJ (1997) A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing. *J Acoust Soc Am* 101:3666–3675.
- Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD (2002) The Processing of Temporal Pitch and Melody Information in Auditory Cortex. *Neuron* 36:767–776.
- Pell MD, Baum SR (1997) The ability to perceive and comprehend intonation in linguistic and affective contexts by brain-damaged adults. *Brain Lang* 57:80–99.
- Penagos H, Melcher JR, Oxenham AJ (2004) A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J Neurosci Off J Soc Neurosci* 24:6810–6815.
- Peng G, Zheng H-Y, Gong T, Yang R-X, Kong J-P, Wang WS-Y (2010) The influence of language experience on categorical perception of pitch contours. *J Phon* 38:616–624.
- Pfurtscheller G, Cooper R (1975) Frequency dependence of the transmission of the EEG from cortex to scalp. *Electroencephalogr Clin Neurophysiol* 38:93–96.
- Pierrehumbert J (1979) The perception of fundamental frequency declination. *J Acoust Soc Am* 66:363–369.
- Plack CJ, Oxenham AJ, Fay RR (2006) *Pitch: neural coding and perception*. Springer Science & Business Media.

- Plante E, Creusere M, Sabin C (2002) Dissociating sentential prosody from sentence processing: activation interacts with task demands. *NeuroImage* 17:401–410.
- Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat Neurosci* 12:718.
- Ray S, Maunsell JHR (2011) Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol* 9:e1000610.
- Rose P (1987) Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Commun* 6:343–352.
- Ross ED (1981) The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. *Arch Neurol* 38:561–569.
- Schönwiesner M, Zatorre RJ (2008) Depth electrode recordings show double dissociation between pitch processing in lateral Heschl's gyrus and sound onset processing in medial Heschl's gyrus. *Exp Brain Res* 187:97–105.
- Shattuck-Hufnagel S, Turk AE (1996) A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res* 25:193–247.
- Sonntag GP, Portele T (1998) PURR—a method for prosody evaluation and investigation. *Comput Speech Lang* 12:437–451.
- Steinschneider M, Fishman YI, Arezzo JC (2008) Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb Cortex N Y N 1991* 18:610–625.

- Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. *Science* 357:797–801.
- Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Netw Comput Neural Syst* 12:289–316.
- Tian B, Rauschecker JP (2004) Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J Neurophysiol* 92:2993–3013.
- Titze IR (1989) Physiologic and acoustic differences between male and female voices. *J Acoust Soc Am* 85:1699–1707.
- Tong Y, Gandour J, Talavage T, Wong D, Dziedzic M, Xu Y, Li X, Lowe M (2005) Neural circuitry underlying sentence-level linguistic prosody. *NeuroImage* 28:417–428.
- Tsang Y-K, Jia S, Huang J, Chen H-C (2011) ERP correlates of pre-attentive processing of Cantonese lexical tones: The effects of pitch contour and pitch height. *Neurosci Lett* 487:268–272.
- van Dommelen WA (1990) Acoustic parameters in human speaker recognition. *Lang Speech* 33 (Pt 3):259–272.
- Vidal JR, Ossandón T, Jerbi K, Dalal SS, Minotti L, Ryvlin P, Kahane P, Lachaux J-P (2010) Category-Specific Visual Responses: An Intracranial Study Comparing Gamma, Beta, Alpha, and ERP Response Selectivity. *Front Hum Neurosci* 4 Available at:

<https://www.frontiersin.org/articles/10.3389/fnhum.2010.00195/full> [Accessed March 23, 2019].

Wang WS-Y (1976) Language Change*. *Ann N Y Acad Sci* 280:61–72.

Warren JD, Uppenkamp S, Patterson RD, Griffiths TD (2003) Separating pitch chroma and pitch height in the human brain. *Proc Natl Acad Sci U S A* 100:10038–10042.

Weintraub S, Mesulam MM, Kramer L (1981) Disturbances in prosody. A right-hemisphere contribution to language. *Arch Neurol* 38:742–744.

Witteman J, van Ijzendoorn MH, van de Velde D, van Heuven VJJP, Schiller NO (2011) The nature of hemispheric specialization for linguistic and emotional prosodic perception: a meta-analysis of the lesion literature. *Neuropsychologia* 49:3722–3738.

Wong PCM, Diehl RL (2003) Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *J Speech Lang Hear Res JSLHR* 46:413–421.

Zatorre RJ, Delhommeau K, Zarate JM (2012) Modulation of auditory cortex response to pitch variation following training with microtonal melodies. *Front Psychol* 3:544.

Zatorre RJ, Evans AC, Meyer E, Gjedde A (1992) Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256:846–849.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

6/12/19

Date