UNIVERSITY OF CALIFORNIA,
IRVINE

Perception of socially-relevant cues from face and body movements: behavioral and neural investigations

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Daniel A. Stehr

DissertationCommittee:
Emily D. Grossman, Chair
Gregory Hickok
Donald Hoffman
Charlie Chubb

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest thanks to my advisor and mentor, Emily Grossman, for all her support, insights, and constructive feedback throughout my graduate studies. Her enthusiasm, attentiveness, and encouragement over the years has inspired me to work with dedication and absorption. The work in this dissertation wouldn't nearly have the definition it does without the consistent exchange of ideas and shared decision making we practiced. I'm so thankful for the foundation she has helped me build and I will continue to emulate her as my best role model.

I would also like to thank my other committee members, Greg Hickok, Donald Hoffman, and Charlie Chubb. I'm particularly grateful of the additional time I've spent around the labs of Don and Greg, imbibing their critical analysis and fearless eclecticism. Other faculty members I'd like to thank for their time and guidance include Louis Narens, Kimberly Jameson, Ted Wright and Jeff Krichmar. Reaching way back to my time as an undergraduate, I would like to extend a sincere thanks to Peter Ross and Nancy Alvarado for being the first to ignite my interest in the study of the mind.

I'm very grateful to my more senior collaborators John Pyles and Javier Garcia for generously sharing their many talents and helpful advice on several of the projects presented here. I am also thankful for all the amazing labmates and friends I've had the pleasure to get to know here at UCI, especially Andrew Burton, Lisa Harvey and Xiaojue Zhou.

It would be a huge mistake to not express my gratitude to one of the most important groups of all - the (approximately) 200 research participants who patiently offered their time, voices, and/or BOLD activity to make the studies in this dissertation possible. I can't thank you all by name but hopefully it's the thought that counts.

I also warmly thank John Sommerhauser and the Departments of Cognitive Science, Education, and Philosophy for their support, both financial and otherwise.

Last but not least, I'd like to thank my wonderful partner for more than ten years, Kate Crocker, for all her love and support. I can't wait to join her in our new home state of New Hampshire.

VITA

Daniel A. Stehr

EDUCATION

Doctor of Philosophy in Psychology **2020**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Philosophy **2012**
California Polytechnic University, Pomona *Pomona, CA*

RESEARCH EXPERIENCE

Visual Perception and Neuroimaging Lab **2015–2020**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2013–2020**
University of California, Irvine *Irvine, CA*

JOURNAL PUBLICATIONS

Stehr, D. A., Zhou, X., Tisby, M., Hwu, P. T., Pyles, J. A., Grossman, E. D. (*under review*). Top-down attention guidance shapes action perception in the pSTS. *Cerebral Cortex*.

Stehr, D. A., Hickok, G., Ferguson, S. F., Grossman, E. D. (*under review*). Examining vocal attractiveness through measures of articulatory working space. *Journal of the Acoustical Society of America*.

Stehr, D. A., Pyles, J. A., Garcia, J. O., Grossman, E. D. (*in preparation*). The impact of trial averaging, mean centering, cost tuning, and data cleaning on multivariate pattern analyses using least squares separate (LSS) beta series. *NeuroImage*.

CONFERENCE PRESENTATIONS

Top-down attention guidance shapes action perception in the pSTS **Oct 2019**
Society for Neuroscience conference in Chicago, IL

ABSTRACT OF THE DISSERTATION

Perception of socially-relevant cues from face and body movements: behavioral and neural investigations

By

Daniel A. Stehr

Doctor of Philosophy in Psychology

University of California, Irvine, 2020

Emily D. Grossman, Chair

People are highly skilled at extracting socially-relevant information from the movements of others. The primary human movements analyzed here involve movements of the articulatory organs (which produce speech sounds capable of transmitting a wealth of talker-specific characteristics) and movements of the entire body (from which we regularly and rapidly infer others' goals and intentions). In the first study, I examine how acoustic correlates of articulatory kinematics shape perceived *attractiveness* of the voice. I test the hypotheses that spectral and temporal correlates of precise articulation are relevant predictors of vocal attractiveness through the roles of sexual dimorphism and processing fluency accounts of preferences. In a sample of talkers producing vowels in carrier words, I find that a high proportion of variance in voice preference ratings for female talkers can be explained by measures related to the acoustic-phonetic distinctiveness of speech. The next study shifts focus to brain regions encoding representations of actions performed by human bodies. Contemporary models of action observation now posit a special role for the posterior superior temporal sulcus (pSTS) in integrating low-level perceptual cues with top-down influences of attention. This implies that action representations in the pSTS are not immutable but instead are dynamic and context-dependent. Multivariate pattern analysis (MVPA) evaluated how task demands shape the specific information in the pSTS during action observation. The

statistical structure of multivariate patterns in the pSTS was found to be highly susceptible to feature-based attention, revealing that the pSTS plays an important intermediary role at the interstices of bottom-up and top-down cues. The last chapter serves as an important guide for researchers seeking to optimize experimental design, data preprocessing and machine learning parameters for rapid event-related MVPA. Here, I evaluate the independent and joint effects of four methodological data processing choices aimed at reducing the effects of trial-, voxel-, scan-, or motion-related noise sources. Two of these choices in particular interacted to produce large increases in classifier performance in cases where there is true signal present, a finding which is consistent across both real and simulated datasets.

Chapter 1

Introduction

Human observers are expertly adept at extracting socially-relevant information from the movements of others. From gestures of the face and body, we regularly and rapidly infer others' goals and intentions, an ability laying the bedrock of the social brain. Furthermore, as a distinctly communicative species, we devote significant attention to movements of the lips and tongue - some of the fastest and most precise human actions possible - which give rise to audible noises capable of transmitting a wealth of information, including not just linguistic content but also health, emotional state, and reproductive fitness among others.

In the first study, I examine how acoustic correlates of articulatory kinematics shape a specific kind of inter-personal impression, that of the *attractiveness* of the voice. The voice is part and parcel of the constellation of cues responsible for driving mate selection, with the vocal apparatus being one of the most highly sexually dimorphic traits in humans. This is interesting because sexual dimorphic traits are often strong predictors of attractiveness by association with sexual selection pressures. More broadly, a bias known as the Halo effect (Dion et al., 1972) demonstrates that one's attractiveness has a direct impact on one's social and professional success. One often overlooked gender difference in speech centers on the

acoustic-phonetic distinctiveness of speech sounds, with females producing speech that is overall more intelligible and acoustically distributed than that of males (Yoho et al., 2019; Hillenbrand et al., 1995; Bradlow et al., 1996). Therefore, I test the hypothesis that temporal and spectral correlates of articulatory precision are relevant predictors of vocal attractiveness. In addition to the potential relevance from sexual dimorphism, this hypothesis follows from theories stating that we derive greater aesthetic response to objects capable of being processed more easily (see Reber et al., 2004). In a sample of talkers producing vowels in carrier words, I find that a high proportion of variance in voice preference ratings for female talkers can be explained by measures related to the acoustic distinctiveness of speech.

The next study investigates socially relevant movements on a much larger scale - attending to the actions of human bodies. Contemporary models of action observation now posit a special role for the posterior superior temporal sulcus (pSTS) in integrating low-level perceptual cues with top-down influences of attention (Patel et al., 2019). This implies that action representations in the pSTS are not immutable but instead are dynamic and context-dependent. Though a handful of fMRI studies have empirically demonstrated that task demands - such as instructing participants to attend to social versus spatial aspects of movement (Tavares et al., 2008a) - modulate activity in the pSTS, these studies are currently limited to univariate mapping approaches; that is, analyzing voxels individually or averaging activity across regions of interest (ROIs). This is problematic because the brain is now best thought to encode complex stimuli in high dimensional representational spaces supported by the collective effort of whole populations of distributed neurons. Consequently, we used multivariate pattern analysis (MVPA) to investigate how task demands shape the specific information in the pSTS (and other nodes of the Action Observation Network) while participants watched action vignettes. We found the statistical structure of multivariate patterns in the pSTS to be highly susceptible to feature-based attention revealing the pSTS plays an important role at the interstices of bottom-up and top-down cues.

Finally, in the last chapter I investigate the relative impact of certain methodological choices aimed at enhancing the discriminability of multivariate activation patterns of the sort analyzed in Chapter 3. Despite the growing appeal of multivariate decoding analyses for exploring the rich multidimensional geometry of mental states, these analyses have been accompanied by a proliferation of methodological choices confronting researchers. In both real and simulated fMRI data, I evaluate the independent and joint effects of four methodological data processing choices aimed at reducing the effects of trial-, voxel-, scan-, or motion-related sources of noise. Two of these choices in particular interacted to produce large increases in classifier performance in cases where there was true signal present and this was consistent across both real and simulated datasets. This final chapter serves as an important guide for researchers seeking to optimize experimental design, data preprocessing and machine learning parameters for rapid event-related MVPA.

Chapter 2

Examining vocal attractiveness through measures of articulatory working space

2.1 Introduction

An attractive-sounding voice bears a host of important social implications for the talker. In a professional context, an appealing voice can be a powerfully advantageous "tool of the trade" for many workers - such as educators, politicians, health professionals, and salespeople - who depend on the use of their voice to inform and persuade others (Titze, 1989). Listeners regularly uphold attractiveness stereotypes (Feingold, 1992; Langlois et al., 2000) and are more likely to attribute socially desirable personality traits to talkers with more attractive voices (Zuckerman and Driver, 1989), a phenomenon known to influence important social outcomes ranging from political elections (Klofstad et al., 2012; Tigue et al., 2012; Gregory JR. and Gallagher, 2002) to job interviews (Dion et al., 1972; Schroeder and Epley, 2015). In a per-

sonal context, there is evidence linking the attractiveness of the voice to the attractiveness of the face and body (Feinberg et al., 2005a; Collins and Missing, 2003; Hughes et al., 2004), making the voice part and parcel of the constellation of cues responsible for driving mate selection and many facets of human sexual behavior (Hughes et al., 2004; Hodges-Simeon et al., 2010a).

Previous studies on physical attractiveness have found that traits honestly signalling physical health, reproductive fitness or membership in a community contribute to judgments of attractiveness (Grammer et al., 2003). Such traits are often highly sexually dimorphic, taken as evidence that males and females differ in part because of gender-specific preferences that promoted reproductive success in the evolutionary past. Quite conspicuously, the human voice is one of the most highly sexually differentiated characteristics there is, with the sex difference in fundamental frequency (F0) - the rate of vibration of the vocal folds during phonation and the acoustic parameter closest to what we perceive as pitch - differing by almost six standard deviations (Puts et al., 2012b). This far exceeds the magnitude of most other commonly studied sexually dimorphic traits such as waist-to-hip ratio, height, weight, and handgrip strength (Puts et al., 2014). The larynx and vocal folds are hormonal target organs, even undergoing histologic changes correlated with cyclic hormone levels (Amir and Biron-Shental, 2004; Abitbol et al., 1999), and therefore vocalizations have the potential to contain acoustic cues that signal biological information relevant to mate selection, such as hormonal profile and reproductive fitness. Not surprisingly, then, previous investigations have linked attractiveness to vocal parameters that exaggerate gender-typical voice features such as F0 (Collins, 2000; Hodges-Simeon et al., 2010b; Feinberg et al., 2005b, 2008a, 2006, 2008b; Collins and Missing, 2003; Jones et al., 2008) and formant (resonant) frequencies (Collins and Missing, 2003; Collins, 2000; Hodges-Simeon et al., 2010b; Pisanski et al., 2014; Sell et al., 2010).

When individuals perform social evaluations on the basis of vocal cues, they almost

always do so in the context of spoken communication (Puts et al., 2014). Gender differences have also arisen in this vein, such that the speech produced by females is overall more intelligible than that of males, using criteria such as percent words correct (Bradlow et al., 1996; Yoho et al., 2019; Hazan and Markham, 2004) and subjective rating scales (Yoho et al., 2019; Kwon, 2010). Gender differences have also surfaced in studies of clear speech - defined as the style of speech produced when one is prompted to speak as though their conversational partner is either hearing impaired or not a native speaker (Ferguson, 2004). Use of clear speech as opposed to conversational speech has regularly been found to improve intelligibility (Bradlow and Bent, 2002; Bradlow et al., 2003; Ferguson, 2004; Krause and Braida, 2002; Payton et al., 1994; Schum, 1996) - a phenomenon referred to as the clear speech benefit. Females exhibit a stronger clear speech benefit compared to males (Ferguson, 2004; Bradlow and Bent, 2002; Bradlow et al., 2003), corroborating colloquial notions that speaking clearly and carefully is a stereo-typically female trait (Babel et al., 2014; Weirich et al., 2016) and mumbling by contrast is 'macho'-sounding (Heffernan, 2010). Like other sexually dimorphic features of the voice, gender differences in vocal clarity may therefore have arisen through adaptive preferences for either clearer vocal characteristics in females or less clear vocal characteristics in males.

The clarity of speech is known to vary by temporal as well as spectral properties, many of which interact systematically with gender. In the temporal domain, perhaps the most robust difference is that clear speech is slower because of longer and more frequent pauses (Bradlow et al., 2003) and longer vowel durations (Picheny et al., 1986; Moon and Lindblom, 1994; Ferguson and Quene, 2014; Liu et al., 2003; Smiljanic and Bradlow, 2009). In the spectral domain, clearer speech is associated with a larger vowel space area (VSA) (Ferguson and Kewley-Port, 2002, 2007; Bradlow and Bent, 2002; Bradlow et al., 2003; Johnson et al., 1993; Picheny et al., 1986). VSA is a long-standing metric in acoustic phonetic research that quantifies the distinctiveness between vowels in the two primary acoustic dimensions: the first and second formant frequencies (F1 and F2 respectively) that relate to the size

and shape of the cavities created by tongue height (F1) and tongue advancement (F2) (Lee et al., 2016; Whitfield et al., 2018). When instructed to speak clearly, talkers exhibit further articulatory excursions, presumably in an effort to make their vowel tokens as acoustically distinct as possible, which manifests as a larger vowel space area. Expanded vowel space is also characteristic of infant-directed (ID) speech (Kuhl, 1997; Burnham et al., 2002), a speaking style which may share the overlapping goal of being highly intelligible for the purpose of teaching infants to discriminate the basic phonetic units of their native language. VSA therefore serves as a measure of the clarity or precision of speech, with larger VSA indicating greater acoustic and articulatory distinctiveness among vowels.

VSA is known to vary by gender cross-linguistically in American English (Whiteside, 1996; Neel, 2008; Hillenbrand et al., 1995; Hay et al., 2006), Canadian English (Hagiwara, 2006), French (Hay et al., 2006), German (Simpson and Ericsdotter, 2007; Hay et al., 2006), Hebrew (Amir and Amir, 2007; Most et al., 2000), and Korean (Yang, 1992). Although females have, on average, higher formant frequencies than males (due to differences in vocal tract size), the male-to-female formant scale factor across different vowel categories is non-uniform. The consequence is that vowel productions by females stake out a larger area in acoustic space than those of males (see Figure 2.1).

VSA also varies *within* genders in ways that may be meaningful to attractiveness. Vowel space area in men is negatively related to body height and acoustic correlates of vocal tract length (Kempe et al., 2013), features which have been identified as signals of mate quality and threat potential (Puts et al., 2012a; Bruckert et al., 2006; Evans et al., 2006). In addition, Heffernan (2010) analyzed the speech of eight male American radio disc jockeys and found that, in a sample of on-air speech recordings from the DJ's, vowel space dispersion correlated highly with a set of subjective ratings loading highly on traits of masculinity.

The purpose of the current investigation is to examine the relationship between vocal attractiveness judgments and temporal and spectral correlates of articulatory behavior. We

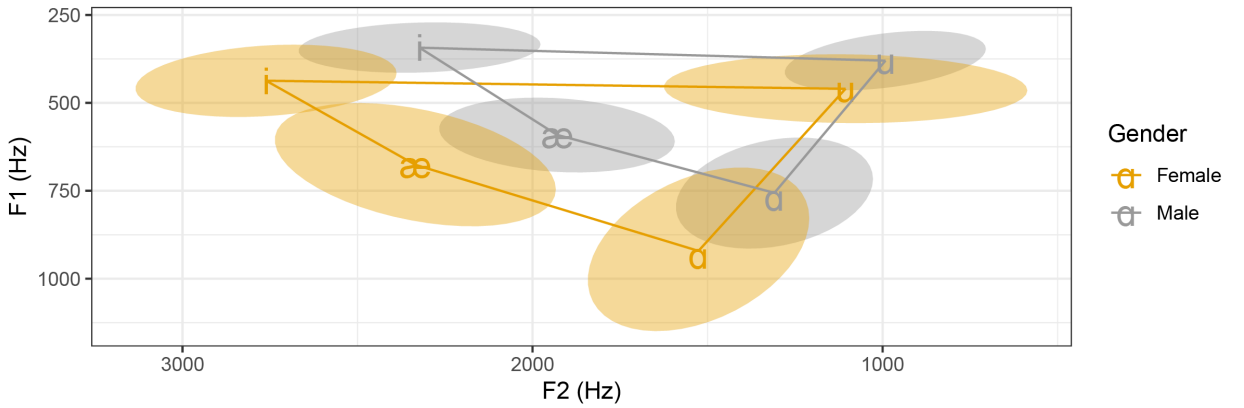

Figure 2.1: Traditionally, vowel space area (VSA) is computed as the 2D area of the triangle or quadrilateral enclosing mean F1 and F2 values measured from the steady state (middle 50%) of either the three point (/i, u, ɑ/) or four corner (/i, u, ɑ, æ/) vowels. When VSA is plotted with both axes reversed and F1 on the y-axis, the result closely resembles the pseudo-articulatory vowel space diagram which mirrors the actual kinematics of the articulators. Shown here is the quadrilateral VSA measured by mean values of the four corner vowels /i, u, ɑ, æ/ from 45 men and 48 women in the Hillenbrand et al. (1995) dataset (freely available at http://homepages.wmich.edu/ hillenbr/)

hypothesize that measures reflecting clearer, more carefully articulated speech will be predictive of attractiveness ratings due to the robust gender differences in acoustic correlates of clearly produced speech - such as VSA and sentence/vowel duration. Second, we predict that, irrespective of talker gender, speech exhibiting *larger* VSA will be more attractive because it serves as an indicator of talker health, with reduced VSA shown in a range of neurological speech motor disorders including Parkinson's Disease (Tjaden and Wilding, 2004; Rusz et al., 2013; Lam and Tjaden, 2016; Hsu et al., 2017; Whitfield and Goberman, 2014; Whitfield and Mehta, 2019), dysarthria (Weismer et al., 2001), and down syndrome (Bunton and Leddy, 2011). Finally, we hypothesize that clearer, more carefully articulated speech is predictive of vocal attractiveness based on perceptual fluency accounts of preferences. By these accounts, the more easily perceivers can encode and analyze a stimulus, the more positive the aesthetic response they derive from it (Reber et al., 2004).

A secondary aim of this study is to examine voice preference ratings in the present context across both word and sentence-length stimuli. To date, most studies on vocal at-

tractiveness have used a relatively narrow range of stimuli consisting mostly of isolated monophthong vowel sounds (e.g., /u/, /a/, or /i/ in English) and occasionally monosyllabic words (for exceptions, see Puts et al. (2006); Lander (2008); Jones et al. (2008); Fischer et al. (2011); Hodges-Simeon et al. (2010a); Puts (2005)). This research decision appears to be prevalent because quantitative metrics of voice quality and formant frequencies are typically easiest to measure from simple vowel sounds and because the use of vowel stimuli enables more tightly-controlled experiments by eliminating contextual factors such as co-articulation, word stress, and semantic meaning.

However, such a restrictive choice of spoken materials may overlook important sources of variation in attractiveness. Proof of this concept comes from Ferdenzi et al. (2013) who found that, on average, word-length stimuli elicited higher attractiveness ratings for the talkers as compared to isolated vowel sounds. The current investigation examines the independent and joint effects of acoustic-phonetic correlates of articulatory behavior on voice preference judgments in both word and sentence-level stimuli.

2.2 Methods

2.2.1 Stimuli

Talkers

A total of forty two talkers (21 females and 21 males) were recruited through the UC Irvine human subjects pool to record speech stimuli for the study. All talkers were native speakers of English and reported normal hearing. Male talkers (mean age = 21.6 years, SD = 3.3, range: 19-33 years) and female talkers (mean age = 22.0 years, SD = 4.7, range: 18-36 years) did not differ significantly in age, $t(35.8)$ = -0.27, p = ns. The majority of the talkers

(86% of males and 90% of females) indicated their hometown was somewhere in California.

Voice recording procedure

Recordings were made inside a large anechoic chamber. In each recording session, the talker stood and spoke into a Røde NT1 cardioid condenser microphone attached to a stand via a shock mount. Talkers maintained a microphone-to-mouth distance of approximately 15 cm and a pop filter was positioned in front of the microphone to attenuate the energy of plosive sounds. The signal from the microphone was digitized at 44.1 kHz and 24-bit quantization by a Focusrite Scarlett 2i2 audio interface connected to a Dell XPS laptop located outside the recording chamber. The microphone gain levels were custom-set for each participant at the beginning of the session while they produced extemporaneous speech. Throughout each session, the experimenter monitored the output of the preamplifier through headphones to verify the quality of the recording.

Speech materials

The recordings made during this phase of the experiment were part of a larger speech corpus comprising five separate speech tasks, two of which were used in the present study: readings of /bVd/ words and corner vowel sentences. The /bVd/ words were constructed by placing ten vowels (/i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u/) into /bVd/ context (Ferguson, 2004; Ferguson and Kewley-Port, 2007; Rogers et al., 2010). Talkers repeated each /bVd/ word four times and were encouraged to produce each word as consistently and evenly as possible. In addition, talkers were recorded speaking four different corner vowel sentences designed to contain at least two tokens of each of the four most peripheral 'corner' vowels (/i, u, ɑ, æ/) in the stressed position. Each talker spoke each sentence three times in a row at a comfortable, conversational pace. The order of /bVd/ words and corner vowel sentences was randomized

for each talker.

Post-processing of speech recordings

All stimuli were extracted from the complete recording session and post-processed using Reaper v5. First, the peak amplitude of each recording was normalized to 0 dB full scale (dBFS). An audio effects chain was then applied to each talker's set of sentences, including: a limiter (to reduce signal peaking from plosives), a compressor (to minimize differences in dynamic volume changes across subjects), and a de-esser (to remove harsh sounds from high frequency sibilants). Finally, the amplitude of each recorded word/sentence was normalized (using MATLAB version R2018b) relative to a signal consisting of only the concatenated vocalic portions of speech to an RMS amplitude of -25 dBFS.

2.2.2 Acoustic Measures

Formant extraction

Continuous F1 and F2 trajectories for vocalic portions of /bVd/ words and corner vowel sentences were estimated using Linear Predictive Coding (LPC) analysis in Praat (Burg method, time step = 1 ms, window length = 25 ms, pre-emphasis = 50 Hz). LPC parameters (maximum formant frequency and number of LPC coefficients) were adjusted manually for each vocalic interval based on visual inspection of the formant trajectories overlaid on the spectographic display using a custom application in R using the Shiny (Chang et al., 2018) and PraatR (Albin, 2016) packages.

Following formant extraction, for each F1 and F2 sample the logarithmic power spectral density of the speech signal (in dB/Hz relative to 2×10^{-5} Pa) was averaged across three frequency bins (binwidth = 43 Hz) and compared. Timepoints where the normalized F1/F2

power ratio exceeded $\pm$ 0.35 were discarded to ensure that the analysis only included samples where F1 and F2 were present with roughly equivalent power. Local outliers in both F1 and F2 time series were identified and removed using the median absolute deviation (MAD, cutoff = 6) calculated over a sliding window of length 30 msecs and missing datapoints were interpolated using a method based on discrete cosine transforms (Wang et al., 2012). The resulting timeseries were finally low-pass zero-phase filtered using a second order Butterworth filter with cutoff frequency of 15 Hz to minimize start-up and ending transients.

Measures of vowel space size

Spectral density The underlying probability density of each talker's formant samples over the entire F2 x F1 plane was estimated using 2 dimensional kernel density estimation (KDE) performed in R using the 'ks' package (Duong, 2018)(see Figure 2.2, C and D). A Gaussian kernel was selected with the kernel bandwidth estimated from each talker's formant data so as to minimize the asymptotic mean integrated squared error (AMISE) criterion. The F2 x F1 space was then discretized into a 500 x 500 linear grid and the density of the kernel smoothed data was sampled at each F2 x F1 gridpoint to add a third dimension (spectral density) to the F2 x F1 space. Regions of the resulting space with the highest density represent the probable locations of distinct vowel nuclei.

Convex hull area The probability contours of the upper 10 to 90% highest density regions were then calculated from the F2 x F1 density (see Figure 2.2C). A convex hull algorithm (R, 'sp' package Pebesma and Bivand (2018)) specified the smallest convex polygon wrapping around all points at a given probability density threshold (Figure 2.2D) and the area of each resulting polygon was computed in kHz^2. Similar area measurements based on continuously sampled formant density distributions have been used to characterize differences in hyper versus hypo-articulated speech (Story and Bunton, 2017) and habitual versus clear speech

in patients with and without Parkinson's Disease (Whitfield and Mehta, 2019).

Standardized general variance The standardized generalized variance (SGV) of a p-dimensional random variable is a scalar measure of overall multidimensional scatter (Wilks, 1932, 1960; SenGupta, 1987), and has recently been applied to time-varying formant data to characterize clear speech production for talkers with and without Parkinson's disease (Whitfield and Goberman, 2014, 2017; Whitfield et al., 2018; Whitfield and Mehta, 2019). The SGV of each talker's productions of the four /bVd/ words and corner vowel sentence was computed by taking the square root of the determinant of the F1 and F2 variance-covariance matrix (see Figure 2.2B). This quantity is interpretable as a bivariate standard deviation and has been shown to correlate highly with measures of the vowel convex hull area (Whitfield and Mehta, 2019).

Measures of vowel space shape

The preceding set of measures quantify the overall size of working vowel space most heavily used by a talker. However, as suggested by Story and Bunton (2017), there may be relevant information contained in the particular *shape* of an individual's formant data distribution not captured by gross *size*.

Circularity The convex hull polygons at each probability contour were first converted to binary masks and fitted with an ellipse (see Figure 2.2E). The roundness or 'circularity' of each convex hull was then computed by finding the ratio of the length of the major axis to the length of the minor axis of the best fitting ellipse. As such, values closer to one characterize convex hulls that are more circular and values that substantially deviate from one characterize convex hulls that are more elongated in one dimension.

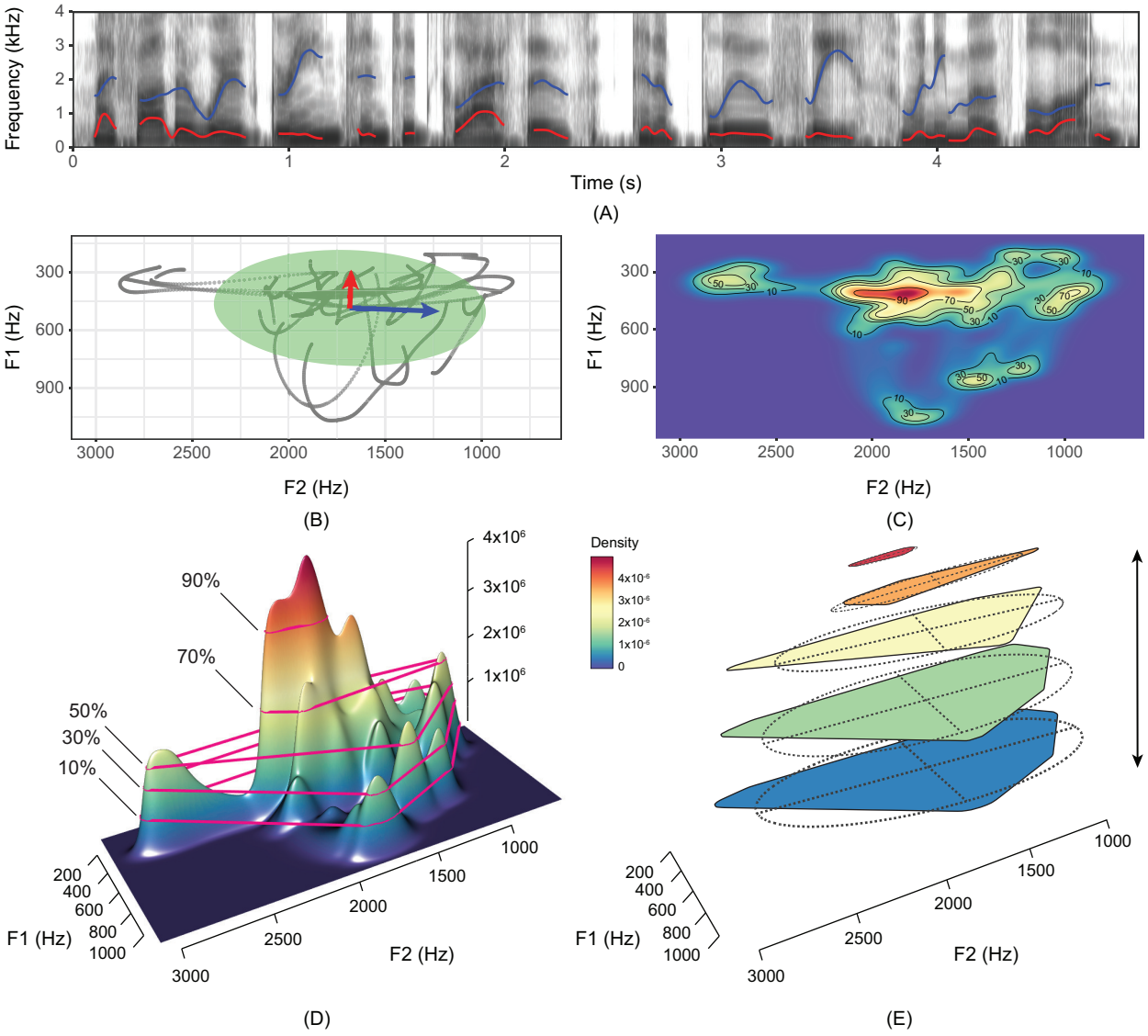

Figure 2.2: Measures of working vowel space size and shape. (A) Processed F1 and F2 timeseries from a single talker speaking the corner vowel sentence, *"My father wears green tinted glasses and a blue fleece when he mows the lawn"* (B) Scatterplot of F1 and F2 samples including error ellipse (1 SD). The red and blue vectors at the center of the ellipse show the directions and magnitudes of greatest variation in the data. (C) Density plot showing 10, 30, 50, 70, and 90 percent probability contours (only 5 out of 81 total contours are shown for simplicity of visualization). (D) Convex hull polygons (pink bands) computed from the formant density distribution at 5 sample probability density contours. (E) An exploded diagram showing the best fitting ellipses, including major and minor axes, computed from convex hull polygons (shaded surfaces).

Orientation The tilt of the convex hull away from the F2 and F1 axes was quantified by finding the angle (in degrees) between the major axis of the fitted ellipse and the F2 (or x) axis. Values ranged from 90 to -90 degrees with 0 degrees representing a convex hull whose major axis is perfectly aligned with the F2 axis.

Blob count Under careful articulation, formants from unique vowel types cluster in different regions of formant space, apparent as distinct 'blobs' in the F2 x F1 formant density distribution (see Figure 2.2C). However, in fast and/or weakly articulated speech, certain vowels may gravitate to more central regions of formant space making what would otherwise be separate 'blobs' indistinguishable from one another. Therefore, in a data-driven manner, we estimated the number of distinct 'blobs' at each probability contour of the formant density distribution. Beginning with a talker's formant density estimate, binary masks were generated quantifying whether each F2 x F1 bin was above or below a given probability threshold. The total number of distinct 'blobs' was then computed using the graph theoretic approach of 8-connected components labeling (MATLAB function 'regionprops') which scans the image and groups pixels into 'blobs' based on their connectivity with other pixels. Counts of distinct 'blobs' were computed at each density threshold for corner vowel sentences only.

Interior angles of /bVd/ vowel quadrilateral The four /bVd/ words stake out a quadrilateral shape in F2 x F1 space, the shape of which varies by talker and can be characterized by four interior angles. This was accomplished by subdividing the /bVd/ vowel quadrilateral into two nested triangular regions and using the Law of Cosines to find the angle (in degrees) of the vertex occupied by each /bVd/ word, as shown in Figure 2.3A.

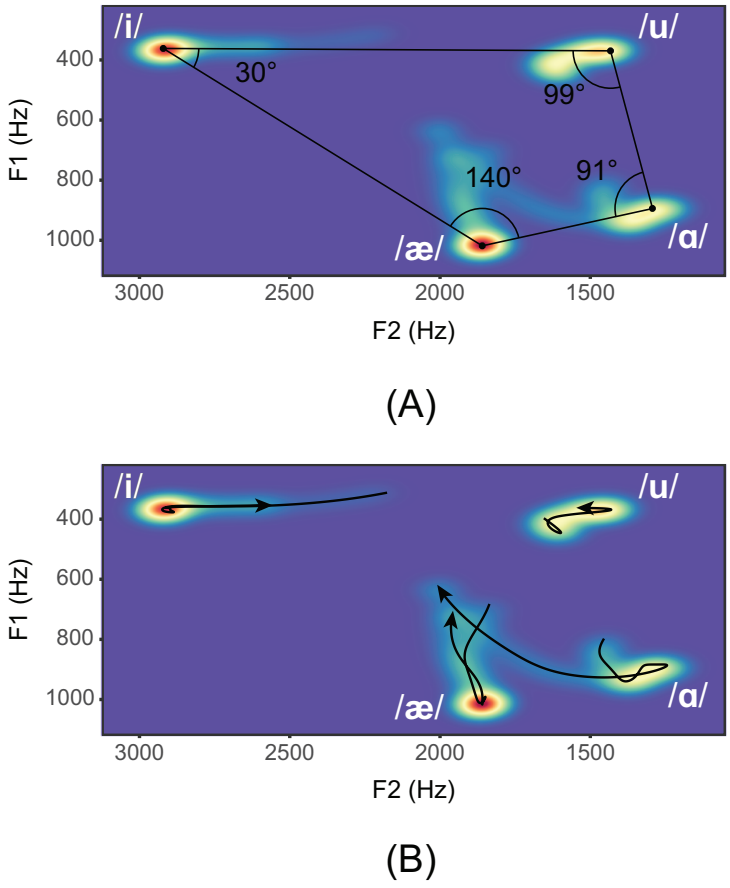

Figure 2.3: (A) The formant density distribution for four /bVd/ words from a single talker with vowel quadrilateral and interior angles superimposed. (B) Time-varying spectral change plotted for each /bVd/ word. Arrowheads indicate offset of each word in time.

Temporal measures of speech

Spectral change over time Time variation in the spectral pattern of each /bVd/ word, denoted by λ, was evaluated by finding the sum of the Euclidean distances between each successive pair of formants, sampled at 5 ms intervals. The total distance traversed in the F2 x F1 plane was then divided by the number of samples to normalize for word duration (see Figure 2.3 B).

Number and duration of stop gaps Each talker's distribution of silent intervals was summarized with the median silent interval duration and total number of silent intervals detected. Silent intervals were identified using Praat with a minimum silent interval duration of 15 ms. Inspection of the distributions of silent intervals for each talker revealed them to be largely unimodal and close to the typical range of articulatory (stop) gaps (Rosen et al., 2010; Whitfield and Gravelin, 2019).

Speech-to-pause ratio Speech-to-Pause ratio was computed for the corner vowel sentences by dividing the summed duration of all sounding intervals by the total duration of each sentence.

Speaking rate Each talker's speaking rate (in syllables per second) was measured by dividing the total number of dictionary syllables in the corner vowel sentence (n=18) by the total sentence duration.

/bVd/ word duration Each /bVd/ word duration was computed as the difference between the offset and onset of voicing determined from the waveform and spectrogram.

Other acoustic measures

Median F0 and interquartile range of F0 were measured in Hertz over the whole utterance produced by each talker in /bVd/ words and sentence conditions (Praat, time step = 0.01 msecs). Pitch floors were 75 and 100 Hz and pitch ceilings were 300 and 600 Hz for men and women respectively.

2.2.3 Listening experiment

Participants

A total of 124 participants were recruited through the UC Irvine human subjects pool to serve as listeners and provide attractiveness ratings for the voice recordings. Separate pools of participants rated the /bVd/ stimuli (N = 64, 32 females) and corner vowel sentence stimuli (N = 60, 30 females). All participants gave their informed consent before beginning the listening experiment and were native speakers of English with no known hearing impairments.

To ensure the attractiveness ratings were not biased by any personal affiliation between listeners and talkers, all listeners first completed an online screening form requiring them to listen to samples of the voices and indicate whether they recognized the voice or not. Subsequently, data from one female participant was excluded.

Procedure

Participants listened to the stimuli at approximately 70 dB SPL in a sound-attenuated booth wearing Sennheiser HD 380 PRO headphones. All stimuli were presented through an Apple MacBook Pro running Matlab R2018b (Mat, 2010) and the PsychToolbox extensions (Brainard, 1997). The study was administered as a paired comparison two-alternative forced

choice design such that, on each trial, speech samples from two talkers were played in succession (500 msecs apart) after which the participant was asked to respond, *"Which talker has the more attractive voice?"* For the /bVd/ stimuli, all four words were concatenated together using one of four different orderings created and administered in a balanced Latin square design with 500 ms of silence inserted between each word.

A complete paired comparison design, using all pair-wise combinations of talkers, would require 420 trials. Therefore, to manage participant fatigue, the number of paired comparisons was reduced by half through the use of an incomplete cyclic design (ICD) which equalizes the frequency with which each stimuli is paired with other stimuli (Burton, 2003; McCormick and Bachus, 1952). In a small pilot study we verified the validity of the ICD by having 7 listeners (2 males, 5 females) complete all 420 trials in the complete design and then extracting a subset of their responses based on 100 psuedorandomly constructed unique ICDs containing 50% of the original responses. Across all listeners, we found very high correlations between attractiveness scores derived from the complete design and incomplete ones ($.86 \leq r \leq .99$). Therefore, all subsequent experiments were administered using an ICD that reduced the number of trials by half for a total of 210 trials.

Scaling of pairwise voice preference judgments

In order to translate the pairwise decisions resulting from our listening study into a continuously ordered vocal 'attractiveness' scale, we applied the Elo rating algorithm (R, "EloChoice package" (Neumann, 2015)). The Elo rating method is a self-correcting system, originally invented to rank chess players, that sequentially updates ability scores for each item/player based on the actual outcome of each trial along with the prior predicted probability of either item/player 'winning'. Separate pools of Elo ratings were generated for each gender of talker and stimulus type (/bVd/ words versus corner vowel sentences). To ensure the sequence of trials did not bias the scores derived by the Elo rating algorithm, Elo ratings were generated

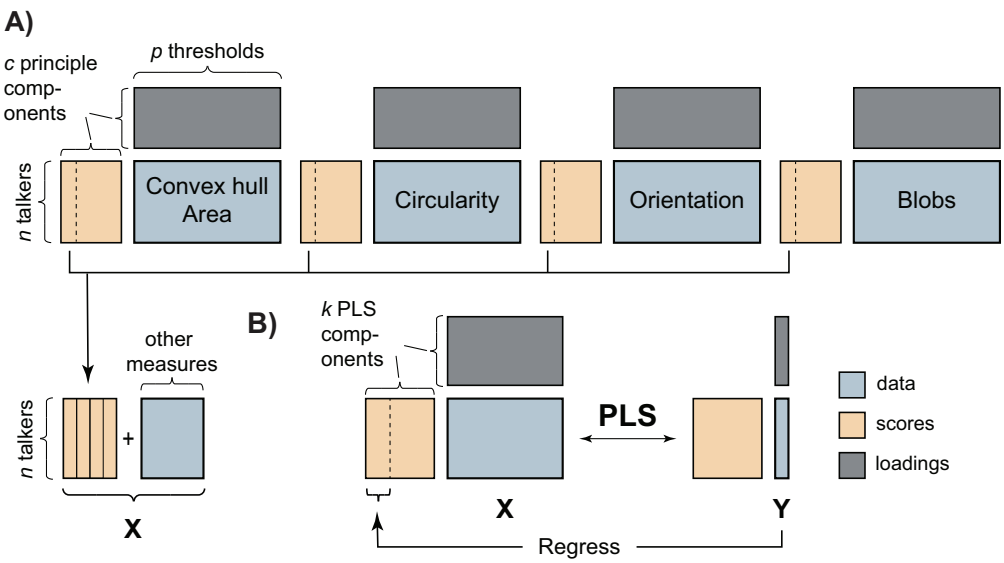

Figure 2.4: Schematic diagram of the two-stage modelling approach evaluating the fundamental relations between acoustic measures and ratings of attractiveness. (A) Dimensionality reduction on blocks of measures using principle components analysis. The optimal number of components from each PCA was chosen visually by looking for an elbow in the cumulative variance explained plot. (B) Partial least squares regression.

for 500 different permutations of experimental trials and then averaged to obtain mean Elo ratings as recommended by Clark et al. (2018).

Agreement among listeners

Each listener's consistency across paired comparison trials was assessed using the weighted Elo consistency index (Clark et al., 2018). Conceptually, the Elo consistency index tracks the frequency with which the outcome of each trial violates the expectation based on the trials that came before it. The weighted form of the Elo consistency index is:

$$R = 1 - \sum_{i=1}^{N} \frac{u_i * w_i}{\sum w}$$

where u is an indicator variable (0 = expectation confirmed, 1 = expectation violated), N is the total number of trials for which an expectation existed, and w is the absolute difference

in preceding Elo scores between talkers in a given trial. The weighted consistency index varies between 0 and 1, with .5 representing chance selection and 1 representing perfect consistency. A weighted Elo consistency index was generated separately for each listener's choices and therefore reflects intra- (as opposed to inter-) rater reliability. For each listener, the weighted consistency index was generated for 500 different permutations of experimental trial orderings and then averaged to ensure that the stimulus order did not unduly influence the consistency measure.

Pearson correlation coefficients were calculated to assess the relationship between ratings from male and female listeners separately for male and female talkers. A Deming regression (which accounts for errors in observations on both variables) was conducted between average ratings from male versus female listeners to identify any systematic differences between the two sets of measurements.

2.2.4 Statistical analyses

To identify the best predictors of vocal attractiveness ratings among our set of acoustic features, we constructed partial least squares regression (PLS-R) models (Vinzi et al., 2010; Geladi and Kowalski, 1986). PLS-R was chosen because it is ideally suited for situations where there are several, highly collinear predictor variables and relatively few observations. PLS methods seek to identify a small number of latent variables (or components) that explain the maximal amount of variance in the predictor variables and the maximal covariance between predictors and responses. In the regression phase, the response variable is regressed not onto the original (highly colinear) measures but onto the first few columns of the PLS scores to generate predictions.

We fit four separate PLS-R models (for each gender of talker in each type of speech material) using the "pls" package in R (Mevik et al., 2019; Mevik and Wehrens, 2007). A two-

stage modelling approach was implemented (see Figure 2.4) such that a principal component analysis (PCA) was first used to reduce the dimensionality of all measures computed across the 81 probability contours of the formant density distributions (separately for convex hull area, circularity, orientation, and blob count). Components from each block of measures explaining the most variance were combined with other unsummarized measures and then submitted as predictors to PLS-R.

For each PLS-R model, we selected the number of PLS components that produced the first local minimum in the root mean square error of prediction (RMSEP) on the basis of leave-one-out cross-validation (LOO-CV). The predictive accuracy of the models was compared using the coefficient of determination (R^2) for the validation results as well as the normalized RMSEP (or NRMSEP) calculated by dividing the RMSEP by the range in attractiveness ratings. The statistical significance of each PLS-R model in predicting left out vocal attractiveness ratings was assessed by permutation resampling. Null hypothesis distributions were constructed by randomly permuting the attractiveness ratings then fitting a PLS regression model and extracting the LOO-CV RMSEP amount. This was repeated 10,000 times for each model and a p-value (P_{perm}) was calculated as the proportion of samples in which the RMSEP from the unpermuted PLS-R model exceeded the RMSEP from the null distribution.

To examine how the acoustic measures themselves vary as a function of the talker's gender and the type of speech stimulus produced, 8 linear-mixed effects models (LMMs) were constructed using as dependent variables the first principle components of convex hull area, circularity, and orientation measures as well as SGV, median F0 and the interquartile ranges of F0, F1, and F2. Talkers were modeled with a random intercept to resolve the non-independence resulting from repeated measurements. Fixed factors were entered sequentially to determine which ones improved the fit of the model best and included (in order) talker gender, stimulus type (/bVd/ words versus corner vowel sentences) as well as the stimulus

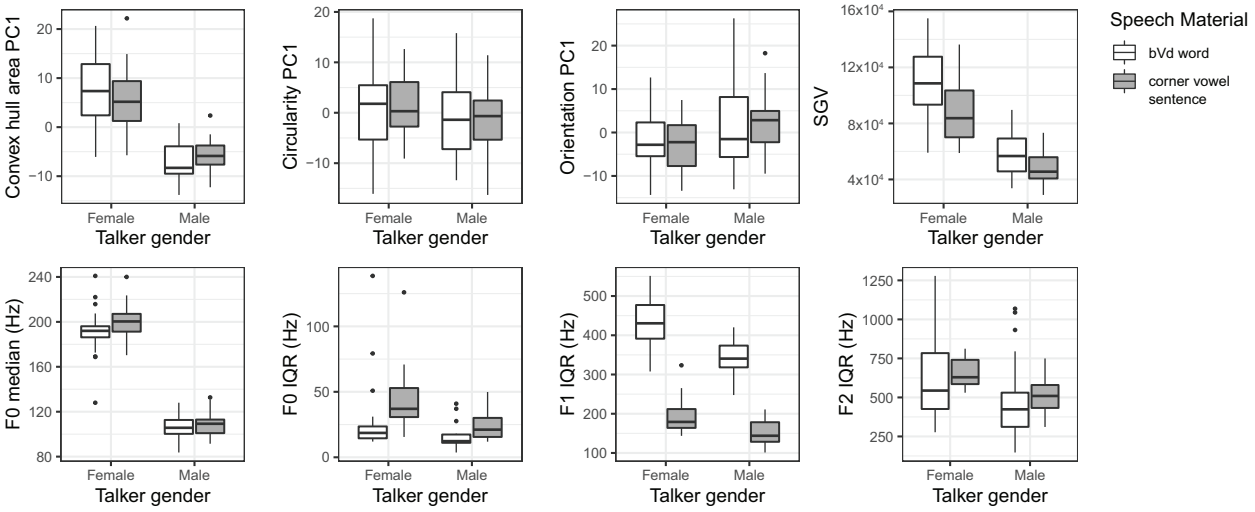

Figure 2.5: Distributions of acoustic measures by talker gender and speech material.

x gender interaction. P values were obtained using Likelihood Ratio Tests comparing each model to reduced models lacking the variable (or interaction) in question. All analyses were performed using the "lme4" package (Bates et al., 2019) implemented in R.

2.3 Results

2.3.1 Effect of gender and stimulus type on acoustic measures

Figure 2.5 shows distributions of the acoustic measures broken down by talker gender and type of speech material produced. The LMM for PC1 of the convex hull area measures revealed a main effect of gender, $\chi^2(5) = 45.31, p < .001$, with males on average 14.19 units lower than females (SE = 1.64). An inspection of the loadings of PC1 (see Figure 2.6) revealed high loadings on all contour levels for the /bVd/ stimuli and most contour levels below 75% probability density for the corner vowel sentences. Therefore, the main effect of gender for PC1 of the convex hull area measure suggests that, across many probability contour levels, speech from males encloses a much smaller area of working vowel space than

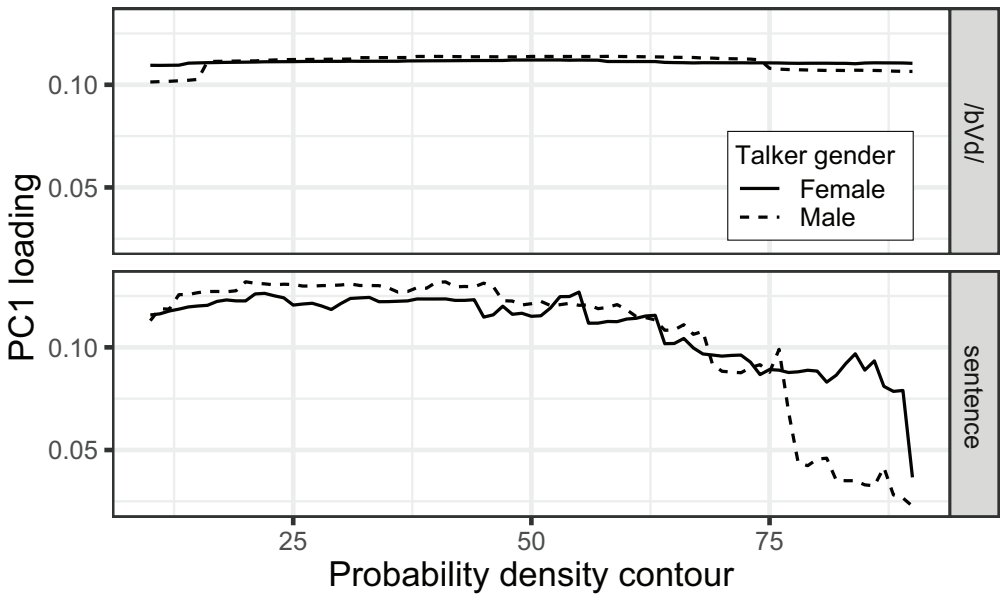

Figure 2.6: Loadings from the first principle component of convex hull areas computed at 10 to 90% probability density contours of the formant density distribution.

that of females. The circularity and orientation measures did not reveal any significant main effects of gender or speech material or the gender x speech material interaction.

The LMM for the standardized general variance (SGV) of formant samples revealed main effects of stimulus type, $\chi^2(4) = 23.81, p < .001$, and gender, $\chi^2(4) = 42.27, p < .001$, and a significant stimulus x gender interaction, $\chi^2(4) = 4.25, p = .04$. The formants from male talkers exhibited an SGV that was on average 49,326.4 Hz^2 (SE = 5,737.2) lower than that of female talkers, and the /bVd/ words had a significantly lower SGV as compared to sentence stimuli for both genders (a difference of 20,388.4 Hz^2, SE = 3,575.8). The decrease in SGV from the sentence to /bVd/ word stimuli is likely attributable to the fact that the sentence stimuli had more variability in formant samples within the central region of vowel space. Further inspection shows that this decrease between speaking conditions was significantly larger for female talkers than it was for male talkers (a difference of 10,777.4 Hz^2, SE = 5,057.0).

The two LMMs with interquartile range of F1 and and interquartile range of F2 as

dependent variables showed a significant main effect of gender, $F1_{IQR}$: $\chi^2(4) = 26.58, p < .001$; $F2_{IQR}$: $\chi^2(4) = 8.07, p = .005$, with male talkers having interquartile ranges that were 92.13 Hz (SE = 13.92) and 164.42 Hz (SE = 65.53) smaller for F1 and F2, respectively, than female talkers. For the interquartile range of F1, this main effect of gender was qualified by a significant stimulus x gender interaction, $\chi^2(6) = 7.50, p = .006$. While both genders exhibited a decrease in F1 interquartile range moving from word to sentence productions, this decrease was significantly smaller for female talkers compared to male talkers.

The LMMs with F0 median and F0 interquartile range as dependent variables both showed significant main effects of speech material, $F0_{median}$: $\chi^2(4) = 7.51, p = .006$; $F0_{IQR}$: $\chi^2(4) = 28.97, p < .001$, as well as gender, $F0_{median}$: $\chi^2(5) = 100.55, p < .001$; $F0_{IQR}$: $\chi^2(5) = 7.32, p = .007$. Males spoke with a lower median F0 and smaller F0 interquartile range than female talkers, and the sentences elicited a higher median F0 and larger F0 interquartile range than the /bVd/ words. The speech material x gender interaction was not significant for either median F0 or F0 interquartile range.

Gender differences in measures computed only for the /bVd/ stimuli were evaluated through Welch two sample t-tests. Compared to male talkers, female talkers made significantly longer productions of /bid/, $t(36.08) = 3.55$, $p = .001$, $r = .51$, and /bɑd/, $t(37.59) = 2.22$, $p = .03$, $r = .34$. Females also exhibited greater λ for /bid/, $t(30.28) = 5.08$, $p < .001$, $r = .68$, /bɑd/, $t(39.08) = 2.18$, $p = .035$, $r = .33$, and /bad/, $t(39.69) = 2.84$, $p = .007$, $r = .41$. Finally, in the context of the vowel quadrilateral, female talkers exhibited a greater interior angle at the vertex occupied by /bɑd/, $t(39.31) = 2.33$, $p = .025$, $r = .35$, and a significantly smaller interior angle at the vertex located at /bad/ ($t(39.81) = -2.36$, $p = .023$, $r = .35$). A separate analysis revealed a strong correlation between the traditional quadrilateral vowel space area (qVSA) computed on the basis of the /bVd/ words and both SGV and PC1 of convex hull area, $.81 \leq r's \leq .92$.

Among measures unique to sentence stimuli (blob count, median pause length, and

syllables per second), only syllables per second was found to differ significantly between genders, $t(39.78)$ = -2.90, p = .006, r = .42, with females producing fewer syllables per second (i.e. talking at a slower rate) than males.

2.3.2 Reliability of listener ratings of attractiveness

Intra-rater reliability, assessed by weighted Elo consistency indices, are shown in Figure 2.7A. Overall individual consistency was high for both male and female listeners' ratings of male and female talkers in both speaking conditions. An LMM with mean Elo rating as dependent variable and a listener-specific random intercept revealed significant main effects of listener gender, $\chi^2(4) = 9.05, p = .003$, and talker gender, $\chi^2(5) = 10.05, p = .002$. Female listeners had .04 higher consistency indices (SE = 0.01) than male listeners irrespective of the gender of talker they were listening to. Female and male listeners had .03 higher consistency indices (SE = 0.01) when rating female talkers compared to male talkers. The listener gender x speech material interaction trended towards, but did not reach, significance, $\chi^2(8) = 3.57, p = .059$, with female listeners, on average, having higher consistency indices than male listeners while rating fellow female talkers in sentence compared to /bVd/ word productions.

2.3.3 Predicting attractiveness ratings

We used partial least squares regression (PLS-R) models to relate attractiveness ratings derived from listeners' paired comparison judgments to the entire set of acoustic measures reflecting articulatory behavior. The results from the four separate PLS regression models are summarized in Table 2.1.

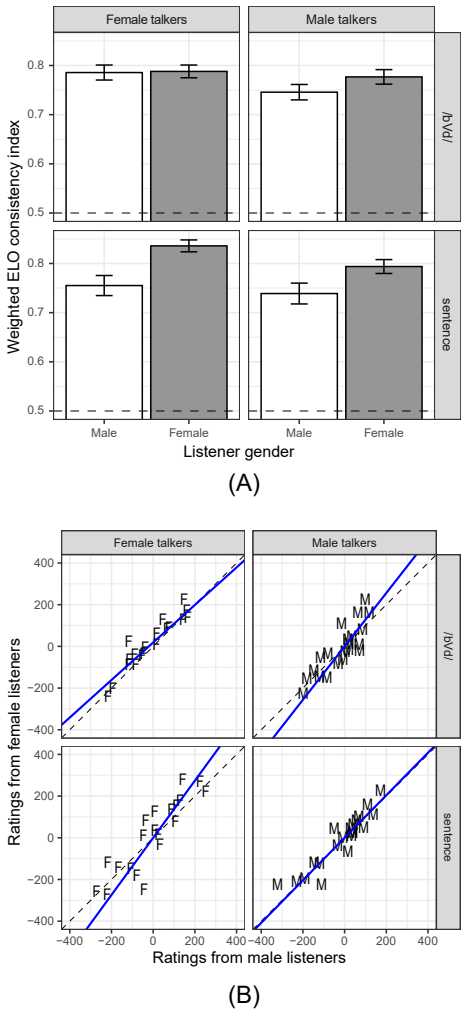

Figure 2.7: (A) Weighted Elo consistency indices broken down by listener gender, talker gender, and type of speech material produced.(B) Agreement between ratings from male and female listeners. Blue lines represent Deming regression lines of best fit.

		Calibration			Cross-Validation			
dataset	NLV	RMSEP	NRMSEP	R^2	RMSEP	NRMSEP	R^2	P_{perm}
/bVd/ words								
Female Talkers	4	44.57	4.98	0.94	93.77	10.47	0.73	0.000
Male Talkers	1	78.86	20.78	0.38	103.01	27.15	-0.06	0.127
Corner vowel sentence								
Female Talkers	1	124.87	22.27	0.40	169.61	30.25	-0.10	0.192
Male Talkers	1	74.88	16.53	0.62	145.35	32.08	-0.43	0.569

Table 2.1: Results from PLS-R models predicting vocal attractiveness ratings. NLV = number of latent variables (or components) chosen based on minimizing cross-validated prediction error. RMSEP = root mean squared error of prediction. NRMSEP = root mean squared error of prediction normalized by the range in mean Elo ratings expressed as a percentage.

/bVd/ word stimuli

For female talkers producing /bVd/ words, the PLS-R model reached the first local minimum in RMSEP with four components (RMSEP = 93.77, or 10.47% of the range in mean Elo ratings). The coefficient of determination calculated on the held out validation data showed the predictive accuracy of the four component model to be extremely strong ($r^2 = .73$). A permutation test revealed the mean of the resulting *null* distribution to be 3.12 standard deviations away from the true LOO-CV RMSEP amount for the four-component model ($p < 0.001$).

Figure 2.8 displays the loading weights of the first two PLS components on the original acoustic measures for female talkers producing /bVd/ words. The first component, explaining the largest proportion of variance in attractiveness ratings ($r^2 = .43$), loaded highly on several measures of vowel working space, including qVSA, SGV, PC1 of the convex hull area, and circularity of the convex hull. All these measures loaded *positively* on the first component indicating that female talkers who produced speech with formants that were more dispersed in F2 x F1 space elicited higher vocal attractiveness ratings.

The second PLS component explained an additional 18% of the variance in attrac-

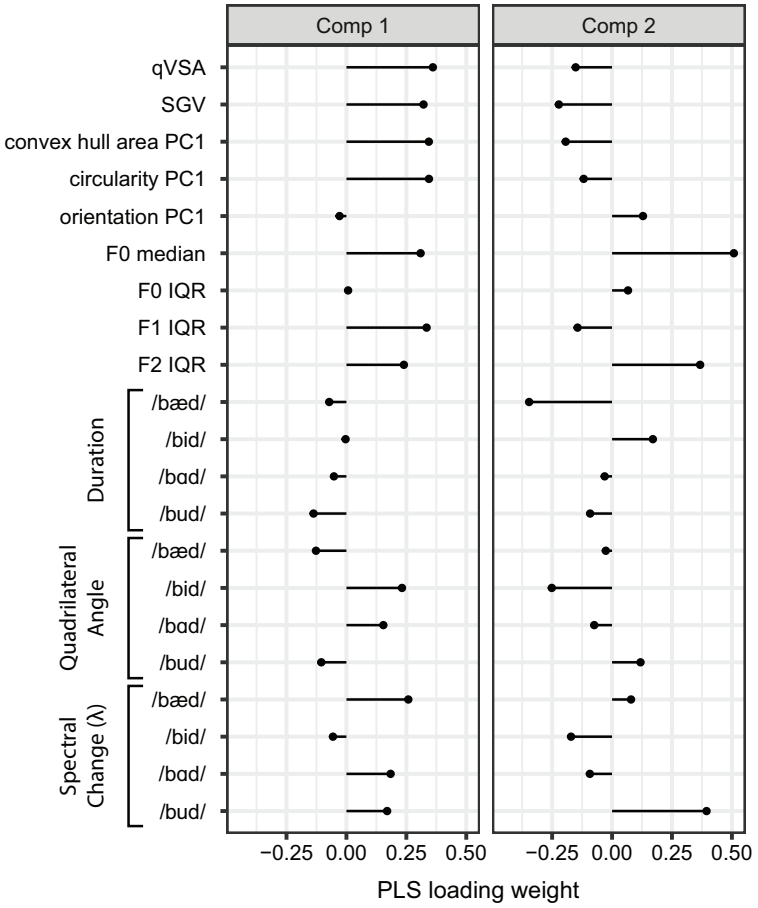

Figure 2.8: Weights from the first two PLS-R components loading on the original acoustic measures from female talkers producing /bVd/ words.

tiveness ratings (cumulative $r^2 = .61$). This component was dominated by large positive loadings on median F0, interquartile range of F2, and λ of /bud/, meaning that females who spoke with higher F0, more range in F2 and greater spectral change in the back vowel /bud/ elicited higher attractiveness ratings. In addition, the second component had a large negative loading on the duration of /bæd/ such that shorter productions of /bæd/ received higher attractiveness ratings. The third and fourth PLS component explained relatively little additional variance in attractiveness ratings (9% and 3% respectively) and therefore are not be discussed further.

For male talkers producing /bVd/ words, the first local minimum in RMSEP occurred for the one component model (RMSEP = 103.01 or 27.15% of the mean Elo rating), however, this was only marginally better than the simplest model containing only an intercept. The permutation test revealed that the one component model for male talkers in /bVd/ context did not predict better than chance, $p = 0.127$.

Corner vowel sentence stimuli

PLS-R models built on the sentence stimuli for both male and female talkers failed to produce any local minimum in RMSEP for any number of components; in all models, the simplest intercept-only model outperformed all others.

2.4 Discussion

The purpose of this study was to examine vocal attractiveness through several measures of articulatory behavior across both word and sentence length stimuli. To this end, samples of four /bVd/ words containing the four corner vowels and one corner vowel sentence (a sentence containing at least two tokens of each corner vowel) were collected from adult

talkers and analyzed acoustically. Measures of the size and shape of working vowel space computed from continuously sampled formant trajectories, along with measures of speech timing, were used to predict vocal attractiveness ratings from a separate group of listeners. Partial least squares regression (PLS-R) models were used to identify which measures most strongly predicted attractiveness ratings in a cross-validated sample of talkers.

For females speaking /bVd/ words, a high percentage of the variance in attractiveness ratings was explained by the first PLS component, which loaded highly on several measures of working vowel space size and shape. Female talkers producing speech with greater convex hull area, convex hull circularity, standardized general variance (SGV), quadrilateral VSA (qVSA), and F1 and F2 range elicited higher attractiveness ratings. The second PLS component loaded mainly on median F0, replicating previous findings showing F0 to relate positively to perceptions of vocal attractiveness in female talkers (Feinberg et al., 2008b; Collins and Missing, 2003; Jones et al., 2008).

The measures of vowel space that predicted female vocal attractiveness reflect, in part, the extent of kinematic displacements during articulation, with the extent of articulatory motion for more dispersed formants in F2 x F1 space contributing to higher acoustic-phonetic distinctiveness among vowel categories. As evidence for this, previous studies of vowel formants have reported the convex hull area (Story and Bunton, 2017; Whitfield and Mehta, 2019), SGV of F1 and F2 (Whitfield and Goberman, 2017; Whitfield and Mehta, 2019), quadrilateral VSA (Lam et al., 2012), and range in F1 and F2 (Ferguson and Quené, 2014; Lam et al., 2012; Bradlow et al., 2003) to be acoustic correlates of clearly produced speech. Our findings extend these features of vowel articulation to perceptions of vocal attractiveness in female talkers.

The origins of the gender differences in VSA are partially, though not entirely, due to physical differences in vocal tract anatomy. VSA is inversely related to physical vocal tract length, with the VSA of children and adolescents declining gradually with age in parallel with

developmental increases in vocal tract length (Flipsen and Lee, 2012; McGowan et al., 2014). In adulthood, the articulatory movements of female German speakers display significantly larger jaw angle openings (and therefor higher F1) than males for low, accented vowels shedding light on a possible physiological component explaining sex-specific differences in vowel distinctiveness (Weirich et al., 2016). On the other hand, preschool children, who do not significantly differ in gross physical measurements, already exhibit gender-specific differences in formant values (Perry et al., 2001). Talkers may therefore adopt speech patterns which either affirm or contrast biologically constrained sex differences as a form of identity construction.

Preferences for larger working vowel space size in the speech of female talkers may also have arisen due to the potentially adaptive implications for infant language acquisition. Acoustic correlates of clearly produced speech such as vowel space size (Kuhl, 1997; Liu et al., 2003; Uther et al., 2007) and a slow speaking tempo (Fernald and Simon, 1984) have been found to be exaggerated in infant directed speech which infants prefer over adult-directed speech. Furthermore, Liu et al. (2003) demonstrated a significant positive correlation between mothers' vowel space area and infants' speech discrimination performance. Preferences for larger working vowel spaces may therefore have arisen because the enhanced acoustic distinctiveness of basic phonetic units accompanying vowel space expansion facilitates infant language acquisition.

The relative ease with which a talker's phonetic units can be distinguished may also be a relevant predictor of vocal attractiveness based on perceptual fluency accounts of preferences. By these accounts, the more easily perceivers can encode and analyze an acoustic object, the more positive the aesthetic response they derive from it (for a review see Reber et al. (2004)).

Another consideration of the current results stems from the role of emotionally valenced facial movements. Smiling behavior, which shortens the length of the vocal tract, has the

effect of expanding the distribution of formants along the F2 dimension (Tartter et al., 1994). Such signs of positive social interest can be a strong driver of perceptions of vocal attractiveness (Jones et al., 2008). Although we instructed our talkers to speak as neutrally as possible, it is impossible to say with complete certainty that the speech we collected entirely lacked all affective coloration. The first PLS component did in fact load highly on the interquartile range of F2 suggesting that an"auditory smile" in some of the /bVd/ productions may have partially driven responses.

Finally, reduced working vowel space has been tied to a range of neurological speech motor disorders including Parkinson's Disease (Tjaden and Wilding, 2004; Rusz et al., 2013; Lam and Tjaden, 2016; Hsu et al., 2017; Whitfield and Goberman, 2014; Whitfield and Mehta, 2019), dysarthria (Weismer et al., 2001), and down syndrome (Bunton and Leddy, 2011). Therefore, preferences for speech exhibiting a larger range of articulatory motion may reflect an overall preference for talker health.

Among the temporal measures computed for each /bVd/ word, greater time-varying spectral change (λ) in several words (/bæd/, /bɑd/, and /bud/) predicted higher attractiveness in female talkers. Greater λ was also linked to overall vowel duration for /bæd/ and /bɑd/. A plausible interpretation of this finding is that talkers who speak more slowly are able to make more precise articulatory excursions in an effort to avoid undershooting their articulatory targets and, for certain vowels, this results in greater dynamic formant movement over time. In support of this, previous studies have found greater dynamic formant movement for certain vowels produced in clear versus conversational speaking styles (Ferguson and Kewley-Port, 2002, 2007; Ferguson and Quene, 2014). These findings are in line with our previous hypotheses linking clearly articulated speech to attractiveness in female talkers.

Contrary to the results for female talkers, none of the PLS components loading on acoustic measures of male talkers were capable of reliably predicting attractiveness ratings

in /bVd/ words. This was surprising given that measures of vowel space size, in males, have been negatively linked to physical indicators of male mate quality such as height and measures of overall vocal tract length (Kempe et al., 2013). This finding in the context of articulatory measures contrasts with previous studies finding that women prefer more masculine vocal characteristics in males such as lower F0 (Collins, 2000; Feinberg et al., 2005b; Hodges-Simeon et al., 2010b; Riding et al., 2006) and lower formant dispersion (Feinberg et al., 2005b; Hodges-Simeon et al., 2010b).

Another aim of the current investigation was to study the relationship between attractiveness and measures of articulation in a more ecological corpus, including sentences of connected speech. Traditionally, working vowel space has been measured based on point estimates of F1 and F2 measured at the vocalic midpoint (or steady state) of target vowels, thus requiring laborious hand segmentation of simple stimuli. However, recent techniques enabling quantification of vowel formant space based on the density distribution of continuously sampled F1 and F2 trajectories, makes it possible to estimate the size of working vowel space for far more complex vocal productions such as those found in connected speech. Although none of the vowel space density measures computed on the formant density distributions of word versus sentence stimuli differed significantly, we were unable to reliably predict attractiveness ratings of sentence length stimuli. This suggests that the failure to predict attractiveness from the sentence stimuli is entirely attributable to other differences in acoustic variables, not quantified in the two types of speech material collected.

2.4.1 Listener ratings by gender

Given that males and females have undergone different selection pressures concerning mate selection, we were careful to consider how attractiveness ratings differed between male and female listeners for the male and female talkers. Previous studies on vocal attractiveness

have found that although males and females largely agree with each other when rating the attractiveness of female talkers, ultimately there is less agreement between genders when rating male talkers, with male listeners tending to give fellow males uniformly lower attractiveness ratings (Babel et al., 2014; Pisanski and Rendall, 2011). Babel et al. (2014) suggests this may be partially due to inexperience among males at ranking other males or a reluctance due to taboos surrounding masculinity and perceived sexuality. With this in mind, we implemented this study using a paired comparison design instead of the more common method of collecting Likert-type ratings. If the disproportionately lower ratings of male talkers by male listeners as previously reported is due to inexperience among males at the task, then this would be reflected in a lower Elo consistency index among male listeners relative to female listeners. Conversely, if males rated other males less attractive because of cultural stereotypes, then the forced choice nature of the paired comparison design will elicit a more honest response leaving reliability unaffected.

Results showed strong correlations between male and female listeners' ratings for speech produced by both male and female talkers, with high mean weighted Elo consistency indices, a measure of intra-rater reliability. This finding is evidence that the uniformly lower ratings of males for other males elicited by Likert-based tasks is not likely caused by inexperience at rating same sex talkers on attractiveness. We recommend that future studies on attractiveness enact a paired comparison design. In addition to eliciting more honest responses, a paired comparison design also simplifies the task and lowers the cognitive load on participants by only requiring them to attend to two stimuli at any one time. By contrast, judging each stimulus on its own using a predefined scale requires conscientious participants to calibrate their response to each item based on their memory of all the items that came before. And because the reference point used to judge each item at various points in the experiment is prone to shifting, this approach lowers the reliability of the results.

2.4.2 Conclusions

To our knowledge, this is the first study investigating vocal attractiveness through novel measures of working vowel space extracted from density estimates of continuously sampled formant trajectories across both word and sentence length productions. Our results add to previous work showing that acoustic measures related to sexual dimorphism, talker health, and the processing dynamics of listeners contribute to perceptions of female vocal attractiveness.

Chapter 3

Top-down attention guidance shapes action encoding in the pSTS

3.1 Introduction

The posterior superior temporal sulcus (pSTS) is linked to the perceptual representations of body actions during action observation. Classically the pSTS is characterized as providing the key sensory input needed to facilitate the interpretation of goals from motor behavior and intentions in social interactions (Lingnau and Downing, 2015; Thompson and Parasuraman, 2012; Pyles and Grossman, 2013). This strictly perceptual characterization of the pSTS, however, fails to account for the influence of high level contextual factors on the neural response. Activation in the pSTS is modulated by recent history of the observed action events (Vangeneugden et al., 2011), whether the viewer is attending to the social dimensions of an event (Tavares et al., 2008b), whether the observed action is consistent with the expectation of the viewer (Jastorff et al., 2011; Maffei et al., 2015; Saygin et al., 2012; Urgen and Saygin, 2019; Wyk et al., 2009), and whether the action is construed as intentional or incidental

(Morris et al., 2008).

Contemporary theories of the action observation network (AON) now emphasize an integrative role of the pSTS rather than strict sensory encoding. In these proposals, specific action features, such as body postures and local kinematics, are encoded in the lateral occipitotemporal regions (LOTC) and subsequently bound into action representations in the pSTS (Giese Poggio, 2001). The action representations are further tuned by top-down modulatory signals that reflect top-down influences imposed by cognitively-derived internal models (Geng and Vossel, 2013; Sokolov et al., 2018). These modulatory influences are derived from higher levels of the AON (i.e. the inferior temporal cortex; IFC) and are proposed to shape action representations so as to facilitate the behavioral goals of the viewer (Patel et al., 2019; Carter and Huettel, 2013). In one special class of these models, predictive coding models, top-down signals bias perceptual encoding in favor of expected actions as determined from prior knowledge of action goals, increasing the efficiency of perceptual encoding of the subsequently observed action (Kilner, 2011; Bach and Schenke, 2017; Koster-Hale and Saxe, 2013).

An important innovation in this new class of theoretical models is the specialized role of the pSTS as the integrator of two information streams: bottom-up sensory encoding of observed actions and top-down cognitively derived context. Unlike strictly representational accounts, integrative models are highly flexible in that they emphasize the encoding of sensory cues dependent on the observer's cognitive state. This integrative role, therefore, provides a new framework by which the local functional heterogeneity of the pSTS can be interpreted (Patel et al., 2019), namely that sensory information may be represented uniquely depending on the attentive goals of the viewer. This is contrast to proposals that characterize the pSTS as host to distinct neural populations for low-level perceptual and high-level social cognitive functions, intermixed and distributed through lateral temporo-occipito cortex (Hein and Knight, 2008; Deen et al., 2015; Bahnemann et al., 2009).

Integrative and predictive coding models are both influential in understanding brain systems that underlie action observation, and moreover are supported by univariate mapping studies showing that the behavioral goals of the observer alter activation maps distributed along the superior temporal sulcus. What is currently lacking, however, is direct evidence that the behavioral goals of the observer impact action representations that are constructed during action observation. One proposed mechanism by which this may occur is the sharpening of neural tuning attended actions, akin to the attention-mediated gain increases observed in early visual cortex during feature-based attention tasks (Treue and Martínez Trujillo, 1999; Saenz et al., 2002; Kok et al., 2012). Feature-based attention gain is a mechanism consistent with all classes of top-down integrative models and has been observed widely throughout sensory systems (Maunsell and Treue, 2006). Alternatively, observer goals have the potential to alter behavior without restructuring action representations directly. This could be achieved through the introduction of bias in the decision-making process, which would manifest in later stages of cortical processing while leaving action representations unadulterated (i.e. Summerfield and Egner, 2009).

A further consideration is the level of abstraction of the top-down influences that may shape perceptual representations. This is particularly important during action observation, in which a specific goal can be achieved through various combinations of an individual's actions, while specific actions may not be diagnostic of an individual's current goals or intentions (Thompson et al., 2019). Thus expectations of upcoming actions could include anticipated kinematic events, action outcomes, or perhaps even abstracted representations of action goals (Kilner, 2011).

In this study we investigate how feature-based attention modifies the statistical structure of action representations embedded within the spatial activation patterns elicited during action observation. We test the hypothesis that directed attention to kinematic aspects of an action vignette sharpens the tuning of these representations, and compare it to when

attention is directed to features not associated with action recognition (namely, the identify of the actor). We also evaluate the efficacy by which directing attention to observer goals (rather than specific actions) facilitates the decoding of action representations. We evaluate this hypothesis in three regions of the AON: the pSTS, the form and motion-selective LOTC (Oosterhof et al., 2010; Wurm and Lingnau, 2015), and the IFC (Ogawa and Inui, 2011; Wurm and Lingnau, 2015)). In a second analysis, we compare connectivity strength within the AON as a function of observer attention state to evaluate if the changes in information are likewise associated with selective strengthening of information through key pathways. Our results are consistent with models of the pSTS as dynamically restructuring action representations depending on the viewer's attentive state, with actions most strongly differentiated when observers attend to the kinematic content. These results are consistent with top-down and predictive coding models that emphasize the role of prior knowledge in shaping action representations.

3.2 Methods

3.2.1 Participants

Twenty-five healthy adults (8 male, 16 female) ranging in age from 21 to 42 years old (mean = 24.7, sd = 3.6) from the UC Irvine campus and surrounding community enrolled in and completed the study. Participants gave written informed consent. All experimental procedures were approved by the University of California Irvine Institutional Review Board. All participants had normal or corrected-to-normal vision. One participant was excluded from the analysis due to excessive motion during scanning.

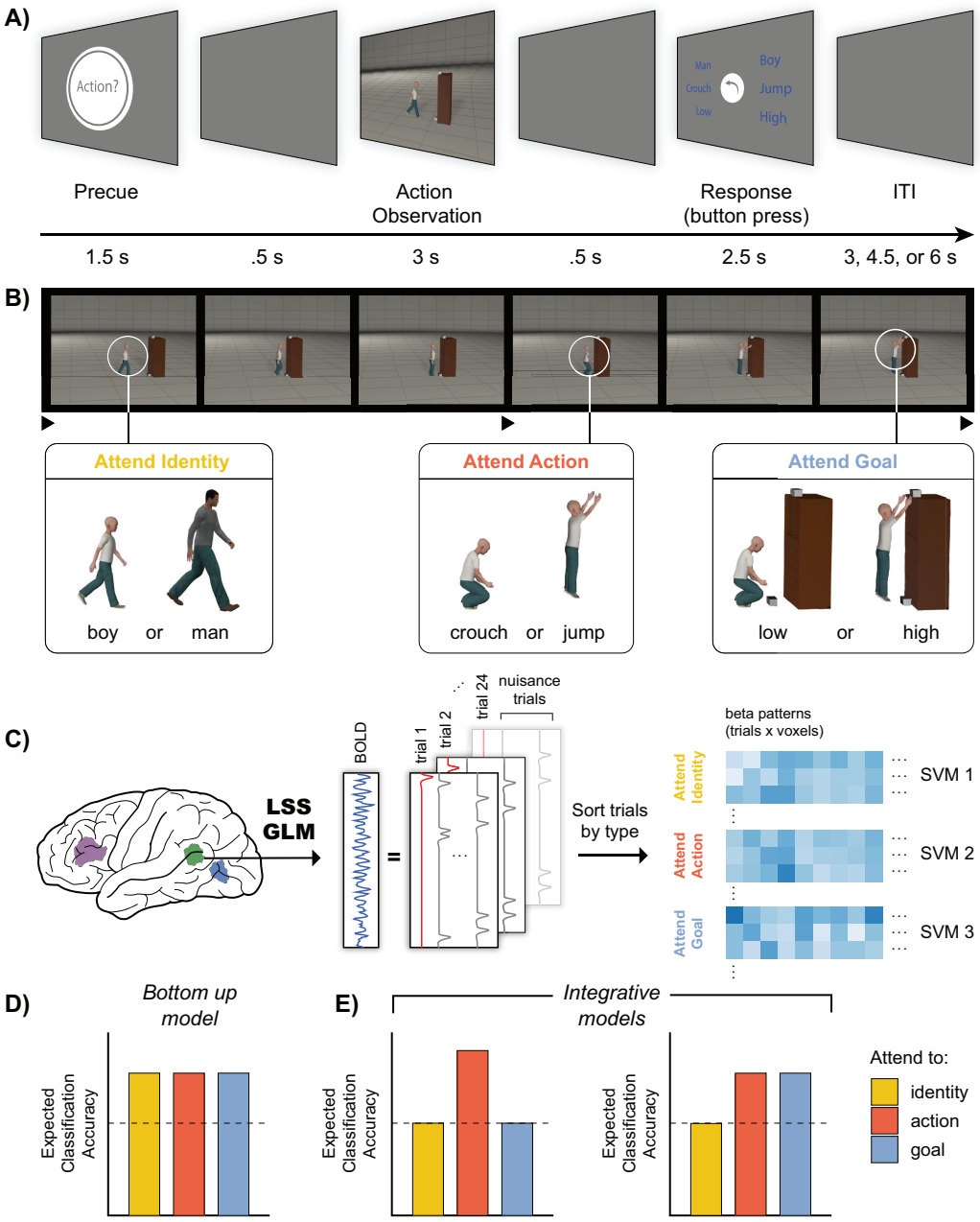

Figure 3.1: A) Filmstrip view of stills from the action vignette showing an avatar jumping with the intention to reach the box on top of the bookshelf. Each sequence depicted an avatar approaching a bookshelf, then making a head movement to indicate intent prior to executing the appropriate action to retrieve the box (either crouching down to reach the box on the floor or jumping up to reach the box on top of the bookshelf). B) Timing of trials in the rapid event-related design. C) The response to each event was estimated by iteratively fitting a linear model that included a separate regressor for each trial and confound regressors for all other trials grouped by type. The resulting matrix of beta coefficients, with trials as rows and voxels as columns, was sorted into three datasets by trial type (attend to action, goal, or identity) and passed on to three separate support vector machine classifiers. D-E) Expected pattern of MVPA results for action classification across different regions of interest.

3.2.2 MR Image Acquisition

Participants were scanned at the Facility for Imaging and Brain Research at the University of California, Irvine on a 3 Tesla Siemens Prisma MRI scanner (Siemens Medical Solutions) equipped with a 32-channel receive-only phased array head coil. High resolution anatomical images were collected using a single T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence (176 sagittal slices; 1 mm isovoxel resolution; field of view = 256 mm; TR = 2000 ms; TE = 1.99 ms; TI = 900 ms: flip angle = 9 degrees; GRAPPA acceleration factor = 2; bandwidth=240Hz/Px).

Two types of functional scans were acquired across two sessions, both using a T2*-weighted gradient recalled echoplanar imaging multi-band pulse sequence (cmrrmbep2dbold) from the University of Minnesota Center for Magnetic Resonance Research (CMRR). Session one consisted of localizer scans designed to identity regions of interest (ROIs) within the AON (69 slices co-planar with the AC/PC; in-plane resolution = 2 2 mm; 106 106 matrix size; 2 mm slice thickness, no gap; interleaved acquisition; field of view=212mm; phase partial Fourier scheme of 6/8; TR = 2000 ms; TE = 30 ms; flip angle = 79 degrees; bandwidth = 1814 Hz/Px; echo spacing = 0.66 ms; excite pulse duration = 8200 microseconds; multi-band factor = 3; phase encoding direction = PA; fat saturation on; advanced shim mode on). Session two comprised the main experiment and therefore incorporated rapid event-related scans that were designed to sample the hemodynamic response more rapidly (68 slices co-planar with the AC/PC; in-plane resolution = 2 2 mm; 106 106 matrix size; 2 mm slice thickness, no gap; interleaved acquisition; field of view = 212 mm; phase partial Fourier scheme of 6/8; TR = 1500 ms; TE = 30 ms; flip angle = 79 degrees; bandwidth = 2144 Hz/Px; echo spacing = 0.57 ms; excite pulse duration = 8200 microseconds; multi-band factor = 4; phase encoding direction = PA; fat saturation on; advanced shim mode on). At the beginning of each session, an additional pair of EPI images with phase-encoding directions of opposite polarity in the anterior to posterior plane were acquired to correct for

susceptibility distortions in each participant's functional data.

3.2.3 Session 1: Functional localizers

In the first session, all participants underwent three functional localizer scans (two repetitions each) to identify the posterior superior temporal sulcus (pSTS), middle temporal complex (hMT+), and extrastriate body area (EBA). Stimuli were displayed on a BOLDScreen32 LCD monitor controlled by MATLAB (The Math Works, Inc.) and the Psychophysics Toolbox extensions (Brainard, 1997) on a Windows desktop. Subjects viewed the animations through a mirror mounted on the head-coil and directed at a screen positioned at the head end of the scanner.

pSTS To localize areas of the brain that respond selectively to biological motion, participants were shown 12 alternating blocks of intact and scrambled point-light biological motion (Grossman et al., 2010). Animations depicted an actor with 12 lights attached to their joints performing 25 unique actions, such as walking, jogging, throwing, kicking, etc.. Scrambled animations were produced by randomizing the starting position of the point-light dots within a region approximating the target figure, then leaving their motion vectors intact. Animations had a duration of 1 second and were separated by a 1 second fixation inter-trial interval (ITI). Participants performed a 1-back task on each animation, indicating by button press whether the current animation was the same or different action as the one immediately prior. The pSTS was identified using a group random-effects GLM that contrasted intact versus scrambled trials, thresholded using a False Discovery Rate (FDR; Genovese et al. (2002)) of $q < 0.005$.

LOTC The LOTC was identified jointly using two localizers, one targeting hMT+ and the other targeting the EBA. Although separable in individuals subjects (Weiner and Grill-

Spector, 2011, 2013), the hMT+ and EBA in group analyses jointly occupy the ascending limb of the posterior inferotemporal sulcus (pITS; Downing et al. (2007)). The LOTC was therefore identified as the union of the hMT+ and EBA (described below), constrained to the dorsal extent by the inferior temporal gyrus.

To isolate the motion-selective hMT+, participants passively viewed alternating 12 sec blocks of optic flow dot motion and stationary dot patterns (Huk et al., 2002). Optic flow was constructed with 500 black dots randomly dispersed within a circular aperture, alternating between expansion and contraction. In the stationary interval, dots remained frozen in position for 12 seconds. The motion-selective responses on the pITS were thresholded using FDR, $q < 0.005$.

To isolate the body-selective EBA, participants viewed images of headless bodies, cars and limbs (hands and feet) (Stigliani et al., 2015). Each image was superimposed on top of a 10.5 degree phase scrambled background generated from a randomly selected image to minimize low-level differences across categories. Images were presented in 12 blocks, with 9 images shown per block. Body and limb selective brain regions were identified as those with higher brain response when viewing bodies and limbs versus images of cars, FDR, $q < 0.005$.

3.2.4 Session 2: Action observation

Action vignettes spanning 3 seconds (see Figure 3.1A) were generated in Poser Pro 11 (Bondware, Inc.), and depicted one of two avatars (a boy or a man) performing the same sequence of actions in which the avatar walked towards a bookshelf, indicated intent to reach one of two boxes, then either crouched down or jumped up to reach the box. The vignette ended after the execution of the action and prior to the avatar making contact with the box. Each vignette was constructed such that it was visualized from 8 unique viewpoints that spanned an 80 deg viewing range on each side (left and right, profile to rear views).

Before beginning the experiment in the scanner, all participants were familiarized with the action vignettes and practiced the task under all three attention conditions: attending to the avatar's identity, the action category, or the proximate goal of the action. To prevent motor response preparation while viewing the action vignette, stimulus-response mappings were obscured until the response interval, during which the labels for the three binary dimensions of the action (identity: boy/man; action: crouch/jump; goal: low box/high box) were randomly assigned to the left and right sides of fixation on each trial. Participant reported the correct label by pressing the button corresponding to the side of the screen correctly displaying the value of the feature they attended. Classification was always conducted on the trials with the action labels jumping versus crouching.

Trials were separated by a 3, 4.5, or 6 sec inter-trial interval (ITI), pseudo-randomized within each run such that, in total, each trial lasted 10.5, 12, or 13.5 seconds. The onset of each trial event was synchronized with the onset of volume acquisition to ensure synchronization with the event-related acquisition. Each run of the experiment contained 8 conditions per attention task from a fully crossed design comprising 2 avatars (boy, man), 2 actions (crouching, jumping) and 2 viewpoints (leftward and rightward walking). The three attention tasks (attend to avatar identity, action category, action goal) were randomly interleaved within each run, resulting in a total of 24 trials per run or approximately 5 minutes of scan time. The experiment was organized into 8 runs for a total of 192 trials.

3.2.5 Imaging analysis

Preprocessing

Preprocessing of functional imaging data was conducted using BrainVoyager QX v20.6 (Goebel et al., 2006). All functional images were slice-time corrected, motion corrected within and between runs, linearly detrended, and temporally high pass filtered (cutoff fre-

quency 0.01 Hz). Session 2 scans were additionally corrected for susceptibility-induced magnetic field distortions using the field map method (Jezzard and Balaban (1995), implemented in BrainVoyager's COPE v1.0 plugin). All functional images were co-registered to each individual's T1-weighted image.

Session 1: Regions of interest

Functional data in session 1 was aligned to a template pilot subject using cortex-based alignment (Frost and Goebel, 2012). Sulcal curvature was constructed on white matter surfaces derived from Freesurfer's recon-all algorithm (http://surfer.nmr.mgh.harvard.edu/), imported into BrainVoyager using custom library functions (https://github.com/tarrlab/Freesurfer-to-BrainVoyager). Regions of interest were identified on the cortical surface and then projected back into native volumetric coordinates by searching along the vertex normal in towards the white matter 1 mm and into the gray matter outward from 3mm.

The IFC (comprising BA44, BA45A, BA45B, BA47 and the inferior frontal sulcus) was identified anatomically in each individuals' native anatomical images using Freesurfer's cortical surface atlas mapping algorithms in conjunction with the 400 atom resolution Schaefer atlas (Schaefer et al., 2018). This atlas emphasizes homogeneity of functional systems within the parcels, coupled with high resolution "atomic" parcellation in approximately equisized units, and therefore higher precision in identifying ROI boundaries.

Session 2: Action observation under attentional instructions

The timeseries from each voxel in the ROIs were first z-scored across time and trial-by-trial patterns of estimated BOLD activation were derived using the least squares separate (LSS) general linear model approach Mumford et al. (2012); Turner et al. (2012)). The LSS procedure uses a separate GLM to estimate the pattern of activity for each trial where the

model for the i^{th} trial is

$$Y = X_{LSS_i}\beta_{LSS_i} + \epsilon_i$$

such that the design matrix for the i^{th} trial, X_{LSS_i}, contains one regressor of interest modeling the stimulus-evoked BOLD response to the i^{th} trial and several other nuisance regressors modeling responses to the remaining trials grouped by trial type. Stimulus-evoked BOLD responses to each event were modeled as boxcar functions convolved with a canonical double-gamma hemodynamic response function (HRF) (Friston et al., 1998; Glover, 1999). In order to account for variability in the latency of the HRF across the brain and across subjects (Steffener et al., 2010), we optimized the time-to-response-peak parameter of the two-gamma function (5 possible values between 5 and 7 secs in steps of .5 secs), with the modeled HRF that produced the highest coefficient of determination (R^2) for all trials within the voxel selected for downstream analysis. Our LSS design matrix contained 6 nuisance regressors, one for each action condition (crouching, jumping) crossed with each of the three attention tasks (attend to identity, action, goal), and additional nuisance regressors capturing the average signal and first derivative measured from the white matter and ventricles over time. Following beta extraction, trials with extreme movement near the peak response (three or more consecutive timepoints of framewise displacement above 2mm) were censored from later analysis. Also, variance in the beta series accounted by repetitions of actions was removed.

Multivariate pattern analysis (MVPA)

Trial betas were separated by attention task (attend to identity, action, goal) for each participant, and mean centered within runs to remove spurious correlations between the estimated activity levels of different trial types across runs (Lee and Kable, 2018). The resulting normalized betas were then averaged within runs to a single activation estimate per action class. The matrix of n activation estimates by k voxels and $1xn$ class labels was then used to train

three separate support vector machine (svm) classifiers, one per attention task, implemented in the e1072 package in R (Meyer et al., 2018). The SVM consisted of a linear kernel and a cost value of 1. Classification was performed within subjects using eightfold leave-one-run out cross validation. Within each fold, two predictions were made from the held-out test set, one from each action class. The final classification accuracy for each subject was computed as the mean accuracy across all eight folds.

To examine task-related differences in MVPA classification accuracy, we constructed a linear mixed-effects model (LMM) using the lme4 package implemented in R (Bates et al., 2019). The LMM predicted classification accuracy based on the fixed effects of attention task, ROI, and their two-way interaction, with participants as random effects. P values were obtained using Likelihood Ratio Tests comparing each model to reduced models lacking the variable (or interaction) in question.

Functional connectivity

Functional connectivity was computed as the Pearson correlation of the beta series between two ROIs, separately for each attention task. Beta series connectivity is based on the assumption that if two brain regions are functionally interacting, then the amount of activity captured by beta estimates should correlate across trials (Rissman et al., 2004). Beta series correlations were calculated from ROI-averaged time series in which volumes with FDR greater than .4mm were excluded. As for the MVPA analysis, trial-wise betas were estimated using LSS GLMs with nuisance regressors including the global signal measured from the white matter and ventricles and the Volterra expansion of all 6 rigid body motion realignment parameters (Fristen et al., 1996). Pearson correlations were computed between each 64 beta timeseries, Fisher r-z transformed. Paired, one-tailed repeated measures t-tests for the planned contrasts of action $>$ identity and goal $>$ identity were conducted on the transformed correlations.

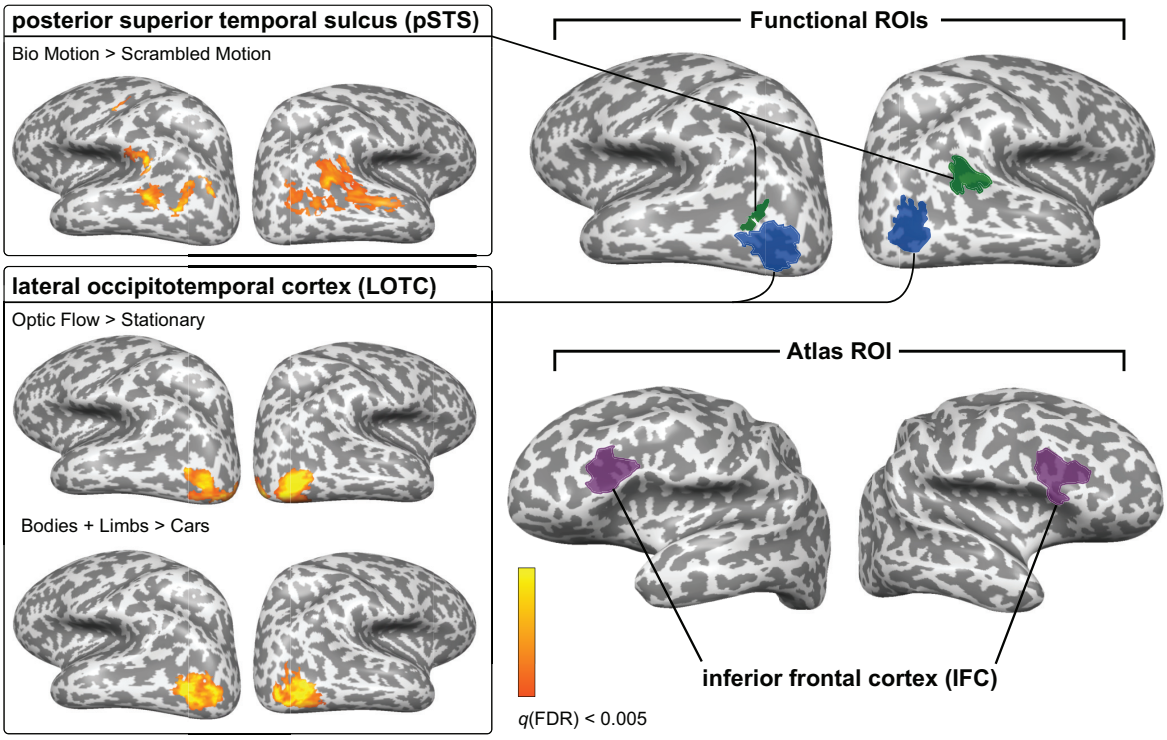

Figure 3.2: Identification of regions of interest. Left: Group activation maps from the three independent functional localizer scans, displayed on inflated cortical surface meshes of a pilot subject. Right: The regions of interest, including the atlas-derived IFC, projected onto a single subject cortical surface.

3.3 Results

Functional Localizer Analysis

Results from the independent localizer scans are shown in Figure 3.2. The biological motion localizer identified large bilateral regions of the pSTS, notably of larger extent in the right hemisphere; whereas the hMT+ and EBA localizers jointly revealed large bilateral co-activation in ventral temporal cortex and LOTC. The spatial overlap between hMT+ and EBA is consistent with reports of functionally distinct neural populations that co-localize to the inferior occipital sulcus when identified in group-based localizers (Downing et al., 2007).

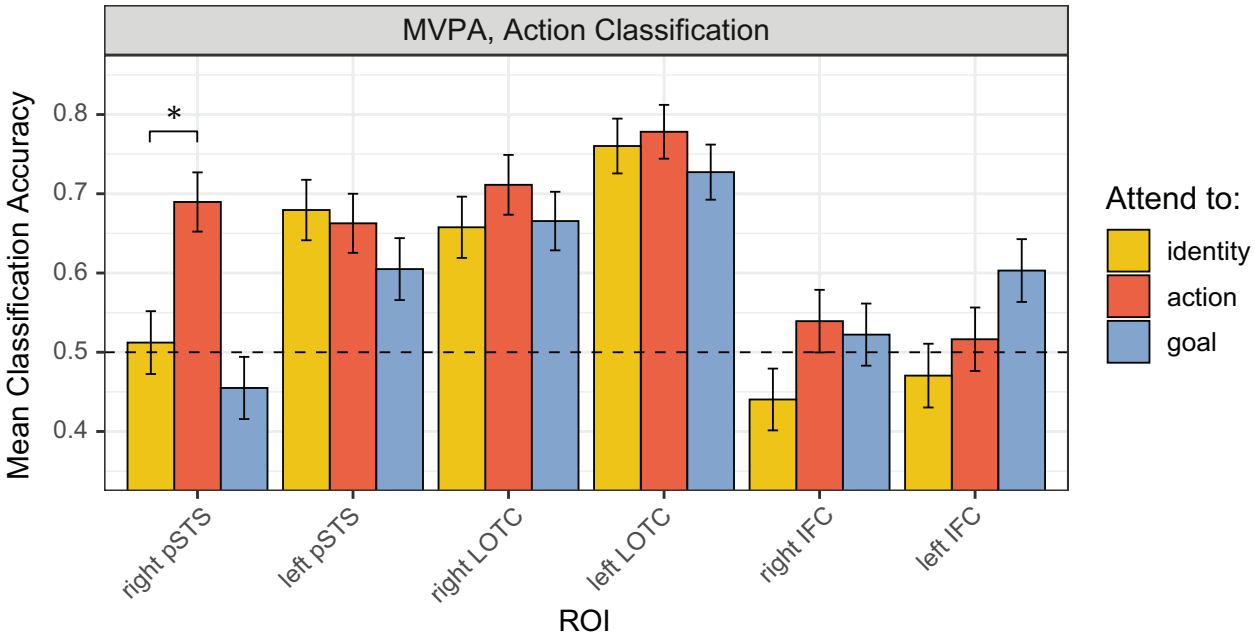

Figure 3.3: MVPA classification accuracies from decoding action class (jumping, crouching) by task demand. Error bars indicate SEM. Asterisks indicate statistical significance (*= $p < 0.008$). Dashed line indicates binary classification accuracy at chance (50%).

Multivariate pattern analysis

To test the hypothesis that the attentional state of the participant sharpens the population tuning of the multivariate informational content during action observation, we evaluated the cross-validated accuracy of action classification (labels: jumping and crouching) from ROI activation patterns (Figure 3.3). A linear mixed effects model with mean classification accuracy as the dependent variable yielded significant main effects of ROI ($\chi^2(5) = 72.83, p < 0.001$) and task ($\chi^2(2) = 9.69, p < 0.007$), and a significant ROI x task interaction ($\chi^2(10) = 21.81, p = 0.016$). Thus, attentional demands influenced the decodability of actions in a subset of ROIs.

To better break down the task x ROI interaction, 6 within-ROI LMMs were constructed evaluating the influence of the attention instruction on classification accuracies. Planned contrasts compared mean classification accuracies during action and goal attention conditions to the identity attention condition (the control condition). Action decoding in the right pSTS

was significantly more accurate when participants attended to the action kinematics versus the identity of the avatar ($b = 0.172, SE = 0.062, t(48) = 2.786, p = 0.008$, uncorrected), consistent with the sharpening hypothesis of the action-tuned neural populations. Action decoding did not, however, differ significantly between trials when the participant attended to the goal of the action versus the identity of the actor ($b = -0.022, SE = 0.060, t(48) = -0.370, p = 0.713$).

In all conditions, the trials were labeled according to the portrayed action and the goal of the actor, which were strictly confounded (i.e. the actor always gazed upwards prior to jumping up, and gazed downwards prior to crouching down). Because these two conditions reflect the same stimulus events, we attribute variations in classification performance to reflect changes in the cortical state of the observer, mediated by attention goals.

No other ROIs revealed significant task-related differences in action decoding.

Univariate Analysis

It could be argued that variations in multivariate decodability of actions as a function of top-down instruction may reflect differences in univariate activation levels across tasks, rather than sharpened neural tuning *per se*. We therefore compared the univariate responses in the ROIs as a function of task (Figure 3.4).

Statistical analysis of the average stimulus-evoked responses revealed a main effect of ROI ($\chi^2(5) = 496.96, p < 0.001$), but no main effect of task instruction ($\chi^2(2) = 1.26, p = 0.533$), nor an interaction between task and ROI ($\chi^2(10) = 496.96, p = 0.992$). Thus the more diagnostic activation patterns in the pSTS when attending to action kinematics versus actor identity cannot be attributed to attentionally-driven increases in the average BOLD response within the region. That multivariate classification accuracy is independent of BOLD activation levels is consistent with previous reports that classification is just as high for non-

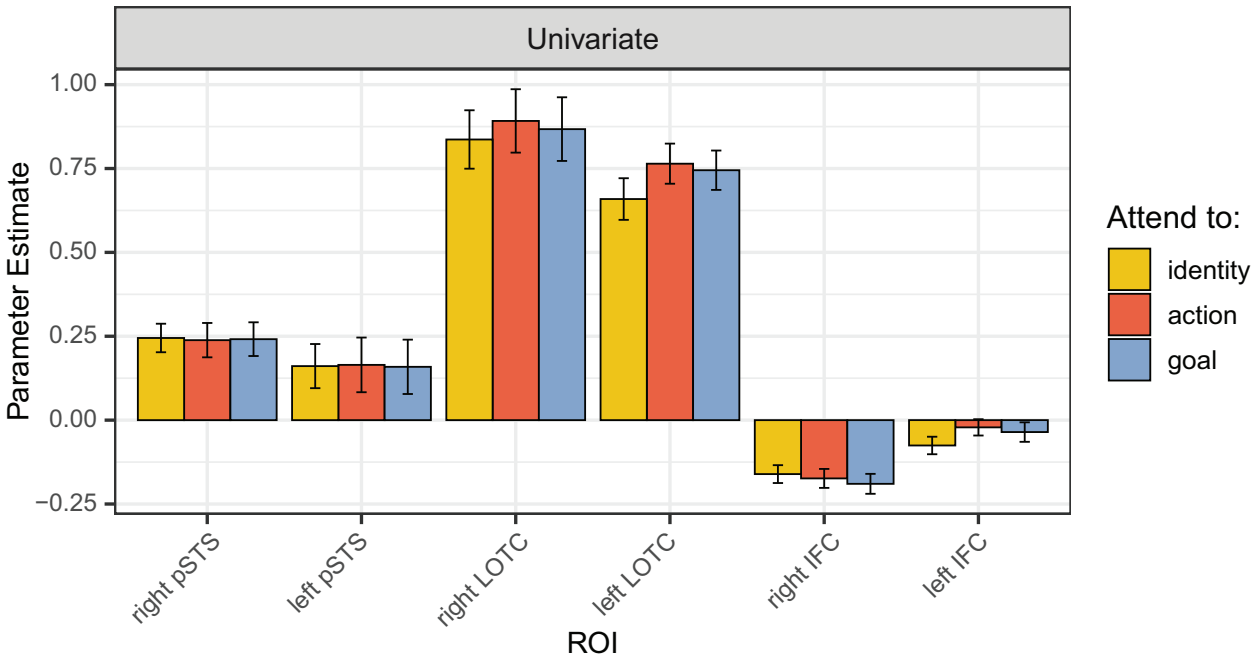

Figure 3.4: Group univariate responses by task demand modeled during the precue and action observation periods. Univariate activity estimates were produced by averaging the trial-by-trial LSS beta coefficients across trials of each task demand and then averaging the data across voxels within the ROI.

preferred categories of visual stimuli as it is for preferred categories, within the same brain region (Haxby et al., 2001).

Task instructions as a means to modulate attention.

In this experimental design, observers were instructed to attend to particular dimensions of an action vignette without knowing in advance which action was upcoming. One could hypothesize that the failure to modulate classification accuracy in the action observation network more broadly (outside of the right pSTS) may reflect a failure of task instructions to guide observer behavior, and therefore to alter brain state.

To evaluate this, we analyzed behavioral performance in the scanner, in which participants were required to properly identify the label for the action, action goal or identity of the actor on each trial. Participants were highly accurate in detecting the features

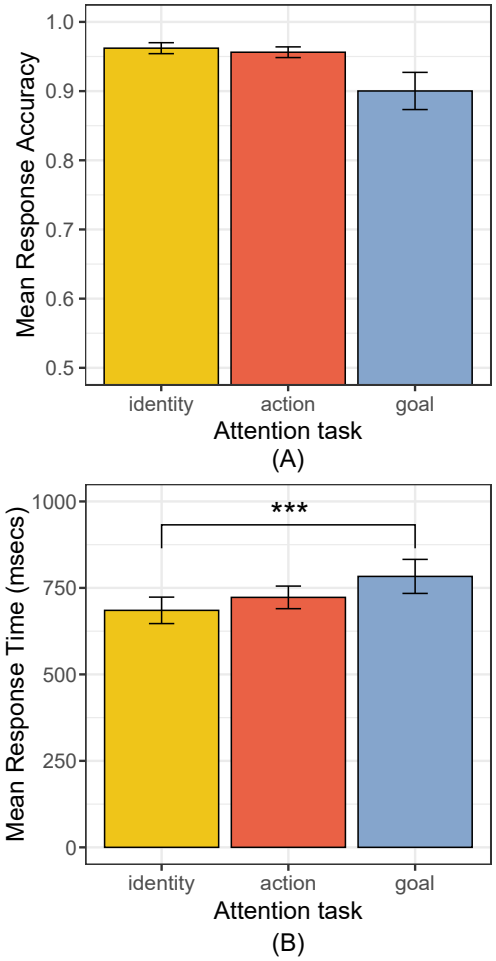

Figure 3.5: Behavioral results from scanner broken down by attention task showing mean accuracy and response latency (msecs) for detecting the feature cued at the beginning of each trial.

of the action vignettes that they were cued to attend (see Figure 3.5). A linear mixed-effects logistic regression model predicting the binary outcome of each trial ("correct" versus "incorrect") for each task revealed a trend, but no significant effect of task on accuracy ($\chi^2(2) = 5.470, p = 0.065$). An LMM on response latencies, however, yielded a significant main effect of task ($\chi^2(2) = 12.068, p = .002$) such that response latencies were longer when participants identified the goal of the action versus the identity of the actor ($b = 92.74, SE = 23.85, t(22.12) = 3.888, p < 0.001$), but not when they identified the action category compared to the identity of the actor ($b = 34.86, SE = 21.09, t(23.18) = 1.653, p = 0.112$).

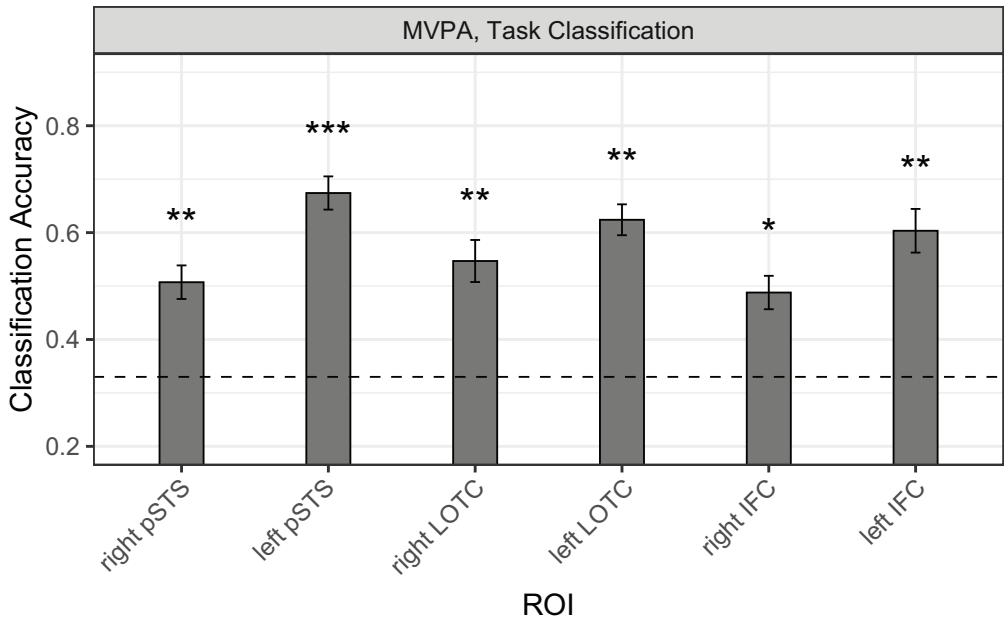

Figure 3.6: MVPA classification accuracies from decoding task instruction (attend to action, goal, or identity). Error bars indicate SEM. Asterisks indicate statistical significance based on nonparametric permutation tests (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). Dashed line indicates three-way classification accuracy at chance (33%).

In a second analysis, we evaluated whether activation patterns in the AON regions were modulated by the task demands imposed by the attention cue. We trained a single classifier per ROI to perform a three-way classification of task instruction (attend to action, goal, or identity). Figure 3.6 displays mean classification accuracies for decoding task. All ROIs classified the task demand reliably higher than what would be expected by chance (randomized permutation tests; right pSTS, $p = 0.01$; left pSTS, $p < 0.001$; right LOTC, $p = 0.009$; left LOTC, $= 0.007$; right IFC, $p = 0.02$; left IFC, $p = 0.004$), evidence that participants differentially allocated their attention, which in turn altered the informational content in each ROI.

Functional connectivity

Theoretical models propose the inferior frontal cortex to function as a biasing agent such that the sensory representations of specific body kinematics consistent with the observer's current

behavioral goals (Kilner, 2011; Koster-Hale and Saxe, 2013). We therefore hypothesized that the signature of such feedback may be reflected through increased functional connectivity between the pSTS and IFC during experimental conditions when participants directed their attention to action features.

Analysis of the Pearson's coefficients on the beta timeseries revealed strong functional connectivity between the right pSTS and the right IFC during action observation under all task instructions (Figure 3.7a). When compared across tasks, only the connection between the right pSTS and right IFC varied as a function of task such that it increased significantly when participants attended to action versus when they attended to the actors' identity ($t = 2.21$. $p = 0.018$, uncorrected) (Figure 3.7b).

3.4 Discussion

The pSTS is increasingly recognized as an integrative hub for decoding social cues that convey essential information for making inferences about actions and intentions (Sokolov et al., 2018; Dasgupta et al., 2016). Contemporary theories propose that the action representations in pSTS are modulated by the attentive state of the observer, and thus identical actions may result in unique representations when viewed with different goals (Patel et al., 2019). In this study, we test the hypothesis that directed attention to action features sharpens the tuning of neural populations in the pSTS for subsequently viewed actions, reflecting a strengthening of top-down influences acting upon the pSTS.

We found that the attentive state of the observer alters the population code in the right pSTS when viewing action vignettes, as demonstrated by a significant effect of the attention instruction on MVPA accuracy. Specifically, the spatial activation patterns for two distinct actions - jumping and crouching - are more easily differentiated in the right

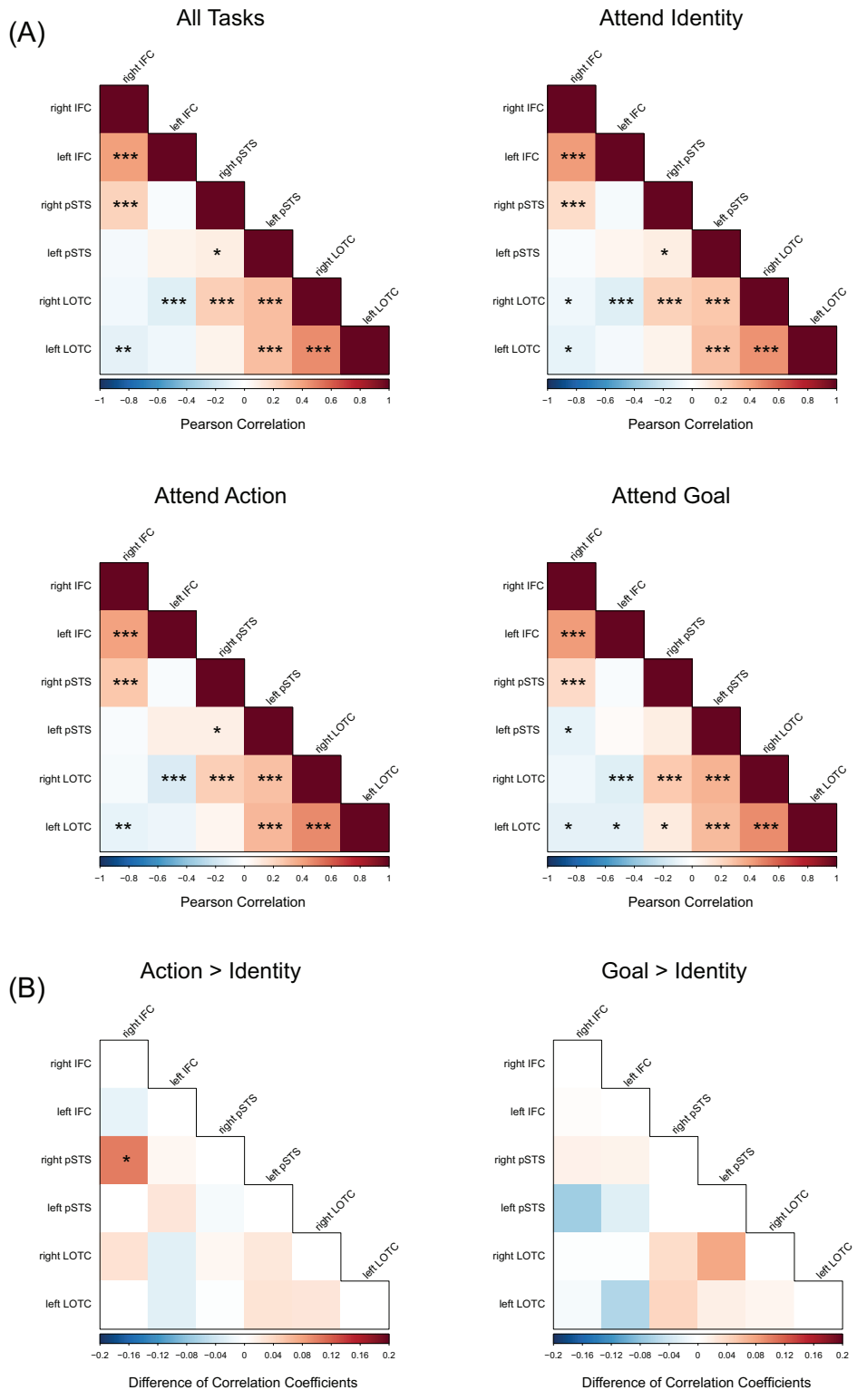

Figure 3.7: Functional Connectivity Results. (a) Task-based functional connectivity correlation matrices for each attention task. (b) Difference of correlation coefficients comparing differences in connection strength while attending to action kinematics and action goals versus actor identity.

pSTS when observers direct their attention to the kinematic features of the vignette. The significant improvement is both higher than expected by chance and higher as compared to the decodability of the same actions when observersreported the actors' identities. Identity and action kinematics are two features associated with unique processing pathways, with action kinematics linked to the pSTS, and body postures associated with encoding in the more posterior LOTC and ventral fusiform body area (FBA; Peelen and Downing (2007)).

Our findings add to the handful of reports in other sensory domains in which directed feature-based attention refines information in the population response, resulting in more distinct activation patterns that facilitate classification (Braunlich and Love, 2019; Kok et al., 2012). A likely mechanism of this attention benefit is the known increased gain in neurons with underlying tuning preferences for the attended features, resulting in an overall sharpening of the population response (for review, see Reynolds and Heeger, 2009). In fMRI activation patterns this has the consequence of warping the representational distinctiveness of the attended items and, when analyzing for information within distributed activation patterns, improving the efficacy of the trained classification algorithm (Çukur et al., 2013; Nastase et al., 2017).

Our finding is consistent with previous fMRI univariate mapping studies that have documented stronger and more widespread activation on the pSTS when attention is directed to social dimensions of an event rather than non-social features (Tavares et al., 2008b; Lee et al., 2014; Safford et al., 2010). Those studies conclude that directed attention to the social aspects of a scene differentially engages neural populations with tuning to features that promote the interpretation on social events. In our study we found no change in the univariate response across our three attention tasks, likely because all tasks focused on highly salient social aspects of the stimulus (identity of the actor, actions being conducted, or the goals of the actor). Instead we have documented a shift in the statistical structure of the information within the pSTS multivariate activation patterns, without an associated increase

(or decrease) in the univariate response. Thus we conclude that our attention manipulation did not recruit new populations of neurons during action observation, but instead altered the information content of the representations constructed during action observation.

In our results, directing attention to the kinematic features of the action vignette improved classification across a wide range of viewing perspectives of the scene, from profile views of the actors with strong lateral movements, to near midline views. Although there is evidence of viewpoint specificity in STS neurons recorded in monkeys (Oram and Perrett, 1994), evidence from fMRI strongly favors viewpoint invariant representations on the human STS (Grossman et al., 2010). In line with this, behavioral research indicates that not all features of an action sequence are equally salient, with key diagnostic features most strongly capturing the attention of the observer (Thurman and Grossman, 2008; Casile and Giese, 2005). Moreover, with practice observers can readily identify those salient features and more easily discern action exemplars, with changes in the univariate pSTS response closely following those improvements in training (Grossman et al., 2004; Jastorff et al., 2009). Our current findings are consistent with both of these observations, namely that attention operates on action representations in a manner that is robust to changes in viewpoint, and therefore likely reflects the enhanced salience of diagnostic features for action templates, or action categories, rather than specific instances themselves.

We did not find evidence for attentional modulation of action representations in other regions of the AON, which is consistent with the proposal that the LOTC is largely sensory-driven and tuned to specific body postures and kinematic features Lingnau and Downing (2015), and that the IFC is largely responsible for mapping action goals to motor sequences (Kilner, 2011; Molnar-Szakacs et al., 2005). It is worth noting that each of these regions was able to accurately classify the current trial instruction condition, indicating the attentive state of the observer altered how information was processed throughout the AON. Changing task instructions did not, however, render the two observed actions more or less easily

decodable from the pattern response outside of the pSTS.

Predictive coding in action observation

Although computational models emphasize bottom-up, feedforward mechanisms of action perception from form and motion features (Mather et al., 1992; Lange and Lappe, 2006; Hoffman and Flinchbaugh, 1982), biologically-inspired models have always noted the top-down influences from prefrontal cortex (Giese and Poggio, 2003; Kilner, 2011). Prefrontal cortex activation is commonly observed when measuring brain activity during action observation (Saygin et al., 2004; Dasgupta et al., 2016), and interruption to the inferior frontal cortex using noninvasive brain stimulation likewise interferes with action recognition (van Kemenade et al., 2012).

There is mounting evidence that the interpretation of actions, including identifying specific actions and their associated goals, follows a predictive coding framework (Kilner et al., 2007b,a; Urgen and Miller, 2015). Empirical studies leveraging the power of dynamic causal modeling (DCM) to infer the direction of causal influence between functionally connected brain regions have revealed both feedforward and feedback connections between IFC and pSTS that are modulated when viewing actions (Sokolov et al., 2018; Maffei et al., 2015; Gardner et al., 2015; Urgen and Saygin, 2019).

An important component of predictive coding models is the error signal that is elicited when the observed events mismatch the predicted sensory signals. This error signal has been repeatedly documented in univariate fMRI studies as an increase in the pSTS response when the observed actions violates expectations (Koster-Hale and Saxe, 2013; Hillebrandt et al., 2014; Marsh et al., 2014). These include situations in which actors perform irrational reaching and grasping movements (Jastorff et al., 2011), when humans engage unexpectedly in robot-like movements (Saygin et al., 2012; Urgen and Saygin, 2019), or when stick-figures

perform actions at reduced versus normal gravity (Maffei et al., 2015), among other similar violations (Gardner et al., 2015; Cardellicchio et al., 2018; Wyk et al., 2009). Moreover, predictive coding is proposed to operate hierarchically such that cognitively-derived internal models can exist in multiple levels of abstraction, from visual kinematic features to the more abstracted action goals (Bach and Schenke, 2017; Kilner et al., 2007b).

The pathways by which error signals propagate through the AON are an active area of investigation. Dual pathway models propose distinct structural and functional pathways for action understanding, with a ventro-dorsal pathway for action identification further split into a caudal route that codes diagnostic action features and a ventral route that processes action goals (Binkofski and Buxbaum, 2013; Buxbaum and Kalénine, 2010). Tracing studies in monkeys supports the notion of dual pathways, with a dorsal route connecting the upper bank of the STS to premotor cortex via parietal connections, and a second ventral route direct between the lower bank of the pSTS and premotor cortex (Nelissen et al., 2011). In humans, undirected functional connectivity analyses reveal strong connectivity between the IFC and pSTS that carries information unique from that in other segments of the AON network (Dasgupta et al., 2016), and dynamic causal modeling shows that this top-down pathway is strongly modulated by viewing biological motion (Sokolov et al., 2018). As further indirect evidence in support of these top-down models, in this study we observe attention to mediate this pathway such that directed attention to actions increases functional connectivity between the right IFC and pSTS. Although functional connectivity does not imply direct structural connectivity, it is nonetheless consistent with a model in which neural information is biased along processing pathways contingent on the attentive state of the observer.

3.5 Conclusions

The pSTS supports the initial perceptual encoding of dynamic body states that underlie particular goals (e.g. hand movements during reaching actions, decoding dynamic facial expressions, and the encoding of limb kinematics during whole-body movements). These perceptual representations are subsequently interpreted by higher-level cognitive systems to support action understanding and intentional states for social interactions. Our findings indicate that the converse is also true: cognitive systems shape the coding of action representations in the pSTS when observers attend to action features. We propose that the putative functional heterogeneity of pSTS may be accounted for, in part, by top-down influences reflecting the observer's goals when engaged in action observation.

3.6 Acknowledgments

This material is based upon work funded by the NSF under grant BCS-1658560 to EG and JP.

Chapter 4

The impact of trial averaging, mean centering, cost tuning and data cleaning on multivariate pattern analyses using least squares separate (LSS) beta series

4.1 Introduction

Multivariate pattern analysis (MVPA or pattern decoding) has become an increasingly preferred analytical tool in functional magnetic resonance imaging (fMRI) studies. In contrast to conventional univariate analyses, which relate the effects of experimental variables to the activity of single voxels or to the average activity within a region of interest (ROI), MVPA leverages modern machine learning (or pattern recognition) algorithms (Hastie et al., 2001;

Vapnik, 1995) to classify (or 'decode') attributes of the experimental stimuli from the distributed pattern of BOLD activity across *many* voxels (Haynes and Rees, 2006; Kriegeskorte, 2011; Pereira et al., 2009; Norman et al., 2006). Successful classification is taken as evidence that the particular collection of voxels contains information relevant to the task at hand. Multivariate analyses have gained wide appeal over univariate approaches for offering improved sensitivity and, at least in principle, the possibility to map regions coding experimental variables in latent multidimensional spaces (Diedrichsen et al., 2013; Naselaris et al., 2011;but see Davis et al., 2014; Popov et al., 2018), thus greatly deepening the richness of informational content available for analysis.

With the advent of multivariate methods, however, has come a proliferation of methodological choices required on the part of the researcher. Some of these choices are image processing decisions that have long been fundamental to all fMRI studies but which nonetheless require re-appraisal in light of the unique needs of MVPA; whereas, other choices are specific to machine learning algorithms and thus are new arrivals to the neuroimaging arena.

Many previous studies have investigated the extent to which individual methodological choices improve the power and reliability of classification algorithms when working with BOLD data. For instance, a common starting point for multivariate classification analyses is the estimation of trial-specific activation patterns in rapid event-related designs. Rapid trials are increasingly favored in fMRI experiments because they allow unique conditions and cognitive events to be closely interleaved in time, and the rapid trial pacing allows more task-related samples to be collected in a scan session than slower ER designs. Rapid ER designs do, however, require special statistical approaches so that hemodynamic responses associated with the individual trial events can be estimated as accurately and unbiasedly as possible (Turner et al., 2012; Mumford et al., 2012). These statistical approaches interact with other design considerations (order, number, and spacing of trials or runs) in determining the best method of pattern estimation (Mumford et al., 2014; Coutanche and Thompson-Schill, 2012).

In addition, fMRI suffers from a problem of high dimensionality, with a typical functional imaging acquisition measuring neural activity across upwards of 100,000 voxels. Thus, much attention has been devoted to dimensionality reduction techniques (e.g. feature selection) that seek to improve classifier performance by selecting the most informative voxels, which may be spatially dispersed and not otherwise captured by more traditional region-of-interest approaches (De Martino et al., 2008; Mourão-Miranda et al., 2006).

Still, other authors have investigated processing decisions related to the statistical decoding of response patterns, such as: the type of classifier used crossed with different kernels and hyperparameter values; the method of cross-validation and data partitioning scheme employed (Etzel et al., 2011; Varoquaux et al., 2017); as well as the effectiveness of various performance measures at quantifying model performance (Dinga et al., 2019).

The outcomes of many of these decisions may interact in complex ways. For example, the Type I error rate of the estimated activation patterns is influenced by many interacting factors including the type of pattern estimator used, study design (condition order and timing of trials) as well as whether similarities are computed using patterns from the same or different functional runs (Mumford et al., 2014). Furthermore, many data processing decisions interact as well, such as that between temporal compression strategy (estimating activation patterns by averaging several timepoints in the middle of experimental events or fitting a model), data partitioning (splitting the data by runs or leaving out a selected number of observations of each class), and detrending (Etzel et al., 2011).

Given the sheer number of these analytic choices available, certain authors have warned of the danger of spurious results arising due to trying out a large number of variations in the processing pipeline directly on experimental data with the hopes of maximizing classifier performance (Etzel et al., 2011). Though no "one size fits all" set of guidelines exists for all experimental questions and designs, the field would benefit from the systematic study of how these processing strategies (or unique combinations thereof) impact diverse data sets.

In this paper, we evaluate the independent and joint effects of certain data processing decisions on linear support vector classifications, with the goal of providing recommendations for rapid event-related designs using the LSS trial-wise beta estimation approach. This analysis focuses on four aspects of the data processing pipeline: preprocessing, condition-based trial averaging, within-run mean centering, and SVM cost parameter selection. Preprocessing steps focus on the impact of commonly used nuisance regressors during estimation of trial-by-trial parameter estimates: the six rigid body realignment parameters from motion correction, the Volterra expansion of the realignment parameters, a spike model removing variance caused by timepoints with high levels of instantaneous motion (framewise displacement), and global signal regression. By condition-based trial averaging, we mean averaging trial-specific activation estimates from each condition of a particular run. This potentially improves the signal-to-noise ratio (SNR) producing a more prototypical activation pattern but comes at the expense of reducing the number of training and testing examples. Within-run mean centering is a recommended technique that centers each voxel's mean trial activation from all trial estimates within each run (Lee and Kable, 2018). Finally we evaluate the benefits of using a fixed cost parameter versus tuning the SVM cost parameter within a nested cross-validated fold, a computationally intensive process particularly when implemented in searchlight or permutation testing. Each of these selections has important implications for both design choices and the implementation of an analytic pipeline.

These approaches were evaluated on two datasets: classification of button responses in somato-motor and auditory cortex and classification of simulated pattern responses. The measured fMRI data were obtained as part of a study in which observers made finger presses in response to discrimination judgements on visual stimuli. The somato-motor and auditory cortex regions were specifically chosen to generate strong *a priori* expectations of how decodable the finger presses would be. Somato-motor was expected to classify strongly whereas auditory cortex was expected to perform at chance. We applied the same data processing steps to simulated data, generated for many crossed levels of trial- and voxel- level noise, in

an additional effort to better understand the underlying dynamics of each method.

4.2 Methods

4.2.1 Human participant fMRI data

The first dataset we analyzed involved real fMRI data acquired from human participants for a study investigating the action observation network (currently under review). Twenty-five healthy adults, ranging in age from 21 to 42 years (mean = 24.7, sd = 3.6), participated in the experiment which was approved by the ethical review board of the University of California, Irvine. One participant was excluded from the study due to excessive head motion. Participants were scanned at the Facility for Imaging and Brain Research at the University of California, Irvine on a 3 Tesla Siemens Prisma MRI scanner (Siemens Medical Solutions) equipped with a 32-channel receive-only head coil. High resolution anatomical images were collected using a single T1-weighted magnetization prepared rapid acquisition gradient echo (MP-RAGE) sequence (repetition time (TR)/echo time (TE) = 2,000/1,990 ms, field of view = 256 x 256 x 176 mm, flip angle = 9 degrees, and spatial resolution = 1.0 mm^3 isometric). Functional images were acquired using T2*-weighted echo-planer imaging (EPI) pulse sequences (TR/TE = 1,500/30 ms, 68 slices with no gap, AP phase encoding direction, multiband factor = 4, field of view = 212 x 212 mm, flip angle = 79 degrees, spatial resolution = 2 mm^3)

This was an event-related study in which participants viewed short (3 second) animations of human avatars performing one of two actions. After viewing the clip, the participant's task was to press a button with the index or middle finger of their right hand reflecting a 2-alternative forced choice judgment on the action depicted. To prevent motor planning during the action vigenette, stimulus-response labels were randomized across the two buttons

and were not displayed until the vignette was completed. The screen cleared as soon as the participant made their response.

Trials were separated by a 3, 4.5, or 6 sec inter-trial interval (ITI), pseudo-randomized within each run such that, in total, each trial lasted 10.5, 12, or 13.5 seconds. The onset of each trial event, including the response interval, was synchronized with the onset of volume acquisition. The experiment was organized into 8 runs containing 24 trials each for a grand total of 192 trials. Of the 24 participants included in the study, 22 completed the full 8 runs while 2 participants only completed 7 runs.

Image preprocessing

Preprocessing was conducted using BrainVoyager QX v20.6 (Goebel et al., 2006). All functional images were slice-time corrected, motion corrected within and between runs, linearly detrended, and temporally high pass filtered (cutoff frequency 0.01 Hz). Scans were additionally corrected for susceptibility-induced magnetic field distortions using the field map method (Jezzard and Balaban (1995), implemented in BrainVoyager's COPE v1.0 plugin). All functional images were co-registered to each individual's T1-weighted image.

ROI definition

For the current investigation, our aim was to classify which button was pressed during the response intervals in a target and control region of interest (ROIs): left somatomotor (SomMot) and right primary auditory cortex (A1), respectively.

SomMot To identify brain areas activated by making button presses, we computed a group random-effects GLM containing a single predictor modelling all button presses within each

scan. Button presses were modelled as 200 msec events starting from the moment the button was depressed, and the predictor was convolved with a two-gamma hemodynamic impulse response function (Friston et al., 1998; Glover, 1999). The functional data from individual subjects were aligned to a template subject (a pilot participant) in surface space using cortex based alignment (Frost and Goebel, 2012). The group-aligned map of β-weights was then thresholded using a false discovery rate (FDR; Genovese et al., 2002) of $q < 1 \times 10^{-6}$ The resulting left hemisphere somatomotor ROI, contralateral to the right hand button presses, was identified on the cortical surface and projected back into native volumetric coordinates for each participant.

A1 Primary auditory cortex served as a control ROI, and was identified anatomically in each individuals' native anatomical images using Freesurfer's cortical surface atlas mapping algorithms in conjunction with the 1,000 atom resolution Schaefer atlas (Schaefer et al., 2018). This atlas emphasizes homogeneity of function within parcels, coupled with high resolution "atomic" parcellation in approximately equisized units, and therefore higher precision in identifying ROI boundaries.

Motion-related nuisance regressors

We evaluated the impact of 4 different types of motion-related nuisance regressors on MVPA classification results. Nuisance regressors included the detrended time series of the 6 rigid-body realignment parameters (R = [X Y Z pitch yaw roll]), estimated from the three-dimensional motion correction (3DMC) procedure performed during preprocessing. The 24 parameter Volterra expansion nuisance model included the 6 rigid body estimates and the preceding timepoint, as well as their first derivatives ((Fristen et al., 1996): [$R = [R\ R^2\ R_{t-1}\ R^2_{t-1}$], where t and $t-1$ refer to the current and immediately preceding timepoint).

Despiking (e.g. volume censoring) is commonly used to reduce variance accounted for by large head jerks (producing large changes in image intensity) which may not be captured well by the 3DMC or Volterra nuisance regressors (Lemieux et al., 2007; Satterthwaite et al., 2013; Yan et al., 2013). Despiking was performed by including in the model a matrix of "scan nulling" regressors (i.e. a heaviside function) targeting each corrupted timepoint identified as volumes with framewise displacement exceed 0.5 mm (Power et al., 2012). We took an additional step to drop all trials where three or more timepoints had been censored near the peak of the expected hemodynamic response, since trial-specific beta estimates in those situations become highly unreliable.

Global signal regression (GSR) is commonly used to remove distributed, non-neural sources of variance contaminating the images. The global signal was computed as the average intensity measured from the white matter and ventricles over time and the first derivative thereof.

All nuisance regressions were performed prior to estimating the trial-by-trial activation estimates, which constitute the data passed on to the classifier. The denoised timeseries was produced by regressing each voxel's timeseries onto the respective matrix of nuisance regressors and collecting the model residuals.

Trial-specific activation estimates

Trial-specific activation estimates for each voxel in the ROIs were derived by z-scoring the timeseries over time and then applying the least squares separate (LSS) approach (Mumford et al., 2012; Turner et al., 2012). The LSS procedure uses a separate GLM to estimate the

activity, $\hat{\beta}$, of each trial, with the model for the i^{th} trial given by

$$Y = X_{LSS_i}\beta_{LSS_i} + \epsilon_i \tag{4.1}$$

such that the design matrix for the i^{th} trial, X_{LSS_i}, contains one regressor of interest modeling the boxcar convolved BOLD response to the i^{th} trial and two other nuisance regressors modelling all other trials grouped by the type of button pressed. The estimate for the first trial is therefore given by

$$\hat{\beta}_{LSS_{i,1}} = c((X_{LSS_i}^{\top}X_{LSS_i})^{-1}X_{LSS_i}^{\top}Y) \tag{4.2}$$

where c is the row vector $\begin{bmatrix}1 & 0 & 0\end{bmatrix}$

4.2.2 Simulated multi-voxel activation patterns

The second type of data to which we applied our set of processing strategies was simulated multi-voxel activation patterns. We begin by discussing the derivations we used to simulate the multivariate patterns.

Analytic framework of simulations

There is wide agreement that BOLD fMRI data contains multiple sources of variability, including trial-, voxel-, and run-level variability (Friston et al., 1994). To generate multivariate response patterns that properly incorporate all these unique variance components, we used a

multilevel modelling approach (for similar models, see Diedrichsen et al., 2013; Davis et al., 2014).

The first level of the model is given by equation 4.3 and describes how activation in voxels, regardless of the type of condition, varies randomly from trial to trial.

$$\begin{aligned} A_{tvs} &= \alpha_{vs} + X_s\beta_{vs} + \epsilon_{tvs}, \\ e_{tvs} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{4.3}$$

Here, we assume that the data are summary statistics (e.g. LSS regression coefficients) representing the activation, A_{tvs}, observed on trial t, voxel v, and scan s. The variable X is a N_{trials} x $N_{covariates}$ design matrix and the observed activation is represented as a linear combination of the baseline activation (or intercept), α_{vs}, plus the product of the beta coefficients, β_{vs}, and X, plus trial-specific deviations, ϵ_{tvs} (Equation 4.3). These trial-level errors are assumed to follow a normal distribution with mean zero and variance σ^2.

The voxel-level model (Eq 4.4), describes how the multivariate patterns constitute repeated measurements across voxels that vary in two important respects: firstly, voxels vary in their mean baseline activation across trials of all types and secondly they vary in the effect of the experimental conditions.

$$\begin{aligned} \alpha_{vs} &= \alpha_s + \epsilon_{\alpha vs}, \\ \beta_{vs} &= \beta_s + \epsilon_{\beta vs}, \\ \begin{pmatrix} \epsilon_{\alpha vs} \\ \epsilon_{\beta vs} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mu_{\alpha s} \\ \mu_{\beta s} \end{pmatrix}, \begin{pmatrix} \tau_\alpha^2 & \rho\tau_\alpha\tau_\beta \\ \rho\tau_\alpha\tau_\beta & \tau_\beta^2 \end{pmatrix} \right) \end{aligned} \tag{4.4}$$

This level of the model characterizes the entire population of voxels as having a mean baseline activity in each scan, α_s, and a mean effect of the experimental contrast in each scan, β_s. Voxel-specific deviations to each of these summary statistics are allowed by the inclusion of error terms, $\epsilon_{\alpha vs}$ and $\epsilon_{\beta vs}$ respectively. The regression parameters in equation 4.3, α_{vs} and β_{vs}, are therefore not fixed but conceptualized as random variables with a multivariate Gaussian probability distribution across voxels and scans. This probability distribution is catalogued by two main structures: the mean vector of coefficients, $\mu_{\alpha s}$ and $\mu_{\beta s}$ and the variance-covariance matrix containing the between voxel variances in baseline (τ_α^2) and effect of the experimental contrast (τ_β^2) as well as their covariance, $\rho\tau_\alpha\tau_\beta$. The parameters τ_α and τ_β are of particular importance as they inherently model voxel-level variability and fit the common understanding that in any ROI there are, to greater or lesser extent, mixtures of both task-relevant and task-irrelevant voxels.

The third, and final, level of our model (Eq. 4.5) accounts for the finding that there are often signal-related shifts in the mean activity of all trials within each run. There may be many causes of these run-level shifts including drifts in attention, changes in physiological arousal, or between-run differences in proportions of trial types.

$$\begin{aligned} &\alpha_s = \gamma + \epsilon_{\alpha s}, \\ &\epsilon_{\alpha s} \sim \mathcal{N}(0, \omega^2) \end{aligned} \tag{4.5}$$

The variance component of interest in this model corresponds to run-by-run variability in the mean activity of all trials across voxels. Like other levels, this is implemented by an error term, $\epsilon_{\alpha s}$, which quantifies each run's deviation from the expected value of all runs, α_s. These errors are also assumed to be normally distributed with mean zero and variance ω^2.

Substituting each of these levels into the others produces the combined equation (Eq.

4.6) demonstrating that activation A on trial t in voxel v for scan s is a combination of fixed effects of the experimental variables as well as trial-, voxel-, and scan-level random effects.

$$A_{tvs} = \gamma + \epsilon_{\alpha s} + \epsilon_{\alpha vs} + X_s\beta_s + X_s\epsilon_{\beta vs} + \epsilon_{tvs} \tag{4.6}$$

Simulation methods

Multi-voxel activation patterns were generated for a simulated ROI containing 200 voxels using a design consisting of two trial types with 12 repetitions each, for a total of 24 trials per run. Each simulated study consisted of 8 runs of data which we generated 30 times, simulating 30 subjects. We did not include any subject-level random effects in our model, though all simulations were generated from independent draws of the model.

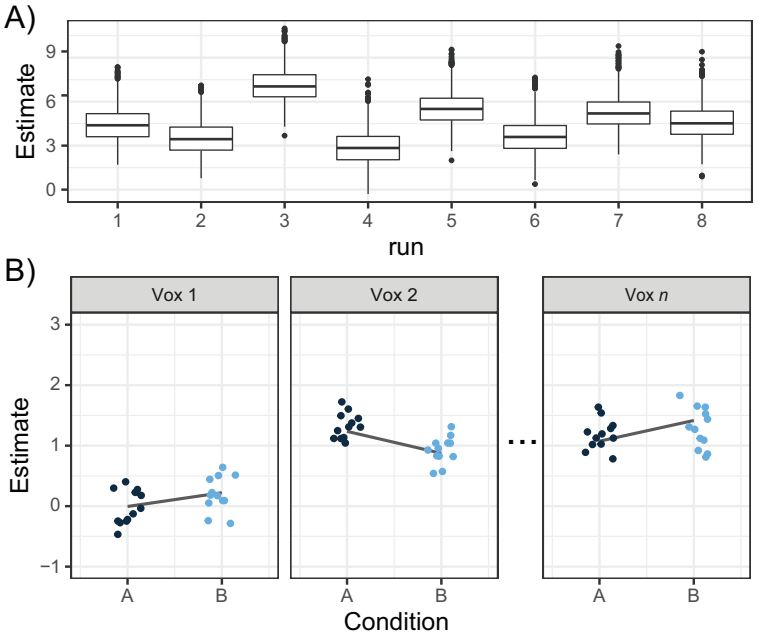

Figure 4.1: A) Illustration of run-level shifts in mean activity across all trial types. B) Three sample voxels illustrating how trial-specific estimates were created by first generating an ideal line, with unique intercept and slope, reflecting each voxel's true response to the two experimental conditions, and then adding normally distributed noise to each trial.

Figure 4.1 illustrates how the pattern data was generated. First, for each run we generated a unique run-level shift in mean activity across all conditions ($\epsilon_{\alpha s}$) through independent draws from a Gaussian distribution with mean zero and standard deviation (ω) equal to 1.5. Next, we generated a unique intercept, (α_{vs}) and slope (β_{vs}) for each voxel by sampling from a bivariate Gaussian distribution with mean intercept set to 1, mean slope set to zero and standard deviation of intercepts set to 1. The voxel-level variability in slope (τ_{β_1}) was varied across simulations over four values (0.01, 0.05, 0.1, and 0.2). The correlation between intercepts and slopes ($\rho\tau_{\alpha}\tau_{\beta_1}$) was set to zero.

Since we coded the two trials in the design matrix as -0.5 and 0.5 (deviation coding scheme) this meant that the voxel-specific intercept represented the voxel's average (baseline) activation for trials of all types and the voxel-specific slope (beta weight) related the voxel's activations to the effect of the experimental contrast (i.e. effect size) in that voxel.

Within each simulated run, the value of the run-specific shift in mean activity ($\epsilon_{\alpha s}$) was added to each voxel's baseline as a single constant. Each voxel's general response to the experimental conditions (β_{vs}), however, was consistent across runs. Then, noise was generated for each trial with mean zero and standard deviation σ, varied over three values (0.7, 1.0, and 1.3) reflecting low, medium, and high trial noise respectively. Finally, trial-by-trial activation estimates were generated by combination of all fixed and random effects (including trial-, voxel-,and scan-related noise).

4.2.3 Classification

MVPA was performed on the acquired and simulated data using a linear support vector machine (svm) with default scaling (all voxels and observations standardized to zero mean and unit variance) as implemented in R using the e1071 package (Meyer et al., 2018). The labels classified in the human subject fMRI data corresponded to which of two response buttons

were pressed during the experiment. In the simulated data, the svm algorithm classified labels corresponding to the two simulated trial types - type A and type B. All analyses were performed within subject and classifier performance was evaluated by computing the mean classification accuracy across subjects.

Trial averaging by condition and run

We investigated the impact of two methods of aggregating ER data for multivariate analyses. The first and most commonly used method involves training and testing the classifier on patterns composed of separate activation estimates from each trial. In these situations, the number of training or testing examples passed to the classifier is equal to the number of trials within the current data partition (i.e. exchangeability block, Winkler et al., 2014). Due to unavoidable temporal dependencies and the fact that trial estimates from the same run are more similar than trial estimates from different runs (Pereira et al., 2009; Etzel et al., 2009) a common choice is to partition the data on runs leading to separate observations for each trial in each run.

The second approach we investigated involves averaging all trial-specific estimates of each type within run (e.g. averaging all Type A trials together within run and averaging all Type B trials together within run, etc.). We refer to this method as "Avg-1" because it results in one averaged observation per class within each run. Averaging trial-specific estimates by run has the potential to reduce trial-variability that could be a major source of noise limiting classifier performance, but comes at the expense of greatly reducing the number of training and testing examples supplied to the classifier. Reducing training observations has the potential to impoverish the fit of the decoder to the data whereas reducing test observations impacts the precision with which the prediction error can be estimated within each cross-validated fold thereby increasing between-subject variance of the final classification accuracy

In a third approach we evaluate a hybrid model to strike a better balance between the opposing effects of improving signal-to-noise ratio (SNR) by trial averaging and maintaining a large enough test set. In this "Avg-2" approach, we produce two summary statistics of activity per class within each run. To do so, we randomly sampled half of the trials from each condition within a run without replacement and averaged each group of trials separately thus producing two averaged activity estimates per condition. For our datasets, both of which involve two trial types, this results in a test set of four observations. To reduce sampling error, this process was repeated ten times within each fold and the resulting classification accuracies were averaged.

Run-wise mean centering

Mean centering was performed by subtracting each voxel's run-level mean beta estimate, for all trial types, from the estimates within that run (Lee and Kable, 2018). Though we subsequently refer to this operation as"mean centering", it should be noted that this type of mean centering is distinct from the default mean centering performed by the majority of svm algorithms since it is performed on a run-by-run basis.

Cost tuning

Many MVPA studies fit linear svm's to multi-voxel response patterns using a fixed cost parameter, C, of 1. However, optimizing C by minimizing the cross-validated test-error has been shown to improve the predictive power of a classifier (Hastie et al., 2001). On both datasets, we evaluate the impact of using a cost parameter that is either fixed ($C = 1$) or tuned over 12 values from 2^{-12} to 2^{1}. This was achieved using a nested cross-validated fold in which an inner second level-split was generated with one run of the original training data left out and used to evaluate the performance of each value of C. This was repeated for

all folds of the nested loop, and the lowest value of C maximizing predictive accuracy of the inner cross-validation test data was then applied to the training and testing data in the external loop.

Statistical analyses

To statistically evaluate the individual and joint impact of the tested methodological decisions on MVPA decodability, we constructed multilevel linear models (MLMs), a form of hierarchical linear regression, for both the real and simulated datasets. MLM was chosen because it allows heteroscedasticity to be specified and explicitly models dependency in the data (i.e., nested structure) which otherwise leads to underestimation of standard errors in ordinary least squares regression models. This was most important for the human subject fMRI dataset, in which dependencies existed between the data from the two ROIs, which were collected from the same participants in the same scan sessions, and thus shared variance.

An MLM was constructed using each participant's cross-validated classification accuracy as dependent variable and nesting scores from the two ROIs within participants. Fixed factors included each methodological decision (type of motion-related nuisance regression, the type of trial averaging, presence or absence of mean centering, and cost parameter choice) along with the type of ROI (somato-motor versus the control region). Participants were modelled as random intercepts only. The data grouping structure of the simulated patterns diverged considerably from the human participant fMRI data and therefore changes were incorporated into its own MLM. In the simulated data, multiple subjects were simulated by independent draws from the model, repeated for 25 different combinations of trial-level noise and voxel-level variability in the overall effect of the conditions. Though the data processing steps were applied at the level of each simulated subject, ultimately we wanted to understand the impact of these decisions across many possible parameterizations of the data-generating model. We refer to each unique combination of trial-level noise and voxel-level variabil-

ity between conditions as the "dataset", with each dataset containing activation patterns from multiple simulated subjects. The random effects structure of the MLM was therefore specified by a random intercept for each dataset (to account for differences in baseline classification accuracy across different parameter settings) with simulated subjects nested within datasets. The fixed effects included the same methodological factors that were tested in the human fMRI data with the exception of timeseries preprocessing (motion-related nuisance regression), which was not evaluated because the data were simulated at the level of trial activation estimates rather than timeseries.

Prior to fitting the MLMs, we checked the assumption of homogeneity of variance across the levels of each fixed factor in both human and simulated data. Results showed strong violations of this assumption for several of the factors, including trial averaging method (human data: $F(2, 2864) = 333.01, p < 0.001$; simulated data: $F(2, 8997) = 887.61, p < 0.001$), run-wise mean centering application (human data: $F(2, 2865) = 127.21 p < 0.001$; simulated data: $F(1, 8998) = 867.54, p < 0.001$), data cleaning approach (human data: $F(4, 2862) = 2.86, p = 0.022$) and cost selection method (simulated data only: $F(1, 8998) = 61.454, p < 0.001$)). Therefore, the heteroscedasticity was included in the model by means of a variance function allowing different variances per stratum, computed as the ratio of each variance to a reference level. For the human participant data, specifying unique variances for fully crossed levels of trial averaging, mean centering and data cleaning was not computationally feasible due to the sheer number of levels and model convergence issues. Therefore, we specified unique variances for the two most critical heteroscedastic factors based on the magnitude of the F statistic from Levene's test. This meant that heteroscedasticity was modelled for trial averaging and mean centering but not the data cleaning approach. For the simulated data, unique variances were modelled for all heteroscedastic factors including trial averaging, mean centering, and cost tuning.

All MLM analyses were conducted in R using the 'nlme' package **?**. The significance of

each factor (or interaction) was assessed using Likelihood Ratio Tests comparing each model to reduced models lacking the variable (or interaction) in question.

4.3 Results

4.3.1 Human participant data

To establish a baseline for comparing the impact of our four methodological choices on mean classification accuracy, we begin by reporting group-level results in each ROI using only the most common processing combinations: no motion-related nuisance regression besides that typically deployed during preprocessing, training and testing on separate estimates for each trial, no run-wise mean centering of trial estimates, and training the SVM with a fixed cost value of 1.

Across all 24 subjects, the left somato-motor region (SomMot) classified the type of button pressed (first finger vs. second finger) with a mean classification accuracy of 56.90% (SE = 1.36), which non-parametric permutation tests revealed to be significantly higher than that expected by chance ($p_{perm} < 0.001$, mean of null distribution = 50.10%, SD = 0.88). The ROI serving as a control region (primary auditory cortex or A1) classified the type of button pressed with a mean accuracy of 51.89% (SE = 0.82) which is 5.01% lower that that obtained in SomMot and slightly higher than that expected by chance ($p_{perm} = 0.029$, mean of null distribution = 50.17%, SD = 0.90).

Figure 4.2 displays mean classification accuracies for all combinations of methodological choices: type of motion-related nuisance regression used (none, 3DMC, Volterra expansion, despiking, or GSR), level of beta averaging (using separate estimates for each trial or "no-avg", averaging half of the exemplars of each class within each run or "2-avg", or averaging all

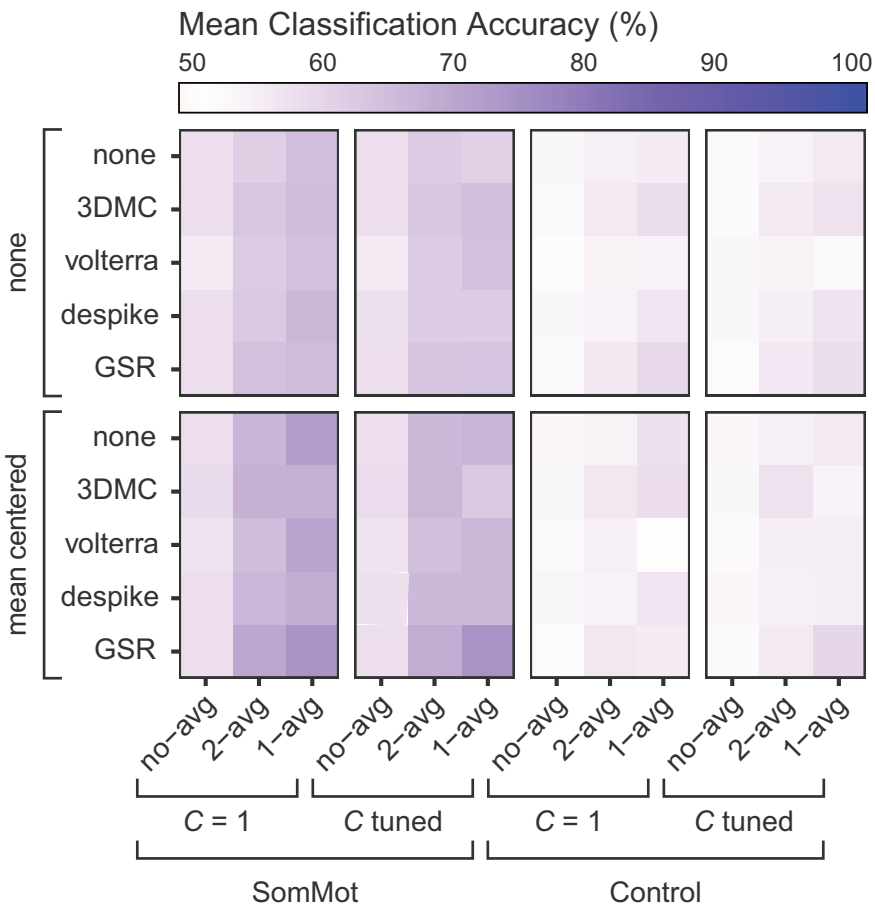

Figure 4.2: Average classification accuracy for all combination of methodological decisions grouped by ROI (SomMot = somatomotor; Control = primary auditory cortex).

exemplars of each class within run or "1-avg"), run-wise mean centering, and cost parameter selection (using a fixed cost value of one or tuning the cost parameter in a nested cross-validation loop).

The multilevel linear model (MLM), including all methodological factors as well as the type of ROI as fixed factors, revealed a significant main effect of ROI on classification accuracies ($\chi^2(10) = 12.208, p < 0.001$) with SomMot classifying the type of button pressed with significantly higher accuracy than the control region ($b = 5.434$, $SE = 1.458$, $t(23) = 3.727$, $p = 0.001$).

However, the main effect of ROI was qualified by a significant three-way interaction

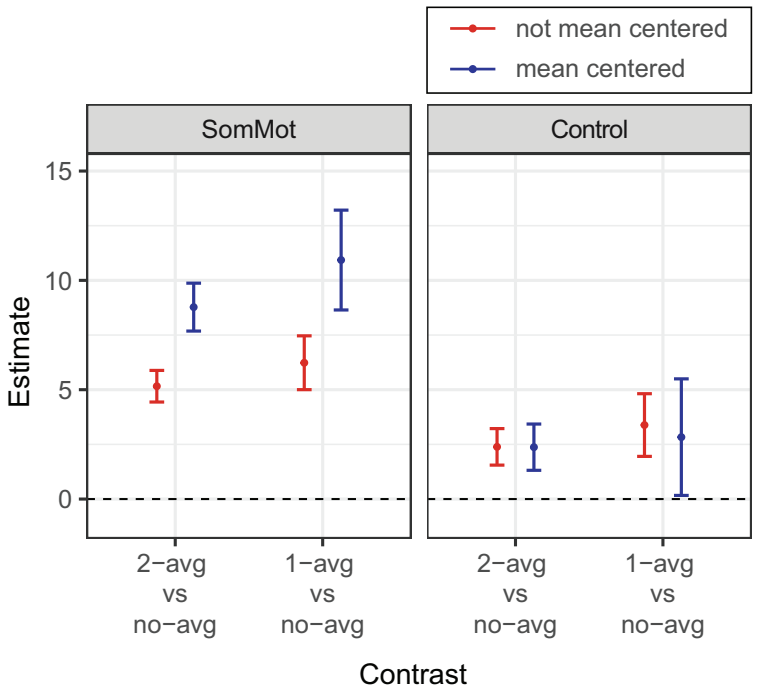

Figure 4.3: Fixed effect parameter estimates from multilevel linear models (MLMs) showing the interaction between ROI, trial-averaging technique, and within-run mean centering in the human subject fMRI dataset. The 95% confidence intervals were computed for the contrasts comparing the two conditions where trials were averaged within runs (2-avg and 1-avg) versus data comprising a separate estimate for each trial. Parameter estimates above zero indicate that averaging trials by run produced higher accuracies than training/testing the classifier on individual trial estimates. Estimates were computed from four separate MLMs fixing the level of ROI and mean centering.

between ROI, the method of trial averaging, and the presence of within-run mean centering ($\chi^2(65) = 19.947, p < 0.001$). To interpret this interaction, planned contrasts compared classification accuracies for the two methods of computing condition-based trial averages (2-avg and 1-avg) to the results obtained by classifying data consisting of separate estimates for each trial (no-avg). Figure 4.3 shows all parameter estimates along with 95% confidence intervals.

First, fixed effect parameter estimates revealed that classification accuracies were higher in both the 2-avg and 1-avg conditions compared to no-avg (2-avg vs no-avg: $b = 4.180$, $SE = 0.235$, $t(2813) = 17.761$, $p < 0.001$; 1-avg vs no-avg: $b = 5.124$, $SE = 0.430$, $t(2813) = 11.930$, $p < 0.001$). Additionally, classification accuracies in the 2-avg and 1-avg conditions

were even higher when the data was also mean centered within each run prior to training and testing the classifiers (2-avg vs no-avg with mean centering: $b = 1.795$, $SE = 0.500$, $t(2801) = 3.591$, $p < 0.001$; 1-avg vs no-avg with mean centering: $b = 2.072$, $SE = 1.030$, $t(2801) = 2.011$, $p = 0.044$). Furthermore, this increase in classification accuracies by averaging, when coupled with run-wise mean centering, was found to exist only in the SomMot region (2-avg vs no-avg with mean centering in SomMot vs control: $b = 3.664$, $SE = 0.960$, $t(2780) = 3.818$, $p < 0.001$; 1-avg vs no-avg with mean centering in SomMot vs control: $b = 5.250$, $SE = 2.034$, $t(2780) = 2.581$, $p = 0.010$).

In order to better understand the effect of trial averaging within each ROI, two separate MLMs were constructed using only data from each ROI. Both models revealed that averaging trials together within each run improved classification accuracies over not averaging any trial estimates (SomMot: $\chi^2(10) = 377.355, p < 0.001$; Control: $\chi^2(10) = 66.382, p < 0.001$). However, trial averaging improved decodability in SomMot (2-avg vs no-avg: $b = 6.025$, $SE = 0.325$, $t(1401) = 18.540$, $p < 0.001$; 1-avg vs no-avg: $b = 7.120$, $SE = 0.570$, $t(1401) = 18.540$, $p < 0.001$) considerably more than it did in the control region (2-avg vs no-avg: $b = 2.312$, $SE = 0.328$, $t(1414) = 7.058$, $p < 0.001$; 1-avg vs no-avg: $b = 3.155$, $SE = 0.638$, $t(1414) = 4.946$, $p < 0.001$). As shown in Figure 4.3, confidence intervals were wider for the 1-avg condition compared to the 2-avg condition, reflecting larger between subject variation in classification accuracies when trial estimates for an entire run are averaged into a single exemplar per condition. This was true of both ROIs, indicating the increase in variance is linked to having fewer observations rather than the presence or absence of true signal embedded in the data.

The impact of the type of motion-related nuisance regression (i.e. data cleaning step) applied prior to pattern estimation can be seen by comparing the rows of Figure 4.2 within each panel, with parameter estimates shown in Figure 4.4. There was a significant main effect of the type of data cleaning step applied ($\chi^2(14) = 17.190, p = 0.002$) as well as a

significant interaction between the data cleaning step and the amount of trial averaging applied within run ($\chi^2(34) = 19.423, p = 0.010$).

This interaction was broken down by comparing classification accuracy resulting from each data cleaning step to accuracies obtained from using no motion-related nuisance regression separately for each of the two contrasts on the trial averaging level (2-avg vs no-avg and 1-avg vs no-avg). These contrasts revealed that applying global signal regression (GSR) to the timeseries before extracting trial estimates significantly improved classification accuracy in the 2-avg condition compared to no-avg (b = 2.222, SE = 0.729, $t(2795)$ = 3.049, p = 0.002) as well as in the 1-avg condition compared to no-avg (b = 2.881, SE = 1.347, $t(2795)$ = 2.140, p = 0.033).

No other data cleaning steps significantly differed by the type of trial averages computed. Parameter estimates from the main effect of data cleaning step revealed that, averaged across all other factors, using the Volterra expansion as nuisance regressor significantly lowered classification accuracy (b = -0.800, SE = 0.285, $t(2815)$ = -2.8100, p = 0.005).

The impact of cost parameter selection can be assessed by comparing the first and last three columns within each group of ROIs in Figure 4.2. Overall, choosing a fixed cost value of one versus tuning the cost parameter did not impact mean classification accuracies nor interact with any of the other three processing decisions (all *p's n.s.*).

4.3.2 Simulated Data

Figure 4.5 shows mean classification accuracies from simulated pattern data for all combinations of methodological factors (level of trial averaging, within-run mean centering, and cost parameter selection). These methodological approaches were applied to several simulated datasets generated with varying levels of trial-level variability (σ^2) and voxel-level variability

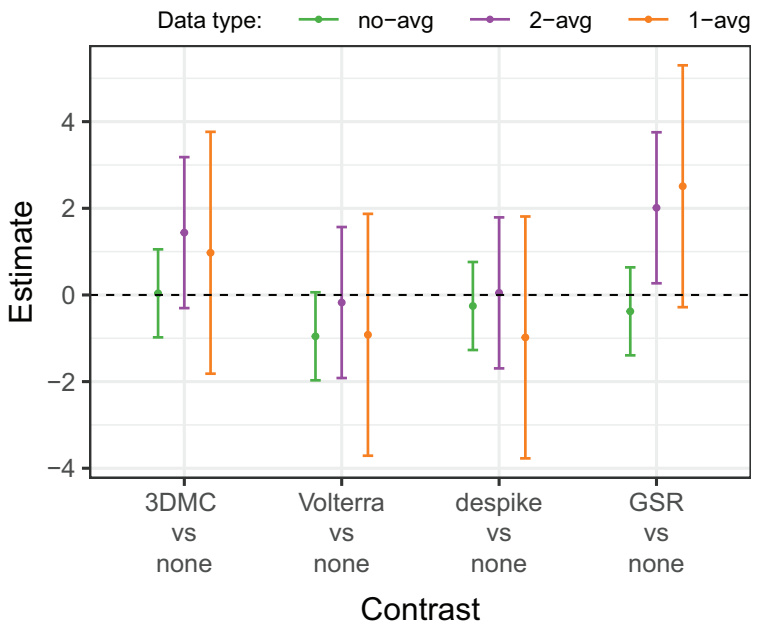

Figure 4.4: The interaction between type of motion-related nuisance regression (data cleaning) and trial averaging within the human subject fMRI dataset. Results show parameter estimates with 95% confidence intervals from a multilevel linear model. Contrasts were set on the type of data cleaning step applied by comparing each data cleaning step to using no nuisance regression at all. Contrasts on the type of trial averaging method compared each method to the baseline approach using a separate activation estimate for each trial. Therefore, estimates above zero indicate that the given data cleaning step produced higher classification accuracies for the given trial averaging method versus using no trial averaging.

in experimental contrast (slope or τ_{β_1}).

Overall, mean classification accuracies varied with both trial- and voxel-level variability. Classification accuracy increased when trial-level variability decreased, indicating that more consistent patterns across trials improved decoding. Moreover, classification accuracy increased as voxel-level variability in the mean effect of the experimental conditions increased, consistent with reports that increased variance in the spatial patterns, even when the fixed effect size is zero, improves classifier performance (Davis et al., 2014).

An MLM was conducted to determine which, if any, of the data processing choices impacted mean classification accuracies across all combinations of model parameter settings used to generate the pattern data. The analysis yielded many significant main effects and

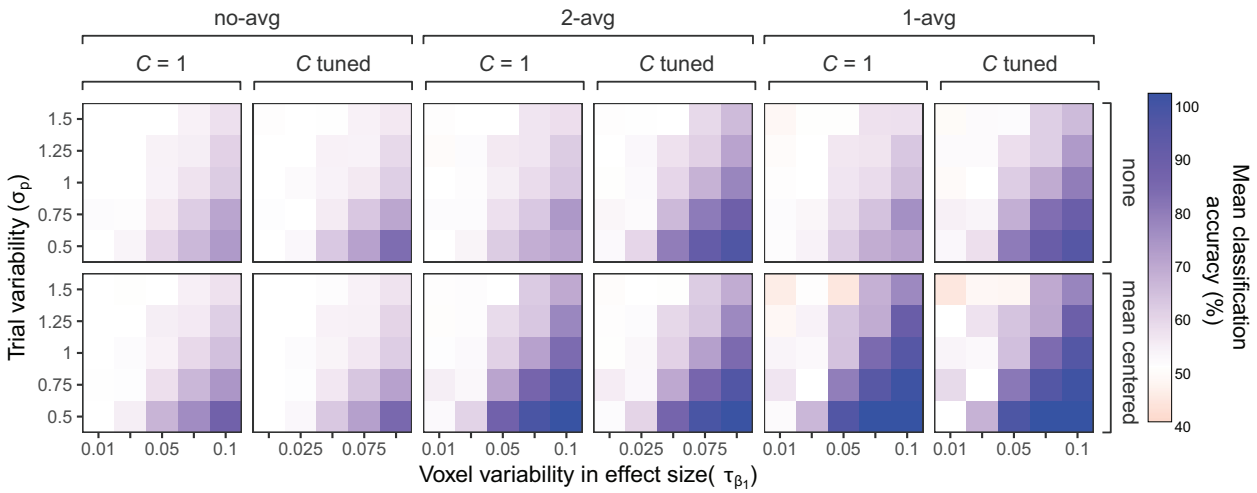

Figure 4.5: Average classification accuracy for each combination of methodological factors applied to simulated pattern data. Pattern data was simulated for several combinations of trial-by-trial variability, σ, and voxel-level variability in the mean difference between trials of each type, τ_{β_1}. Each colored square displays mean cross-validated classification accuracy computed across 30 different simulations.

interactions, therefore we focus on the highest order interaction which occurred between all three processing choices ($\chi^2(20) = 67.104, p < 0.001$). Figure 4.6 displays this interaction graphically by plotting the MLM parameter estimates along with confidence intervals for the two contrasts on trial averaging from four simpler MLMs holding mean centering and cost choice constant.

Results from the omnibus MLM revealed that both methods of trial averaging improved classification accuracies over using separate trial estimates (avg-2 vs no-avg: $b = 5.173$, $SE = 0.182$, $t(8248) = 28.368$, $p < 0.001$; avg-1 vs no-avg: $b = 5.369$, $SE = 0.249$, $t(8248) = 21.567$, $p < 0.001$). Furthermore, this improvement from averaging trials (both 2-avg and 1-avg) was made significantly better when the data were also mean centered within runs (avg-2 vs no-avg: $b = 4.842$, $SE = 0.358$, $t(8248) = 13.524$, $p < 0.001$; avg-1 vs no-avg: $b = 8.530$, $SE = 0.588$, $t(8248) = 14.497$, $p < 0.001$). Finally, tuning the cost parameter improved classification accuracies for both trial averaging methods but much less so when the data had been mean centered within each run (avg-2 vs no-avg: $b = -5.189$, $SE = 0.701$, $t(8239) = -7.397$, $p < 0.001$; avg-1 vs no-avg: $b = -4.865$, $SE = 1.167$, $t(8239) = -4.169$, p

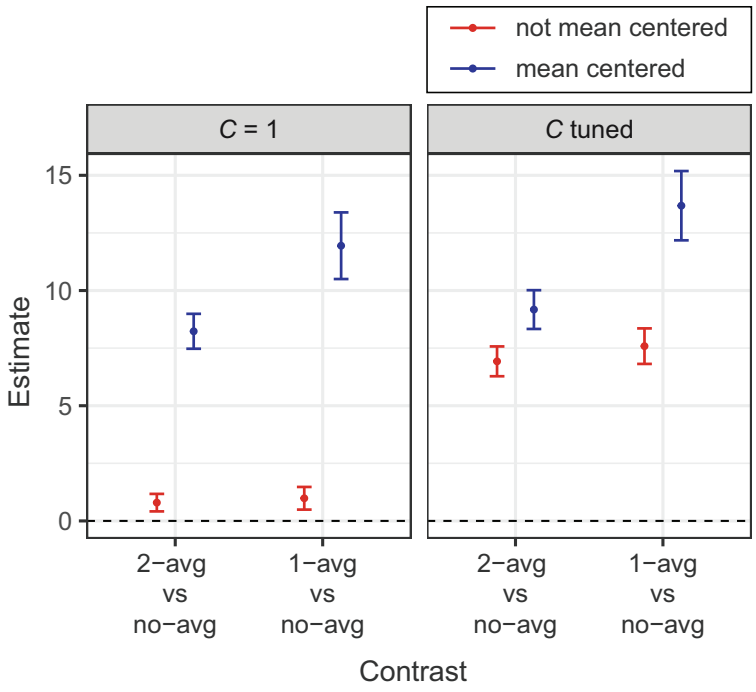

Figure 4.6: The three way interaction between trial averaging, cost tuning, and mean centering present in the simulated pattern data. Fixed effect parameter estimates and 95% confidence intervals were computed from four multilevel linear models contrasting the trial averaging method versus using separate trial estimates while fixing the method of mean centering and cost parameter selection method. An estimate above zero indicates that the trial averaging technique deployed improved mean classification accuracy versus training/testing on separate trial activation estimates.

< 0.001).

4.4 Discussion

We evaluated the impact of four methodological decisions on MVPA-decoded classification accuracies in both real and simulated fMRI data. Several of the methodological considerations were selected, in part, because of their common use in fMRI univariate and functional connectivity analyses, with the potential benefits when implementing them for multivariate pattern analysis unclear. This analysis is intended to serve as a practical guide for researchers wishing to optimize multivariate classification analyses without the risk of causing spurious

results by testing each method directly on experimental hypotheses of interest.

Some general observations across these analyses warrant attention. First, methodological approaches leading to large improvements in classifier performance did so in the context of both real and simulated datasets. In this analysis, that is most prominently the case with run-wise trial averaging coupled with mean-centering. The benefits of these approaches for the classification of both real and simulated data is evidence that the potential to improve classification is not limited to highly specific characteristics of either dataset. With that said, future studies should test the effectiveness of these methods across a wider range of experimental designs and regions of interest.

Secondly, the methods producing significant improvements often interacted in complex ways, highlighting the complex dynamics inherent to SVM analyses of multivariate pattern data. For example, the benefit of global signal regression for classification accuracy was only apparent for trial averaged data, with no improvement observed for MVPA conducted on individual trial exemplars. It is for precisely these interactions that motivated this evaluation of processing pipelines.

Lastly, while classifying button presses in real human participant data, the improvements brought about by these decisions were much larger in a region which we had strong *a priori* expectations for highly accurate classification (somato-motor) versus a control region (primary auditory cortex). This is very reassuring, as one does not want to unintentionally introduce bias to the classifier algorithm, as has been observed with some feature-reduction approaches (Ambroise and McLachlan, 2002).

4.4.1 Trial averaging

It is generally advised to use as many observations for training the classifier as possible (Pereira et al., 2009; Etzel et al., 2009). Therefore classifying based on separate estimates for each trial may be thought to give better results because it maximizes the training set size. Alternatively, averaging trials by condition and run could reduce one of the main sources of noise in fMRI data - trial-level noise or noise arising from repeated measures even for the same conditions - thus enhancing the discriminability of the multivariate patterns by improving the signal-to-noise ratio (SNR). We found that reducing noise by trial averaging produced one of the largest gains in classification accuracies among the methods we tested and this result was consistent for both real and simulated data.

Given the trade-offs anticipated from trial averaging (reduced number of training/test exemplars versus trial-level noise reduction), two findings from this analysis are particularly surprising. The first is the magnitude of improvement caused by trial averaging. In the human participant data when classifying button presses in somatomotor cortex, the improvement in mean classification accuracy brought about by averaging all trials of each type within runs was 6.3% and when coupled with within-run mean centering (discussed below) the improvement climbed to 10.9%.

Another important finding is that trial averaging causes a marked increase in the between-subject variability of the classification accuracies. One possible explanation for the increased variance may be the reduced size of the test set used to assess the prediction error of the classifier at each split of cross-validation. When estimating classification accuracy using the more traditional trial-based approach, the algorithm is tested on an entire run of samples, which in this study consisted of 24 exemplars (twelve from each condition). When all trial estimates are averaged within run to one per condition (1-avg), the cross-validated test error is evaluated with only two observations per split, constraining the test

error to only a few possible values. It has been theoretically shown that with training sets of the same size, having more data for validation decreases the variance of the estimated accuracy (Arlot and Celisse, 2010). Therefore, we conclude that it is important to strike a balance between maintaining a large enough test set to yield a stable estimator of classifier performance and reducing trial-level noise through trial averaging.

As expected, doubling the number of items in the test set nearly halved the between-subject variability in classification accuracy (the 2-avg condition compared to the 1-avg condition). However, this was also associated with a reduction in mean classification accuracy, which we interpret as due to a higher SNR from having fewer trials included in each average. Therefore, when planning an MVPA study, researchers should carefully weigh any knowledge they have about the amount of trial-level noise inherent to the region(s) under study versus the increased test-set variance brought about by limiting that noise through averaging trials of variously sized subsets.

4.4.2 Run-wise mean centering

It is widely recognized that each scan in a session is associated with a unique shift in the mean activity across all trial types. These shifts may be due to drifts in attention, changes in physiological arousal, or between-run differences in the proportions of different trial types. Whereas condition-based trial averaging was used to reduce trial-by-trial variability, the variance component that run-wise mean centering aims to reduce is run-level variation in the baseline activation for all trials within each run.

The mechanism by which this improves classification is intuitive: Since the cross-validation procedure for most MVPA studies is partitioned on runs, training a classifier using exemplars from run-shifted distributions introduces artificial clusters within the training data. This in turn, should be anticipated to impair the classifier's ability to find a stable

separating hyperplane between blocks of training data from different runs or to generalize to test data from new runs. Our data confirm this hypothesis in both real and simulated datasets replicating other studies (Lee and Kable, 2018).

Furthermore, we show that mean centering interacts with the method of trial averaging. With separate trial estimates, mean centering did not make a significant difference to mean classification accuracy, perhaps because of the increased variance associated with the noisy trial exemplars in effect masks the partitioning effect of run-wise variance. However, with the inclusion of run-wise trial averaging, trial variance is reduced and large improvements are seen when mean centering is included.

4.4.3 Cost selection

Tuning the SVM cost parameter, C, within a nested cross-validated loop is a computationally intensive process, particularly when conducted over many regions of interest (as in a searchlight MVPA analysis) or when implemented as part of permutation testing. Our analysis shows that tuning C versus using a fixed value of 1 depends on the statistical structure of the underlying dataset.

Cost tuning did not have a significant impact on classification performance using the human participant data in either ROI. In contrast, cost tuning significantly interacted with trial averaging and mean centering in the simulated datasets such that cost tuning improved classification accuracies when trials were run-averaged and mean centered prior to classification. One explanation for this finding is that setting C high, such as when $C = 1$, leads to a higher likelihood of overfitting the classifiers (Hastie et al., 2001), a significant disadvantage when the classifier is trained and tested on data composed of blocks with distinct shifts in mean activity. Thankfully, our results show that cost tuning can be omitted from MVPA pipelines without penalty by simply mean centering the data within each run prior

to classification, which is a computationally simpler and faster operation.

4.4.4 Motion-related nuisance regression

It has long since been recognized that head movements severely compromise the quality of fMRI data (Fristen et al., 1996; Hajnal et al., 1994), sparking many endeavors to denoise the BOLD signal through reference time series capturing motion-related fluctuations (for a review see Caballero-Gaudes and Reynolds, 2017). These reference signals are sometimes added as nuisance regressors to the design matrix that is fit to the voxel time series and therefore constitute additional data cleaning above and beyond the volume registration performed during normal preprocessing. Though it is now standard to use such nuisance regressors to denoise BOLD data in preparation of functional connectivity analyses, no studies to date have examined their impact on multivariate decoders.

Prior to estimating trial-specific activation estimates, we denoised the raw timeseries using four different types of motion-related nuisance regressors: the 6 rigid-body realignment parameters (3DMC), the 24 parameter Volterra expansion, a despiking model using a FD threshold of 0.5 mm, and the average signal from the white matter and ventricles (GSR). Of all four nuisance regressors, we only found GSR to lead to a significant improvement to mean classification accuracy. This improvement, though significant, was small and present only for analyses where trials were averaged (both 2-avg and 1-avg conditions). In no cases did any of the motion-related nuisance regressors significantly lower accuracies. This reveals that, though use of GSR may improve decodability in some contexts, multivariate classifiers are resilient to motion-related sources of noise.

4.4.5 Conclusions

Though hard and definitive guidelines regarding the tested methods cannot be drawn for all designs and tasks, the current investigation reveals that across real and simulated datasets mvpa-decodability can be significantly improved through trial averaging, mean centering, and inclusion of Global Signal Regression.

Chapter 5

Conclusion

The experiments in this dissertation have focused on how observers perceive the effects of human motion across diverse modalities. The first study investigated the relationship between several measures of speech articulation and subjective ratings of vocal attractiveness in both male and female talkers when producing vowels in /bVd/ context and sentences containing the four most peripheral 'corner' vowels. Multiple measures of working vowel space were computed from continuously sampled formant trajectories and combined with measures of speech timing known to co-vary with careful articulation. Partial least squares (PLS) regression modelling predicted ratings of vocal attractiveness for male and female talkers based on the acoustic measures. PLS components that loaded on size and shape measures of working vowel space and measures of speech timing were highly successful at predicting attractiveness in female talkers producing /bVd/ words, explaining 73% of the variance in a cross-validated sample.

These findings build on existing work demonstrating that acoustic features related to sexual dimorphism, indicators of apparent talker health, and the processing dynamics of observers, contribute to interpersonal perceptions of attractiveness. Future work should try

to more directly test the relationship between articulatory movements and attractiveness. One potential fruitful approach would be to artificially and systematically manipulate the size and shape of working vowel space in synthesized speech and evaluate the relationship to attractiveness. Though we observed a high correlation between measures of working vowel space from individual words and sentences, ultimately our model failed to predict attractiveness ratings using measures computed from sentence length stimuli. One reason for this may be that the sentences contained higher level cues, such as prosody, signs of social interest (or disinterest) or indicators of personality characteristics. The true source of this variation in sentence length stimuli cannot be determined based on our data.

The second study investigates how movements of the entire body are processed by several nodes of the action observation network, specifically the posterior superior temporal sulcus (pSTS). The pSTS is a key brain region linked to encoding perceptual representations of human body actions. Increasingly, however, action observation is recognized as being strongly shaped by the expectations of the observer (Patel et al., 2019; Kilner, 2011; Koster-Hale and Saxe, 2013). Therefore to test for the influence of top-down influences on perceptual encoding, we evaluated the statistical structure of the multivariate activation patterns in the pSTS while observers attended to different dimensions (the action kinematics, the goal, or the identity) of an avatar engaged in two different actions. Multivariate pattern decoding accuracy varied as a function of attention instruction in the right pSTS, but not in the other regions of the AON, with the highest classification occurring when observers attended to the action kinematics themselves. Furthermore, functional connectivity between the pSTS and inferior frontal cortex (IFC) was stronger when observers attended to the action kinematics portrayed in the vignettes. Our findings are evidence that the attention goals of the viewer modulate sensory representations in the pSTS. These results depict the pSTS as an interstitial zone mediating top-down context and bottom-up perceptual cues during action observation.

The final chapter follows directly from the preceding study and investigates the impact of four methodological choices on multivariate pattern decoding using trial-specific activation estimates in the context of event-related fMRI experimental designs. Support vector classifiers were trained and tested on data that had undergone four different data processing steps including: trial averaging (using separate trial estimates versus run and condition averaged estimates), within-run mean centering (centered or not), cost selection method (using a fixed cost value or tuning the cost parameter within an inner cross-validated fold), and motion-related denoising techniques (regressing out different reference signals capturing motion-related noise). The impact of these decisions was evaluated on real fMRI data from two ROIs as well as simulated pattern data computed over many trial- and voxel-level noise settings. The largest improvements occurred when trial-specific estimates were first mean centered and then averaged within runs. This was observed both for real data in ROIs most likely to contain signal as well as for simulated data across many different levels of noise. The convergence of results across these two datasets suggest that these methods are effective across a wide range of experimental designs and regions of interest.

Although averaging trial-specific estimates by run lead to one of the largest improvements in classification accuracy seen in our data, it was also accompanied by a large increase in between-subject variability. One explanation for the increased variance may be the reduced size of the test set used to assess the prediction error of the classifier at each split of cross-validation. When all trial estimates are averaged within run, the cross-validated test error is constrained to only a few possible values leading to highly variable accuracy estimates. Therefore, to strike a better balance between reducing trial-by-trial variability and maintaining ample test set size, we tested a new technique which involved averaging many smaller subsets of trials from multiple random samples of the data. This technique offered the best middle road between increasing the signal-to-noise ratio by trial averaging and maintaining a sufficient test set size.

Bibliography

(2010). MATLAB.

Abitbol, J., Abitbol, P., and Abitbol, B. (1999). Sex Hormones and the Female Voice. *Journal of Voice*, 13(3):424–446.

Albin, A. (2016). PraatR: An architecture for controlling the phonetics software 'Praat'.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6566.

Amir, N. and Amir, O. (2007). Novel measures for vowel reduction. *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 849–852.

Amir, O. and Biron-Shental, T. (2004). The impact of hormonal fluctuations on female vocal folds. *Current opinion in otolaryngology & head and neck surgery*, 12.3:180–184.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Babel, M., McGuire, G., and King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS one*, 9(2):1–10.

Bach, P. and Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, 11(7):1–17.

Bahnemann, M., Dziobek, I., Prehn, K., Wolf, I., and Heekeren, H. R. (2009). Sociotopy in the temporoparietal cortex: Common versus distinct processes. *Social Cognitive and Affective Neuroscience*, 5(1):48–58.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2019). lme4: Linear Mixed-Effects Models using 'Eigen' and S4.

Binkofski, F. and Buxbaum, L. J. (2013). Two action systems in the human brain. *Brain and Language*, 127(2):222–229.

Bradlow, A. R. and Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, 112(1).

Bradlow, A. R., Kraus, N., and Hayes, E. (2003). Speaking Clearly for Children With Learning Disabilities : Sentence Perception in Noise. *Journal of Speech Language and Hearing Research*, 46(February):80–97.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine grained acoustic-phonetic talker characteristics. *Speech communication*, 20(96):255–272.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, 10:433–436.

Braunlich, K. and Love, B. C. (2019). Occipitotemporal Representations Reflect Individual Differences in Conceptual Knowledge. *Journal of Experimental Psychology: General*, 148(7):1192–1203.

Bruckert, L., Lienard, J.-s., Lacroix, A., Kreutzer, M., and Leboucher, G. (2006). Women use voice parameters to assess men 's characteristics. *Proceedings of the Royal Society B: Biological Sciences*, 273(November):83–89.

Bunton, K. and Leddy, M. (2011). An evaluation of articulatory working space area in vowel production of adults with Down syndrome. *Clinical Linguistics and Phonetics*, 25(4):321–334.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's New, Pussycat? On Talking to Babies and Animals. *Science*, 296:1435–1435.

Burton, M. (2003). Too Many Questions? The Uses Of Incomplete Cyclic Designs for Paired Comparisons. *Field Methods*, 15(2):115–130.

Buxbaum, L. J. and Kalénine, S. (2010). Action knowledge, visuomotor activation, and embodiment in the two action systems. *Annals of the New York Academy of Sciences*, 1191(1):201–218.

Caballero-Gaudes, C. and Reynolds, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage*, 154(December 2016):128–149.

Cardellicchio, P., Hilt, P. M., Olivier, E., Fadiga, L., and Ausilio, A. D. (2018). Early modulation of intra-cortical inhibition during the observation of action mistakes. *Scientific Reports*, (January):1–9.

Carter, R. M. K. and Huettel, S. A. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, 17(7):328–336.

Casile, A. and Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, 5(4):348–360.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2018). shiny: Web Application Framework for R.

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., and Neumann, C. (2018). Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLoS ONE*, 13(1):1–16.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, 60:773–780.

Collins, S. A. and Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, 65(5):997–1004.

Coutanche, M. N. and Thompson-Schill, S. L. (2012). The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. *NeuroImage*, 61(4):1113–1119.

Çukur, T., Nishimoto, S., Huth, A. G., and Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770.

Dasgupta, S., Tyler, S., Wicks, J., Srinivasan, R., and Grossman, E. (2016). Network connectivity of the right STS in three social perception localizers. *Journal of cognitive neuroscience*, 29(2):221–234.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., and Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, 97:271–283.

De Martino, F., Valente, G., Ashburner, J., Goebel, R., Formisano, E., and De Martino, F. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43(1):44–58.

Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, 25(11):4596–4609.

Diedrichsen, J., Wiestler, T., and Ejaz, N. (2013). A multivariate method to determine the dimensionality of neural representation from population activity. *NeuroImage*, 76:225–235.

Dinga, R., Penninx, B. W., Veltman, D. J., Schmaal, L., and Marquand, A. F. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv*, page 743138.

Dion, K., Berscheid, E., and Walster, E. (1972). WHAT IS BEAUTIFUL IS GOOD. *Journal of Personality and Social Psychology*, 24(3):285–290.

Downing, P. E., Wiggett, A. J., and Peelen, M. V. (2007). Functional magnetic resonance imaging investigation of overlapping lateral occipitotemporal activations using multi-voxel pattern analysis. *Journal of Neuroscience*, 27(1):226–233.

Duong, T. (2018). ks: Kernel Smoothing.

Etzel, J. A., Gazzola, V., and Keysers, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain Research*, 1282:114–125.

Etzel, J. A., Valchev, N., and Keysers, C. (2011). The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. *NeuroImage*, 54(2):1159–1167.

Evans, S., Neave, N., and Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology*, 72(2):160–163.

Feinberg, D. R., DeBruine, L. M., Jones, B. C., and Little, A. C. (2008a). Correlated preferences for men's facial and vocal masculinity. *Evolution and Human Behavior*, 29(4):233–241.

Feinberg, D. R., DeBruine, L. M., Jones, B. C., and Perrett, D. I. (2008b). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception*, 37(4):615–623.

Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Law Smith, M. J., Cornwell, R. E., Tiddeman, B. P., Boothroyd, L. G., and Perrett, D. I. (2005a). The voice and face of woman: One ornament that signals quality? *Evolution and Human Behavior*, 26(5):398–408.

Feinberg, D. R., Jones, B. C., Law Smith, M. J., Moore, F. R., DeBruine, L. M., Cornwell, R. E., Hillier, S. G., and Perrett, D. I. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and Behavior*, 49(2):215–222.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., and Perrett, D. I. (2005b). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, 69(3):561–568.

Feingold, A. (1992). Good-Looking People Are Not What We Think. *Psychological Bulletin*, 111(2):304–341.

Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D., and Delplanque, S. (2013). Voice attractiveness: Influence of stimulus duration and type. *Behav Res*, 45(2):405–13.

Ferguson, S. and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *Journal of speech, language, and hearing research : JSLHR*, 50(5):1241–1255.

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 116(2365):2365–2373.

Ferguson, S. H. and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 112:259–271.

Ferguson, S. H. and Quene, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 112(1):259–271.

Ferguson, S. H. and Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 135(6):3570–3584.

Fernald, A. and Simon, T. (1984). Expanded Intonation Contours in Mothers' Speech to Newborns. *Developmental Psychology*, 20(1):104–113.

Fischer, J., Semple, S., Fickenscher, G., Jürgens, R., Kruse, E., Heistermann, M., and Amir, O. (2011). Do women's voices provide cues of the likelihood of ovulation? The importance of sampling regime. *PLoS ONE*, 6(9):1–8.

Flipsen, P. and Lee, S. (2012). Reference data for the American English acoustic vowel space. *Clinical Linguistics & Phonetics*, 26(11-12):926–933.

Fristen, K. J., Frith, C. D., Fletcher, P., Liddle, P. F., and Frackowiak, R. S. (1996). Functional topography: Multidimensional scaling and functional connectivity in the brain. *Cerebral Cortex*, 6(2):156–164.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7(1):30–40.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

Frost, M. A. and Goebel, R. (2012). Measuring structural-functional correspondence: Spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage*, 59(2):1369–1381.

Gardner, T., Goulden, N., and Cross, X. E. S. (2015). Dynamic Modulation of the Action Observation Network by Movement Familiarity. *Journal of Neuroscience*, 35(4):1561–1572.

Geladi, P. and Kowalski, B. (1986). PARTIAL LEAST-SQUARES REGRESSION: A TUTORIAL. *Analytica Chimica Acta*, 185:1–17.

Geng, J. J. and Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: Contextual updating? *Neuroscience and Biobehavioral Reviews*, 37(10):2608–2620.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 878:870–878.

Giese, M. A. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192.

Glover, G. H. (1999). Deconvolution of Impulse Response in Event-Related BOLD fMRI. *NeuroImage*, 9:416–429.

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of FIAC data with BrainVoyager QX: From single-subject to cortically aligned group GLM analysis and self-organizing group ICA. *Human Brain Mapping*, 27(5):392–401.

Grammer, K., Fink, B., Møller, A. P., and Thornhill, R. (2003). Darwinian aesthetics : sexual selection and the biology of beauty. *Biol. Rev.*, 78:385–407.

Gregory JR., S. W. and Gallagher, T. J. (2002). Spectral Analysis of Candidates ' Nonverbal Vocal Communication : Predicting U .S . Presidential Election Outcomes. *Social Psychology Quarterly*, 65(3):298–308.

Grossman, E. D., Blake, R., and Kim, C. Y. (2004). Learning to see biological motion: Brain activity parallels behavior. *Journal of Cognitive Neuroscience*, 16(9):1669–1679.

Grossman, E. D., Jardine, N. L., and Pyles, J. A. (2010). fMR-adaptation reveals invariant coding of biological motion on the human STS. *Frontiers in Human Neuroscience*, 4(March):1–18.

Hagiwara, R. (2006). Vowel Production in Winnipeg. *The Canadian Journal of Linguistics / La revue canadienne de linguistique*, 51(2/3):127–141.

Hajnal, J. V., Myers, R., Oatridge, A., Schwieso, J. E., Young, I. R., and Bydder, G. M. (1994). Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magnetic resonance in medicine*, 31(3):283–291.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Haxby, J. V., Gobbini, I. M., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293:2425–2430.

Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., and Diehl, R. L. (2006). Enhanced contrast for vowels in utterance focus: A cross-language study. *The Journal of the Acoustical Society of America*, 119(3022):3022–3033.

Haynes, J. D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

Hazan, V. and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116(5):3108–3118.

Heffernan, K. (2010). MUMBLING IS MACHO: PHONETIC DISTINCTIVENESS IN THE SPEECH OF AMERICAN RADIO DJs. *American Speech*, 85(1):67–90.

Hein, G. and Knight, R. T. (2008). Superior temporal sulcus - It's my area: Or is it? *Journal of Cognitive Neuroscience*, 20(12):2125–2136.

Hillebrandt, H., Friston, K. J., and Blakemore, S. J. (2014). Effective connectivity during animacy perception - Dynamic causal modelling of Human Connectome Project data. *Scientific Reports*, 4:1–9.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.

Hodges-Simeon, C. R., Gaulin, S. J., and Puts, D. A. (2010a). Voice Correlates of Mating Success in Men: Examining "Contests" Versus "Mate Choice" Modes of Sexual Selection. *Archives of Sexual Behavior*, 40(3):551–557.

Hodges-Simeon, C. R., Gaulin, S. J. C., and Puts, D. A. (2010b). Different Vocal Parameters Predict Perceptions of Dominance and Attractiveness. *Human Nature*, 21(4):406–427.

Hoffman, D. D. and Flinchbaugh, B. E. (1982). The Interpretation of Biological Motion. *Biological Cybernetics*, 204:195–204.

Hsu, S.-C., Jiao, Y., McAuliffe, M. J., Berisha, V., Wu, R.-M., and Levy, E. S. (2017). Acoustic and perceptual speech characteristics of native Mandarin speakers with Parkinson's disease. *The Journal of the Acoustical Society of America*, 141(3):EL293–EL299.

Hughes, S. M., Dispenza, F., and Gallup, G. G. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution and Human Behavior*, 25(5):295–304.

Huk, A., Dougherty, F., and Heeger, D. (2002). Retinotopy and Functional Subdivision of Human Areas MT and MST. *The Journal of Neuroscience*, 22(16):7195–7205.

Jastorff, J., Clavagnier, S., Gergely, G., and Orban, G. A. (2011). Neural mechanisms of understanding rational actions: Middle temporal gyrus activation by contextual violation. *Cerebral Cortex*, 21(2):318–329.

Jastorff, J., Kourtzi, Z., and Giese, M. A. (2009). Visual learning shapes the processing of complex movement stimuli in the human brain. *Journal of Neuroscience*, 29(44):14026–14038.

Jezzard, P. and Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic resonance in medicine*, 34(1):65–73.

Johnson, K., Flemming, E., and Wright, R. (1993). THE HYPERSPACE EFFECT: PHONETIC TARGETS ARE HYPERARTICULATED. *Language*, 69(3):505–528.

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., and Vukovic, J. (2008). Integrating cues of social interest and voice pitch in men's preferences for women's voices. *Biology Letters*, 4(2):192–194.

Kempe, V., Puts, D. A., and Cárdenas, R. A. (2013). Masculine Men Articulate Less Clearly. *Human Nature*, 24(4):461–475.

Kilner, J. M. (2011). More than one pathway to action understanding. *Trends in Cognitive Sciences*, 15(8):352–357.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007a). Predictive coding : an account of the mirror neuron system. *Cognitive Processing*, 8:159–166.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007b). The mirror-neuron system : a Bayesian perspective. *NeuroReport*, 18(6).

Klofstad, C. A., Anderson, R. C., and Peters, S. (2012). Sounds like a winner : voice pitch influences perception of leadership capacity in both men and women. *Proc. R. Soc. B*, 279(March):2698–2704.

Kok, P., Rahnev, D., Jehee, J. F., Lau, H. C., and De Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex*, 22(9):2197–2206.

Koster-Hale, J. and Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, 79(5):836–848.

Krause, J. C. and Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America*, 112(5):2165–2172.

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, 56(2):411–421.

Kuhl, P. K. (1997). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, 277(5326):684–686.

Kwon, H.-B. (2010). Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. *The journal of advanced prosthodontics*, 2(3):71–6.

Lam, J. and Tjaden, K. (2016). Clear Speech Variants: An Acoustic Study in Parkinson's Disease. *Journal of Speech, Language, and Hearing Research*, 24(2):1–14.

Lam, J., Tjaden, K., and Wilding, G. (2012). Acoustics of Clear Speech: Effect of Instruction. *Journal of Speech Language and Hearing Research*, 55(6):1807.

Lander, K. (2008). Relating visual and vocal attractiveness for moving and static faces. *Animal Behaviour*, 75(3):817–822.

Lange, J. and Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience*, 26(11):2894–2906.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hauam, M., Smoot, M., Bigler, R., Buss, D., Cohen, D., Feingold, A., Holden, G., Kalick, D., Miller, P., and Swann, W. B. (2000). Maxims or Myths of Beauty? A Meta-Analytic and Theoretical Review. *Psychological Bulletin*, 126(3):390–423.

Lee, J., Shaiman, S., and Weismer, G. (2016). Relationship between tongue positions and formant frequencies in female speakers. *The Journal of the Acoustical Society of America*, 139(1):426–440.

Lee, S. and Kable, J. W. (2018). Simple but robust improvement in multivoxel pattern classification. *PLoS ONE*, 13(11):1–15.

Lee, S. M., Gao, T., and McCarthy, G. (2014). Attributing intentions to random motion engages the posterior superior temporal sulcus. *Social Cognitive and Affective Neuroscience*, 9(1):81–87.

Lemieux, L., Salek-Haddadi, A., Lund, T. E., Laufs, H., and Carmichael, D. (2007). Modelling large motion events in fMRI studies of patients with epilepsy. *Magnetic Resonance Imaging*, 25(6):894–901.

Lingnau, A. and Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in Cognitive Sciences*, 19(5):268–277.

Liu, H. M., Kuhl, P. K., and Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6(3):1–10.

Maffei, V., Indovina, I., Macaluso, E., Ivanenko, Y. P., Orban, G. A., and Lacquaniti, F. (2015). Visual gravity cues in the interpretation of biological movements : neural correlates in humans. *NeuroImage*, 104:221–230.

Marsh, L. E., Mullett, T. L., Ropar, D., and Hamilton, A. F. C. (2014). Responses to irrational actions in action observation and mentalising networks of the human brain. *NeuroImage*, 103:81–90.

Mather, G., Radford, K., and West, S. (1992). Low-level visual processing of biological motion. *Proceedings of the Royal Society B: Biological Sciences*, 249(1325):149–155.

Maunsell, J. H. and Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, 29(6):317–322.

McCormick, E. J. and Bachus, J. A. (1952). Paired Comparison Ratings. I. The Effect on Ratings of Reductions in the Number of Pairs. *Journal of Applied Psychology*, 36(2):123–127.

McGowan, R. W., McGowan, R. S., and Denny, M. (2014). A longitudinal study of very young children's vowel production. *Jouranal of Speech Langugae and Hearing Research*, 57(1):1–15.

Mevik, B.-H. and Wehrens, R. (2007). The pls Package: Principle Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2).

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2019). pls: Partial Least Squares and Principal Component Regression.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Molnar-Szakacs, I., Iacoboni, M., Koski, L., and Mazziotta, J. C. (2005). Functional segregation within pars opercularis of the inferior frontal gyrus: Evidence from fMRI studies of imitation and action observation. *Cerebral Cortex*, 15(7):986–994.

Moon, S.-j. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. 40(1994):40–55.

Morris, J. P., Pelphrey, K. A., and McCarthy, G. (2008). Perceived causality influences brain activity evoked by biological motion. *Social Neuroscience*, 3(1):16–25.

Most, T., Amir, O., and Tobin, Y. (2000). The Hebrew Vowel System: Raw and Normalized Acoustic Data. *Language and speech*, 43(Pt 3):295–308.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., and Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, 33(4):1055–1065.

Mumford, J. A., Davis, T., and Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, 103:130–138.

Mumford, J. A., Turner, B. O., Ashby, F. G., and Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–2643.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Di Oleggio Castello, M. V., Gors, J., Gobbini, M. I., and Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, 27(8):4277–4291.

Neel, A. T. (2008). Vowel Space Characteristics and Vowel Identification Accuracy. *Journal of Speech Language and Hearing Research*, 51(3):574.

Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., and Orban, G. A. (2011). Action Observation Circuits in the Macaque Monkey Cortex. *Journal of Neuroscience*, 31(10):3743–3756.

Neumann, C. (2015). EloChoice: Preference Rating for Visual Stimuli Based on Elo Ratings.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Ogawa, K. and Inui, T. (2011). Neural representation of observed actions in the parietal and premotor cortex. *NeuroImage*, 56(2):728–735.

Oosterhof, N. N., Wiggett, A. J., Diedrichsen, J., Tipper, S. P., and Downing, P. E. (2010). Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. *Journal of Neurophysiology*, 104(2):1077–1089.

Oram, M. W. and Perrett, D. I. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to 'biological motion' stimuli. *Journal of Cognitive Neuroscience*, 6(2):99–116.

Patel, G. H., Sestieri, C., and Corbetta, M. (2019). The evolution of the temporoparietal junction and posterior superior temporal sulcus. *Cortex*, 118(xxxx):38–50.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 95(3):1581–1592.

Pebesma, E. and Bivand, R. (2018). sp: Classes and Methods for Spatial Data.

Peelen, M. V. and Downing, P. E. (2007). The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–S209.

Perry, T. L., Ohde, R. N., and Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America*, 109(6):2988–2998.

Picheny, M., Durlach, N., and Braida, L. (1986). Speaking Clearly for the Hard of Hearing II. *Journal of Speech Language and Hearing Research*, 29(4):434.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Roder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., and Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, 95:89–99.

Pisanski, K. and Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, 129(4):2201–2212.

Popov, V., Ostarek, M., and Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, 174(November 2017):340–351.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154.

Puts, D. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior*, 26(5):388–397.

Puts, D., Doll, L. M., and Hill, A. K. (2014). Sexual Selection on Human Voices. In Weekes-Shackelford, V. and Shackelford, T., editors, *Evolutionary Perspectives on Human Sexual Psychology and Behavior*, pages 69–86. Springer, New York.

Puts, D., Jones, B. C., and DeBruine, L. M. (2012a). Sexual Selection on Human Faces and Voices. *The Journal of Sex Research*, 49(2-3):227–243.

Puts, D. A., Apicella, C. L., and Cárdenas, R. A. (2012b). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences*, 279(1728):601–609.

Puts, D. A., Gaulin, S. J. C., and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, 27(4):283–296.

Pyles, J. A. and Grossman, E. D. (2013). Neural Mechanisms for Biological Motion and Animacy. In *People Watching: Social, Perceptual, and Neurophysiological Studies of Body Perception*, chapter 17, pages 304–317. Oxford University Press, New York.

Reber, R., Schwarz, N., and Winkielman, P. (2004). Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience? *Personality and Social Psychology Review*, 8(4):364–382.

Reynolds, J. H. and Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, 61(2):168–185.

Riding, D., Lonsdale, D., and Brown, B. (2006). The Effects of Average Fundamental Frequency and Variance of Fundamental Frequency on Male Vocal Attractiveness to Women. *Journal of Nonverbal Behavior*, 30(2):55–61.

Rissman, J., Gazzaley, A., and D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, 23(2):752–763.

Rogers, C. L., DeMasi, T. M., and Krause, J. C. (2010). Conversational and clear speech intelligibility of /bVd/ syllables produced by native and non-native English speakers. *The Journal of the Acoustical Society of America*, 128(1):410–423.

Rosen, K., Murdoch, B., Folker, J., Vogel, A., Cahill, L., Delatycki, M., Corben, L., Rosen, K., Murdoch, B., Folker, J., Vogel, A., Vogel, A., Cahill, L., and Delatycki, M. (2010). Automatic method of pause measurement for normal and dysarthric speech. *Clinical Linguistics & Phonetics*, 24(2):141–154.

Rusz, J., Cmejla, R., and Tykalova, T. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3):2171–2181.

Saenz, M., Buracas, G. T., and Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, 5(7):631–632.

Safford, A. S., Hussey, E. A., Parasuraman, R., and Thompson, J. C. (2010). Object-based attentional modulation of biological motion processing: Spatiotemporal dynamics using functional magnetic resonance imaging and electroencephalography. *Journal of Neuroscience*, 30(27):9064–9073.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64(1):240–256.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, 7(4):413–422.

Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., and Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *Journal of Neuroscience*, 24(27):6181–6188.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114.

Schroeder, J. and Epley, N. (2015). The Sound of Intellect : Speech Reveals a Thoughtful Mind , Increasing a Job Candidate ' s Appeal. *Psychological Science*, 26(6):877–891.

Schum, D. J. (1996). Intelligibility of Clear and Conversational Speech of Young and Elderly Talkers. *Journal of the American Academy of Audiology*, 7(3):212–218.

Sell, A., Bryant, G., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., Krauss, A., and Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B*, 277(1699):3509–18.

SenGupta, A. (1987). Tests for Standardized Generalized Variances of Multivariate Normal Populations of Possibly Different Dimensions. *Journal of Multivariate Analysis*, 23(2):209–219.

Simpson, A. and Ericsdotter, C. (2007). Sex-specific differences in f0 and vowel space. *Proceedings of the XVIth ICPhS, Saarbrücken*, (August):6–10.

Smiljanic, R. and Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass*, 3(1):236–264.

Sokolov, A. A., Zeidman, P., Erb, M., Ryvlin, P., Friston, K. J., and Pavlova, M. A. (2018). Structural and effective brain connectivity underlying biological motion detection. *Proceedings of the National Academy of Sciences of the United States of America*, 115(51):E12034–E12042.

Steffener, J., Tabert, M., Reuben, A., and Stern, Y. (2010). Investigating hemodynamic response variability at the group level using basis functions. *NeuroImage*, 49(3):2113–2122.

Stigliani, X. A., Weiner, X. K. S., and Grill-spector, X. K. (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *Journal of Neuroscience*, 35(36):12412–12424.

Story, B. H. and Bunton, K. (2017). Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):EL458–EL464.

Summerfield, C. and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9):403–409.

Tartter, V. C., Braun, D., and Tartter, V. C. (1994). Hearing smiles and frowns in normal and whisper registers. 96:2101–2107.

Tavares, P., Lawrence, A. D., and Barnard, P. J. (2008a). Paying attention to social meaning: An fMRI study. *Cerebral Cortex*, 18(8):1876–1885.

Tavares, P., Lawrence, A. D., and Barnard, P. J. (2008b). Paying attention to social meaning: An fMRI study. *Cerebral Cortex*, 18(8):1876–1885.

Thompson, E. L., Bird, G., and Catmur, C. (2019). Conceptualizing and testing action understanding. *Neuroscience and Biobehavioral Reviews*, 105:106–114.

Thompson, J. and Parasuraman, R. (2012). Attention, biological motion, and action recognition. *NeuroImage*, 59(1):4–13.

Thurman, S. and Grossman, E. D. (2008). Temporal "Bubbles" reveal key features for point-light biological motion perception. *Journal of Vision*, 8:1–11.

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., and Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3):210–216.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707.

Tjaden, K. and Wilding, G. E. (2004). Rate and Loudness Manipulations in Dysarthria. *Journal of Speech, Language, and Hearing Research*, 47(4):766–783.

Treue, S. and Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579.

Turner, B. O., Mumford, J. A., Poldrack, R. A., and Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage*, 62(3):1429–1438.

Urgen, B. and Saygin, A. (2019). Effective connectivity in the action observation network. *bioRxiv*, pages 1–26.

Urgen, B. A. and Miller, L. E. (2015). Towards an empirically grounded predictive coding account of action understanding. *Journal of Neuroscience*, 35(12):4789–4791.

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49(1):2–7.

van Kemenade, B. M., Muggleton, N., Walsh, V., and Saygin, A. P. (2012). Effects of TMS over premotor and superior temporal cortices on biological motion perception. *Journal of Cognitive Neuroscience*, 24(4):896–904.

Vangeneugden, J., De Mazière, P. A., Van Hulle, M. M., Jaeggli, T., Van Gool, L., and Vogels, R. (2011). Distinct mechanisms for coding of visual actions in macaque temporal cortex. *Journal of Neuroscience*, 31(2):385–401.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145(August 2015):166–179.

Vinzi, V., Chin, W., Henseler, J., and Wang, H. (2010). *Handbook of Partial Least Squares: Concepts, Methods and Applications*. Springer, Berlin.

Wang, G., Garcia, D., Liu, Y., de Jeu, R., and Johannes Dolman, A. (2012). A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations. *Environmental Modelling and Software*, 30:139–142.

Weiner, K. S. and Grill-Spector, K. (2011). Not one extrastriate body area: Using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. *NeuroImage*, 56(4):2183–2199.

Weiner, K. S. and Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. *Psychological Research*, 77(1):74–97.

Weirich, M., Fuchs, S., Simpson, A., Winkler, R., and Perrier, P. (2016). Mumbling: Macho or Morphology? *Journal of Speech Language and Hearing Research*, 25(November):1–15.

Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R. D., and Kent, J. F. (2001). Acoustic and Intelligibility Characteristics of Sentence Production in Neurogenic Speech Disorders. *Folia Phoniatrica et Logopaedica*, 53(1):1–18.

Whiteside, S. (1996). Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26(01):23.

Whitfield, J. A., Dromey, C., and Palmer, P. (2018). Examining Acoustic and Kinematic Measures of Articulatory Working Space: Effects of Speech Intensity. *Journal of Speech Language and Hearing Research*, 61(May):1104–1117.

Whitfield, J. A. and Goberman, A. M. (2014). Articulatory – acoustic vowel space : Application to clear speech in individuals with Parkinson ' s disease. *Journal of Communication Disorders*, 51:19–28.

Whitfield, J. A. and Goberman, A. M. (2017). Articulatory-acoustic vowel space: Associations between acoustic and perceptual measures of clear speech. *International Journal of Speech-Language Pathology*, 19(2):184–194.

Whitfield, J. A. and Gravelin, A. C. (2019). Characterizing the distribution of silent intervals in the connected speech of individuals with Parkinson disease. *Journal of Communication Disorders*, 78(January 2018):18–32.

Whitfield, J. A. and Mehta, D. D. (2019). Examination of Clear Speech in Parkinson Disease Using Measures of Working Vowel Space. *Journal of Speech, Language, and Hearing Research*, 62(7):2082–2098.

Wilks, S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, 24(3/4):471–494.

Wilks, S. (1960). Multidimensional statistical scatter. *Contributions to Probability anb Statistics*, 2:486.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92:381–397.

Wurm, M. F. and Lingnau, A. (2015). Simultaneously learning at different levels of abstraction. *The Journal of Neuroscience*, 35(20):7727–7735.

Wyk, B. C. V., Hudac, C. M., Carter, E. J., Sobel, D. M., and Pelphrey, K. A. (2009). Action Understanding in the Superior Temporal Sulcus Region. *Psychological Science*, 20(6):771–777.

Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., Li, Q., Zuo, X. N., Castellanos, F. X., and Milham, M. P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage*, 76:183–201.

Yang, B. (1992). An acoustical study of Korean monophthongs produced by male and female speakers. *The Journal of the Acoustical Society of America*, 91(4):2280–2283.

Yoho, S. E., Borrie, S. A., Barrett, T. S., and Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception, and Psychophysics*, 81(2):558–570.

Zuckerman, M. and Driver, R. E. (1989). WHAT SOUNDS BEAUTIFUL IS GOOD: THE VOCAL ATTRACTIVENESS STEREOTYPE. *Journal of Nonverbal Behavior*, 13(2):67–82.