Journal of Sociolinguistics 20/5, 2016: 677–711

THEME SERIES

Interaction: Talk and beyond
Edited by Devyani Sharma

This article is part of a running Theme Series entitled 'Interaction: Talk and Beyond', which explores non-verbal dimensions of talk in interaction. A full Series introduction can be found in issue 20/3 (2016).

Cans and cants: Computational potentials for multimodality with a case study in head position[1]

Rob Voigt, Penelope Eckert, Dan Jurafsky and Robert J. Podesva

Stanford University, California, U.S.A.

As the study of embodiment and multimodality in interaction grows in importance, there is a need for novel methodological approaches to understand how multimodal variables pattern together along social and contextual lines, and how they systematically coalesce in communicative meanings. In this work, we propose to adopt computational tools to generate replicable annotations of bodily variables: these can be examined statistically to understand their patterning with other variables across diverse speakers and interactional contexts, and can help organize qualitative analyses of large datasets. We demonstrate the possibilities thereby with a case study in head cant (side-to-side tilt of the head) in a dataset of video blogs and laboratory-collected interactions, computationally extracting cant and prosody from video and audio and analyzing their interactions, looking at gender in particular. We find that head cant indexes an orientation towards the interlocutor and a sense of shared understanding, can serve a 'bracketing' function in interaction (for speakers to create parentheticals or asides), and has gendered associations with prosodic markers and interactional discourse particles.

具身性和多模态在话语交际研究中愈来愈受到重视。为了更好理解多模态变量在不同交际场景中的变异及意义,本文引入一种新的研究方法。通过具有普适性的计算方法对具身变量的标注,我们得以运用统计分析来理解具身变量与其它语言学变量的互动,从而理解具身变量跨个体、跨情境的变异,并进一步对大型语料库进行定性分析。我们采集了播客视频以及在实验室中采集的交际录像作为语料,并侧重分析了侧头、语言韵律及性别的互动。研究结果显示侧头与说话者对于听话者的关注与共情相关,并提示话语中的附加信息(类似于书面语中括号的功能)。此外,侧头还与韵律与话语标记有相关性,这种相关性同时受到说话者性别的影响。[Chinese]

KEYWORDS: Embodiment, computer vision, multimodality, head cant, body positioning, prosody, gender, interaction

1. INTRODUCTION

Language is a multilayered, multimodal system; in spoken talk, meanings – and particularly social meanings – are conveyed not only by phonetics, syntax, and pragmatics, but also by facial expression, gesture, and movement. A growing body of research takes seriously the consequences of this fact by addressing the central issue of *embodiment*: the complex ways in which the meaning-making capacity of language is tied to the physical bodies of those who use language.

In the view from the cognitive sciences, this implies that the full sensory experience of any event is deeply intertwined with, and even to an extent may inescapably constitute, the mental representations of that event (Glenberg and Kaschak 2003; Matlock, Ramscar and Boroditsky 2003; Barsalou 2008). In linguistics, this line of analysis takes form in the concept of multimodality, whereby the production of meaning is always in progress and can recruit resources from diverse semiotic modes including but not necessarily privileging spoken language (Kress and Van Leeuwen 2001). Multimodality in linguistics has a long history, even if the term is relatively new. At least as early as Birdwhistell's *kinesics* (1952, 1970) linguistic anthropologists have recognized the rich communicative capacity of the body, and later researchers such as McNeill (1992, 2008) and Kendon (1995, 2004) argued for the integration of gesture and spoken language as two parts of one system.

Though numerous experimental studies have provided convincing pieces of evidence for the claim that speech and bodily movements and postures are tightly connected (see, for instance, Mendoza-Denton and Jannedy 2011; Loehr 2012; Voigt, Podesva and Jurafsky 2014), our understanding of how they interact moment to moment and coalesce into meaningful signs is based primarily on observational study. Scholars of conversation analysis (CA) in particular have explored such moment-to-moment multimodality. The first article in this Series (Mondada 2016) employs just such a CA approach to consider embodiment and interactional multimodality as related to the 'ecology of the activity' taking place in interaction, looking at full-body physical positioning as a crucial resource for meaning-making. Indeed, as Mondada (2016: 341) notes, in linguistic communication, 'potentially every detail can be turned into a resource for social interaction.'

But with this unlimited potential comes a set of daunting analytical challenges. Kendon (1994) provides an early review of observational studies, noting the difficulty of accurate and consistent transcription of bodily gestures. In a review of this 'embodied turn,' Nevile (2015) notes that despite the

importance of preserving the structure of the original interaction via holistic qualitative analysis of every detail, a consistent methodological framework for the analysis of embodiment remains elusive. Given the detailed analysis that qualitative work requires, analyses are based on a small number of relatively brief interactions in a limited range of contexts.

We propose that computational methodologies, by allowing for the study of particular interactional details on a large scale, will afford robust comparisons of elements of bodily movement and the interaction of movement with speech, across diverse communicative contexts. Linguists have long embraced the utility of computational tools for analyzing sound and text in interactional and non-interactional domains. In this paper we develop ways to advance the linguistic study of embodiment by drawing on tools from computer vision to augment sound and text with a third modality, video data.

Computer vision technologies are reaching a point of maturity sufficient for the analysis of embodiment from video data. These technologies can carry out some annotation tasks with high accuracy, including extracting broad characteristics of the context from video, like recognizing objects (Viola and Jones 2001; Krizhevsky, Sutskever and Hinton 2012; Girshick et al. 2014; Russakovsky et al. 2015) or tracking people in a scene (McKenna et al. 2000; Pellegrini, Ess and Van Gool 2010; Tang, Andriluka and Schiele 2014), which may help establish the setting of a given dataset. They can also extract person-level features like finding boundaries for faces and tracking head movements (Kim et al. 2008; Murphy-Chutorian and Trivedi 2009), identifying smiling and other emotive expressions (Shan 2012; Dhall et al. 2014), and tracking hands and categorizing hand gestures (Suarez and Murphy 2012; Rautaray and Agrawal 2015), which may identify axes along which meaningful variation may occur.

Since such models can generate reliable annotations of real-world phenomena, there is a sense in which they are consonant with the aims of qualitative approaches; they can help us to holistically identify and capture meaningful axes of variation. Annotations derived from computational tools, moreover, have a number of advantages. They are replicable, allowing scholars to repeat prior measures and apply them to obtain information about new speakers or contexts. And they offer scale and statistical power for descriptive observation: since we expect social and ideological structures to show themselves in aggregate as well as in individual expression, automatic tools that can operate on more data than could be analyzed by hand allow us to test many hypotheses about variation across multiple contexts and speakers.

But beyond the possibility for statistical results, in this work we aim to demonstrate that we can view computational tools as a powerful complement to quantitative and qualitative analysis of smaller and more localized data sets. If we accept the premises that first, every detail of an interaction is potentially recruitable for meaning-making, and second, that variation may reflect large-scale social and ideological structures, then a broad view of the possibilities of

computational methodologies is inevitably a step forward. Any interactional feature that can be recorded and defined cleanly is potentially available for computational modeling, and such modeling allows us to put such features under the microscope and uncover something about how these features combine to produce social meaning.

We demonstrate the possibilities of such an analysis in this paper by analyzing one such interactional variable – head cant (colloquially, side-to-side tilt of the head) – in a multimodal dataset of 65 different speakers. We use computational tools to examine how visual, textual, and acoustic properties combine in interaction, and how these interactions correlate with social and interactional factors. Of course, a statistical association does not directly reveal social meaning, but indicates that meaning may be at work at the local level. Thus, we allow our statistical analysis to guide us in our choice of specific examples for qualitative analysis. Our analysis confirms head cant's role as an interactional variable, its robust connection to prosodic variation, and its participation in communicative and social meanings having to do with floor management and with a frame of shared understanding between the speaker and interlocutor.

In section 2, we explain our methodology in detail, as well as the dataset to which we apply it, which includes 65 speakers across two distinct interactional contexts: YouTube video blog (henceforth 'vlog') monologues; and experimentally-collected laboratory dialogues. We take advantage of a computer vision algorithm to calculate head cant annotations automatically and use these annotations to both generate statistical results and guide a qualitative analysis, exploring the interactional functions of head cant in three stages. In section 3, we consider the simple question of the distribution of head canting: is cant more prevalent when an interlocutor is physically present? Is head cant a listening gesture? In section 4, we explore high-level statistical connections between head canting and prosodic features indicative of conversational engagement. Then, in section 5, we draw upon those connections to engage in a quantitatively-guided qualitative analysis of head cant. This involves identifying particular functions of head cant, discussing them in context, and providing statistical support for these where possible.

1.1 Head movement and posture as interactional variables

Language researchers have long known that movements of the head can participate in a diverse field of meanings. McClave (2000) provides a comprehensive review, cataloguing an extensive list of functions of head movement: as signals for turn-taking; as semantic and syntactic boundary markers; to locate discourse referents; or to communicate meanings like inclusivity, intensification, and uncertainty. Kendon (2002) looks at the functions of head shakes in particular, suggesting they participate in implied

negation and superlative expressions, and noting in particular their common usage in 'multimodal constructions' in which head shaking co-occurs with particular linguistic features to jointly build an 'expressive unit.' Cvejic, Kim and Davis (2010) use optical markers on the heads of participants to show that head movement tends to co-occur with prosodically focused words.

In this study, we focus on head cant as a resource for the production of meaning. By head cant, we mean left-right lateral displacement of the head. Head cant is distinct from head tilt, the term in the literature for up-down (raised vs. lowered) displacement, which has been shown to be associated with perceived dominance (Mignault and Chaudhuri 2003; Bee, Franke and André 2009).

Head cant as an interactional posture is not as richly studied as head movements more broadly. However, researchers such as Goffman (1979) have identified gendered ideological associations of head cant in depictions of women in advertising, noting that it 'can be read as an acceptance of subordination, an expression of ingratiation, submissiveness, and appeasement' (1979: 46). This association is likely not new, as evidenced by art historical work from Costa, Menzani and Ricci Bitti (2001) finding that women were depicted in postures with head cant more often than men in a large-scale historical analysis of paintings. Moreover, these patterns are consistent: a systematic analysis by Kang (1997) found few differences between advertisements in 1979 and 1991, and even today gendered associations of the type identified by Goffman can be easily found in advertising from around the world.

Folk tellings, as well, tend to draw explicit links between head cant and both gender and sexuality, as in this excerpt from *Body Language for Dummies*:

> Although men tilt their heads in an upward movement, mostly as a sign of recognition, women tilt their heads to the side in appeasement and as a playful or flirtatious gesture. When a woman tilts her head she exposes her neck, making herself look more vulnerable and less threatening. (Kuhnke 2012)

In fact, head cant has been shown to be gendered in its distribution in multiple experimental settings. Mills (1984) found women used more head cant in self-posed photographs than men, in conjunction with increased smiling and postures oriented away from the camera. Grammer (1990) manually coded head cant among a number of other posture and movement variables, and found that when used by women – but not men – it functioned in part as an indicator of romantic interest between different-sex strangers.

Since the gender binary abstracts over a wide range of local practices, a binary gender finding is never the end of the story, but indicates that some kind of meaning associated with gender is at work at the local interactional level. Thus, head cant's meaning-making potentials are not by any means limited to associations with gender and sexuality.

In this work we propose that many of the gendered associations of head cant may stem from a deeper relationship between head cant and what Tannen and Wallat (1987), building on the work of Goffman (1974), call the 'interactive frame' – or the definition of what is taking place at a given interactional moment – as well as the entailed alignment or orientation to one's interlocutor, or what Goffman (1981) calls 'footing.' In particular, head cant appears to participate in communicating orientation towards the interlocutor and a sense of shared understanding, in some cases even serving a relatively explicit 'bracketing' function which speakers use to create parentheticals, asides, and confessions.

2. CASE STUDY METHODOLOGY

In this study we investigate head cant as an interactional feature and a semiotic resource. In this section we describe the selection of data, preprocessing to prepare the data for analysis, and our computational methodology for extracting head cant measurements.

2.1 Data

We compare two interactional contexts: two-person dialogues between friends recorded in a laboratory setting; and video blog monologues on YouTube with no apparent physically present interlocutor. We refer to these settings throughout the paper as 'Lab' and 'Vlog,' respectively. The two settings allow us to compare speakers who are anticipating and getting immediate feedback from an interlocutor with those who are not. Our dataset in total from these sources includes more than 18 hours of speech from 65 speakers.

Laboratory dialogues. The first interactional context is dyadic interactions between familiars recorded in the Interactional Sociophonetics Laboratory at Stanford University in California. The lab has the acoustical specifications of a sound-proof recording booth to ensure high quality audio recordings, but is staged as a living room to facilitate less self-conscious interactions. In addition to being audio recorded via lavalier wireless microphones, interactants were videorecorded by concealed video cameras (though their presence was known to all participants) positioned to capture head-on images. As many computer vision algorithms have been developed for video blog data, it was imperative that speakers not be positioned at a significant angle to the camera lens.

Participants engaged in two conversational tasks. First, speakers discussed their answers to a variety of 'would you rather . . .' questions, such as 'Would you rather always be overdressed, or always be underdressed?' This task, which lasted approximately five minutes, gave participants an opportunity to relax into the recording environment and enabled the researcher to adjust audio recording levels as needed. For the remainder of the approximately

30-minute recording session, speakers asked each other a variety of questions presented on a large rolodex on a coffee table positioned between the interactants. Questions, like 'How has the way you dress changed since high school?', were chosen to encourage speakers to reflect on identity construction without asking them about it explicitly. Participants were informed that they could use questions as prompts as desired, but that their conversation did not need to stick to the prompts at all. Following the recording session, participants filled in online surveys designed to collect demographic information as well as assessments of the interaction.

Data for 33 speakers are considered here. Of these, 22 were women, and 11 men. The great majority of the dyads were between friends or close friends (according to participant characterizations of the relationship), with a handful between romantic partners or family members. The majority of speakers were undergraduates aged 18–22; the remainder of speakers were mostly in their mid to late twenties. Although the results below focus on gender, the corpus was reasonably diverse with respect to several other variables. Speakers represented a range of racial groups. The majority self-identified as White, a sizeable minority (of nine) as multiracial, and the remainder as African American, Asian American, or Latinx. The majority of speakers were from the West Coast of the U.S.A., though a significant group (of eight) were from the South; the remainder were from the Northeast and Midwest.

Data were recorded directly onto a Mac Pro located in a room outside the living room space. Each speaker was recorded onto separate audio and video tracks. Each audio track was orthographically transcribed in Elan (Lausberg and Sloetjes 2009) and force-aligned using FAVE to automatically determine the timing for each word in the transcript based on its alignment with the audio file (Rosenfelder et al. 2011).

Video blog monologues. Video blogs ('vlogs') are a form of computer-mediated communication in which people record videos of themselves discussing their lives or other topics of interest, to be shared with close friends or the public at large. For this study, we manually collected a dataset of 32 vlogs from different speakers. Since vlogs can be about a wide variety of topics, for the greatest comparability with our laboratory data we focused on vlogs about three emotive topics tied up in identity: high school students discussing their first day of school; students discussing their experiences studying for and taking the MCATs; and pregnancy vlogs in which pregnant women discuss various stages and milestones of their pregnancies. Vlogs on such topics by women are far more prevalent than those by men; therefore, in this study our Vlog dataset is composed entirely of women. The dataset consists of mostly White speakers (with a handful of Asian American speakers and one African American speaker) ranging in age from mid-teens to approximately 40 years old.

Web video data is in an important sense 'naturalistic.' YouTube has over one billion users and hundreds of millions of hours of video watched per day,[2] and individual vloggers may post often and over a long period of time. This makes vlogs an everyday speech event, part of vloggers' regular repertoire. So, while the language may be highly performative, a YouTube performance is naturally and regularly occurring in the world rather than elicited by a researcher.

Digital communications researchers and anthropologists have theorized about social phenomena like the construction of identity in such public online spaces, including vlogs in particular (Kollock and Smith 2002; Griffith and Papacharissi 2009; Biel and Gatica-Perez 2013; Burgess and Green 2013). Linguistic phenomena in such data, however, are generally underexplored. On the computational side, researchers have investigated tasks such as multimodal sentiment prediction (Wöllmer et al. 2013; Poria et al. 2016), but these tend to focus on predictive tasks and binary judgements like positive versus negative.

On the linguistic side, Androutsopoulos (2010) discussed some parameters of the unique 'sociolinguistic ecology' and multimodal 'spectacle' of Web 2.0 environments, emphasizing the importance of interaction. Frobenius (2014) considered audience design in vlogs, showing that the unique circumstance of a monologue with no feedback from an audience leads to interesting interactional phenomena; for example, the observation that prosodic shifts such as differences in volume can distinguish different intended audiences for particular utterances. The interactional component of vlogging is dynamic, sometimes going so far as to produce an actual asynchronous conversational context in a back-and-forth of videos (Harley and Fitzpatrick 2009).

Indeed, the vlogging world is permeated with the idea of interaction. Comments, shares, and 'likes' are often explicitly mentioned by vloggers as a crucial means of building a dialogue between the vlogger and the audience. Duman and Locher (2008) explored this 'video exchange is conversation' metaphor in detail in Hillary Clinton and Barack Obama's 2008 campaign clips on YouTube.

2.2 Preprocessing

The Vlog data sometimes presents potential problems for the computer vision algorithm to be used, due to sections of excessive cuts or additional visual effects such as introductory splash screens. Therefore, we manually determined appropriate start and end times for each video, clipping them to extract the largest possible contiguous sections without such effects.

To perform our computational analyses we first needed to define our units of analysis, and in this study we used pause-bounded units, henceforth referred to as simply 'phrases.' Since we have manual transcripts for data from the Lab setting, we performed forced alignment (Rosenfelder et al. 2011) to obtain

boundaries for each spoken word from each speaker. We then used a transcript-based method to extract phrases, defining a phrase as any continuous set of words such that no word is more than 100 milliseconds apart from the words surrounding it.

We did not, however, need manual transcripts to carry out many of the analyses we were interested in. Since we did not have manual transcripts for the Vlog data, we used an automatic heuristic based on the silence detection function in *Praat* (Boersma and Weenink 2015) to extract phrases. We generated phrases by running silence detection on the audio channel of each video, defining sounding portions as phrases. The more accurate phrases in our Lab data, extracted by the forced alignment method above, had an average length of 1.50 seconds. We approximated this in the Vlog data by setting the same 100 millisecond minimum boundary between sounding portions used above and starting with a silence threshold of -25dB. We iteratively ran silence detection, increasing or decreasing the silence threshold by 1dB and re-running, until the average phrase length was as close as possible to 1.50 seconds.

While this procedure may have smoothed over some individual variation in phrasal pacing, our primary need was for consistent units of analysis, which we defined using phonetic rather than intonational, syntactic, or discursive criteria for delineating phrase boundaries. In the analyses to follow we used the transcript-based phrases for the Lab data and the silence-detection-based phrases for the Vlog data; however, the results presented in the following sections held even if we also used silence-detection-based units for the Lab data, further suggesting that these units of analysis are roughly equivalent.

2.3 Head cant feature extraction

We calculated head tilt by adapting the shape-fitting algorithm of Kazemi and Sullivan (2014), as implemented in the open-source machine learning library *dlib* (King 2009). This algorithm is relatively computationally efficient and robust to differences in video quality, lighting, and occlusion, which made it feasible for the contextual diversity of our data (Figure 1).

For each frame of video in the dataset, we first used the standard face-detection implementation in *dlib* to find the speaker's face. We then used the aforementioned shape-fitting algorithm on the detected face, with a model pre-trained on the facial landmark data from the 300 Videos in the Wild dataset (Shen et al. 2015) which outputs locations of 68 facial landmark points per frame.

We could then calculate head cant using the points for the far corner of the left and right eyes by triangulation (Figure 1). Assuming (as we do in this dataset) a speaker roughly facing the camera, the cant angle is the arctangent of the vertical displacement of these eye corner points over their horizontal distance. We took the absolute value of these measurements as in this work we

Figure 1: Left, shape-fitting output from Kazemi and Sullivan (2014), showing robustness to occlusion. Right, visualization of our adaptation for the calculation of head cant angle on a vlog from our dataset, calculated by first fitting a shape model of the face to find landmark points as on the left, and then triangulating cant angle from the corners of the eyes

were interested in head cant primarily as displacement from an upright posture.

This method allowed us to generate a continuous estimate of head cant throughout all the videos in our dataset, analogous to measures of acoustic prosody like pitch and loudness, albeit at a more coarse resolution of once per frame (30Hz for a video at 30 frames per second). This method inevitably suffers from some limitations, since by the nature of large-scale automatic modeling we expect the model to introduce noise. At moments of severe occlusion – such as if a speaker turns fully away from the camera – or due to peculiarities of the algorithm's classification process, we may have failed to detect a face in a given frame or failed to accurately fit the shape model. We handled this by simply keeping track of these failures, and found that they occurred in approximately six percent of frames in the dataset. In the statistical analyses correlating head cant in prosody in section 4, we removed phrases from the analysis where more than half of the video frames that occurred during the phrase constitute classification failures of this type and as such have no accurate measurement.

A related limitation lies in the fact that head cant is naturally implicated in other bodily movements and postures, and our measurements may have been affected by this. Body cant, in particular, where the speaker's entire body is tilted and thus necessarily the head as well, presents an interesting difficulty in this regard. Nevertheless, in our qualitative analyses of the data we found this phenomenon to be relatively rare, and indeed this challenge is perhaps inherent to the study of embodiment. Even if we were hand-labeling the entire dataset, it is unclear whether a body cant of 20 degrees with a relative head

cant of 0 degrees should be labeled as a head cant of 0 or 20, since the head is straight relative to the body but at a 20 degree angle relative to the floor.

This difficulty becomes even more stark when we consider the potential for perceptual entanglements. If one speaker's head is canted their spatial coordinates are necessarily rotated, so should their interlocutor's head cant best be conceived of relative to that rotated perception, or relative to some 'objective' standard like the floor or other contextual grounding? All of the above likely constitutes a direction for future research in its own right, so in this work we sidestepped the issue by taking our computational method at face value.

3. HEAD CANT IN AND OUT OF INTERACTION

In framing the importance of head cant as an object of study in section 1, we postulated it to be an 'interactional variable,' playing a role in functions such as turn management between interlocutors. For example, head cant could function as a listening posture, signaling the listener role, or it could also signal interest in what the interlocutor is saying. We expected that neither of these functions would be present in the Vlogs, which have no explicit interlocutor, but that either or both could be present in the Lab data.

To explore this potential difference between datasets, we randomly sampled 5,000 individual frames of video from each speaker in the dataset, and determined whether the head cant measured in that frame occurred during a spoken phrase or not. As shown in Figure 2, speakers in the laboratory setting used more head cant overall than those in vlogs, with a mean cant of 6.4 degrees as compared to vloggers' mean cant of 4.5 degrees (two-sided t-test, $t = -105.4$, $df = 323{,}430$, $p < 0.001$).

We observed no statistical difference between speech and non-speech segments in the Vlogs, while the Lab participants used more head cant while not speaking than while speaking (two-sided t-test, $t = -21.425$, $df = 135{,}860$, $p < 0.001$). Moreover, we saw gender effects within the laboratory data. While men and women appeared to use nearly the same mean head cant of around six degrees during speech segments, an ANOVA analysis revealed a significant interaction effect with gender: men in our dataset used more head cant while not speaking than did women ($F = 192.5$, $p < 0.001$).

The relative low amount of cant in the Vlogs suggests that the movements that people make while speaking and listening in the Lab dialogues have an interactive signaling effect. It supports an association between listening and head cant, and it may suggest that cant is playing a role in floor management. It could also, though, reflect the importance of an interlocutor in supporting whatever other functions cant is playing.

Our results provide an interesting contrast with the results of Hadar et al. (1983), who used a polarized-light goniometer technique to measure head movements during conversation, finding evidence for constant movement during talk, while listening was marked by the absence of head movement.

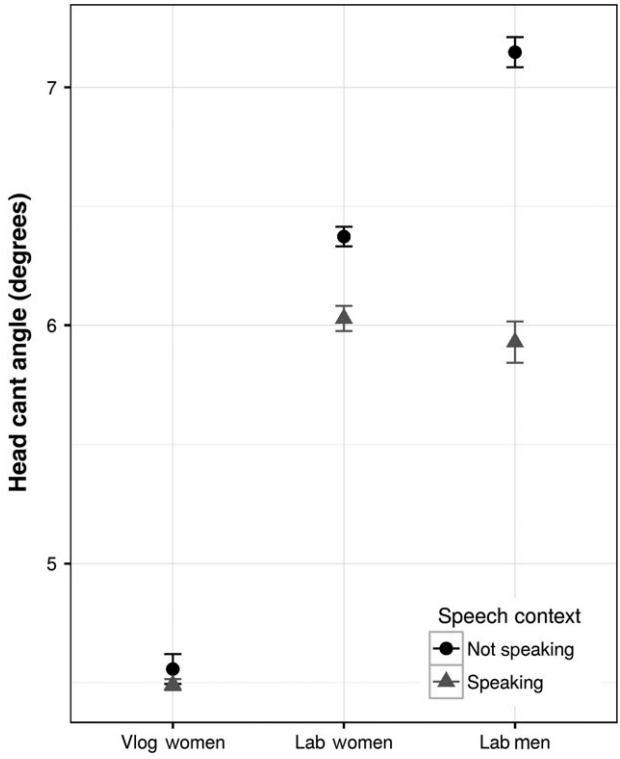

Figure 2: Distribution of head cant by gender and interactional context, distinguishing between speech context, that is, whether the speaker is currently speaking or not. Error bars represent 95 percent confidence intervals; these intervals are small since the number of observations is very high

Together, these results suggest that listening may be marked more by static but perhaps meaningful *postures* (such as head cant), while speaking may be marked by dynamic *movements*.

These findings also begin to challenge the gendered associations of head cant mentioned previously. In our data, men use more head cant overall, an effect driven by their use during non-speech portions of the interaction. Given this, we raise the question of whether women and men are doing more or less of the same thing, or whether they are actually using cant differently. We will explore these questions in the following sections.

4. HEAD CANT AND PROSODY

In the previous section we established a relationship between head cant and the simple fact of speaking, showing that this relationship is affected by

in-person interaction. In this section, we delve further by exploring not just incidence of speech but also some aspects of the nature of the speech produced. Beyond the straightforward distribution of cant, we can join our computational methodology in the visual modality for measuring head cant with existing common computational methodology in the acoustic modality: the automatic extraction of measurements of prosodic features of speech like pitch and loudness.

Prosodic variation on its own, of course, can communicate a number of social meanings, such as emotion (Scherer 2003) and attitude (Krahmer and Swerts 2005). Most relevant for our work, a number of previous studies show strong correlations between increased values of various prosodic variables and a speaker's general level of interest (Jeon, Xia and Liu 2010; Schuller et al. 2010; Wang and Hirschberg 2011), or engagement and excitement; for instance, Trouvain and Barry (2000) show that horse race commentators speak with higher pitch and pitch range, greater loudness, and increased speech rate as the races reach their peak of excitement in the finale.

Like head cant, pitch and loudness are continuously varying signals that are inevitably implicated in speech: by comparing these signals on a large scale we can aim to uncover cross-modal synchrony in the sense of McNeill's view of gestures and speech as fundamentally part of the same system. Existing studies have detailed particular elements of the strong relationship between the two modalities:

- Mendoza-Denton and Jannedy (2011) show evidence for the co-occurrence of pitch accents and gestural apices;
- Loehr (2012) similarly shows that gestural phrases align with intermediate phrases;
- Voigt, Podesva and Jurafsky (2014) show that increased overall body movement in a phrase predicts greater pitch and loudness mean and variability.

Our automatic annotations allow us to compare these signals – head cant and acoustic prosody – statistically on a large scale, which will allow us to understand both their overall relationship and how that relationship may differ or interact with contextual variables like the gender of the speaker or the context.

4.1 Methodological setup

In order to understand the joint influence of prosodic features and contextual factors, we built statistical models in which these features act as independent variables predicting head cant. We used *Praat* (Boersma and Weenink 2015) to extract F0 (hereafter 'pitch') and intensity (hereafter 'loudness') measurements throughout the audio track of each video. We then z-scored (subtracted the speaker's mean and divided by their standard deviation) all pitch, loudness, and head cant measurements, to convert them into an equivalent scale of units

of standard deviations across speakers. We did this so that genders and speakers are comparable: variation is only relevant with reference to some perceived baseline, which in this case we modelled as being speaker-specific. We then calculated the mean pitch and loudness, z-scored by speaker, for every phrase in the dataset (automatically identified as described in section 2.2). Our preprocessing resulted in a total of 17,533 usable phrases, representing more than 7.5 hours of continuous speech from our 65 speakers.

We modelled the interaction between these variables with a linear mixed-effects regression as implemented in the lme4 package in R (Bates et al. 2015); reported p-values were calculated with Satterthwate's approximations using the lmerTest package (Kuznetsova et al. 2013). Our regression used phrases as observations, including speakers as random effects. We modelled mean z-scored head cant in the phrase as the dependent variable, with independent variables of the z-scored pitch and loudness mean in the phrase, as well as the gender of the speaker, source of the video (Vlog or Lab), and the log duration of the phrase. We also included interaction effects between the prosodic (pitch, loudness) and contextual (gender, source, phrase duration) features in the model. Visual inspection of diagnostic plots confirmed that our fitted model met the model assumptions. We discuss particular results inline; a full regression table is given in an appendix.

4.2 Results

We found associations between head cant and both pitch and loudness in a phrase, again differing by both gender and interactional context, visualized in Figure 3.

Overall, we found that the higher the pitch in a phrase, the higher the degree of head cant (estimate 0.058, $p < 0.001$). This effect was modulated by gender with a significant interaction effect (estimate -0.042, $p = 0.002$), such that the trend held for women but substantially less for men in our dataset. For women an increase of one standard deviation in mean pitch during a phrase predicted an increase in head cant of nearly a degree. Pitch is quite regularly invoked as a fundamental gender difference, and one might be tempted to connect pitch and cant as jointly expressing something like femininity. If pitch in this case were directly associated with gender, though, one would expect the men's pitch to decrease rather than to simply increase less than the women's; this recalls the body of research suggesting that linguistic features do not map directly onto aspects of identity but rather that the relationship is complex and indirect (for instance, Ochs 1992).

While women and men differed only in significance in the pitch pattern, they showed opposite effects in the relation between cant and loudness. We found no overall fixed effect for loudness (estimate -0.088, $p = 0.58$), but there were strong interactions between loudness and both gender (estimate 0.083, $p = 0.003$) and context (estimate -0.109, $p < 0.001$). To some degree, these

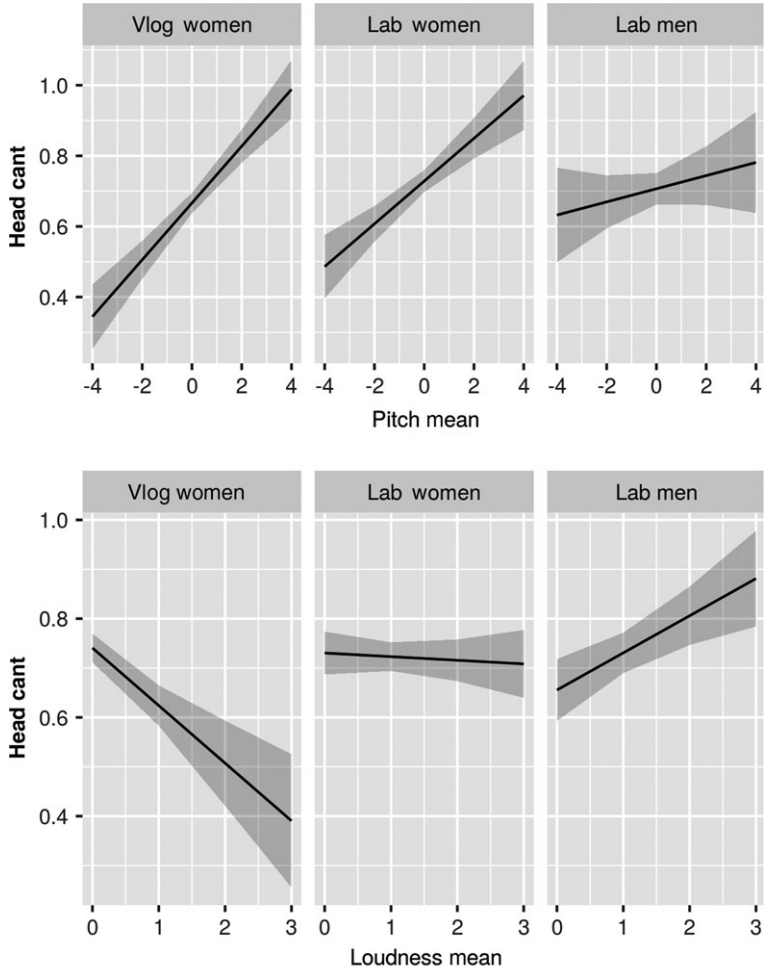

Figure 3: Marginal effects of pitch and loudness on head cant across genders and interactional contexts, holding other factors constant. All variables are z-scored by speaker, and observations in the model are silence-bounded phrases. Ribbons represent estimated 95 percent confidence intervals around the trend line

findings are in accord with prior work comparing the degree to which prosodic variables like pitch and loudness are 'socially loaded' differentially by gender; for example, in a study of speed dates Ranganath, Jurafsky and McFarland (2013) found that perceived friendliness is marked by pitch maximum and variability in women as compared to loudness variability in men. Women in vlogs displayed a strong negative relationship between loudness and head cant: for speakers in this category a decrease in loudness of one standard deviation

during a phrase predicted an increase in head cant of more than a degree. Recall that these speakers are also using significantly less cant overall than those in the Lab data. One interpretation of this result might be that, while cant as a socially meaningful variable for these speakers is, in general, rarer, it may have a more marked effect in the times it is used.

Women in the Lab interactional context showed a slight negative relationship between loudness and canting, but one far less strong than for Vlog women. The men in the Lab context, on the other hand, showed a strong positive relationship between loudness and head canting.

In summary, for both men and women, and for both Vlog and Lab settings, head cant was associated with higher pitch, although the effect was weaker for men. Head cant was generally associated with quieter speech in women, but louder speech in men.[3] These results suggest that women and men were doing quite different things with cant, since the gender difference is not a matter of degree but a reversal of effect.

While these statistical correlations cannot tell the whole story, they point out the tight connection across these modalities. To truly understand the meanings we have to delve deeper.

5. QUANTITATIVELY-GUIDED EXPLORATION

In the previous section, we identified several high-level relationships between head cant and prosodic features, demonstrating how these differed by gender and interactional context. To explore these in greater detail, in this section we use larger-scale 'distant reading' as a guide to facts on the ground. With this method we uncover a relationship between head cant and meanings having to do with shared understanding, and further show its gendered distribution with particular discourse particles for women (*you know* and *I mean*) as compared to conversational acknowledgements for men (*mmhmm* and *yeah*).

We postulate that the statistical relationships between prosody and cant discovered in the previous section are in part an indication that these features are participating in 'multimodal Gestalts' (Mondada 2014). That is, it is because these features are independently important as resources for the generation of social meaning that they tend to cluster in meaningful and consistent ways across speakers. This clustering may reflect sites of particularly rich generation of social meanings. But what is happening at these moments? To answer this question, in this section we use cross-modal quantitative measurements to directly observe these moments of meaningful co-variation.

In what we call a quantitatively-guided exploratory analysis, we developed a qualitative analysis by observing selections of randomly sampled phrases in context from the dataset that met some conditions observed in our broader statistical trends. This is an analytic move analogous to Labov's (2001) search for the social forces in sound change by using the data on variation to identify

the leaders in change, and looking close up at those leaders for commonalities in their social characteristics.

Since we found women to combine greater head cant with high pitch and low intensity, and men to combine head cant with high pitch and high intensity, we extracted phrases high in head cant co-occurring with high pitch and low intensity on the one hand, and with high pitch and high intensity on the other. We defined 'high' and 'low' as the top and bottom 30 percent, respectively, and considered five categories: high pitch alone; high intensity alone; low intensity alone; high pitch with high intensity; and high pitch with low intensity. For each category we randomly sampled and qualitatively examined at least 100 of these central exemplars.

One of the strongest trends we observed in this examination was that head cant is implicated both in floor management and in processes of signaling shared understanding – and that the two cannot be easily separated. Head cant appears to frequently be called on to establish that the speaker and the interlocutor share (or ought to share) some pre-existing knowledge about the discourse at hand. In this way we can view head cant as participating in shifts in footing, in the sense of Goffman (1981); that is, head cant may subtly modify the alignment or 'interactional frame' (Tannen and Wallat 1987) taken up by a speaker in a given utterance.

5.1 The framing of shared understanding

The interactional frame of 'shared understanding' in which head cant participates can take many forms. It appears to carry overtones of friendship, a sense of obviousness, or a taking into confidence, and can appear in the context of repetition or restating. In turn, it may be used for many purposes: to induce the interlocutor to interpret a claim as properly belonging to a shared understanding; to propose a presupposition of such understanding that softens an utterance for stylistic purposes; or to indicate dismissiveness of the obvious thing.

For example, in the Vlog context, consider a YouTuber named Nat in a video entitled '5 Weeks Pregnancy Vlog.' Nat's vlog records the journey of her pregnancy, discussing physical and emotional changes throughout and chronicling milestones along the way. Considering the audience design which might influence her linguistic choices, it's worth noting that Nat's channel is surprisingly popular, with over 40,000 subscribers, and as of this writing the video in question has had more than 28,000 views; however, the video in question is only the third published by her channel, so perhaps it had far less viewership when it was made.

In Example 1, below, Nat is ten minutes into the video, and is talking about telling her two best friends about her new pregnancy, both of whom have children of their own, as well as her husband Weston telling his friends. This moment follows a long and detailed account of telling her parents, and their

excited reactions. In contrast, she gives the story of telling those friends in a few brief sentences, ending with:

Example 1

1 Nat: and they're of course very excited
2 and very supportive and Weston told his two best friends

During this segment, Nat uses head cant in alternating directions with reduced loudness and variable pitch (Figure 4). The overall effect is to create a sense of obviousness but gratefulness in describing the reactions of her friends to hearing of her new pregnancy, which is strengthened by co-occurrence with the explicit 'of course.' Given the excited reactions of her parents she just described in detail, and the knowledge Nat expects to share with her imagined interlocutor that friends are generally excited about pregnancies, these head cants contribute to framing the content of her utterance as almost going without saying. We note that cant here is combined with semi-closed eyes and a smile. While we cannot comment authoritatively on eye and mouth features since we do not have equivalent data on them, it may be these features that contribute intimacy and positive affect. We note that these features can also be measured automatically, and ultimately an understanding of body movement is going to require careful analysis of multiple and co-occurring gestures, or constructions.

Moments later, Nat uses head cant again as she reiterates a point made earlier in the discourse: the pregnancy is still meant to be kept a secret to everyone but the couple's parents and very best friends. Earlier Nat has mentioned this fact several times, but tags it on with a clearly conspiratorial stylistic move generated by not only her words but the near-whispered tone, head cant combined with forward tilt, sly smile, wide open eyes, and a finger to the lips (Figure 5). We note that unlike the earlier uses of cant, here it

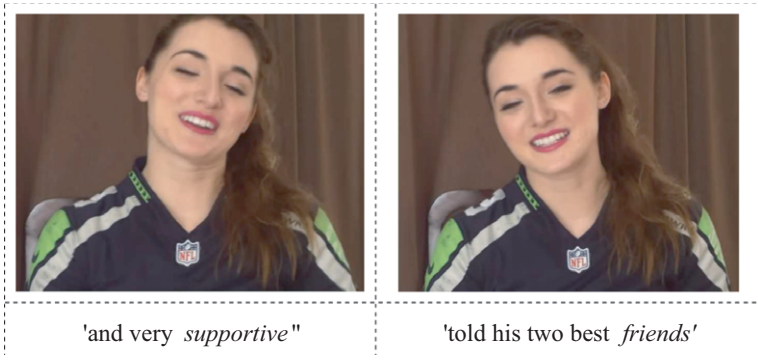

Figure 4: Nat's two head cants from 10:43–10:50 in Example 1 (https://youtu.be/fLS8RFnCcII?t=10m43s)

Figure 5: Nat's cant from 11:01–11:05 (https://youtu.be/fLS8RFnCcII?t=11m1s)

participates in a combination of gestures constituting a highly enregistered or conventionally 'iconic' sign. One question we might ask is whether complex enregistered signs like this one occur more frequently in monologues than in face-to-face interaction, which would support the hypothesis that the lack of an interlocutor calls for less subtle gestures.

Nat is invoking a set of shared beliefs about the social bonds associated with pregnancy. Example 2 from the Lab setting is a little more risky, as two interlocutors jointly confirm shared knowledge that might be face-threatening. Two friends, a White female (speaker A) and a Hispanic male (speaker B), are discussing the question of whether they have ever been mistaken for a person of another race. The conversation has turned to talking about racial diversity in AP ('advanced placement' or college-level) classes, as the Hispanic male describes being mistaken for Asian by virtue of being in those classes, and in other circumstances being mistaken for 'every race except White.' After a brief joking digression about how the White female speaker could never be mistaken for Black, she responds to the issue by bringing up the case at her high school:

Example 2

1 A: it was really weird at our school, cause, like
2 my school was like,
3 B: mostly. . .
4 A: a hundred percent White pretty much
5 B: White. . . yeah.

Immediately after the first 'like' in line 1, above, the speaker makes a shift to a head-canted posture, and simultaneously her loudness decreases, her speech rate increases, and her voice gets very creaky. These conditions hold through the end of line 4, and her head cant holds in the canted posture as well. Her cant marks a particular type of almost conspiratorial side comment, as if she is making an overly obvious confession, the content of which her interlocutor already knows (indeed, he produces simultaneous speech conveying the same proposition), intensified by the exaggerated 'a hundred percent' (Figure 6).

Figure 6: A head cant and its overlapping, softer-spoken response as both speakers discuss the ethnic diversity of their high schools in Example 2

At the same time, the information in the canted clause is highly relevant to the following discourse and is by no means obvious. After this comment, the speaker goes on to discuss more detailed specifics about the diversity of her high school and childhood community overall before returning to the topic of AP classes, suggesting this is information of which her interlocutor was not previously aware even though they are friends.

This example illustrates how head cant is used to establish footing for an interactional frame of shared understanding, as opposed to an indication of actual common ground in the sense of Clark and Brennan (1991). The speaker is drawing upon head cant as an interactional resource to frame the revelation of the lack of diversity at her high school in a particular way: as obvious, expected, and perhaps even somewhat embarrassing.

This is further evidenced by her interlocutor's reaction during the phrase, wherein he bobs his head in a minor mirroring head cant immediately following the speaker's initial cant, and speaks overlapping with her during the phrase in a low and creaky voice: in line 3 canting on 'mostly. . .', and in line 5 saying 'White' aloud almost exactly in time with speaker A, smiling gently at the end (Figure 6). Her head cant marks an invitation to the frame, and he participates; he in fact already knows the point at hand and doesn't have to wait for her to say 'White' but instead speaks it in time with her.

While head cant may enable the speaker and her interlocutor to orient to, more specifically distance herself from, the lack of diversity at her school, growing evidence suggests that creaky voice may function similarly. Lee (2015) observes that creak is often used to produce parenthetical speech, and that in much the same way that creak distances parenthetical speech from the primary thread of discourse, so too can speakers use it to distance themselves from their interlocutors or the topics about which they are talking. Similar claims have been made by D'Onofrio, Hilton and Pratt (2013), who show that two adolescent girls used creak regularly to distance themselves from statements that made them potentially vulnerable, and Zimman (2015), who showed that a transmasculine speaker telling a narrative about visiting family used more creak when referencing familial tensions. Thus, our argument about the interactional function of head cant is independently supported by the use of creaky voice in a completely different modality (speech).

5.2 Discourse particles

The hypothesis we've been exploring is that head cant functions in part to establish or index an interactional frame of shared understanding. We might, therefore, expect that cant would co-occur with discourse particles serving the same function. Since we have time-aligned transcripts for the Lab data, we can expand our quantitatively-guided investigation beyond cant and prosody to include the words used in the interactions. Schiffrin (1987) describes an interactional frame of shared understanding in the discourse particles *you know* and *I mean*. She suggests that *you know* directly acts as a mechanism for reaching a state of shared understanding where the speaker knows that the interlocutor has knowledge of the topic at hand. *I mean* focuses on other-orientation in the adjustment of footing towards the production of talk that the interlocutor will understand.

However, these particles can also serve as fillers, markers of upcoming delay or hesitation, or floor holders analogous to particles such as *like*, *um*, and *uh* (Clark and Fox Tree 2002; Fox Tree 2007). To approach this distinction, we compared the use of high cant in conjunction with *I mean* and *you know* on the one hand, and *um, uh*, and *like* on the other. We defined 'high cant' as phrases with a mean cant in the top 30 percent of all phrases.

Table 1: Odds ratios for discourse particles appearing in the top 30 percent of phrases with the highest head cant as compared to the bottom 70 percent*

Discourse particle	Women	Men
Shared understanding: *I mean, you know*	1.58 (p=0.03)	1.30 (p=0.29)
Hesitation/floor: *um, uh, like*	0.83 (p=0.01)	0.99 (p=1.00)

*Values higher than 1.0 indicate a positive association with canted phrases, while those lower than 1.0 indicate association with less canted phrases. P-values with Fisher's exact test are given in parentheses; values for women are significantly different while for men there is no association.

Across all 9,038 phrases in the Lab data, Fisher's exact test (Fisher 1922) shows (Table 1) that women are significantly more likely to use *you know* and *I mean* in phrases with high head cant, and less likely to use *um*, *uh*, and *like* in those phrases. *Like*, in particular, is strongly associated with phrases with low head cant. Andersen (1998) compiles an extensive review of research on *like*, finding that overall it acts as a 'loose talk marker' from a relevance-theoretic perspective – that is, the speaker is opting to signal a pragmatic 'discrepancy between the propositional form of the utterance and the thought it represents.'

To look at a particular example, the following section of speech occurs during a discussion of finding one's 'true passions,' where the speaker is expressing her surprise at finding that a set of activities in high school she originally participated in to pad her resume turned into a more sincere passion. During this extended turn (Example 3), the speaker starts with a relatively low cant, initiating a slight cant at the first 'actually' (line 3); however, she moves to a large head cant directly upon the phrase containing *you know* (line 4), spoken with a somewhat heightened pitch. This phrase marks the beginning of a conversational aside not constituting the content of her own story, but rather more as an attempt to gain 'meta-knowledge,' in Schiffrin's terms, that her interlocutor is also familiar with the background against which she was making her decisions.

Example 3

1 A: I kinda felt like I was doing it for the resume
2 like in high school to be honest
3 like, but then like I actually really liked it and then like –
4 like you know how like when –
5 when you wanna like fill your transcript up with like –
6 I mean resume up with like a bunch of like activities.
7 Like you get to choose what activities you want
8 B: mmhm
9 A: and all the activities I chose were

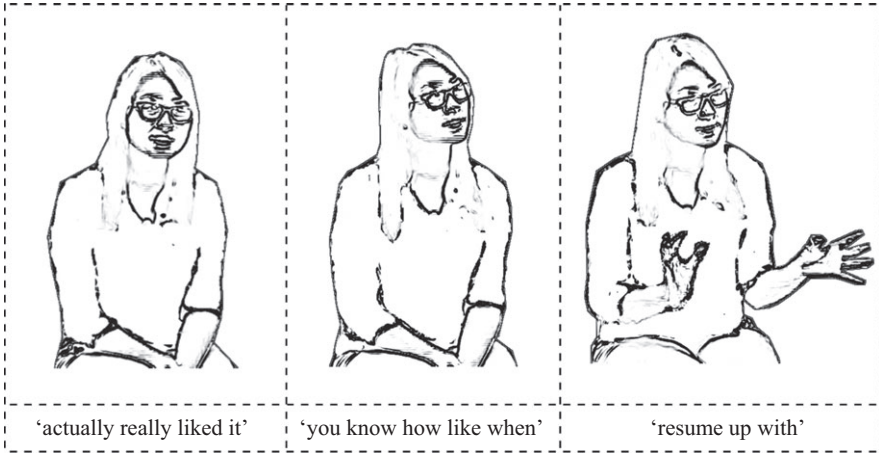

Figure 7: A speaker canting as she makes an aside about a shared experience of resume-padding in Example 3

Both the speaker and her interlocutor are students at an elite undergraduate institution: the speaker's use of *you know* helps to signal that she has made the very reasonable assumption that her interlocutor, too, knows about needing to bolster one's resume in high school. Her head cant pointedly marks the sentence and a half that follow as an almost redundant aside, helping to put this decision-making context into a frame of shared understanding that will allow her interlocutor to empathize with the experiences to follow (Figure 7).

The speaker continues to cant her head back and forth lightly during those phrases, and as she finishes saying 'choose what activities you want' (line 7) her interlocutor responds to the frame by smiling, nodding her head up and down, and backchanneling 'mmhm.' Precisely as the speaker returns to talking about her own experiences (line 9) her head cant returns to a neutral upright position, suggesting the bracketing function of the cant has come to an end.

5.3 Conversational acknowledgements

In the preceding section, we found a statistical relation confirming the link between other-oriented discourse markers such as *you know* and *I mean*, but these results held only for women in our dataset. One crucial difference across genders in our prosodic findings from section 4 was that men's increased head cant is associated with increased loudness, precisely the opposite relationship from that found in women.

In our analysis of phrases in the data matching the statistical trend – in this case, louder phrases spoken by men with high head cant – we found that this may accompany men's backchannels, acknowledgements, and affirmative responses. Recall from section 3 that men in our dataset used more head cant

Table 2: Odds ratios for conversational acknowledgements appearing in the top 30 percent of phrases with the highest head cant as compared to the bottom 70 percent*

Conversational acknowledgements	Women	Men
mm, mmhm, yeah	0.88 (p= 0.37)	2.19 (p<0.01)

*Values higher than 1.0 indicate a positive association with canted phrases, while those lower than 1.0 indicate association with less canted phrases. P-values with Fisher's exact test are given in parentheses; values for men are significantly different while for women there is no association.

while not speaking than did women; in investigating the data qualitatively we repeatedly encountered cases where men were listening for extended periods while holding a head cant posture, and then responded to statements from the interlocutor while maintaining that cant, often speaking more loudly. It may be that the move out of head cant towards a neutral posture accompanied by a (potentially louder) affirmation sets into motion a footing shift that brings an interactant out of a 'listener' role and into the 'speaker' role by catching on to moments of shared understanding. In this case, floor management merges with shared understanding.

This hypothesis recalls previous findings about conversational acknowledgements and backchannels such as *mm*, *mmhm*, and *yeah* that signal conversational engagement and which can also function as floor-grabbers (Jefferson 1984; Lambertz 2011). We again used Fisher's exact test on all 9,038 phrases in the Lab data to calculate the odds ratios for the occurrence of these words in phrases with high head cant as opposed to those without. We found that for the men in our dataset, conversational acknowledgements were highly associated with phrases with high head cant, while for women there was no relationship (Table 2).

Example 4 illustrates this phenomenon with an interaction between two men in the dataset, discussing what they would do with unlimited money. Both are undergraduates studying computer science, and clearly share a disdain for 'start-up accelerators' – companies that provide small amounts of early funding to start-ups in exchange for a substantial portion of equity in the company. They're discussing a particular such accelerator, anonymized here as 'Company X,' in overtly positive words but with a tone of dripping sarcasm that they know they both share.

Example 4

1 A: pay people to to think of ideas for start-ups for you
2 B: yeah that's th th that's –
3 A: yeah
4 B: that's exactly what um the Company X does

5 A: no, I yeah I know {laughter}
6 it's really funny.. oh man . . .
7 my teacher was like yeah Company X's like supported all this stuff {laughter}
8 B: supported {laughter}
9 A: supported, yeah it's like we like your idea here's some money
10 and now we get a . . .
11 like chunk of your company {laughter}

The sarcastic aside begins with speaker A's marked head cant on 'for you' (line 1) after the preceding part of the phrase was spoken with no cant and his eyes half-closed, looking downwards (Figure 8). He simultaneously raises his

Figure 8: Two speakers trade cants and conversational acknowledgements in an extended moment of sarcastic aside (Example 4)

gaze to meet eyes with his interlocutor as he cants his head, explicitly handing off the turn as his head cant helps frame the statement within their shared understanding as sarcastic. As they continue the sarcastic aside they trade turns repeatedly starting with 'yeah' or repetitions of their interlocutor's words ('supported') and ending with brief moments of laughter, throughout canting their heads to various degrees.

In this section we have explored several instances where head cant is participating in multimodal Gestalts having to do with a frame of shared understanding. This frame is broad and can contain numerous related overtones such as obviousness, embarrassment, and sarcasm. Coupled with quantitative evidence from the distribution of particular words in the transcripts, we identified a potential gender difference in how this frame is expressed, more commonly for women taking shape in interactional particles involving meta-knowledge and checks on the interlocutor's understanding, while for men in conversational acknowledgements and affirmative responses.

Nonetheless, the big question remains for future work: why are certain patterns of using cant more prevalent among men than women and vice versa? For instance, in the case of conversational acknowledgements, women do also show this pattern, simply less commonly, so the question is which women and when? In other words, what is the social significance of this kind of interactional move, and how does it enter into the construction of gender and other aspects of identity?

6. DISCUSSION

In this study, we have presented evidence that head cant is a robust interactive resource. It was more prevalent in our face-to-face Laboratory data than in YouTube monologues, suggesting that it plays an important signaling role to one's interlocutor. This seems to be related both to floor management and to footing in relation to conversational content, at times serving to bracket off particular frames.[4]

We also found that, in the dialogue but not the monologue context, head cant was more prevalent during times when the interactant was not speaking, suggesting an association with listening. This could be a simple signal that one is listening, yielding the floor, or it could communicate the listener's orientation in relation to the content. We found high-level statistical correlations between elements of engaged prosody and head cant. There was an overall positive relationship between increased cant in a phrase and increased pitch, and a complex relation between cant and loudness.

All of these correlations showed important gender effects. Men canted while listening more than women, suggesting that the traditional gendered associations that link head cant to hegemonic femininity are likely not telling the whole story. Increased cant correlated robustly with higher pitch among women, but appeared only as a trend among men. Finally, while men's loudness

increased with cant, women's decreased, particularly in the Vlog setting. The latter points to a qualitative gender difference, in which cant appears to be playing a more important role in floor management for men than for women.

This appears to be supported by the relation between cant and discourse particles in the Lab data. We found that women's phrases with high head cant were associated with discourse particles having to do with shared understanding like *you know* and *I mean*. This did not hold for men. Conversely, for men but not for women, phrases with high head cant were associated with conversational acknowledgements like *mmhm* and *yeah*, suggesting more of a floor management function.

The set of gender differences we uncovered at every stage – across speaking contexts, in prosodic correlations, and in particular lexical items – suggests that the distribution of the communicative uses of head cant is gendered to some extent. However, the relation of this feature to gender is neither simple nor direct. We note that binary gender is low hanging fruit, as very little information is required to assign speakers to the male or female category. Our attention to gender in this study emerged initially from the previous literature, but it is possible that equally interesting patterns may emerge with other macro-social categorization schemes, such as class, ethnicity, or age. Ultimately, the meaning of cant is not 'male' or 'female,' but qualities and orientations that differentiate among and between the binary gender categories.

More broadly, we have shown that head cant is an interactional resource, and in this capacity it interacts with both sound and text on the one hand and other body movements on the other, to build higher level structures, or interactional signs. Much work is needed to uncover the nature of gestural signs, and their combinations, a challenge that is shared by current work in variation in speech (e.g. Eckert 2016). Ultimately, this adds an entirely new medium to the study of variation, and challenges us to integrate body movement into our theories of variation.

6.1 Moving forward

Through an extended exploration of head cant, we hope this paper has illustrated the value of taking a computational approach to embodiment. Computational methods facilitate the analysis of larger datasets than are typically employed in research examining the role of the body in interaction. While micro-analyses of interaction have been and continue to be instrumental to understanding the complex orchestration of multimodal interactional resources in communication, large-scale analyses enable researchers to consider other types of questions.

First, beyond simple generalizability, analyses of larger datasets enable us to identify the broader interactional functions for individual embodied resources. For example, we have observed here that, across a relatively large number of interactions, interactants cant their heads to a greater degree when listening.

This reveals that, even though head cant can be used to take up a variety of rather different interactional positionings, as detailed in section 5, it simultaneously serves a common function. It remains to be seen whether other embodied resources pattern similarly, but large-scale analyses can help determine the extent to which specific forms of embodiment are interactionally meaningful in and of themselves, irrespective of the particulars of a given interaction.

While it is important to identify the general interactional functions that embodied resources might serve, we emphasize that such general functions only scratch the surface of the meaning potentials for these resources. Any study of head cant would be incomplete without a discussion of how its meaning is mediated by the other features with which it occurs. Quantitative analyses facilitate the identification of collocations between embodied resources (e.g. head cant) and social (e.g. gender) and linguistic factors (e.g. discourse particles, prosody). We can therefore uncover trends like 'women produce higher head cant with higher pitch and lower loudness, particularly when producing discourse markers like *you know* and *I mean*.' In addition to the methodological ability to identify collocations, we gain an important theoretical insight: that the meaningful interactional resource (or, put another way, sign object) is not simply head cant, but the 'multimodal Gestalt' of head cant, gender, prosody, and discourse particles.

To take the approach advocated in this paper, sociolinguists must have access to both computer vision methods of the sort used here and audiovisual data. Regarding computational methodologies, we call on the research community to share newly developed methods for analyzing embodiment. So that future researchers can replicate our results and study head cant and other visual features in new datasets moving forward, we release all corresponding code at this url: nlp.stanford.edu/robvoigt/cans_and_cants

Although sociolinguistic work is carried out predominantly on audio corpora, audiovisual data greatly expand the kinds of considerations that analysts can take into account. Lab data like those used in this study afford perhaps the largest variety of explorations, given the angle of video capture and the high quality of audio recordings. Yet, field recordings can, in principle, provide many of the same opportunities, given the right setup; though audio quality would surely suffer, the ecological validity would likely be improved. We also underscore the power of web data, particularly video blog data. The majority of videos on YouTube are posted publicly, which allows researchers to share datasets and increase replicability. In such cases, since users choose to post the videos publicly, privacy concerns are less of an issue than with experimental participants. However, users are also free to remove videos, so as a community the organized archiving of large-scale datasets of this type presents an important opportunity and challenge moving forward, as some researchers have begun to explore (Cieri 2014).

While incorporating computational approaches to embodiment surely introduces a number of methodological challenges, we hope this paper has

shown these challenges are surmountable and that the payoff – the ability to attend to the physical world in quantitative analyses of interaction – makes it worth confronting them. As Sharma (2016: 335) notes in the introduction to this Series, 'sociolinguists focus on linguistic form but have always known that interaction does more than bring voices into contact. It creates momentary alliances of bodies, strategies, geographies, and various other signals and positionings.' By gaining insight into interaction through the movement and orientation of the body instead of the voices issued from it, we both arrive at a better understanding of interaction and have more solid footing on which to make claims about the ways linguistic behavior varies as a function of interaction.

NOTES

1. Many thanks to Martin Kay for insightful comments on an early version of this paper, the members of the Interactional Sociophonetics Lab for help with data collection and preprocessing, and Robert Xu and Ciyang Qing for editing of the Chinese abstract. The first author graciously acknowledges the support of the Michelle and Kevin Douglas Stanford Interdisciplinary Graduate Fellowship. This research was supported in part by the NSF via Award #IIS-1159679 and by the DARPA Communicating with Computers (CwC) program under ARO prime contract no. W911NF-15-1-0462. Data collection and annotation was supported by a grant from the Roberta Bowman Denning Initiative in the Digital Humanities, awarded to the last author. We also thank two anonymous reviewers for their useful feedback. Any remaining errors are our own.
2. https://www.youtube.com/yt/press/en-GB/statistics.html
3. In spite of this robust statistical evidence at a high level, these trends are not necessarily universally generalizable to every speaker. To check for individual variability, we tried building separate models for each speaker in the dataset; indeed, we found that for each result a small minority of speakers appeared to buck the trend (for instance, a few speakers showed small negative coefficients for pitch, suggesting lower pitch in phrases with higher head cant). Nevertheless, these effects were non-significant; in no case did we find any speaker with a statistically significant effect opposing the results presented here, so without more data we cannot be certain if this possible variability is due to inherent noise in the data or if these speakers are true outliers.
4. We note that the functions of cant discussed throughout this paper are not meant to be taken as exhaustive; indeed, head cant can likely serve a number of other interactional functions such as conveying skepticism, as pointed out by an anonymous reviewer.

REFERENCES

Andersen, Gisle. 1998. The pragmatic marker *like* from a relevance-theoretic perspective. In Andreas H. Jucker and Yael Ziv (eds.) *Discourse Markers: Descriptions and Theory*. Amsterdam, The Netherlands: John Benjamins. 147–170.

Androutsopoulos, Jannis. 2010. Localizing the global on the participatory web. In Nikolas Coupland (ed.) *The Handbook of Language and Globalization*. Chichester, U.K.: John Wiley and Sons. 203–231.

Barsalou, Lawrence W. 2008. Grounded cognition. *Annual Review of Psychology* 59: 617–645.

Bates, Douglas, Martin Maechler, Ben Bolker and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Bee, Nikolaus, Stefan Franke and Elisabeth André. 2009. Relations between facial display, eye gaze and head tilt: Dominance perception variations of virtual agents. Paper presented at the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 10–12 September, De Rode Hoed, Amsterdam, The Netherlands.

Biel, Joan-Isaac and Daniel Gatica-Perez. 2013. The YouTube lens: Crowd-sourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia* 15: 41–55.

Birdwhistell, Ray L. 1952. *Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture*. Louisville, Kentucky: University of Louisville.

Birdwhistell, Ray L. 1970. *Kinesics and Context*. Philadelphia, Pennsylvania: University of Pennsylvania Press.

Boersma, Paul and David Weenink. 2015. Praat: Doing phonetics by computer [computer program]. Version 6.0.08. Available at http://www.praat.org/

Burgess, Jean and Joshua Green. 2013. *YouTube: Online Video and Participatory Culture*. Chichester, U.K.: John Wiley & Sons.

Cieri, Christopher. 2014. Challenges and opportunities in sociolinguistic data and metadata sharing. *Language and Linguistics Compass* 8: 472–485.

Clark, Herbert H. and Susan E. Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition* 13: 127–149.

Clark, Herbert H. and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84: 73–111.

Costa, Marco, Marzia Menzani and Pio Enrico Ricci Bitti. 2001. Head canting in paintings: An historical study. *Journal of Nonverbal Behavior* 25: 63–73.

Cvejic, Erin, Jeesun Kim and Chris Davis. 2010. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication* 52: 555–564.

Dhall, Abhinav, Roland Goecke, Jyoti Joshi, Karan Sikka and Tom Gedeon. 2014. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*. New York: Association for Computing Machinery. 461–466.

D'Onofrio, Annette, Katherine Hilton and Teresa Pratt. 2013. Creaky voice across discourse contexts: Identifying the locus of style for creak. Paper presented at New Ways of Analyzing Variation 42, 10–14 October, Carnegie Mellon University, Pittsburg, Pennsylvania.

Duman, Steve and Miriam A. Locher. 2008. 'So let's talk. Let's chat. Let's start a dialog': An analysis of the conversation metaphor employed in Clinton's and Obama's YouTube campaign clips. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication* 27: 193–230.

Eckert, Penelope. 2016. Variation, meaning and social change. In Nikolas Coupland (ed.) *Sociolinguistics: Theoretical Debates*. Cambridge, U.K.: Cambridge University Press. 68–85.

Fisher, Ronald A. 1922. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85: 87–94.

Fox Tree, Jean E. 2007. Folk notions of *um* and *uh*, *you know*, and *like*. *Text & Talk – an Interdisciplinary Journal of Language, Discourse Communication Studies* 27: 297–314.

Frobenius, Maximiliane. 2014. Audience design in monologues: How vloggers involve their viewers. *Journal of Pragmatics* 72: 59–72.

Girshick, Ross, Jeff Donahue, Trevor Darrell and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE Computer Society Conference Publishing Services. 580–587.

Glenberg, Arthur M. and Michael P. Kaschak. 2003. The body's contribution to language. *Psychology of Learning and Motivation* 43: 93–126.

Goffman, Erving. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Cambridge, Massachusetts: Harvard University Press.

Goffman, Erving. 1979. *Gender Advertisements*. New York: Harper & Row.

Goffman, Erving. 1981. *Forms of Talk*. Philadelphia, Pennsylvania: University of Pennsylvania Press.

Grammer, Karl. 1990. Strangers meet: Laughter and nonverbal signs of interest in opposite-sex encounters. *Journal of Nonverbal Behavior* 14: 209–236.

Griffith, Maggie and Zizi Papacharissi. 2009. Looking for you: An analysis of video blogs. *First Monday* 15.

Hadar, Uri, T. J. Steiner, E. C. Grant and F. Clifford Rose. 1983. Kinematics of head movements accompanying speech during conversation. *Human Movement Science* 2: 35–46.

Harley, Dave and Geraldine Fitzpatrick. 2009. Creating a conversational context through video blogging: A case study of Geriatric1927. *Computers in Human Behavior* 25: 679–689.

Jefferson, Gail. 1984. Notes on a systematic deployment of the acknowledgement tokens 'yeah'; and 'mm hm'. *Papers in Linguistics* 17: 197–216.

Jeon, Je Hun, Rui Xia and Yang Liu. 2010. Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Japan: International Speech Communication Association. 2806–2809.

Kang, Mee-Eun. 1997. The portrayal of women's images in magazine advertisements: Goffman's gender analysis revisited. *Sex Roles* 37: 979–996.

Kazemi, Vahid and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE Conference Publications. 1867–1874. doi:10.1109/CVPR.2014.241

Kendon, Adam. 1994. Do gestures communicate? A review. *Research on Language and Social Interaction* 27: 175–200.

Kendon, Adam. 1995. Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics* 23: 247–279.

Kendon, Adam. 2002. Some uses of the head shake. *Gesture* 2: 147–182.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge, U.K.: Cambridge University Press.

Kim, Minyoung, Sanjiv Kumar, Vladimir Pavlovic and Henry Rowley. 2008. Face tracking and recognition with visual constraints in real-world videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CPVR 2008)*. New York: IEEE Conference Publications. 1–8.

King, Davis E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10: 1755–1758.

Kollock, Peter and Marc Smith (eds.). 2002. *Communities in Cyberspace*. London/NewYork: Routledge.

Krahmer, Emiel and Marc Swerts. 2005. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech* 48: 29–53.

Kress, Gunther R. and Theo Van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. New York: Oxford University Press.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25: 1106–1114.

Kuhnke, Elizabeth. 2012. *Body Language for Dummies*. Chichester, U.K.: John Wiley & Sons.

Kuznetsova, Alexandra, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen. 2013. lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version, 2(6).

Labov, William. 2001. *Principles of Linguistic Change, II: Social Factors*. Malden, Massachusetts: Blackwell.

Lambertz, Kathrin. 2011. Back-channelling: The use of *yeah* and *mm* to portray engaged listenership. *Griffiths Working Papers in Pragmatics and Intercultural Communication* 4: 11–18.

Lausberg, Hedda and Han Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41: 841–849.

Lee, Sinae. 2015. Creaky voice as a phonational device marking parenthetical segments in talk. *Journal of Sociolinguistics* 19: 275–302.

Loehr, Daniel P. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3: 71–89.

Matlock, Teenie, Michael Ramscar and Lera Boroditsky. 2003. The experiential basis of meaning. In Richard Alterman and David Kirsh (eds.) *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey: Lawrence Erlbaum. 792–797.

McClave, Evelyn Z. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32: 855–878.

McKenna, Stephen J., Sumer Jabri, Zoran Duric, Azriel Rosenfeld and Harry Wechsler. 2000. Tracking groups of people. *Computer Vision and Image Understanding* 80: 42–56.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago, Illinois: University of Chicago Press.

McNeill, David. 2008. *Gesture and Thought*. Chicago, Illinois: University of Chicago Press.

Mendoza-Denton, Norma and Stefanie Jannedy. 2011. Semiotic layering through gesture and intonation: A case study of complementary and supplementary multimodality in political speech. *Journal of English Linguistics* 39: 265–299.

Mignault, Alain and Avi Chaudhuri. 2003. The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior* 27: 111–132.

Mills, Janet. 1984. Self-posed behaviors of females and males in photographs. *Sex Roles* 10: 633–637.

Mondada, Lorenza. 2014. Bodies in action: Multimodal analysis of walking and talking. *Language and Dialogue* 4: 357–403.

Mondada, Lorenza. 2016. Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics* 20: 336–366.

Murphy-Chutorian, Erik and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31: 607–626.

Nevile, Maurice. 2015. The embodied turn in research on language and social interaction. *Research on Language and Social Interaction* 48: 121–151.

Ochs, Elinor. 1992. Indexing gender. In Alessandro Duranti and Charles Goodwin (eds.) *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge, U.K.: Cambridge University Press. 335–358.

Pellegrini, Stefano, Andreas Ess and Luc Van Gool. 2010. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision*. Heidelberg, Germany: Springer Berlin Heidelberg. 452–465.

Poria, Soujanya, Erik Cambria, Newton Howard, Guang-Bin Huang and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174: 50–59.

Ranganath, Rajesh, Dan Jurafsky and Daniel A. McFarland. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language* 27: 89–115.

Rautaray, Siddharth S. and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review* 43: 1–54.

Rosenfelder, Ingrid, Joe Fruehwald, Keelan Evanini and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. Available at http://fave.ling.upenn.edu

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma and Zhiheng Huang. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115: 211–252.

Scherer, Klaus R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40: 227–256.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge, U.K.: Cambridge University Press.

Schuller, Björn, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A. Müller and Shrikanth S. Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Japan: International Speech Communication Association. 2795–2798.

Shan, Caifeng. 2012. Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing* 21: 431–436.

Sharma, Devyani. 2016. Series introduction. *Journal of Sociolinguistics* 20: 335.

Shen, Jie, Stefanos Zafeiriou, Grigorios G. Chrysos, Jean Kossaifi, Georgios Tzimiropoulos and Maja Pantic. 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1003–1011.

Suarez, Jesus and Robin R. Murphy. 2012. Hand gesture recognition with depth images: A review. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. 411–417.

Tang, Siyu, Mykhaylo Andriluka and Bernt Schiele. 2014. Detection and tracking of occluded people. *International Journal of Computer Vision* 110: 58–69.

Tannen, Deborah and Cynthia Wallat. 1987. Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview. *Social Psychology Quarterly* 1: 205–216.

Trouvain, Jürgen and William J. Barry. 2000. The prosody of excitement in horse race commentaries. In R. Cowie, E. Douglas-Cowie and M. Schröder (eds.) *Proceedings of the International Speech Communication Association Workshop on Speech and Emotion*. Belfast, Ireland: Textflow. 86–91.

Viola, Paul and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *CVPR 2001: Proceedings of the 2001 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. 511–518.

Voigt, Rob, Robert J. Podesva and Dan Jurafsky. 2014. Speaker movement correlates with prosodic indicators of engagement. *Speech Prosody* 7.

Wang, William Yang and Julia Hirschberg. 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proceedings of the SIGDIAL 2011 Conference*. Stroudsburg, Pennsylvania: Association for Computational Linguistics. 152–161.

Wöllmer, Martin, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28: 46–53.

Zimman, Lal. 2015. Creak as disengagement: Gender, affect, and the iconization of voice quality. Paper presented at *New Ways of Analyzing Variation 44*, 22–25 October, Toronto, Canada.

APPENDIX: *Cant and prosody regression results*

Full regression results table for the mixed-effects model described in section 4. Head cant, pitch, and loudness are phrase-level means of values z-scored by speaker. Reference levels are a gender of female and a Laboratory interactional context.

	Head cant angle (z-scale)		
Fixed effect	***B***	***CI***	***p***
(Intercept)	0.73	0.69 – 0.78	**<.001**
Pitch mean (z-scale)	0.06	0.04 – 0.08	**<.001**
Loudness mean (z-scale)	−0.01	−0.04 – 0.02	.584
Gender (M)	−0.08	−0.15 – 0.00	**.049**
Context (Vlog)	0.01	−0.04 – 0.06	.674
Log duration	−0.01	−0.03 – 0.00	.057
Pitch mean * gender (M)	−0.04	−0.08 – 0.00	**.032**
Pitch mean * context (Vlog)	0.02	−0.01 – 0.05	.169
Pitch mean * log duration	0.03	0.02 – 0.05	**<.001**
Loudness mean * gender (M)	0.08	0.03 – 0.14	**.003**
Loudness mean * context (Vlog)	−0.11	−0.17 – −0.05	**<.001**
Loudness mean * log duration	0.02	−0.00 – 0.03	.107
Random effect			
σ^2		0.317	
$\tau_{00,\ case}$		0.003	
N_{case}		67	
ICC_{case}		0.010	
Observations		17,533	
R^2 / $\Omega_0{}^2$		.019 / .019	

Address correspondence to:

Rob Voigt
Stanford University – Linguistics Department
Margaret Jacks Hall
Building 460
Stanford, CA 94305
U.S.A.

robvoigt@stanford.edu