THE CONNECTION BETWEEN SMILING AND GOAT FRONTING: EMBODIED AFFECT IN SOCIOPHONETIC VARIATION

Robert J. Podesva, Patrick Callier, Rob Voigt, and Dan Jurafsky

Stanford University
podesva@stanford.edu, pcallier@stanford.edu, robvoigt@stanford.edu, jurafsky@stanford.edu

ABSTRACT

This study examines the effect of smiling on GOAT fronting, a sound change common to many varieties of American English. The data are audiovisual recordings of ten speakers of American English recorded in dyadic conversations in an interactional sociophonetics laboratory. We applied an existing computer vision algorithm for smile detection to the video recordings to identify smiling intervals. A mixed-effects linear regression reveals that higher F2 (i.e., auditorily fronter GOAT) positively correlates with whether speakers are smiling while articulating the vowel and their self-reported comfort levels in the interaction. The latter factor does not correlate with whether vowels were smiled. Together, the findings suggest that GOAT fronting is not only a phonetic consequence of smiling, but also serves an affective, interactional function. While sociophonetic studies typically analyze audio recordings alone, patterns of variation are better explained by also attending to embodied practices observable only in the visual domain.

Keywords: affect, embodiment, GOAT fronting, smiling, sociophonetics

1. INTRODUCTION

In the study of sociophonetic variation, researchers typically examine the audio channel independently from, and to the exclusion of, the visual channel. On the one hand, this practice makes sense, given that the relevant movements of speech articulators can be inferred from the acoustic signal, and that most body parts need not move to produce speech. On the other hand, by ignoring the visual channel, sociophoneticians may be neglecting the ways in which non-speech-articulators co-vary or work in tandem with sociophonetic variation, and perhaps imbue it with social meaning. This paper examines one case of embodied affect—smiling—and its connection to the fronting of the GOAT vowel.

By the term *embodiment*, we refer to the range of ways that the body can be positioned and moved, including gesture, posture, how interactants situate their bodies in relation to one another, and physical displays of affect. Eckert [1] has suggested that affect plays a more fundamental role in structuring variation patterns than is typically recognized. She shows that preadolescents in California produce fronter PRICE and LOT to convey positive affect and retract the same vowels in negative interactional contexts. Wong [9] reports a similar pattern among Chinese American youths in New York City.

Attending to affect in a sociophonetic paradigm presents a significant challenge. The coding of affect can be subjective or too time-consuming to apply to large datasets. Although the objective, empirically rigorous coding of affect is possible using the tools of discourse anlaysis, these methods may be too time-intensive to employ on datasets that are sufficiently large for identifying sociophonetic patterns. In this paper, we investigate the sociophonetic consequences of smiling, a form of embodied affect that can be straightforwardly coded in a large dataset.

That smiling is an inherently affective phenomenon was established in Ekman et al.'s [2] classic research in social psychology. They used instrumental techniques to measure spontaneous facial expressions produced by subjects who watched a video and later reported on their emotional state. Those who smiled reported being happier, and the magnitude of smiles correlated with the intensity of happiness. While there are many kinds of smiles, with distinct meanings and interactional functions, we assume that all are displays of affect.

Smiling also has phonetic consequences for the auditory impression of fronting. Advancement of the tongue body during the articulation of a vowel shortens the front cavity of the vocal tract, resulting in a relatively higher second formant (F2). Smiling achieves a similar auditory effect; the lips are spread and to some extent retracted, which also shortens the front cavity, giving rise to a higher F2.

We focus in particular on the relationship between smiling and a single vowel class, GOAT. We consider a single vowel category because the acoustic effect of smiling differs across vowel classes [3]. We examine a back vowel because fronting of these vowels is prevalent across several varieties of American English [5]. Finally, we consider GOAT, to the exclusion of the other back vowels, because it occurs more frequently in unscripted speech (and the lexicon) than GOOSE and FOOT.

In the following section, we describe the methods and laboratory space used for collecting the audio-visual corpus, as well as analysis procedures. We then present the results of the study (section 3), and discuss their implications (section 4). We conclude by discussing the implications of our findings for previous work on GOAT fronting and by offering suggestions for ways that embodiment might be further incorporated into sociophonetic research.

2. METHODS

2.1. Data collection

We report on an analysis of 10 speakers of American English, half female and half male, recorded in dyadic interactions with friends. Within each sex class, 3 speakers were white, 1 African American, and 1 Latina/o. All speakers were in their 20s and grew up in parts of the United States in which back vowels are fronting (California and Texas). Speakers were paid $10 for their participation and consented to being video- and audio-recorded.

Participants were recorded in an interactional sociophonetics laboratory. The room in which recordings were made has the acoustical specifications of a sound booth. In contrast to most sound booths, however, the space is staged like a living room, with a sofa, armchairs, tables, and bookshelves set up to facilitate natural social interaction in a less overtly experimental context. Acoustical wall panels are covered with fabric to resemble wallpaper, video cameras are housed inconspicuously in decorative boxes, and data cables are routed through the walls and under the floor to an observation station in an adjacent room.

Participants wore wireless lapel microphones, with separate audio files recorded for each speaker. Interactants were also video-recorded in their own frames to facilitate video analysis, described below.

Speakers were first asked to discuss a few "would you rather" questions (e.g., Would you rather always be overdressed or always be underdressed?), during which audio recording levels were adjusted. Answering these questions, presented on a rolodex-style binder, also gave interactants time to become accustomed to each other and the recording environment. Speakers were then asked to engage in about 30 minutes of conversation, with the aid of prompts (also presented on a binder) if desired, though all participants were told that they should feel free to go off topic as they choose. Prompts (e.g., How has the way you dress changed since high school?) were chosen to encourage conversation, extended turns at talk, and reflection about identity. After the conversation, participants were asked to complete an electronic survey and provide demographic information and subjective assessments of the interaction and their interlocutors (e.g., How comfortable did you feel?, recorded on a slider bar).

2.2. Acoustic analysis

We consider the conversation data only here. Recordings were transcribed in ELAN,[1] and forced alignments were generated using FAVE.[2] No phone boundaries were corrected, as the high audio quality gave rise to rather accurate alignments. A variety of acoustic measures were taken via Praat[3] script every 10 ms, including F0, F1-F3, intensity, duration, and a variety of voice quality measures of spectral tilt and periodicity. The set of measurements comprising each vowel was then reduced to the median value, for each acoustic measure. Formant values were normalized using the Lobanov [6] method to facilitate inter-speaker comparisons.

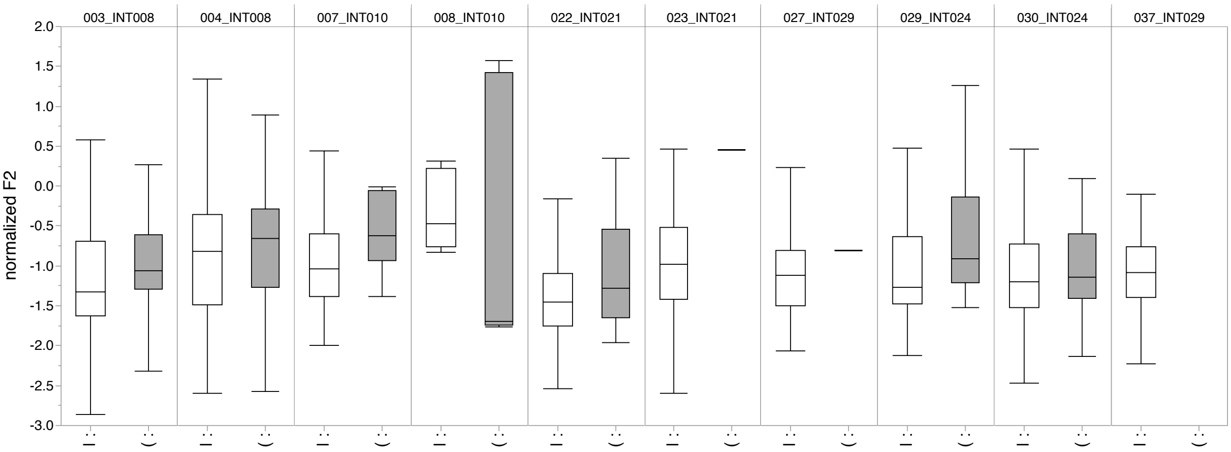
2.3. Smile detection

To identify smiling intervals, we used open source data[4] (annotated for smiling presence/absence) to train a Haar cascade classifier [7] in OpenCV.[5] The smile detector was trained on 13,165 images and performs at 88.8% accuracy (when tested on its own training data). Smiles were misclassified as non-smiles 19.1% of the time, while non-smiles were classified as smiles at only 8.2%, indicating a conservative standard for what constitutes a smile. As smile detection is optimized for head-on video, each interactant was recorded in a separate video file, which served as input for smile detection. GOAT tokens were counted as smiled only if smiles were detected in 40% or more of the video frames comprising the vowel's duration.

2.4. Statistical analysis

A data table integrating the acoustic measures, the output of smile detection, and survey data was assembled. A mixed-effects linear regression model was then built for the response variable, normalized F2 of GOAT (N=1,305). Lexical item and preceding and following sound were included as random intercepts, and speaker was included as a random slope, as explained in section 3. Fixed effects included several lingusitic factors—F0, standard deviation of F0 over the IP (as a prosodic variability measure), and intensity—and the social factors gender (speaker and interlocutor) and whether the speaker was smiling during the vowel. Finally, the regression incorporated elements of the survey, such as the speaker's self-reported comfort level and their assessment of how much they clicked with their interlocutor.

Figure 1: Effect of smiling on GOAT F2 (normalized), for individual speakers.



3. RESULTS

Figure 1 shows the normalized F2 for each speaker (labeled by speaker number and interaction number), depending on whether they were smiling during the production of GOAT (smiley face) or not (neutral expression). A few patterns are evident. First, the predicted pattern—that F2 would be higher for smiled vowels than for vowels that were not smiled—is evident for the majority of speakers (with speaker 008 as the only exception). Second, some speakers smiled infrequently (023, 027), and in one case not at all (037), during the production of GOAT. Finally, even though most speakers exhibit a higher F2 when they smile, individuals vary with respect to the magnitude of this difference. Speaker was therefore included as a random slope in the regression model, based on the by-speaker size of difference between smiled and non-smiled GOAT F2.

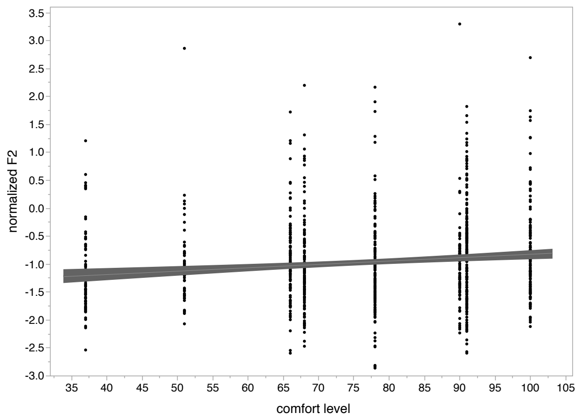
Table 1 summarizes the fixed effects found to significantly affect GOAT F2 (while also factoring in the influence of random effects). The regression confirms that the pattern observed in Figure 1, that GOAT exhibits higher F2 when it coincides with smiling, is generalizeable to the 10 speakers in the sample. Comfort level was also found to have a significant effect on F2.

Table 1: Summary of regression model on GOAT F2 (normalized).

Term	Est.	Std. Err.	DF Den.	t Ratio	prob >|t|
Intercept	-1.4568	0.155	30.06	-9.38	<0.0001 *
smiling	0.0831	0.036	21.68	2.25	0.0346 *
comfort level	0.0042	0.002	18.04	2.51	0.0218 *

The effect of comfort level is shown in Figure 2. As self-reported comfort levels increase, GOAT is produced with higher F2. Put another way, speakers produce auditorily fronter GOAT in situations where they feel more comfortable.

Figure 2: Effect of comfort level on GOAT F2 (normalized).



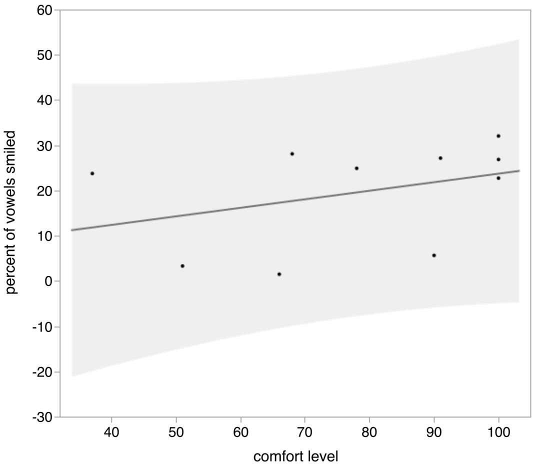
No other factors (including gender) were found to predict GOAT F2 in the regression model.

4. DISCUSSION

To summarize, we find robust effects of smiling, as GOAT is produced with higher F2 when speakers are smiling. One could argue that the effects of smiling are merely an inevitable acoustic consequence of the physical act of smiling. We reject this argument, however, given first that individuals are quite variable in whether and the extent to which smiling influences F2, and second our finding that speakers produce fronter variants of GOAT in interactions in which they feel more comfortable. Together, these findings suggest that fronting is motivated at least in part by the expression of affect.

An alternative interpretation of the comfort level effect is that there is a greater tendency to smile in situations where speakers feel comfortable (i.e., the comfort level effect is attributable to the smiling effect). To evaluate whether this is the case, we examined the potential correlation between the percentage of vowels smiled in an interaction and the speaker's reported comfort level in the same interaction. As shown in Figure 3, there is no clear relationship between how much speakers smile and their comfort level. While the slope of the trend line is consistent with the hypothesis under consideration, the confidence of prediction interval is wide, and the correlation does not reach statistical significance [F(1, 9) = 1.24, p < 0.299]. It appears, then, that the comfort level effect is not an epiphenomenon of the smiling effect.

Figure 3: Effect of comfort level on percent of vowels smiled in interaction.



It is worth noting that the effects of smiling and comfort level emerged when social factors like the gender of the speaker and interlocutor did not. While we caution against drawing conclusions about gender, given that we have only five representatives per gender group here, it is striking that interactional, affective factors have an observable effect in the current dataset.

5. CONCLUSION

In conclusion, this study shows that a form of embodied affect structures sociophonetic patterns. This effect raises several questions that should be addressed in future work. First, how does the strength of the smiling effect compare to that for social factors more commonly considered in sociophonetic work, like gender and age? The collection of a larger, socially diverse corpus in the interactional sociophonetics laboratory will enable the investigation of this question, and the methods here easily scale up to larger audiovisual corpora. Second, to what extent does smiling influence other sociophonetic variables? Our methods could be extended to examine any acoustically quantifiable variable (e.g., prosody, voice quality). Finally, can other forms of embodiment be straightforwardly coded using video analysis tools? As shown by Voigt et al. [7], the magnitude of a speaker's overall movement can be captured and correlated with prosody. Non-invasive methods for detecting the movement of specific joints can also be accomplished with video game technology, like the Xbox Kinect.

While several directions for future research remain, we can nonetheless draw two important conclusions. First, sociophonetic variation is an interactional practice rooted in the performance of affect. As Goodwin et al. [4] argue, emotion is best viewed as an interactional accomplishment rather than an inner psychological state. Operationalizing affect in interactional terms requires that, minimally, we collect information about how speakers experience interactions. One such assessment (comfort level) was shown to have a significant effect here. Second, sociophonetic variation takes place in the visual world. By limiting sociophonetic analysis to audio data, previous studies have been able to observe effects only for a small set of macrosocial categories for which speakers can be easily coded. While factors like gender and age strongly influence sociophonetic variation, the recordings in which the speech of women and men, young and old, have been analyzed are embedded in social interactions where affect is displayed. Technological advances in computer vision facilitate incorporating at least some dimensions of embodied affect into sociophonetic analysis, thus enabling us to improve our explanations of sociolinguistic variation.

6. REFERENCES

[1] Eckert, P. 2010. Affect, sound symbolism, and variation. *University of Pennsylvania Working Papers in Linguistics* 15.2, 70-80.

[2] Ekman, P., Friesen, W.V., Ancoli, S. 1980. Facial signs of emotional experience. *Journal of Personality and Social Psychology* 39.6, 1125-1134.

[3] Fagel, S. 2010. Effects of smiling on articulation: Lips, larynx, and acoustics. In: Esposito, A. et al. (eds), *COST 2102 Int. Training School 2009*. Berlin: Springer-Verlag, 294-303.

[4] Goodwin, M., Cekaite, A, Goodwin, C. 2012. Emotion as stance. In: Sorjonen, M., Perakyla, A. (eds), *Emotion in Interaction.* Oxford: Oxford University Press, 16-41.

[5] Labov, W., Ash, S., Boberg, C. 2006. *The Atlas of North American English*. New York: Mouton de Gruyter.

[6] Lobanov, B. 1971. Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49, 606–608.

[7] Viola, P., Jones, M. 2001 Rapid object detection using a boosted cascade of simple features. *Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1.

[8] Voigt, R., Podesva, R., Jurafsky, D. 2014. Speaker movement correlates with prosodic indicators of engagement. *Proceedings of Speech Prosody* 7.

[9] Wong, Amy Wing-mei. 2014. GOOSE-fronting among Chinese Americans in New York City. *University of Pennsylvania Working Papers in Linguistics* 20.2, 209-218.

[1] http://tla.mpi.nl/tools/tla-tools/elan

[2] http://fave.ling.upenn.edu

[3] http://www.fon.hum.uva.nl/praat

[4] http://github.com/hromi/SMILEsmileD

[5] http://opencv.org