The Voice Embodied: Bringing the Quantitative Analysis of Embodiment into the Study of Phonation

Robert J. Podesva
Stanford University

Patrick Callier
Lab41

Rob Voigt
Stanford University

Katherine Hilton
Stanford University

Body movement and facial expression predict F0 and use of creaky voice as strongly as established linguistic and social factors.

Introduction

Scholars of gesture and bodily hexis recognize the centrality of the body to speech (Bourdieu 1984, McNeill 1992, Kendon 1997).

The body as a predictor of variation

- Pitch accents often coincide with gestural apices (Mendoza-Denton & Jannedy 2011).
- Degree of body movement correlates with fundamental frequency (F0) and intensity variability at the phrase level (Voigt, Podesva & Jurafsky 2014).
- American English speakers produce fronter GOAT when smiling (Podesva, Callier, Voigt & Jurafsky 2015).
- Dutch speakers exhibit higher F2 (for /o:/) and intensity when smiling (Barthel & Quené 2015).

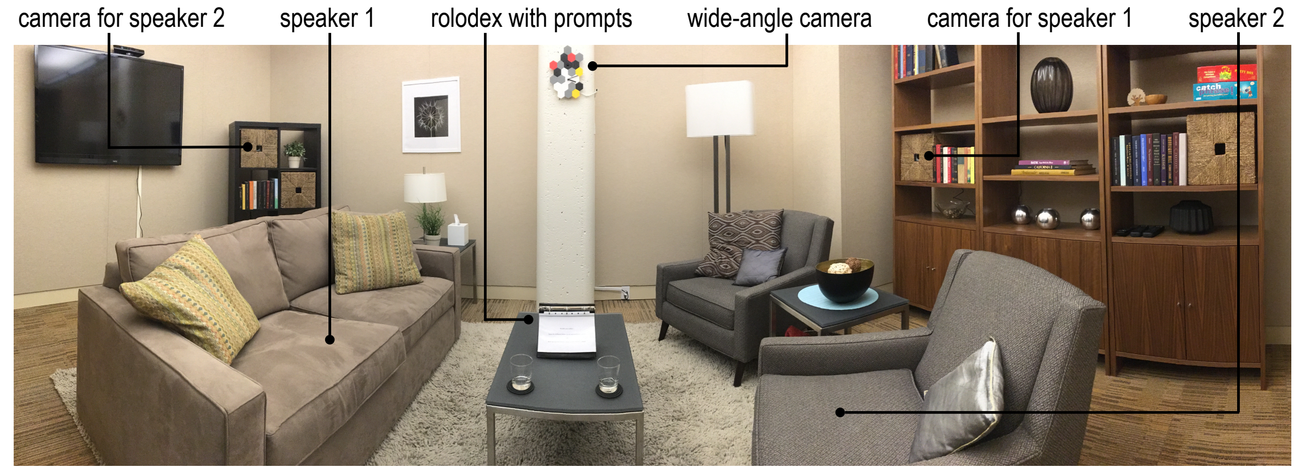
Hurdles to examining the body-variation connection

1. Large-scale analysis of body movement ➡ Using computer vision methods for body movement and smiling
2. Collection of high-quality audio-visual recordings in a relaxed environment ➡ on an audio-visual corpus of friendly interactions

we examine two voice features (**F0** and **creaky voice**).

Methods

Interactional Sociophonetics Laboratory

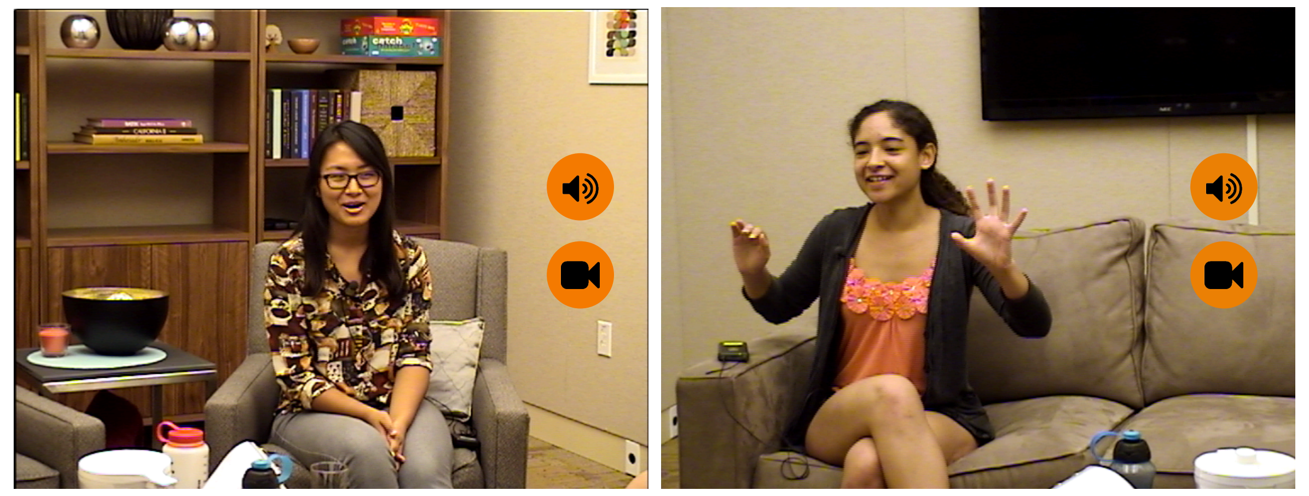


Acoustical specifications of sound booth, staged as living room

Recording: dyadic conversation (30 minutes) between familiars
Survey: demographic info; assessment of interaction, interactant

Sample (35 speakers, about 18 hours of recordings)

RELATIONSHIP	15 close friend, 11 friend, 6 partner, 3 family
SEX CLASS	21 female, 14 male
AGE	23 18-22, 8 23-29, 4 30+
ETHNICITY	18 white, 9 multiracial, 3 black, 2 Pacific Islander, 1 Asian, 1 Latin@ , 1 South Asian
REGION	19 West, 8 South, 5 Northeast, 2 Midwest, 1 Intl



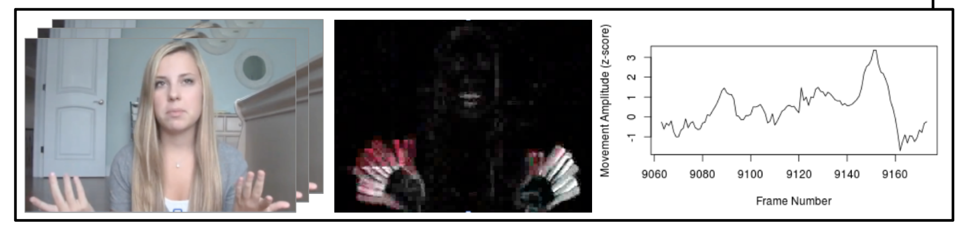
Separate audio and video recordings for each speaker

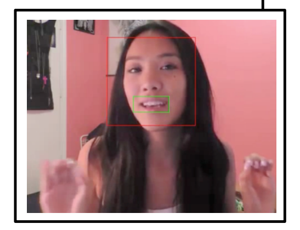
Acoustic analysis

- Transcriptions in ELAN (Lausberg & Sloetjes 2009)
- Forced alignments using FAVE (Rosenfelder et al. 2011)
- F0 measurements every 10 ms via Praat (Boersma & Weenink 2015) script, reduced to median value/vowel
- Each vowel classified as ±creaky using neural network model (Kane et al. 2013)

Computer vision analysis

- Each vowel coded as ±smiled using Haar cascade classifier trained on open source data (Podesva et al. 2015)
- Body movement amplitude (Voigt et al. 2014) based on frame-to-frame changes in pixel value





Mixed-effects regression models

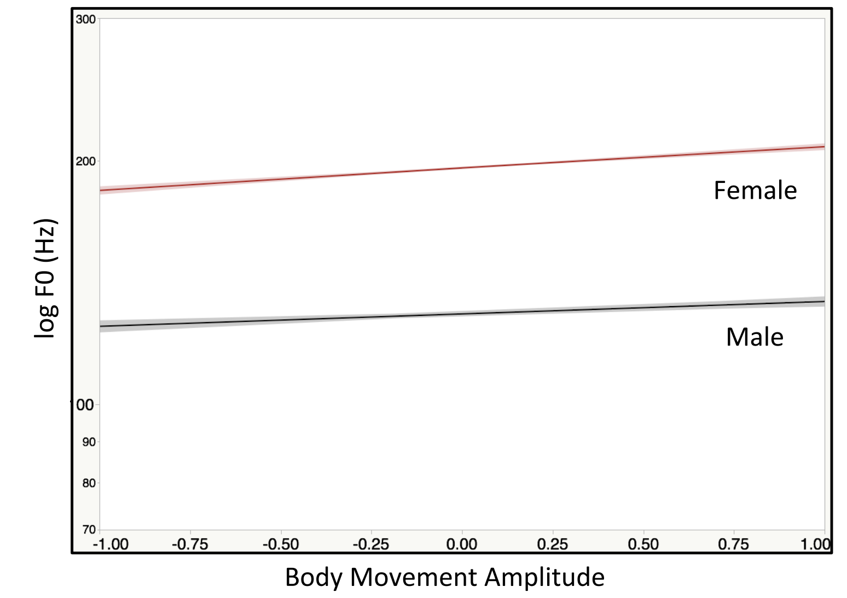
- Observation: vowel segment (N = 104,249)
- Response: log F0 (Hz) [linear model], +creaky [logistic model]
- Random intercepts: speaker, word, (pre/fol) segment
- Established factors: phrase position, segment and phrase duration, lexical stress, sex, age, ethnicity, region
- Embodiment factors: movement amplitude, smiling (at segment and phrase levels)
- Interactional factors: self-reported comfort, degree "clicked"

Fundamental Frequency

Term	Estimate	Std Error	DFDen	t Ratio	P-value
Intercept	5.034	0.0223	75.86	225.55	<.0001*
sex[F]	0.223	0.0176	32.83	12.67	<.0001*
phrase_position	-0.127	0.0047	95206	-27.12	<.0001*
movement_amplitude(phrase)	0.042	0.0041	95380	10.35	<.0001*
stress[secondary]	-0.023	0.0053	92967	-4.31	<.0001*
movement_amplitude(phrase)*sex[F]	0.019	0.0036	95374	5.22	<.0001*
smiling(phrase)[True]	0.017	0.0017	94584	-9.79	<.0001*
stress[primary]	0.015	0.0031	83114	4.97	<.0001*
smiling(segment)[True]	0.012	0.0020	95437	-6.14	<.0001*
segment_duration(log)	-0.006	0.0021	79912	-2.96	0.0031*
phrase_duration(log)	-0.006	0.0021	94117	-2.75	0.0060*
movement_amplitude(segment)	0.002	0.0019	95352	1.12	0.2625ns

Embodiment factors

- Speakers use higher F0 in phrases during which they move more, a pattern women exhibit more strongly than men.
- Speakers use higher F0 in phrases during which they smile.
- F0 is predicted by body movement and smiling at the phrase level, less strongly (smiling) or not at all (movement amplitude) at the segment level.



log F0 (Hz) as a function of the mean movement amplitude during the phrase in which the observation occurred, by sex

Established factors

- F0 is higher among women, at the beginnings of phrases, in syllables carrying primary stress, for shorter segments, in shorter phrases.

Discussion

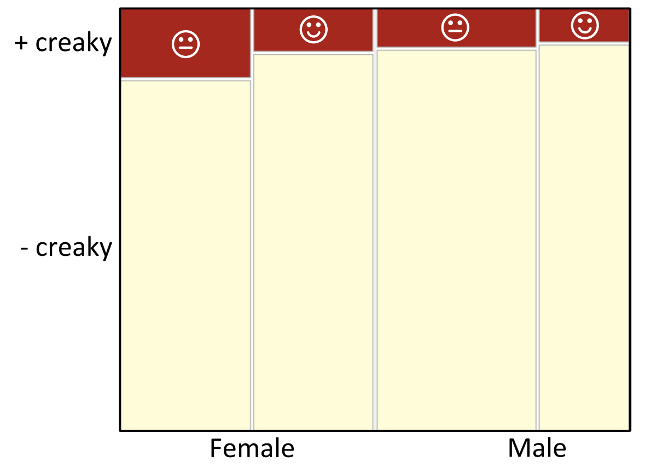
- Embodiment factors (both smiling and body movement) predict F0, and as strongly as established factors.
- Body movement and smiling appear to interface with the linguistic system at the phrasal (vs. segmental) level.

Creaky Voice

Term	Estimate	Std Error	ChiSquare	P-value
Intercept	-0.944	0.0535	311.03	<.0001*
phrase_position	1.435	0.0379	1435.2	<.0001*
segment_duration(log)	0.402	0.0138	851.27	<.0001*
sex[F]	0.203	0.0112	328.17	<.0001*
smiling(phrase)[True]	-0.198	0.0131	230.83	<.0001*
movement_amplitude(phrase)	-0.168	0.0309	29.64	<.0001*
phrase_duration(log)	-0.139	0.0144	93.12	<.0001*
smiling(phrase)[True]*Sex[F]	-0.125	0.0112	124.25	<.0001*
smiling(segment)[True]	0.040	0.0176	5.14	0.0234*
comfort_level(reported)	-0.008	0.0004	345.57	<.0001*
movement_amplitude(segment)	-0.008	0.0165	0.25	0.6164ns

Embodiment factors

- Women creak more in phrases during which they do not smile, driving a main effect of smiling. Men's use of creak is not influenced by whether they smile smile during the phrase.



Number of creaky observations for females vs. males, by whether the phrase in which the observation occurred contained a smiled vowel

- Speakers creak more in phrases during which they move less.
- Creak is predicted by smiling and body movement at the phrase level, less strongly (smiling) or not at all (movement amplitude) at the segment level.

Interactional factors

- Speakers creak more in interactions where they reported feeling less comfortable.

Established factors

- Creak is more common at the ends of phrases, for longer segments, in shorter phrases, and among women.

Discussion

- Embodiment factors as strong as established factors
- Creaky voice can convey negative affect (not smiling) and disengagement (less movement), resonating with claims that creak can distance speakers from what they take stances toward (Grivičić & Nilep 2004, Zimman 2014, Lee 2015).
- Embodiment has scope over the phrase, not the segment.

Conclusions

Importance of incorporating embodiment in variation analysis

- Body movement and facial expression constrain variation as strongly as well established linguistic and social factors.
- Different social groups exhibit different patterns of embodiment (Kendon 1997). These differences may underlie correlations between social category and linguistic variation.
- Focusing on body movement and facial expression may facilitate the operationalization of stance (Kiesling 2009) and affect (Eckert 2010). These types of social meaning may more directly drive variable language use in practice than social category membership.

Variationists should attend to embodiment. Speakers use their bodies in non-random ways to structure linguistic variation in all interactions, including those recorded and analyzed by sociolinguists.

Future directions

- Other forms of embodiment, computer vision technologies
- Additional variables (vowel quality, non-creaky phonation)
- Interactional factors (assessments of interaction, interactant)
- Larger, more diverse sample, interactions between strangers
- Role of embodiment in sound change

References

Barthel, Helen & Hugo Quené. 2015. Acoustic-phonetic properties of smiling revisited. *Proceedings of ICPhS 18.*
Boersma, Paul & David Weenink. 2015. Praat: Doing phonetics by computer. Version 5.4.
Bourdieu, Pierre. 1984. *Distinction*. Cambridge: Harvard University Press.
Eckert, Penelope. 2010. Affect, sound symbolism, and variation. *University of Pennsylvania Working Papers in Linguistics* 15.2: 70-80.
Grivičić, Tamara & Chad Nilep. 2004. When phonation matters: The use and function of *yeah* and creaky voice. *Colorado Research in Linguistics* 17.1: 1-11.
Kane, John, Thomas Drugman & Christer Gobl. Improved automatic detection of creak. *Computer Speech and Language* 27: 1028-1047.
Kendon, Adam. 1997. Gesture. *Annual Review of Anthropology* 26: 109-128.
Kiesling, Scott. 2009. Style as stance. In Alexandra Jaffe, ed. *Stance: Sociolinguistic Perspectives*. Oxford: Oxford University Press, pp. 171-194.
Lausberg, H. & H. Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers* 41.3: 841-849.
Lee, Sinae. 2015. Creaky voice as a phonational device marking parenthetical segments in talk. *Journal of Sociolinguistics* 19: 275-302.
McNeill, David. 1992. *Hand and Mind*. Chicago: University of Chicago Press.
Mendoza-Denton, Norma & Stefanie Jannedy. 2011. Semiotic layering through gesture and intonation. *Journal of English Linguistics* 39.3: 265-299.
Podesva, Robert J., Patrick Callier, Rob Voigt, & Dan Jurafsky. 2015. The connection between smiling and GOAT fronting. *Proceedings of ICPhS 18.*
Rosenfelder, Ingrid, Joe Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) program suite.
Voigt, Rob, Robert J. Podesva, and Dan Jurafsky. 2014. Speaker movement correlates with prosodic indicators of engagement. *Proceedings of Speech Prosody 7.*
Zimman, Lal. 2014. The larynx. Paper presented at the American Anthropological Association. Washington, DC.

This work was supported by a grant from the Roberta Bowman Denning Initiative in the Digital Humanities. Many thanks to members of the Interactional Sociophonetics Lab for feedback and assistance in research: Yesid Castro, Annette D'Onofrio, Sunwoo Jeong, Gabriella Johnson, Hollis Kool, Kyuwon Moon, Jamie O'Keefe, Kelsey Piper, Teresa Pratt, Anita Szakay, Robert Xu, and Lal Zimman.