Reconstructing the perception of gender identity, sexual orientation, and gender expression in American English

Benjamin Lang

University of California, San Diego
blang@ucsd.edu

ABSTRACT

Previous research has identified associations between acoustic cues, such as mean f0 and center of gravity (COG) of /s/, and a speaker's gender identity (GI), sexual orientation (SO), and gender expression (GE). This study investigates how acoustic features combine to influence the perception of GI, SO, and GE from speech. 197 listeners rated the speech of 66 speakers of American English along scales of GI, SO, and GE. Measurements of acoustic features such as f0, vowel formants, fricative spectra, and creaky voice were correlated with listener judgments using random forest and effect size analyses. Results indicate that well-known features such as f0 and COG of /s/, and less-studied ones like creaky voice and diphthong formants, matter for perception of GI, SO, and GE. Results further suggest that listeners reconstruct a speaker's identity along traditional binaries such as masculine and feminine, and outside those binaries for queer-sounding voices.

Keywords: sociophonetics, speech perception, gender, sexual orientation

1. INTRODUCTION

The current work seeks to investigate the ways in which multiple acoustic features combine to influence perception of speakers' gender identity (GI), sexual orientation (SO), and gender expression (GE). Sociophonetic studies of speech production have identified many correlates of GI, SO, and GE. For American English, these include vowel formants and dispersion [1–2], fundamental frequency (f0) [3–8], acoustic characteristics of voiceless alveolar fricative /s/ [4, 5, 9], and creaky voice [9–10]. This body of work suggests a strong association between certain acoustic cues and a speaker's identity.

During speech perception, studies have further shown that listeners can readily infer the sexual orientation of a speaker from acoustic cues [11–15], again presenting a similar association between acoustic cues and a speaker's (perceived and veridical) identity. Mack and Munson [14] report on one specific feature and its acoustic characteristics, /s/, which was associated with the perception of more homosexual-sounding speech, following a common cultural stereotype about "the gay lisp." Importantly, the effect was strong enough that the actual sexual orientation of the speakers did not matter; when heterosexual speakers' speech was modified to include /s/ with a higher acoustic center of gravity (COG) or as [θ], their speech was rated as more homosexual-sounding. Additionally, f0 has been shown to cue listeners' perceptions of gender identity, with higher f0 values associated with the perception of woman-like individuals and lower f0 values associated with the perception of man-like individuals [7–8]. Furthermore, the perception of gender broadly includes the perception of gender expression as well as gender identity. Similar to the perception of sexual orientation, a number of studies [3, 16–17] show that acoustic characteristics of f0 and /s/ influence the perception of a speaker's voice as more or less masculine-sounding or feminine-sounding, again regardless of the speaker's veridical gender identity.

Presently, there is a need to characterize the interaction of multiple acoustic features with the perception of a speaker's multidimensional social identity. Prior work has often focused on "one-to-one" associations between a single feature and a single dimension (for example, the association between /s/ and sexual orientation in homosexual men) [14]. These one-to-one associations can also be shared across multiple identities such that /s/ cues the perception of sexual orientation in both men [14] and women [5]. However, many-to-one associations are common in speech perception, and the combinations of cues may influence reconstruction of multifaceted identities. The current study seeks to investigate how a listener reconstructs a speaker's gender identity, sexual orientation, and gender expression from speech.

2. METHOD

2.1 Stimuli

Stimuli consisted of short excerpts drawn from interview-format podcasts from American media. Speakers were selected primarily according to their gender identity and sexual orientation, as drawn from as many public sources (articles, videos, profiles, etc.) of a speaker's identity as possible. Excerpts were

controlled for socially-meaningful lexical content. For example, the excerpt for one speaker was "learn more and subscribe."

2.2 Participants

Data was collected from 210 undergraduate students at UCSD via an online survey designed on Qualtrics and recruited from the UCSD undergraduate experimental subjects pool. Each participant self-identified as a native English speaker. All participants were required to use headphones in order to proceed with the survey. Data from 197 participants are reported in the Results section (13 participants were removed due to blank responses in the survey).

2.3 Procedure

Participants were instructed to listen to short speech excerpts of 66 speakers of American English and provide Likert scale judgments of the gender identity (GI), sexual orientation (SO), and gender expression (GE) of each speaker, followed by a demographic survey. Participants were given a practice phase with one speaker not included in analysis, and then completed the main block. All 3 Likert scale dimensions were presented simultaneously for each speaker on a 7-point continuous scale. For GI, the question was "I think this person identifies as…" and responses could be recorded as Male (1), non-binary, gender-non-conforming, genderfluid, etc. (4), and Female (7). For SO, the question was "Would this person seek relationships with people of only the same gender identity, any gender identity, or only an opposite gender identity?" and responses could be recorded as Only same gender (1), Any gender identity (4), and Only opposite gender identity (7). Finally, for GE, the question was "Does this person's voice sound more traditionally masculine or feminine?" and responses could be recorded as Masculine (1), Neither/Both (4), and Feminine (7). Scales were continuous such that participants could provide responses anywhere between the landmarks, e.g. 2.58 or 6.43.

2.4 Feature Extraction

Each speaker excerpt was resampled to 16kHz and automatically segmented via textless forced alignment using Charsiu [18], and segments were hand-corrected by the author. The following acoustic features were then extracted from each excerpt:

1. f0: mean fundamental frequency (f0), f0 standard deviation (std), f0 90th percentile, f0 10th percentile, f0 range, from across the entire utterance
2. Monophthongal Formants: mean, maximum, and minimum F1-F4 of /ɑ æ ʌ ɔ ə ɛ ɪ i ʊ u/
3. Diphthongal Formants: mean, maximum, and minimum F1-F4 for the first and last third of segments of /a͡ʊ a͡ɪ e͡ɪ o͡ʊ ɔ͡ɪ/
4. Speaker Vowel Duration: mean vowel duration across utterance
5. Individual Consonant and Vowel Durations: mean individual segment duration of all segments
6. Vowel Dispersion: distance of mean formant measure of /ɑ æ ʌ ɔ ə ɛ ɪ i ʊ u/ by speaker from group mean formant values
7. Fricative Spectral Features: spectral center of gravity (COG), standard deviation (SD), skew, intensity, duration, and kurtosis of /s z f v ʃ ʒ/
8. Creaky Voice: percentage of creaky voice across utterance

For feature groups 1-6, measurements were derived and calculated from VoiceSauce [19]. For feature group 7, measurements were derived from Praat scripting [20–21], and for feature group 8, measurements were derived via creaky voice detection from COVAREP [22]. In total, 370 individual acoustic features were identified. However, not all excerpts contained a given feature since they were extracted from spontaneous speech in podcasts. For example, only 38 out of 66 speaker excerpts contained an /s/ token. Any missing acoustic value was imputed from the mean of the distribution of that acoustic value across excerpts.

3. RESULTS

3.1 *K*-Means Clustering

Participant ratings of speakers were clustered using *k*-means [23], such that each cluster represents speakers whose GI, SO, and GE vary along similar dimensions in perception. Plotting distortions and using the 'elbow' method indicated that 5 clusters were the best fit for the model. Figure 1 shows each of the 5 clusters and their relative position along the three perceived dimensions: GI, SO, and GE. Cluster labels were assigned post-hoc, based on their relative position along the 3 dimensions, GI, SO, and GE. For example, in Figure 1, the cluster with the most man-like rating for GI, the most same gender rating for SO, and the most masculine-sounding, was categorized as SM, loosely representing the commonly-perceived perceptual category of straight men. The clusters are SM (13 speakers), QM (14 speakers), QE (9 speakers), QW (10 speakers), and SW (20 speakers).

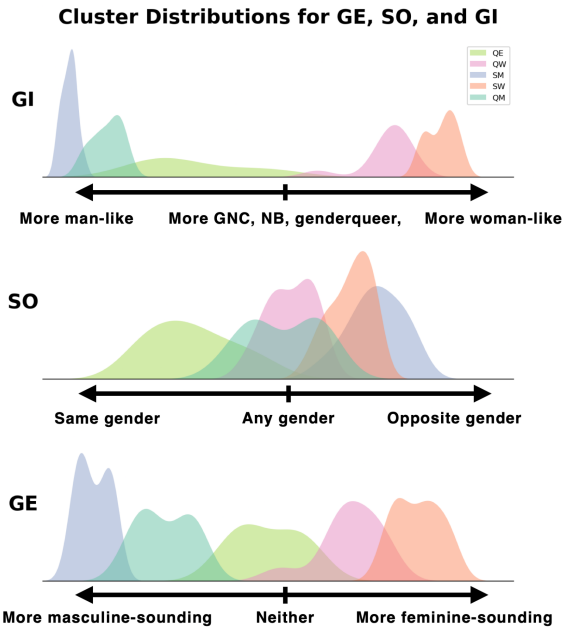

Figure 1: Perceived clusters of speakers according to GI, SO, and GE. Cluster labels are indicated for each of the individual distributions. The clusters are SM (13 speakers), QM (14 speakers), QE (9 speakers), QW (10 speakers), and QE (20 speakers).

Results of the perceptual rating task shown in Figure 1 indicate the emergence of 5 commonly-perceived categories: queer men (QM), queer women (QW), generally queer (non-binary) people (QE), straight men (SM), and straight women (SW).

3.2 Random Forest Analysis

Following clustering, a random forest classifier [23] was trained and tested on the 370 acoustic features with SM, QM, QE, QW, and SW as classes. The random forest classifier establishes the relative importance of the acoustic features for a given cluster, and thus provides a list of features that contribute to the perception of a certain class. The random forest classifier was accurate up to 100% for each cluster on train and test data, and performed at chance (20%) when clusters were randomly-assigned (rather than using the *k*-means clustering results). Top features for each cluster of the 370 features in the model are shown in Table 1. The reported features are truncated for brevity and due to the fact that relative importance values for features begin to plateau within the top 20 features for each cluster.

In an effort to reduce dimensionality further and investigate the apparent similarity in the GI and GE dimensions (see GI and GE in Figure 1), a Pearson correlation analysis [24] was conducted between the participant ratings along the two dimensions and revealed a significant strong correlation ($r = 0.85$, $p < 0.05$). Subsequent analyses are focused on GI and SO only.

Cluster	Features
SM	f0, Monophthongal Formants, Diphthongal Formants, /s/: COG
QM	Diphthongal Formants, Monophthongal Formants, Dispersion
QE	Creaky Voice, Vowel Duration, Monophthongal Formants
QW	f0, Dispersion, Monophthongal Formants, /s/: Duration
SW	f0, Monophthongal Formants, Diphthongal Formants

Table 1: Top features for each cluster via random forest classification.

3.3 Effect Size

In order to investigate the size of the effect of a given acoustic feature on the overall perceptual dimensions GI and SO, the effect sizes of each feature were measured for GI and SO. Linear regressions built with statsmodels [24] using the OLS method were conducted for each feature of interest and the two perceptual dimensions, GI and SO. For example, GI ~ mean f0 and SO ~ mean f0. The coefficient and t-values for select models are reported in Table 2 below.

Results for the effect size analysis show that an ensemble of acoustic features, some expected and others unexpected, contribute significantly to the perception of the GI and SO dimensions. These significant acoustic features were then compared to the 5 clusters to establish what feature or combination of features predict the perception of one cluster identity over another.

4. DISCUSSION

Summarized in Figure 2, certain features remain important to the perception of a single dimension. Replicating some of the results from studies like Mack and Munson [14], voices perceived as belonging to queer men and queer non-binary people can be distinguished from those of perceived straight men, supported by the acoustic characteristics of /s/. Similarly, perceived straight men and straight women's voices can be distinguished from each other on the basis of f0, replicating the effect of f0 on the perception of gender identity (GI) [6–8].

Unexpected features that emerged as important to the perception of identity include creaky voice and diphthongal formants. Drawing attention to the importance of creaky voice in perception as well as production [9–10], queer men, women, and non-

binary people are more easily distinguished from straight men and women via an increase in creaky voice across an entire utterance. Reflecting previous investigations into the idea that queer men share similar perceived speech styles with women [2–4, 12, 14], queer men are distinguished from straight men via an increase in diphthongal formant values, where increases in diphthongal formant values contribute to the perception of more woman-like characteristics along the perceived gender identity (GI) dimension; meanwhile, increases in diphthongal formant values contribute to distinguishing straight women from queer women.

Feature	GI coeff (t-value)	SO coeff (t-value)
f0	0.018 (108.896)	0.002 (12.875)
/ɪ/: F1	0.005 (40.092)	-
/ɪ/: F2	0.001 (48.109)	-
/ɪ/: F3	0.002 (65.992)	0.000 (8.803)
/ɪ/: F4	0.002 (65.78)	0.000 (6.492)
/a͡ɪ/: F1, 1st seg.	0.002 (36.764)	0.001 (10.597)
/a͡ɪ/: F2, 1st seg.	0.001 (24.569)	-
/a͡ɪ/: F3, 1st seg.	0.000 (3.647)	-
/a͡ɪ/: F4, 1st seg.	0.002 (43.149)	0.000 (3.294)
/a͡ɪ/: F1, 3rd seg.	0.002 (36.142)	0.001 (10.69)
/a͡ɪ/: F2, 3rd seg.	0.001 (25.318)	-
/a͡ɪ/: F3, 3rd seg.	0.000 (2.818)	-
/a͡ɪ/: F4, 3rd seg.	0.002 (43.192)	0.000 (3.283)
/ɪ/: F1 distance	0.002 (11.753)	0.000 (-2.715)
/ɪ/: F2 distance	0.001 (22.453)	0.000 (4.043)
/ɪ/: F3 distance	0.000 (5.683)	0.000 (5.672)
/ɪ/: F4 distance	0.000 (-2.867)	0.000 (11.289)
Avg. vowel dur.	-0.001 (-3.297)	-0.004 (-12.34)
/s/: COG	0.000 (34.096)	0.000 (-11.691)
/s/: skew	-0.494 (-32.476)	0.122 (10.006)
/s/: duration	-11.359 (-34.7)	-
Creak (%)	-0.008 (-6.577)	-0.017 (-18.95)

Table 2: Top features with significant effects on GI and SO. Mean feature values were used to compute models. Each cell contains the model coefficient and t-value. "-" indicates non-significance.

Regarding the perceptual dimensions, an asymmetry arises where the variation along perceived sexual orientation (SO) best explains the perception of the man-like individuals; smaller variation along perceived gender identity (GI) explains variation in woman-like individuals. Listeners are overall better at reconstructing different sexual orientations for men

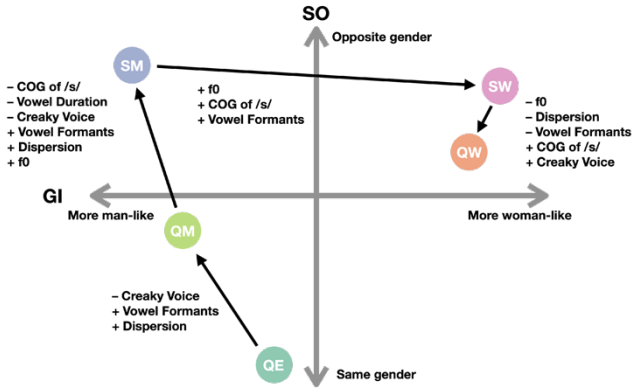

Figure 2: Acoustic features listed alongside arrows show the listener's reconstruction of GI and SO as perception travels between clusters. "+/-" indicate the direction of the effect on acoustic feature values. Vowel formants include both monophthongal and diphthongal formants.

than for women. Hazenburg [5] and Zimman [9] both report relatively constrained ranges of acoustic values produced by straight, man-like voices in comparison to queer men and women, and straight women. Any deviation from those ranges suggested the man-like individual is not straight, which aligns with the observed wider variation in perceived sexual orientation (SO) for man-like individuals. Additionally, Hazenburg [5] and Willis [15] show that any variation in acoustic values that would normally cue the perception of sexual orientation can be attenuated due to an interaction with gender identity. Thus, the more minute differences between the perception of straight and queer women are explained by the perception of their status as woman-like individuals along perceived gender identity (GI).

5. CONCLUSION

The purpose of this study was to investigate the multitude of acoustic features that contribute to a nuanced perception of gender identity, sexual orientation, and gender expression in speakers of American English. The reported associations between speaker identity and expected acoustic features provide further support for existing evidence on the perception of identity, while unexpected acoustic features such as creaky voice and diphthongal formants suggest that the perception of speaker GI, SO, and GE involves more than one acoustic feature. These results illustrate a multidimensional perceptual space in which listeners use many acoustic cues to construct multiple possible speaker identities along traditional binaries, but also create space for speaker identities outside of those binaries, particularly for queer-sounding voices.

6. REFERENCES

[1] J. B. Pierrehumbert, T. Bent, B. Munson, A. R. Bradlow, and J. M. Bailey, "The influence of sexual orientation on vowel production (L)," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 1905–1908, Oct. 2004, doi: 10.1121/1.1788729.

[2] R. Smyth and H. Rogers, "Do gay-sounding men speak like women?," *Toronto Working Papers in Linguistics*, vol. 27, p. 16, 2008.

[3] R. P. Gaudio, "Sounding Gay: Pitch Properties in the Speech of Gay and Straight Men," *American Speech*, vol. 69, no. 1, pp. 30–57, 1994, doi: 10.2307/455948.

[4] S. E. Linville, "Acoustic Correlates of Perceived versus Actual Sexual Orientation in Men's Speech," *Folia Phoniatr Logop*, vol. 50, no. 1, pp. 35–48, 1998, doi: 10.1159/000021447.

[5] E. Hazenberg, "Walking the straight and narrow: linguistic choice and gendered presentation," *GENL*, vol. 10, no. 2, pp. 270–294, Feb. 2015, doi: 10.1558/genl.v10i2.19812.

[6] N. Houle and S. V. Levi, "Acoustic differences between voiced and whispered speech in gender diverse speakers," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 4002–4013, Dec. 2020, doi: 10.1121/10.0002952.

[7] N. Houle and S. V. Levi, "Effect of Phonation on Perception of Femininity/Masculinity in Transgender and Cisgender Speakers," *Journal of Voice*, vol. 35, no. 3, p. 497.e23-497.e37, May 2021, doi: 10.1016/j.jvoice.2019.10.011.

[8] B. Merritt and T. Bent, "Revisiting the acoustics of speaker gender perception: A gender expansive perspective," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 484–499, Jan. 2022, doi: 10.1121/10.0009282.

[9] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men," *JLS*, vol. 2, no. 1, pp. 1–39, Feb. 2013, doi: 10.1075/jls.2.1.01zim.

[10] R. J. Podesva, "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," *Journal of Sociolinguistics*, vol. 11, no. 4, pp. 478–504, 2007, doi: 10.1111/j.1467-9841.2007.00334.x.

[11] R. Smyth, G. Jacobs, and H. Rogers, "Male voices and perceived sexual orientation: An experimental and theoretical approach," *Lang. Soc.*, vol. 32, no. 3, pp. 329–350, Jun. 2003, doi: 10.1017/S0047404503323024.

[12] E. Levon, "Hearing 'Gay': Prosody, Interpretation, and the Affective Judgments of Men's Speech," *American Speech*, vol. 81, no. 1, pp. 56–78, Feb. 2006, doi: 10.1215/00031283-2006-003.

[13] B. Munson, S. V. Jefferson, and E. C. McDonald, "The influence of perceived sexual orientation on fricative identification," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2427–2437, Apr. 2006, doi: 10.1121/1.2173521.

[14] S. Mack and B. Munson, "The influence of /s/ quality on ratings of men's sexual orientation: Explicit and implicit measures of the 'gay lisp' stereotype," *Journal of Phonetics*, vol. 40, no. 1, pp. 198–212, Jan. 2012, doi: 10.1016/j.wocn.2011.10.002.

[15] C. Willis, "Bisexuality and /s/ production," *Proc Ling Soc Amer*, vol. 6, no. 1, p. 69, Mar. 2021, doi: 10.3765/plsa.v6i1.4942.

[16] J. D. Avery and J. M. Liss, "Acoustic characteristics of less-masculine-sounding male speech," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3738–3748, Jun. 1996, doi: 10.1121/1.414970.

[17] B. Munson, "The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation," *Lang Speech*, vol. 50, no. 1, pp. 125–142, Mar. 2007, doi: 10.1177/00238309070500010601.

[18] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A Semi-supervised Approach," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[19] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A Program for Voice Analysis," *Hong Kong*, p. 5, 2011.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2022. [Online]. Available: www.praat.org

[21] C. DiCanio, "Spectral Moments of fricative spectra script in Praat." 2021.

[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.