Can Algorithms Hear Queerness?

Acoustic Classification of Sexual Orientation in Gen Z Male Speech

Jinfan Frank Hu · Yale University · frank.hu@yale.edu

NWAV 54
Montréal · Oct 22–24, 2026
Project Launch

Background

- **"Gay voice":** perception of men's sexual orientation from speech has received empirical attention since Gaudio (1994); acoustic correlates include sibilant frequency, pitch range and vowel quality (Smyth et al. 2003; Munson et al. 2006).
- **Binary framing:** this literature largely compares gay and straight men, rendering bisexual and pansexual speakers invisible, mirroring the "disbelonging" bi+ people report in both straight and queer communities (Yoshino 2000).
- **Open question:** listeners perceive orientation above chance, but the acoustic basis, and whether perceptions match actual identity, remain underexamined.

Previous research

- **Gaudio (1994):** 8 speakers read a monotone and a "dramatic" passage; 13 undergraduates judged them. Listeners were largely accurate, and F0 range (max/min, max/mean, max/median) correlated with gay–straight and masculine–feminine ratings in the .6s. The dramatic text was far stronger than the monotone one.
- **Travis (1981), Wolfe (1990):** men judged feminine had higher mean F0; in a multiple regression only mean F0 predicted masculinity/femininity, with more upward intonation.
- **Smith (1985):** listeners' masculinity/femininity ratings are reliable across listeners and match speakers' self-image, but sexual orientation is not considered and rejecting the masc–fem binary is not allowed for.
- **Rogers & Smyth (2003):** no significant difference in overall F0 mean or range. Distinctive cues are segmental: /s/ and /z/ duration, aspiration duration, /l/ fronting and vowel quality.
- **Jessee (2025):** 90 listeners heard the same voices, half told the speakers were transgender; the label shifted how sibilants were categorized, even more for listeners who had never talked to a trans person. Ideological expectation can shape what listeners "hear" (cf. McGurk).

Gap: every study above is gay/straight or masc/fem. Do these cues generalize to bi+ speakers, and does perception track who speakers actually are?

Phase 1 pilot (F24)

36	85	.964	0.11
speakers	listeners	ICC	listener d′

- **Design:** 36 Gen Z male speakers, who rated themselves on a spectrum of sexual attraction, read an identical 1-minute passage. 85 undergraduates rated each speaker's perceived sexuality from audio alone.
- **Listeners agree:** ICC(2,k) = .964. But mean rating vs. self-reported attraction: ρ = .277, p = .102 (n.s.).
- **PCA:** PC1 (31% of variance) tracks mean rating almost perfectly (r = .998) yet not self-report (ρ = .260, p = .126).

Takeaway: listeners share a coherent but inaccurate schema mapping men's voices to their attraction.

Phase 2: questions & predictions

1 What acoustic features drive listener ratings?

2 Do those features distinguish speakers by orientation: self-identification (gay, straight, bi) and self-reported attraction?

3 Do classifiers replicate, exceed, or diverge from the human baseline (mean d′ = 0.11 in Phase 1)?

Pre-registered: 7 /s/, vowel and diphthong measures (Group A) should move perceived and actual orientation together; 6 pitch and voice-quality measures (Group B) should diverge.

The bi+ question

22	15	13
Kinsey 1	Kinsey 2–4 (bi+)	Kinsey 5

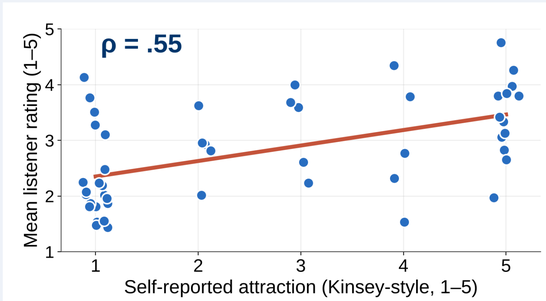
- Do features that predict gay/straight perception generalize to mid-scale speakers? Only 15 speakers sit at Kinsey 2–4.
- **Power:** n = 50 detects only ρ ≥ .39 (80% power); an actual-orientation effect near ρ ≈ .30 needs n ≈ 100.

Methodology

Record → **Align** → **Extract** → **Model**

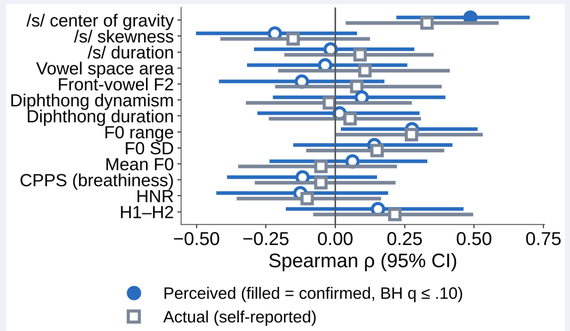
- **Speakers:** 50 Gen Z men (second, independent dataset) read the same scripted ~1-min passage ("So this morning, I woke up late and immediately stubbed my toe on the dresser…"). Self-reported attraction on a modified Kinsey scale (1 = women only … 5 = men only) plus self-identification.
- **Listeners:** 289 raters, 3,450 ratings (69 per speaker), 1 = sounds straight to 5 = sounds gay.
- **Align:** hand-transcribed, hand-checked Praat TextGrids; forced alignment (MFA) yields phone-level measures of /s/, vowels, diphthongs.
- **Extract:** 88 eGeMAPSv02 features via openSMILE (pitch, formants, spectral energy, voice quality) plus 13 pre-registered segmental measures.
- **Model:** Ridge and Elastic Net with leave-one-speaker-out CV against a 1,000-run permutation null; frozen WavLM layer probes, layer chosen inside CV (max-statistic null). Feature tests: Spearman ρ, Benjamini–Hochberg FDR.

Listeners track self-report



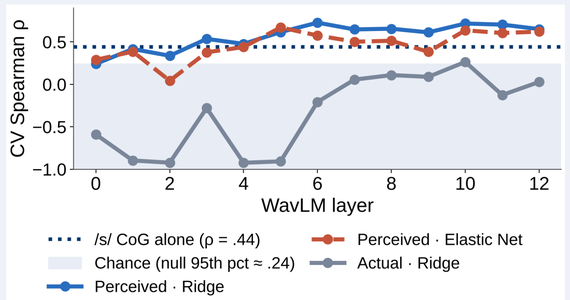
50 speakers, ρ = .55 (p < .001). A moderate association with wide scatter: several Kinsey-1 speakers are rated above 3 and several Kinsey-5 speakers below 3. Unlike the pilot (ρ = .28), the larger sample shows listeners partly tracking self-report.

Early result: /s/ fronting



/s/ center of gravity is the one confirmed cue for perceived orientation (ρ = .49, q = .002), not for self-report (ρ = .33, q = .12). No eGeMAPS feature survives FDR.

Early result: WavLM



Perceived is predictable (ridge L6 ρ = .73, p = .001) and beats /s/ alone (ρ = .44); **actual** is not (p = .14–.20).

Beyond /s/

- Dropping the /s/ block collapses perceived prediction; no other block does.
- WavLM's vowel readout tracks /aɪ/ F2 fronting (ρ = +.51) and /æ/ backing (ρ = −.50), independent of /s/ (its predictions share only R^2 = .11 with /s/ CoG).
- Perceived and actual formally diverge on 0 of 13 measures.

What we ask of NWAV

Feedback on the feature set and classifier architecture for a dataset this size, on the theoretical framing of algorithmic vs. human perception, and on generalizing to bi+ speakers.

Next steps

Audio augmentation to control for confounders; listener ratings on the new dataset; targeted extraction of sibilant frequency, vowel quality, /l/ frontness and aspiration duration; a larger, balanced sample.

Data, code & contact

github.com/jinfanfrankhu/perceived-gay-voice-study (F25)
Jinfan Frank Hu · Yale University · frank.hu@yale.edu

References: Gaudio 1994 · Jessee 2025 · Munson et al. 2006 · Rogers & Smyth 2003 · Smith 1985 · Smyth et al. 2003 · Travis 1981 · Wolfe 1990 · Yoshino 2000