TOWARDS A MODEL OF THE LISTENER IN PERCEPTION OF QUEER SPEECH

Cooper Bedin

University of California, Santa Barbara
cbedin@ucsb.edu

ABSTRACT

A fast-growing body of research is examining the relationship between the acoustic properties of a speaker's voice and sexual orientation ascribed to that speaker by listeners, however, these studies by and large only examine variation between speakers. In this paper I ask: what considerations are necessary to develop a model of how listeners integrate acoustic information to make sexuality judgements? This is done by conducting a perception experiment, and modeling sexuality judgements from two listeners of different sociodemographic backgrounds. I focus on modeling architecture, comparing exemplar models against logistic regression and random forest models. I find that the cognitively-motivated exemplar model best captures the behavior of each listener, indicating that future work on queer speech would benefit from engaging with cognitive computational modeling. Ultimately, a model of how listeners integrate acoustic information to make perceived sexuality judgements will contribute towards a more complete understanding of how queer speech is constructed.

Keywords: Sociophonetics, computational modeling, queer speech, cognitive modeling

1. BACKGROUND

A wide collection of experimental studies have drawn connections between acoustic variation in speakers' voices and a speaker's perceived or actual sexual orientation. Early work finds that listeners can be relatively "accurate" in distinguishing gay vs. heterosexual cisgender male speakers from audio alone [1], even with prompts as short as a single word [2]. However, the picture of how phonetic variation contributes to (perceived) sexual orientation remains murky. Take the phoneme /s/, which is one for which effects have been found in many studies. The way /s/ variation is operationalized is inconsistent, where any of center of gravity [3, 4, 5, 6], peak frequency [7], skewness [8, 4], duration [9, 10, 4], or a combination thereof may be analyzed. Additionally, for any one measure, one study may find a relationship between skewness of /s/ and sexual orientation [8], while another finds none [4]. Studies of other acoustic variables such as pitch have been even less definitive [4, 5].

Furthermore, research on sexuality and the voice often focuses on contrasting cisgender gay and heterosexual men. When researchers have examined other LGBTQ+ groups, such as bisexual [4] or transmasculine [5, 6] individuals, individual speakers have patterned with respect to /s/ in ways that are different than in literature on gay men, indicating that there are ways to sound queer beyond the dominant stereotype of the "gay lisp" [3, 11].

Little such attention has yet been paid to variation in listeners. Given that socially-grounded expectations of listeners inform phonetic perception [12, 13] and the interpretation of sociolinguistic variables [14], it would be reasonable to expect that listeners with different experiences perceive sexuality in the voice differently. Indeed, one study suggests that gay and straight speakers may respond to experimental sexuality judgement tasks differently [9], and another finds that listener variability significantly accounts for variability in elicited sexuality judgements [2].

So, a complete model of queer speech needs to account for complex sociophonetic patterns, a diversity of queer speech styles, and differential expectations of listeners. This paper focuses on what such a model—from the perspective of the listener—would look like.

2. EXEMPLAR MODELING

I argue that missing from current literature on sexuality and the voice is cognitively-motivated model of how listeners go from phonetic input to sexuality judgements. Thus, the central focus of this paper is to evaluate the effectiveness of different modeling strategies' abilities to account for the results of an experiment devised based on prior work [2, 7]. In particular, this paper evaluates exemplar models, which are foundational to a class of modern theories of speech perception and have

been applied to other domains in phonetics and phonology [15, 16, 17]. Exemplar theory posits that, when classifying, humans will compare a new stimulus with experiences, or "exemplars," already stored in memory. The more similar the new stimulus is to the exemplars of a category, the more likely a person will judge that the new stimulus is of that category. An exemplar model operationalizes this theory in computational form. For this paper, exemplar modeling is compared against logistic regression and random forests as baselines.

Exemplar modeling makes a different set of assumptions about its data than the other two models considered. In logistic regression, a change in an input variable linearly affects model output (e.g. higher center of gravity directly predicts a more gay-sounding voice). Random forest models are non-linear, but still require imposing the structure of a decision tree onto variables. Exemplar models, by contrast, directly compare new data to data stored in memory, which make them both non-linear and motivated by a class of theories about how humans do categorization.

Given the complex objective of this line of investigation, I take a highly zoomed-in approach to my data for this preliminary approach. I model the behavior of two individual listener subjects chosen for their vastly different sociodemographic backgrounds, and focus on acoustic variation in sibilant phonemes.

3. EXPERIMENTAL METHODOLOGY

Data was collected in an experiment [2, 7] consisting of two tasks, each of which involved a separate group of subjects: a speaker task, which generated audio to be used in a listener task.

3.1. Speaker task

14 speakers were recruited for this study, all graduate or undergraduate students 18–27 years old. Speakers self-identified as "male, masculine, masc, male-presenting, or male-coded" as a prerequisite to participating in the study. After completing the experimental task, speakers described their sexual orientation and gender in open-responses boxes—categorized gender and sexual orientation responses are given in Table 1. Speakers were intentionally recruited to represent a range of sexual orientations and gender identities, to include a diversity of queer voices; it is worth noting that only 2 speakers identify as straight or heterosexual.

Speakers were administered a list of 25 sentences selected from the MOCHA-TIMIT [18] corpus, and recorded reading each sentence aloud. Due to the COVID-19 pandemic, subjects recorded themselves in a quiet part of their home rather than in a lab. Speakers read through the list of sentences three times to ensure that there would be at least one quality recording per sentence per speaker, yielding 350 recordings (14 speakers $\times$ 25 sentences).

Gender	**Count**
Male	11
Non-binary or gender-nonconforming	3
Sexual orientation	**Count**
Straight or heterosexual	2
Gay	5
Bisexual or bi/pansexual	4
Queer	2
Demisexual/asexual panromantic	1

Table 1: Categorized gender and sexual orientation responses for speakers.

3.2. Listener task

Listeners were recruited and administered their task through Prolific [19]. After completing the experimental task, subjects filled out a multiple-choice survey on demographics and connections with the LGBTQ+ community. This paper focuses on modeling the behavior of two listeners—which will be referred to throughout as listeners A and B—who were chosen because they provided maximally different answers to many survey questions, especially those on exposure to LGBTQ+ communities. Answers to relevant questions on the survey are provided in Table 2.

For the listener task, the 350 speaker recordings were shuffled and divided into three lists of 116–117 recordings each. Each listener subject who participated heard one of these lists—listeners A and B both heard the same list (116 recordings total), and so provided judgements for the same recordings presented in the same order. For each recording a listener heard, they would provide an ordinal 1–7 rating for how "queer/gay" the reading sounded to them (1="not at all queer/gay," 7="very queer/gay"). Spearman's rho indicates that listener A and B's ratings in the generated dataset (described below) are significantly—although not strongly—correlated ($\rho = 0.323$, $p < 0.0001$), so there is variation in how they responded to the task, but their ratings are not entirely independently motivated.

Question	Listener A	Listener B
How old are you?	18–20	61–70
How do you describe yourself with respect to gender?	Male	Female
How do you describe yourself with respect to race and ethnicity?	White or European	White or European
How do you describe yourself with respect to sexual orientation?	Bisexual or Pansexual	Heterosexual or Straight
Do you self-identify as LGBTQ+?	Yes	No
How many of your close friends are LGBTQ+?	Most, more than half	Some, less than half
How frequently do you go to spaces that are explicitly LGBTQ+?	About once a month	Never or almost never
How accurately do you believe you are able to determine if someone is queer or gay, upon meeting them?	Very accurately	Not at all

Table 2: Selected survey responses for listeners A and B.

4. DATASET FORMATION

Existing research suggests that acoustic information from multiple segments and types of segments is simultaneously integrated to form judgements about perceived sexual orientation [2, 6, 14, 20]—this study focuses on the sibilants /s/, /z/, and /ʃ/. Sibilants were chosen because they are a relatively small class of phonemes within which acoustic variation is easily measured, and because there already exists extensive literature connecting /s/ to queer speech. Given that prior work has considered several different acoustic measures of variation, center of gravity, skewness, and duration were all measured and given as features to the models. Typically studies consider specifically /s/, however, /ʃ/ and /z/ were also considered for this study, which allowed for the ability to evaluate models' abilities to handle multiple phonemes, and to implicitly test whether future work should continue to focus on /s/.

TextGrids were generated for all recordings using the Montreal Forced Aligner [21], and alignments for all sibilants were then corrected manually. Acoustic measurements were collected via the Python library Parselmouth [22]. Any sibilant of duration $\leq$ 50ms was excluded from the dataset, in addition to any sibilant which occurred in a function word (in this data, the words 'is' and 'was'). This was done to exclude any sibilant whose articulation was highly reduced. This yielded a dataset of 176 sibilant tokens, and three features for every token (center of gravity, skewness, and duration). All three features in the dataset were normalized to a 0–1 range as a preprocessing step for the models.

Ratings given by listeners were simplified by converting them to binary category labels. To convert ratings to labels, all ratings above the listener's mean rating (listener A: 4.2, listener B: 3.5) were converted to YES, and all below to NO. Every sibilant token was matched with the label the listener assigned to the recording in which the token occurred. This yielded two alternative labelings of the same set of sibilant tokens.[1]

5. EVALUATING MODEL PERFORMANCE

All models were evaluated using the Scikit-learn package in Python [23]. Implementations of logistic regression and random forests were also taken from the Scikit-learn library, and exemplar models were implemented based on formulations used by Johnson [17]. Hyper-parameter tuning was done by five-fold cross validation,[2] and model performance was valued by average accuracy[3] across ten-fold cross validation. Models were trained, tuned, and tested separately on the labelings given by listeners A and B. Additionally, architectures were evaluated over four subsets of the data: the entirety of the data, and subsets composed of each individual phoneme (/s/, /z/, and /ʃ/). This yields 7 different scores for each architecture: 2 speakers $\times$ 4 different subsets of the data, minus the the /ʃ/ subset for listener B, because it only contained three YES labels, making it unsuitable for evaluation.

6. RESULTS

All scores are given in Table 3. The logistic architecture performs at or near chance for nearly all of the data subsets. The forest model fared worse, consistently performing well below chance. The exemplar architecture almost entirely outperformed the others, although confidence intervals indicate that there is considerable variance in its performance.

The results suggest that a purely linear treatment of this data is insufficient to capture the full story of listener behavior. With this in mind, consideration was also given to improving the linear model by preprocessing of the data. Model architecture was kept exactly the same, however, each feature was transformed by generating many univariate B-spline

Listener	Subset	# tokens	Chance	Logistic (Raw)	Logistic (Splined)	Forest	Exemplar
A	All	176	0.52	0.52 ± 0.01	0.55 ± 0.13	0.44 ± 0.10	$\mathbf{0.58 \pm 0.10}$
	/s/	81	0.53	0.53 ± 0.03	$\mathbf{0.64 \pm 0.09}$	0.42 ± 0.09	$\mathbf{0.64 \pm 0.16}$
	/z/	73	0.53	0.54 ± 0.09	0.53 ± 0.03	0.41 ± 0.07	$\mathbf{0.61 \pm 0.20}$
	/ʃ/	22	0.55	0.58 ± 0.13	0.59 ± 0.10	0.50 ± 0.07	$\mathbf{0.77 \pm 0.31}$
B	All	176	0.55	0.55 ± 0.01	0.58 ± 0.13	0.56 ± 0.12	$\mathbf{0.69 \pm 0.08}$
	/s/	81	0.53	0.53 ± 0.03	0.59 ± 0.14	0.48 ± 0.23	$\mathbf{0.68 \pm 0.10}$
	/z/	73	0.55	0.62 ± 0.07	0.63 ± 0.13	0.64 ± 0.13	$\mathbf{0.74 \pm 0.05}$

Table 3: Model performance across all data subsets, listeners, and architectures, with 95% confidence intervals. Also given are the number of tokens in each data subset, and expected performance for each subset if a model were to perform at chance. The score of the best performing model for each subset is bolded. "Logistic (Raw)" is the logistic architecture's performance on the data without additional preprocessing, and "Logistic (Splined)" is logistic architecture performance after projecting raw features onto spline bases.

bases for the feature, and projecting the raw features onto these bases. 51 cubic B-spline bases—50 knots—were generated for each of the 3 features, resulting in 53 basis features.[4] This gives the logistic architecture the ability to learn non-linear fits to the data, and have many more degrees of freedom. Average cross-validation scores are given in the "Logistic (Splined)" column in Table 3. For most data subsets an improvement in performance is seen with the transformed data. However, performance does not quite reach that of the exemplar model, and improvements are not seen on every data subset. This indicates that underlying patterns in the data are broadly non-linear, but also that scaffolding a linear architecture to perform as well as the exemplar model would require considerably more machinery.

7. DISCUSSION

In this study, several modeling architectures were compared for their ability to capture patterns in sexuality judgements given by listeners, based on acoustic information from sibilants. Given the central goals of this paper, discussion will focus on identifying relevant considerations for future iterations of modeling work.

Most importantly, treating the data non-linearly improved performance, and exemplar modeling performed best. This indicates that future work should continue to engage with cognitive science research and cognitive modeling architectures. Follow-up work should consider phonemes beyond sibilants—especially simultaneous information from multiple phonemes, and should consider the ordinal ratings given by listeners (instead of simplified binary labels) to create a more detailed model of listener responses. Additionally, while this paper did not investigate the parameters learned by models, such analysis will be crucial in future work.

Results from modeling for this study do not indicate that /s/ is the only sibilant phoneme with an effect on sexuality judgements—in fact, the best, most consistent performance across models is seen on the $\{$/z/, listener B$\}$ data subset. A similar jump is also seen over the $\{$/ʃ/, listener A$\}$ subset, however, the wide confidence interval make it difficult to draw as strong of conclusions. This suggests that effects previously observed on /s/ may extend at least as far as these two other phonemes, and encourages future work to consider how to parametrize information from many different phonemes and types of phonemes.

Lastly, it was observed that models were, across data subsets, able to fit the labelings given by listener B more accurately than listener A. In fact, only for listener B does the entire confidence interval of all exemplar model scores lie above chance performance. This suggests that, in fact, the two listeners did bring different expectations about queer speech to the given task. Table 2 indicates that listener B did not self-identify as LGBTQ+, where listener A did, and listener B also indicated less connection with the LGBTQ+ community than listener A. One possible explanation for why B's ratings were able to be better-described by the model could come from this (lack of) experience—if the ability of a model to fit data is used as a proxy for how "systematic" the data is, then listener A may be basing his judgements on a larger and more diverse set of experiences with queer and trans individuals, where listener B is referencing a more limited set of experiences that may be better-characterized by dominant stereotypes such as the "gay lisp" [3, 11]. This indicates the explanatory ability of modeling to explore differences between listeners, and invites expanding models to compare an arbitrary number of listeners.

8. ACKNOWLEDGEMENTS

I would like to thank Keith Johnson for his support as I collected the data for this project, and Simon Todd for his support on the writing of this paper, as well as the CPLS lab at UCSB for their feedback.

9. REFERENCES

[1] 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.

[2] E. C. Tracy, S. A. Bainter, and N. P. Satariano, "Judgments of self-identified gay and heterosexual male speakers: Which phonemes are most salient in determining sexual orientation?" *Journal of Phonetics*, vol. 52, pp. 13–25, Sep. 2015.

[3] J. Calder, "From sissy to sickening: The indexical landscape of /s/ in SoMa, San Francisco," *Journal of Linguistic Anthropology*, vol. 29, no. 3, pp. 332–358, 2019.

[4] C. Willis, "Bisexuality and /s/ production," in *Proceedings of the Linguistic Society of America*, vol. 6, no. 1, 2021, pp. 69–81.

[5] L. Zimman, "Hegemonic masculinity and the variability of gay-sounding speech," *Journal of language and sexuality.*, vol. 2, no. 1, 2013-2-27.

[6] ——, "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/," *Language in Society*, vol. 46, no. 3, pp. 339–370, Jun. 2017.

[7] K. Johnson and E. Tracy, "Acoustic and perceptual characteristics of vowels produced by self-identified gay and heterosexual male speakers," *The Journal of the Acoustical Society of America*, vol. 136, pp. 185–214, 2014, 10.1121/1.4899862.

[8] B. Munson, E. C. McDonald, N. L. DeBoe, and A. R. White, "The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech," *Journal of Phonetics*, vol. 34, no. 2, pp. 202–240, Apr. 2006.

[9] G. Jacobs, R. Smyth, and H. Rogers, "Language and sexuality: Searching for the phonetic correlates of gay- and straight-sounding male voices," *Toronto Working Papers in Linguistics*, vol. 18, Jan. 2000.

[10] E. Levon, "Sexuality in context: Variation and the sociolinguistic perception of identity," *Language in Society*, vol. 36, no. 04, p. 533, Oct. 2007.

[11] 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.

[12] 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, publisher: Acoustical Society of America.

[13] E. A. Strand and K. Johnson, "Gradient and visual speaker normalization in the perception of fricatives." in *KONVENS*, 1996, pp. 14–26.

[14] K. Campbell-Kibler, "Accent, (ING), and the social logic of listener perceptions," *American Speech*, vol. 82, no. 1, pp. 32–64, Feb. 2007.

[15] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency," *Frequency and the emergence of linguistic structure*, vol. 45, no. 137, pp. 10–1075, 2001.

[16] K. Johnson, "Speech perception without speaker normalisation," *Talker variability in speech processing*, pp. 145–65, 1997.

[17] ——, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of Phonetics*, vol. 34, no. 4, pp. 485–499, Oct. 2006.

[18] A. Wrench, "The MOCHA-TIMIT articulatory database," https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, 1999.

[19] "Prolific," https://www.prolific.co, 2014.

[20] K. Campbell-Kibler, "Intersecting variables and perceived sexual orientation in men," *American Speech*, vol. 86, no. 1, pp. 52–68, Feb. 2011.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[22] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[1] Ratings were bisected along the listener average, although one could instead have converted all ratings ≥ 5 to YES, ≤ 3 to NO, and dropped all 4's. Splitting along listener average was done instead because it preserves *all* tokens for analysis, without having to shrink the dataset; and it generates labelings that are balanced between YES and NO labels, reducing the risk that the balance of the dataset affects model performance.

[2] Cross-validation is a standard method for evaluating machine-learning models. In five-fold cross-validation, the data is divided into five equal parts, and for each fifth the model is trained on the other four-fifths of the data, and then tested for its ability to predict the labels of the remaining fifth. This allows one to test not just how well a model learned patterns in its training data, but also how well it can generalize to novel data.

[3] Simple accuracy was used, as all datasets were relatively balanced between YES and NO labels.

[4] Several bases were tested, with 2, 5, 10, 50, and 100 knots, and it was found that 50 knots struck the best balance against over- and under-fitting.