MODELING THE INFLUENCE OF CONFIDENCE IN SOCIAL CUES DURING SPEECH PERCEPTION USING GAUSSIAN MIXTURE MODELS

Eric Wilbanks

University of California, Berkeley
wilbanks_eric@berkeley.edu

ABSTRACT

Experimental results have demonstrated that listeners are sensitive to the co-variation of linguistic and social cues, and use this information to make predictions about speech based upon social information they attribute to the speaker. Such results have motivated models of speech perception in which listeners perform statistical inference over the joint distribution of phonetic and social cues in order to infer the speaker's intended utterance. However, initial explorations of these models have assumed that social cue information is known with certainty. I extend such models to situations in which the social cue information is subject to uncertainty and replicate qualitative patterns from the experimental literature. I model the confidence-weighted integration of social cue information using Gaussian mixture models and examine their predictions about how listeners should incorporate social cue information based on their level of certainty in those cues.

Keywords: Speech Perception, Variation, Statistical Inference, Sociophonetics

1. INTRODUCTION

In recent years, models of speech perception have begun to treat the variability of speech as a critical component of perceptual processes, rather than an unfortunate hindrance. Empirical evidence demonstrates that listeners can utilize the co-variation of social cues and phonetic features in order to make inferences about the intended linguistic content of an utterance (for reviews of such socio-phonetic phenomena, see [1, 3, 13]). Listeners' sensitivity to the structured variability of phonetic cues across speakers, contexts, and other social groups has led researchers to posit that listeners' statistical knowledge of joint distributions of phonetic cues and social cues represents a critical part of their phonetic knowledge. Such approaches often incorporate Bayesian inferential models in order to capture the inherent uncertainty and belief-updating nature of human speech perception (e.g., [8]).

Research is beginning to move towards quantitative explorations of such models through the computational implementation and then experimental validation of the predictions of such inference-based models of speech perception. For example, the ideal-adapter model outlined in [8] is found to show comparable adaptation to novel speakers as is found in experimental studies. A similar model is applied to social-contextual cues above the level of speaker in [9], in which it was found that such an inferential model can accurately predict the gender of novel speakers based on their phonetic cues, using knowledge of the joint distributions of phonetic and social cues. Similarly, [7] explores the ways in which an ideal observer might balance the competing forces of informativity and utility when determining the optimal level of specificity of social-cue distributions they should learn.

The behavior of such models has only been explored in situations in which the social cue information is known with absolute certainty. While in many interactions it can be assumed that social characteristics of speakers are known by listeners with high or complete certainty, this is not true of all interactions. There are many situations in which such social cues may be unknown to listeners, for example in interactions with an unknown speaker, or with a speaker whose social cue values may be non-stereotypical and subject to uncertainty.

There is empirical evidence to suspect that incorporating listeners' level of confidence in the social attributes of speakers is well-motivated. In a series of vowel identification tasks, [6] provide listeners with both stereotypical and non-stereotypical male and female voices. The categorization functions for non-stereotypical voices are found to skew systematically towards the opposite gender category boundary: non-stereotypical female voices elicit a category boundary closer to the "male" end of the continuum than stereotypical female voices. Under the assumption that non-stereotypical voices reflect greater degrees of uncertainty about speaker gender, such results appear to indicate that the influence of social contextual cues may be modulated by the degree of certainty the listener has in those cues. Similar results are found in [5], in which listeners had different category boundaries for tokens from a resynthesized fricative continuum for stereotypical male and female voices. Non-stereotypical male and female voices showed intermediate category boundary locations. Such behavior could indicate that an appropriate model would involve the gradient mixture of beliefs, allowing for variable confidence in the exact values of the social cues.

Further experiments have demonstrated that the strength of social cue priming, and perhaps listeners' degree of confidence in those cues, may be conditioned by the specific nature of the priming. [11] find that explicit priming of gender (through face guises) elicits a significantly stronger perceptual effect than does implicit priming of gender, as instantiated through the use of gendered carrier phrases in a grammaticality judgment task. These results suggest that implicit gendered carrier phrases provided less certain evidence of speaker gender than did explicit gendered face guises.

Drawing upon work in visual perception [10], I outline a framework for the incorporation of confidence-weighted social cue information in ideal-observer models. Then, I replicate the effect reported in [5]: gender information influences fricative categorization for all voices, but this effect is stronger for stereotypically gendered voices than for non-stereotypical voices. These results extend previous models of ideal-observer speech perception models by allowing the social cue information attributed to a speaker to be subject to some degree of uncertainty. The modeling approach outlined here provides a quantitative framework with which to make explicit predictions about the categorization behavior of listeners in situations in which social cue information about the speaker is uncertain.

2. BASIC MODELING ASSUMPTIONS

Following the work in the phonetic [8] and visual [2, 10] inference literature, I make the following modeling assumptions. During speech perception, an ideal-observer hears some token with a value, x, along some phonetic continuum (e.g., fricative with Center of Gravity (COG) of 5500Hz). Given this observation, they must determine the speaker's most likely intended phonemic category. This problem simplifies to calculating the posterior probability of that category given the observed value, following Bayes' Rule, as in (1). "Obs." indicates an observation of some acoustic value.

(1) $$\underbrace{p(Phone|Obs.)}_{posterior} = \frac{\overbrace{p(Obs.|Phone)}^{likelihood} \times \overbrace{p(Phone)}^{prior}}{\underbrace{p(Obs.)}_{normalizing\ constant}}$$

Comparing the posterior probability under different possible intended categories will allow the listener to determine which category is most probable given the evidence and their prior beliefs. Let's consider the case of an English listener who observes a coronal fricative with a COG of 5500Hz. If we assume that the likelihood functions $p(x|\text{/s/})$ and $p(x|\text{/ʃ/})$ are known, then the ideal-listener may determine the most probable category by comparing the posterior probabilities of each category based on the observed data point, as shown in Equation 2.

(2)

$$\underbrace{p(s|5500Hz)}_{posterior} = \frac{\overbrace{p(5500Hz|s)}^{likelihood} \times \overbrace{p(s)}^{prior}}{\underbrace{p(5500Hz|\text{ʃ})p(\text{ʃ}) + p(5500Hz|s)p(s)}_{normalizing\ constant}}$$

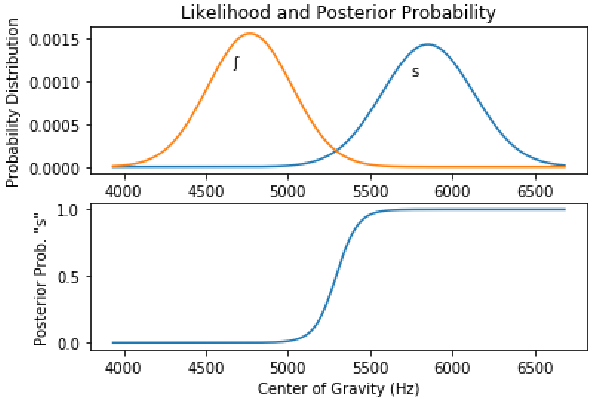
Figure 1 shows the relationship between the likelihood functions and posterior probability of /s/ for values of COG between 3000-8000Hz, assuming a uniform prior. As previously noted by [8], ideal-observer models comparing two categories generate posterior probability curves that approximate categorical perception identification curves in experimental tasks.

Additionally, I make the assumption that the likelihood functions of phonetic categories are normal distributions across phonetic cues. This assumption is not critical to the modeling framework (in theory, any probability distribution may be used), but rather a computational convenience.

3. CONFIDENCE-BASED CUE INTEGRATION

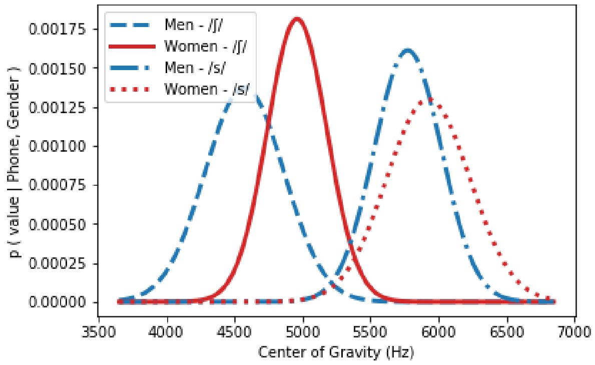
The precise nature of the likelihood function (e.g., what is an appropriate model of /s/ COG) varies according to many linguistic and non-linguistic factors. I choose to represent how social cues (e.g., gender, age, region of origin) condition phonetic variability through a Gaussian mixture model, decomposing the likelihood function by marginalizing across the various values of each social cue. For example, while English listeners will have experience with the distribution of appropriate values of COG for /s/, they will also have knowledge about

Figure 1: Relationship between likelihood functions and posterior probability of various COG values assuming uniform priors ($p(/\text{s}/) = p(/\text{ʃ}/) = 0.5$). Likelihood functions are normal distributions: /ʃ/ $\sim \mathcal{N}(4768.5, 255.25)$, /s/ $\sim \mathcal{N}(5851.25, 277.5)$. Acoustic values calculated from [12].



the structured variability of /s/ COG conditioned by gender, as shown in the reproduction of values from [12] in Figure 2.

Figure 2: COG Distributions for English /s/ and /ʃ/ for Men and Women. Values are from [12].

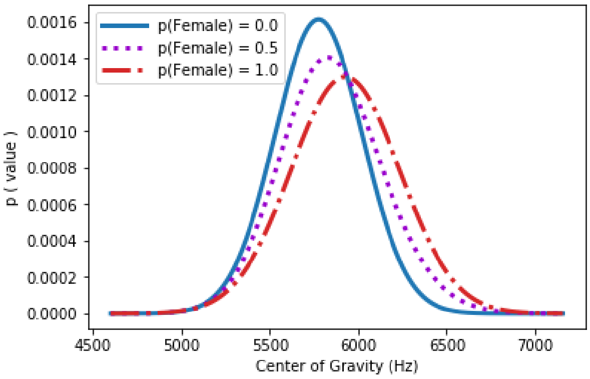


Following [10], I model the overall likelihood function for the social cue, $C = (value_1, value_2, ...value_c)$, as an additive mixture of likelihood functions for each value of the social cue, $p(x|C = c, Phone)$, weighted by the probability that the given value applies to the speaker (π_c, also referred to as *mixing proportions*).

$$(3)\quad \overbrace{p(x|Phone)}^{likelihood} = \Sigma_C \pi_c p(x|C = c, Phone)$$
$$= \Sigma_C \pi_c \mathcal{N}(\mu_c, \sigma_c^2)$$

Assuming that these likelihood functions are normal distributions with means μ_c and variances σ_c^2 leads to the formulation on the second line of (3) as a sum of the confidence-weighted normal distributions. For example, if we use the values from [12] in Figure 2 as our likelihood functions for men and women, we can model the likelihood function $p(x|/\text{s}/)$ for situations in which the listener has varying levels of confidence in the gender of the speaker. These confidence-weighted likelihood functions are presented in Figure 3. In situations in which the listener has complete confidence in the gender of the speaker (e.g., $p(\text{Female}) \in \{0, 1\}$), the resulting mixture reduces to the base likelihood functions for just men or just women.

Figure 3: Weighted mixture likelihood function for /s/ under different levels of confidence in speaker gender



4. APPLICATION

I turn now to a demonstration of how the confidence-weighted mixture approach may be used to model changes in behavioral measures in perception experiments. While much experimental work has been devoted to demonstrating that social information can influence perceptual behavior, less work has explored changes in this influence under conditions of greater or lesser certainty. An exception is the fricative perception experiment of [5]. In this experiment, listeners shifted their categorization of a synthetic /s/ - /ʃ/ continuum according to the gender of the speaker used to frame the synthesized token. Tokens with lower spectral energy were more likely to be categorized as instances of /s/ when spliced into the speech of a male than when they occurred in the female frame. This corresponds to listeners' knowledge that men's productions of /s/ are more likely to have lower frequency components than women's productions of /s/.

Critically, the authors demonstrate that this socially-driven perception effect is conditioned by the gender stereotypicality of the voices involved. A gender effect was also obtained for male and female voices judged to be non-stereotypical, but this effect was much weaker than the effect obtained for the stereotypical male and female voices. The categorization functions for the non-stereotypical male and female voices were intermediate between the stereotypical male and female categorization functions.

Using the model outlined above, I replicate the qualitative pattern of categorization functions observed in [5]. To do so, I provide simulated confidence values for the stereotypical and non-stereotypical voices in Table 1. These confidence values, or mixing proportions, may be interpreted as the relative confidence listeners have in the gender of the speaker (π_c in Equation 3). Without direct access to the quantitative results of [5], the exact values of the mixing proportions cannot be estimated. The values presented in Table 1 were chosen because they replicate the qualitative pattern reported in [5]. Slight adjustments to these mixing proportions would not result in qualitative shifts, but rather a more accurate fit to the behavioral data.

Table 1: Simulated confidence parameters for various speakers. $\pi_{male} = 1 - \pi_{female}$.

Speaker	π_{female}
Stereotypical Male	0.0
Non-Stereotypical Male	0.3
Non-Stereotypical Female	0.6
Stereotypical Female	1.0

The full form of the calculation of the posterior probability for the gender confidence /s/-/ʃ/ scenario of [5] is shown in (4). Note that I have assumed equal priors for /s/ and /ʃ/. The likelihood functions for men, $p(x|/s/,m)$ and women, $p(x|/s/,f)$, are normal distributions with means and variances equal to those reported in [12] and repeated in Figure 2.

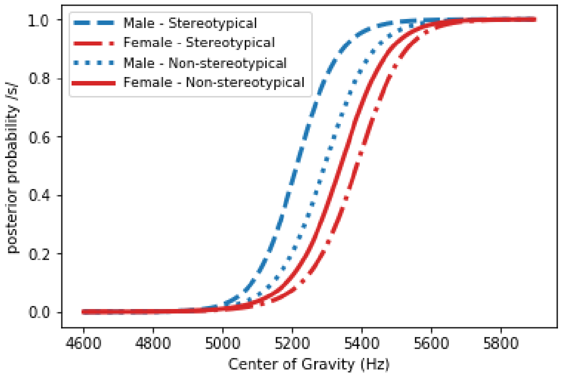
(4)

$$\overbrace{p(/s/|x)}^{posterior} = \frac{\overbrace{p(x|/s/)}^{likelihood} \times \overbrace{p(/s/)}^{prior}}{\underbrace{p(x)}_{normalizing\ constant}}$$

$$= \frac{\overbrace{(\pi_f p(x|/s/,f) + \pi_m p(x|/s/,m))}^{likelihood\ (decomposed)} \times \overbrace{p(/s/)}^{prior}}{\underbrace{p(x|/s/)p(/s/) + p(x|/\text{ʃ}/)p(/\text{ʃ}/)}_{normalizing\ constant}}$$

The posterior probabilities of /s/ under the four different speaker conditions are shown in Figure 4. We successfully replicate the qualitative results of [5]. The gender effect is weaker, but present for the non-stereotypical voices, and the model exhibits categorical perception type behavior in all cases. The response of listeners to stereotypical and non-stereotypical speakers can be successfully modeled as confidence-weighted cue-integration mixtures.

Figure 4: Posterior probability of /s/ under different speaker conditions.



5. CONCLUSION

There is evidence to suggest that listeners may gradiently incorporate information about socio-phonetic cue distributions based on their certainty in the social characteristics of the speakers. Using values from a corpus phonetic study [12], I extend the ideal-observer model outlined in [8] and make predictions about the changes in listener classification behavior of /s/ based on varying levels of certainty in speaker gender. In doing so, I replicate the qualitative findings of gender voice typicality conditioned shifts in /s/-/ʃ/ categorization behavior as reported in [5]. These results extend previous Bayesian inference models in speech perception to more experimental conditions. Additionally, this formulation of ideal-observer models provides a quantitative framework with which to make predictions about how listeners should incorporate gradient beliefs about social cues if they are carrying out optimal statistical inference. For example, such a framework allows us to make predictions about how overlapping social cues (e.g., speaker's age and region of origin) may interact to influence perceptual behavior. Going beyond macro-social variables, Bayesian-inference based approaches to speech perception provide a framework with which to make predictions about how listeners learn what social cues are salient in the first place (see, e.g., the discussion of Structural Learning in [4]). Understanding the ways in which listeners develop and incorporate social cues during speech perception is a critical step in extending our knowledge of sociophonetic processes in general.

6. ACKNOWLEDGEMENTS

I would like to express my gratitude to Keith Johnson, Susan Lin, and Terry Regier for their feedback and fruitful discussions on this project. Additionally, many thanks to Kathleen Gao for her assistance in generating the visualizations in this paper.

7. REFERENCES

[1] Drager, K. 2010. Sensitivity to grammatical and sociophonetic variability in perception. *Laboratory Phonology* 1, 93–120.

[2] Ernst, M. O., Banks, M. S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870), 429–433.

[3] Foulkes, P., Hay, J. B. 2015. The emergence of sociophonetic structure. In: MacWhinney, B., O'Grady, W., (eds), *The handbook of language emergence*. John Wiley & Sons 292–313.

[4] Jacob, R. A., Kruschke, J. K. 2011. Bayesian learning theory applied to human cognition. *WIREs Cognitive Science*.

[5] Johnson, K., Strand, E. 1996. Gradient and visual speaker normalization in the perception of fricatives. In: Gibbon, D., (ed), *Natural language processing and speech technology: results of the 3rd KONVENS conference*. Mouton de Gruyter 14–26.

[6] Johnson, K., Strand, E., D'Imperio, M. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27(4), 359–384.

[7] Kleinschmidt, D. F. frth. Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience*.

[8] Kleinschmidt, D. F., Jaeger, T. F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122(2), 148–203.

[9] Kleinschmidt, D. F., Weatherholtz, K., Jaeger, T. F. frth. Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*.

[10] Knill, D. C. 2007. Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision* 7.7(5), 1–24.

[11] Munson, B., Ryherd, K., Kemper, S. 2017. Implicit and explicit gender priming in english lingual sibilant fricative perception. *Linguistics* 55(5), 1073–1107.

[12] Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., Guenther, F. H. 2004. The distinctiveness of speakers' /s/-/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research* 47, 1259–1269.

[13] Sumner, M., Kim, S. K., King, E., McGowan, K. B. The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology* 4(1015), 1–15.