Tyler Kendall* and Charlotte Vaughn

Exploring vowel formant estimation through simulation-based techniques

https://doi.org/10.1515/lingvan-2018-0060
Received March 7, 2019; accepted June 12, 2019

Abstract: This paper contributes insight into the sources of variability in vowel formant estimation, a major analytic activity in sociophonetics, by reviewing the outcomes of two simulations that manipulated the settings used for linear predictive coding (LPC)-based vowel formant estimation. Simulation 1 explores the range of frequency differences obtained when minor adjustments are made to LPC settings, and measurement timepoints around the settings used by trained analysts, in order to determine the range of variability that should be expected in sociophonetic vowel studies. Simulation 2 examines the variability that emerges when LPC settings are varied combinatorially around constant default settings, rather than settings set by trained analysts. The impacts of different LPC settings are discussed as a way of demonstrating the inherent properties of LPC-based formant estimation. This work suggests that differences more fine-grained than about 10 Hz in F1 and 15–20 Hz in F2 are within the range of LPC-based formant estimation variability.

Keywords: Sociophonetics; vowels; formant estimation; linear predictive coding; methods

1 Introduction

It is regularly the case in sociophonetic research that small but statistically significant differences in vowel formant values are taken as indicating important differences between groups of speakers (e.g. from different regions or points in time) or between different contexts (e.g. casual vs. careful speech). However, sociophonetic research as a whole has not paid careful attention to the question of how to assess the meaningfulness of small formant value differences. In this paper, we propose that one useful practice for assessing the potential meaningfulness of vowel differences is to consider findings in relation to expected ranges of imprecision in the vowel measurement process. To this end, we explore in some detail one major source of variability in vowel measurements, the formant estimation process itself. We expand on a prior pilot project (Kendall and Vaughn 2015; hereafter K&V) and apply simulation methods to extrapolate what range of formant measurement variability should be expected from "reasonable" to "less reasonable" analysts.

Most current methods for vowel analysis, including the formant measurement features in the predominant software tool Praat (Boersma and Weenink 2019), rely on a few basic underlying technologies, including linear predictive coding (LPC; Atal and Hanauer 1971; Markel and Gray 1976), a technique which has known limitations (e.g. Vallabha and Tuller 2002).[1] Although formant measurement using LPC is always estimation, in practice formant values are often treated and reported as if they were precise facts about vowels, down to a fraction of 1 Hz. Thus, although prior work – mostly in the domains of forensic phonetics (Harrison 2004, Harrison 2013; Duckworth et al. 2011; Zhang et al. 2013) and acoustic phonetics (e.g. Atal and Hanauer 1971; Vallabha and Tuller 2002) – has noted the existence of variability introduced by LPC analysis, and most analysts are likely aware that spurious formant estimation can come about in LPC, in this paper we demonstrate

1 LPC is a parametric, model-based signal processing procedure. Its limitations lie in the fact that it requires parameterization (i.e. settings) that corresponds well to the qualities of the signal for accurate results (see Atal and Hanauer 1971; O'Shaughnessy 1988; Vallabha and Tuller 2002).

***Corresponding author: Tyler Kendall,** University of Oregon, Department of Linguistics, Eugene, OR, USA, E-mail: tsk@uoregon.edu. https://orcid.org/0000-0002-0989-1765
Charlotte Vaughn: University of Oregon, Department of Linguistics, Eugene, OR, USA. https://orcid.org/0000-0003-0528-6899

how these known concerns can impact sociophonetic work, and more generally seek to raise awareness about this important methodological consideration for the field of sociophonetics.

In practical terms, when an analyst uses a system reliant on LPC to measure formant frequencies for a vowel, they must answer a relatively small set of important questions: Where exactly is the vowel (i.e. what are its bounds) and at what timepoint(s) should it be measured? And, what LPC settings should be used? Following our prior work (K&V), in this paper we continue to focus on the two LPC settings most commonly manipulated by analysts: the number of LPC poles and the maximum formant frequency (set by Praat users in Praat's "Formant Settings…" menu). Our first simulation also manipulates the time point of measurements, although we focus less in this paper on outcomes resulting from timepoint because the impact of timepoint choice on formant measurements is more regularly discussed by sociophoneticians (though we note that vowel dynamics are important for a wider range of vowels than has traditionally been assumed, see Farrington et al. 2018). While the manipulation of other LPC settings is possible in Praat (e.g. time step, window length, preemphasis), we do not consider them here. The number of LPC poles (which we discuss as number of formants (NumF), which is equal to half the number of LPC poles) and maximum frequency (MaxHz) together are generally accepted as crucial adjustments in the formant estimation procedure, and are recommended to be made on a per-speaker and per-token basis (Vallabha and Tuller 2002; Di Paolo et al. 2011; Thomas 2011; Harrison 2013).

Lower MaxHz values literally reduce the upper range of the LPC analysis. NumF controls the number of formants which LPC returns. Fewer formants searched for over a wider Hz range will result in more spread-out formant estimates, likely leading to higher formant estimates. More formants searched for over a smaller Hz range will result in more compressed formant values and will also lead to the LPC analysis providing values for "false formants", i.e. intermediate regions of the spectrum between two real formants. The influence of LPC settings on formant estimates are known to differ for different formants and different vowels. This has to do with factors like the relationship between F0 and its harmonics and vocal tract resonances, and formant bandwidths (see e.g. Vallabha and Tuller 2002; Harrison 2013).

In recent years, new computational tools have increasingly enabled large-scale vowel analyses. For instance, tools like FAVE (see Labov et al. 2013; Rosenfelder et al. 2014), DARLA (Reddy and Stanford 2015), and ISCAN (McAuliffe et al. 2019) make analysis of huge amounts of vowel data accessible without the researcher needing knowledge of the principles behind LPC. Thus, today's tech-savvy sociophonetician may argue that, with the advent of these automated methods, an understanding of LPC settings is obsolete. We would contend that knowledge of LPC and its assumptions is important for *any* analyst, whether measuring individual tokens by hand, with a Praat script, or using an automated tool. We note that the primary difficulty that automatic vowel extraction programs in general use (like FAVE) attempt to "solve" is in fact the precise issue of determining which LPC settings should be used for a given token for a given speaker. Thus, in terms of actually measuring (i.e. estimating) vowel formant values for a given timepoint in a given vowel, all variability, unreliability, and inaccuracy – whether the measurement is by a human or by an automated system – arise from the parameterization of LPC.[2]

As mentioned, much of the existing work investigating the reliability of LPC formant estimation is based in forensic phonetics. In such settings, formant measurements are often used for forensic speaker comparison (see Nolan and Grigoras 2005; Morrison 2008; Watt 2010), with high-stakes outcomes that have naturally led to serious consideration of analytic reliability. This work has not had as much uptake in sociophonetics as it should have. Duckworth et al. (2011) and Zhang et al. (2013) each provide rare close examinations of the formant estimates obtained by different researchers for the same speech samples. Duckworth et al. (2011), for instance, tested three different research laboratories' formant measurements before and after researchers agreed on a shared set of procedures. Their results indicate that even when procedures were shared, differences between labs were as high as about 80 Hz in F1 and 100 Hz in F2. Harrison's (2004, 2013) work is of

2 Non-LPC-based approaches to vowel formant estimation have been explored, increasingly incorporating alternative approaches to signal processing and techniques from areas like deep learning (see Dissen et al. 2019). However, until such techniques are widely available to analysts, LPC remains the state-of-the-art.

direct relevance to the present paper, and we refer readers to it for additional prior literature and comprehensive analyses of how formant analysis procedures across software tools perform on both real and synthesized (where the "true" formant values are prespecified) vowel tokens. Harrison finds extensive support for the importance of parameterizing LPC by-speaker and by-token, noting that "unfortunately, the speaker attributes of sex, mean fundamental frequency and vowel space position do not provide strong indicators for suitable LPC orders [i.e. poles]" (2013: 227). But, although he finds it is not possible to predict the "right" settings based on characteristics of the vowels or speakers, he also identifies systematic biases in LPC outcomes, such as patterned errors in estimation around the vowel space (something we return to in §3.1).

However, even within the works just discussed, few guidelines are available regarding the range of error we should expect among "reasonable" (i.e. well-trained) sociophonetic analysts, and thus what magnitude of differences between groups of interest should be considered meaningful.

1.1 The present paper

In K&V, we applied bootstrap simulation techniques to examine how slight differences from "gold standard" settings (i.e. those selected by a trained human analyst) affected formant estimations for four speakers, in terms of three factors: measurement timepoint and two LPC parameters (NumF and MaxHz). Our analysis, comparing simulation results to the measurements by the trained human analyst, suggested that small frequency differences (about ∼6 Hz in F1 and ∼14 Hz in F2) should be expected as reasonable fluctuations due to differences in measurement techniques that are still within the range of reasonable best practices, and therefore should not be interpreted as meaningful.

In the current paper, we expand our earlier simulation work to test whether the findings in K&V hold with more speakers and vowel tokens, and to examine the extent of variation across speakers, vowel categories, and vowel tokens (individual instances of speech). Then, in a second simulation, we consider how on- or off-"target" formant estimates are when a similar approach is applied, centered around combinations of default LPC settings, removing input from human LPC settings in the simulation.

The data used in our simulations are recorded word list elicitations from audio recordings collected for a project on regional differences in U.S. English (Kendall and Fridland 2012, Kendall and Fridland 2017). Speakers were recorded with a Tascam digital recorder and a Shure WH30XLR head-mounted microphone in a quiet setting, with just the fieldworker and participant present. In this paper, we measure a single timepoint for each vowel, using 1/3 of the vowel's duration following the practices used in the original analysis (see Kendall and Fridland 2012).

2 Simulation 1: bootstrap, seeded with human analyst settings

In Simulation 1, we use the same bootstrapping (random sampling with replacement) algorithm as in K&V, seeding it with the parameters used by the human analysts as input, but including more speakers and more tokens per vowel. The algorithm works by iterating the following parameters around the seed values according to a normal probability distribution: the timepoint of measurement (Time), NumF, and MaxHz. Following K&V, time is varied around the seed timepoint within a distribution of ±10% of each vowel's duration, NumF is varied within ±1.5 formants (±3 LPC poles) of the seed setting, and MaxHz ranges from ±1000 Hz from the seed. While NumF and MaxHz can interact in their effects on the formant estimation (described in greater detail in Simulation 2), for simplicity we allow them to vary independently. The simulation runs 1000 times for each token sampling from the Time, NumF, and MaxHz distributions. The simulation can thus be thought of as representing 1000 "reasonable" analysts who measure each token, each picking a timepoint and using LPC settings similar to the seed values. Some of these simulated analysts' decisions will be more reasonable than others (some will be practically identical to the seed values), and some will be fairly bad choices of settings and timepoints.

We run the simulation over 10 vowel categories (/i/, /ɪ/, /e/, /ɛ/, /æ/, /ɑ/, /ɔ/, /ʌ/, /o/, and /u/) for each speaker. Our earlier work focused on just one token per vowel category and on four speakers, a male and

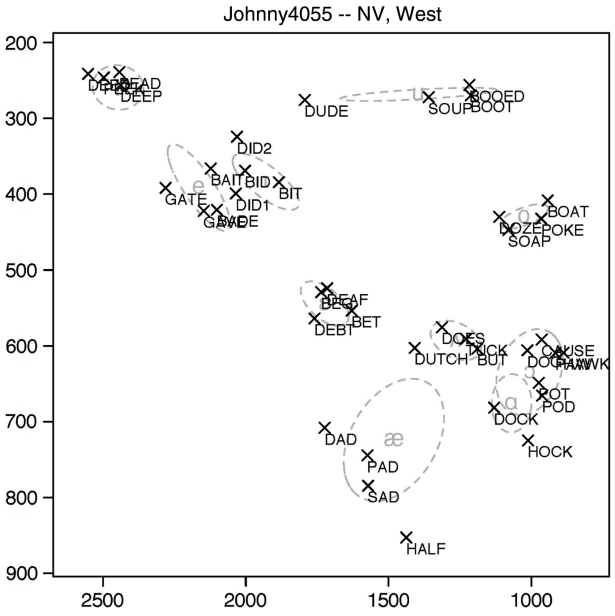

Figure 1: Vowel plot for one male speaker, Johnny4055, showing the seed values (black Xs with word labels) and original means and distributions from the larger (human-analyst) dataset (gray IPA symbols with ellipses).

a female from the Western U.S. and from the Southern U.S. Here, we examine 10 speakers, five males and five females, all European Americans from the Western U.S. (including the two Western speakers from K&V). Instead of a single token as in K&V, we select four tokens per vowel category, choosing the four tokens from the seed dataset closest to each speakers' category means in F1 and F2 space. Figure 1 displays the original vowel plot for one of the speakers, Johnny4055, a male from Reno, Nevada, whom we use to exemplify the process and our results.

2.1 Results

The simulation results in 40,000 measurements for each speaker, 1000 randomly sampled estimates per vowel token ($\times$4 tokens $\times$10 categories). Figure 2 displays the simulation results for the Figure 1 speaker. Not surprisingly, some measurements are clearly outliers. As our goal in this first simulation is to model "reasonable" analysts, we trim obviously bad measurements, defined in a simple way as those farther than 2 standard deviations (SDs) from the mean for each token. This removes a total of 28,250 measurements (7.06%) from the dataset, leaving 371,750 measurements.

Full results of the simulation for individual speakers are presented in the Appendix. Focusing on how the simulation's estimates compare to the gold standards, Table 1 presents these results as average absolute differences from gold standard estimates by vowel category, summarized by speaker sex. Generally, the simulation yields fairly similar patterns as K&V identified for a smaller set of tokens and speakers. With more speakers and tokens here we find more variability, and the differences between the average simulation estimates and gold standards are a bit larger. The current simulation shows average absolute differences from gold standards of 9.1 Hz F1 for males and 12.8 Hz for females. For F2, the difference from the gold standards is 14.4 Hz for males and 21.9 Hz for females. F3 differences (unexamined in our previous work) are higher still, with a 25.8 Hz difference for males and 26.6 Hz for females.[3]

3 The present paper focuses on differences in raw Hz values, since we are most interested in the implications of variability on the estimates themselves rather than on downstream issues in sociophonetic analysis. We note however that some of the scale differences uncovered are mitigated, though not completely eliminated, through normalization procedures. While we do not provide a full treatment here for sake of space, a comparison of differences between simulation results and gold standards

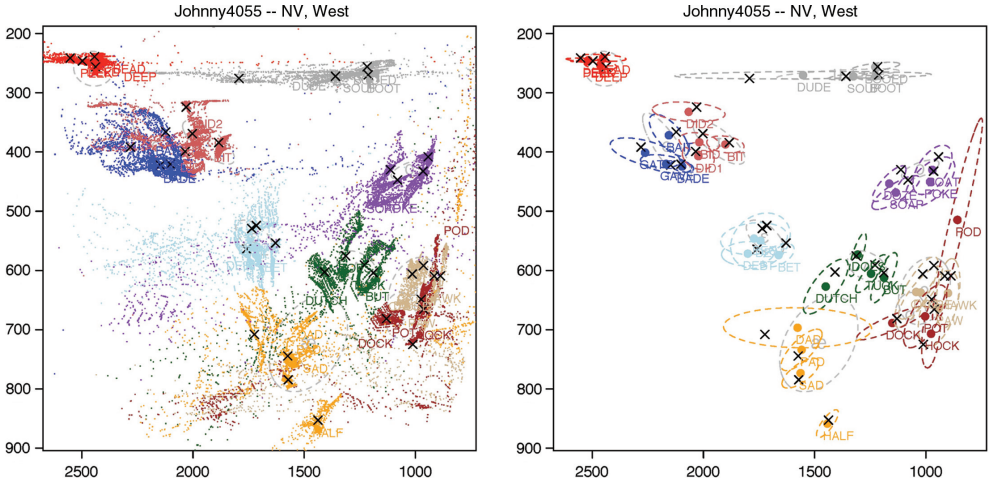

Figure 2: Vowel plots for one male speaker, Johnny4055, showing Simulation 1's individual estimates as colored dots before outlier trimming (left plot) and means and standard deviation ellipses after outlier trimming (right plot). Black Xs show human-measured, seed locations with seed distributions from larger (human-measured) dataset as gray ellipses.

Table 1: Average absolute differences between Simulation 1 results and gold standard estimates by vowel category, summarized by speaker sex.

	Males			Females		
Vowel	**F1 Diff.**	**F2 Diff.**	**F3 Diff.**	**F1 Diff.**	**F2 Diff.**	**F3 Diff.**
/i/	2.4	13.2	61.0	6.1	14.8	28.1
/ɪ/	4.3	10.5	14.5	11.5	9.8	20.4
/e/	4.6	16.3	32.0	10.8	20.3	39.5
/ɛ/	9.5	8.6	24.2	11.6	14.5	19.8
/æ/	6.2	10.2	35.9	18.7	29.8	34.5
/ɑ/	17.8	16.9	20.1	17.6	10.0	21.8
/ɔ/	23.0	17.9	22.3	19.4	20.4	24.9
/ʌ/	8.3	10.0	21.9	14.8	36.4	26.0
/o/	10.8	15.4	13.1	11.9	28.9	30.5
/u/	4.3	25.3	13.0	5.5	33.8	20.3
Mean	*9.1*	*14.4*	*25.8*	*12.8*	*21.9*	*26.6*
Std. Dev.	*6.6*	*5.0*	*14.5*	*4.8*	*9.8*	*6.6*

We use these average differences from gold standard as acceptable ranges (or "bins") of variation for assessing the accuracy of our next simulation; we deem these values to be the range of variation expected based on "reasonable" analysts' settings.

3 Simulation 2: combinatorial, seeded with default settings

Simulation 1 assessed how (un)stable formant estimates are across a range of reasonable LPC settings. To do this, we sampled settings from a distribution around seeds selected by trained human analysts. We now

based on Lobanov-normalized vowels (Lobanov 1971; Kendall and Thomas 2010) finds that average F2 differences from gold standards are actually smallest, with the mean absolute difference for males: F1 = 0.042, F2 = 0.004, F3 = 0.066 normalized units, and females: F1 = 0.066, F2 = 0.028, F3 = 0.062 normalized units. This would also suggest that best-practices for considering (and reporting) meaningful differences in Lobanov normalized data should assume that differences of about ±0.05 normalized units are within the range of measurement variability. In other words, LPC settings-based measurement variability does not get completely "normalized out."

explore whether a simpler simulation, a combinatorial walk through candidate settings based only around commonly accepted defaults, can also obtain reasonable estimates. Thus, this approach is a step toward an unsupervised method, and one that samples evenly across the parameter space of manipulated settings. We use this second simulation as an opportunity to look more closely at how settings impact formant estimates, and in doing so evaluate the results in terms of the extent to which they fall within an "acceptable" range of variability based on the findings from Simulation 1.

Simulation 2 walks through a grid of possible NumF and MaxHz settings. NumF settings are not based on the original seed settings, but instead set to a distribution around Praat's default of 5 formants (10 LPC poles), walking through a set of 9 possible values, from 3 to 7 in 0.5 formant increments (6–14 LPC poles). We center the space of MaxHz values around widely used defaults, 5000 Hz for male voices and 5500 Hz for female voices.[4] The simulation walks through 21 MaxHz values, in 100 Hz increments, ranging from $\pm$1000 Hz centered around the default. This simulation thus gathers 189 estimates for each token (9 NumF $\times$ 21 MaxHz). Since the simulation tests all possible combinations of NumF and MaxHz, it is expected that some combinations will be particularly bad. To limit the complexity here, we do not manipulate timepoints, focusing only on the contribution of the two LPC settings.

Based on the ranges obtained in Simulation 1, each of the simulation's estimates is assessed regarding whether it falls within the mean difference from gold standard estimates, what we refer to below as the "target" bin, or outside the target by $\pm$1, 2, or 3 spans of the mean difference from gold standard estimates. Again, the "target" (the mean difference from gold standard estimates) for F1 is 9.1 Hz for males and 12.8 Hz for females; and for F2 is 14.4 Hz for males and 21.9 Hz for females. Thus, any simulation estimate landing within 9.1 Hz for an F1 value from a male speaker, for example, is considered on target.

3.1 Results

In order to exemplify the kinds of patterns that emerge with different NumF and MaxHz combinations, and the fact that patterns vary widely across vowels (and speakers), Figure 3 displays grids of settings and the estimates obtained for a sample of 6 vowel tokens for the same speaker, Johnny4055, used for exemplification earlier. These tokens were selected simply to illustrate the range of patterns that are obtained. In the figure, green circles indicate on-target formant estimates for a given combination of NumF (settings given on x-axis) and MaxHz (settings given on y-axis), red – indicate lower than target estimates and yellow + indicate higher estimates.

As expected, some combinations of NumF and MaxHz are particularly injudicious. For example, when MaxHz is high and NumF is low (which occurs in the lower left corner in the plots in Figure 3), the resultant formant estimates will tend to be misleadingly high (likely the result of LPC yielding false formants). When low MaxHz settings are coupled with high NumF settings (upper right corner in the plots), formant estimates will tend to be misleadingly low (since the space between poles is compressed). Precisely this systematic and logical relation emerges for the F2 of BEAD and BADE, with overestimates in the lower left corner and underestimates in the upper right, with a band of on-target estimates (green circles) in-between. Another pattern evident is that some formant estimates (often F1 measures for high vowels, which have low frequencies) are relatively stable regardless of the settings. The sample speaker's BEAD and BOOED vowels illustrate this pattern, which indicates that, even with injudicious combinations of settings, the algorithm is best able to accurately estimate lower frequencies (see also Harrison 2013). Finally, in other cases the outcomes are less interpretable. For example, the plots for DEAF and DID (labeled DID1) F1 and F2 and BUT F2, for instance, show a relationship between settings and estimate accuracy that is less predictable.

One other observation is that quite a number of combinations of settings yield reasonable estimates. In Figure 3, it is clear that many pairings of MaxHz and NumF led to on-target estimates (as shown by the bands

4 These are the same default MaxHz settings recommended in Praat (Boersma and Weenink 2019) and used by FAVE (Rosenfelder et al. 2014).

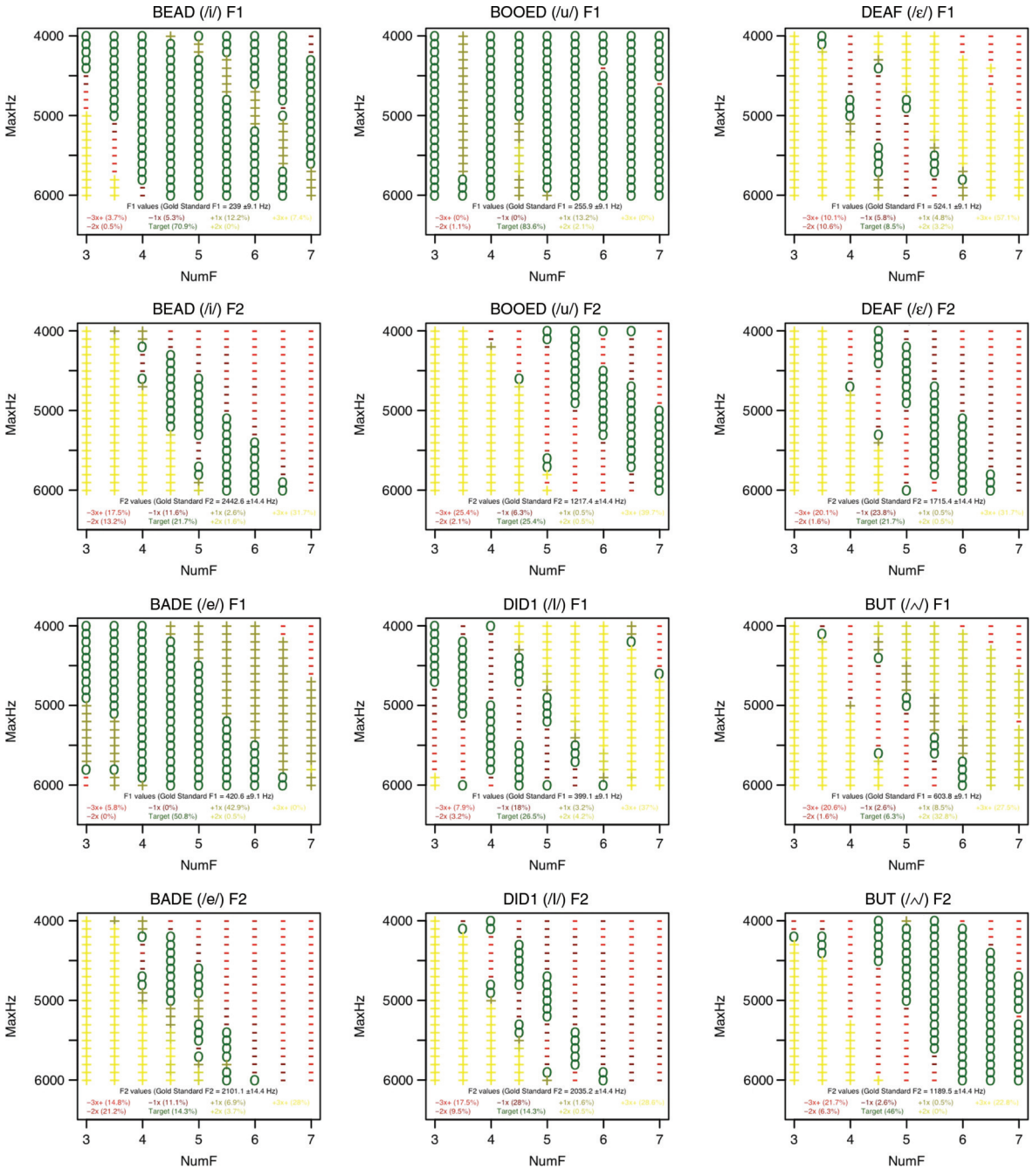

Figure 3: Examples of six tokens from Simulation 2 for one male speaker, Johnny4055, showing F1 and F2 estimates as binned, across all measurement settings.

of green circles). In other words, although certain combinations of settings are clearly "wrong", there is no such thing as a priori "correct" or "best" settings (again see Vallabha and Tuller 2002; Harrison 2013).

Tables 2 (F1) and 3 (F2) summarize the simulation's accuracy and variability across all vowels and speakers, displaying the overall results of Simulation 2 in terms of the percentage of estimates that fall into each of the 7 bins (in the row labeled Mean). The tables also report standard deviations across the 10 vowel categories and 10 individual speakers to show differences across speakers and vowel classes. Results show that F1 values are on target 40.7% of the time, substantially more often than F2 values, for which the simulation estimates are on target only 27.8% of the time. This finding should be taken as further evidence that settings matter.

Table 2: For F1, percentage of simulation estimates that were on target (within the range from Simulation 1) and off target by units of the same bin width.

F1 bin width = 9.1 Hz (males), 12.8 Hz (females)	**$<-3\times$**	**$<-2\times$**	**$<-1\times$**	**On Target**	**$>+1\times$**	**$>+2\times$**	**$>+3\times$**
Mean	16.5	4.7	7.3	40.7	8.6	3.9	18.4
Vowel SD	7.4	2.1	3.3	12.7	2.5	1.8	12.2
Speaker SD	8.6	3.5	3.1	7.4	2.9	1.0	7.3

Table 3: For F2, percentage of simulation estimates that were on target (within the range from Simulation 1) and off target by units of the same bin width.

F2 bin width = 14.4 Hz (males), 21.9 Hz (females)	**$<-3\times$**	**$<-2\times$**	**$<-1\times$**	**On Target**	**$>+1\times$**	**$>+2\times$**	**$>+3\times$**
Mean	23.6	2.9	5.9	27.8	6.5	3.9	29.4
Vowel SD	5.8	1.0	1.6	4.8	2.4	1.4	6.4
Speaker SD	6.8	1.2	1.5	3.8	1.5	0.8	5.3

About 1/3 of F1 estimates fall far from the on target bin from the gold standard values (16.5% in the $<-3\times$ range away, 18.4% in the $>+3\times$ range away) compared to just over half of F2 estimates (23.6% and 29.4%, respectively). Again, these findings of greater instability for F2 are in line with prior research (Vallabha and Tuller 2002; Harrison 2013).

4 Conclusion

These two simulations add to the small body of work exploring how differences in formant analysis procedures affect formant estimates. As noted in our introduction, such work has rarely been of focus in the sociophonetics literature, despite a close attention to other aspects of vowel analysis (such as normalization; see Thomas and Kendall 2007; Watt et al. 2011) and despite some important work in forensic phonetics. The results of Simulation 1 reinforced the fact that small differences in analysts' settings yield variability in formant estimates. Based on these results, F1 differences of 9.1 Hz for males and 12.8 Hz for females, and F2 differences of 14.4 Hz for males and 21.9 Hz for females should be expected. We have reported measurements to one decimal point throughout this paper for across-simulation comparability, but a related take-away is that formant differences on the order of a fraction of 1 Hz are well below the range of meaningful precision. That is, in general, we suggest that distinctions more fine-grained than about 10 Hz differences in F1 and 15–20 Hz differences in F2 are likely due to measurement processes and should not be assumed to be meaningful.

The ranges of estimation variability suggested here are notably on the order of the difference limen, or just noticeable differences, in discrimination of vowels by listeners. Work investigating thresholds of vowel formant discrimination, a domain that we argue is underexplored in sociophonetics, finds that "under optimal listening conditions humans can discriminate vowel formants [for isolated vowels] with a threshold of about 10–15 Hz in the F1 region and a Weber ratio of about 1.5%–2.0% in the F2 region" (Kewley-Port and Neel 2006: 56). Discriminating vowels in the context of longer utterances – more relevant to the present study and sociophonetic work in general – increases the difference limen for discrimination by up to a factor of 4 (Kewley-Port and Zheng 1999). Sociophonetic work taking small differences seriously should make sure to acknowledge both the likelihood of error in the measurements and, depending on the goals of the study, the question of whether the differences are perceptible.

The results of Simulation 2 help to illustrate that settings need to be customized on a per-speaker and per-token basis. While space does not permit a longer treatment, the combinatorial approach taken in Simulation 2 lays out the first steps for an unsupervised, automated approach to formant estimation. We leave fuller treatment of such an idea to future work, but note that such methods can be thought of as changing our approach to formant measurement, from attempting to obtain the *best* estimate for a given formant to attempting to obtain the *most likely* estimate. We suggest that thinking of vowel formant analysis along these lines would be fruitful, for both manual and automatic approaches.

While we have implemented some ways to think about formant estimation as being on or off "target" with respect to a gold standard, we invite the broader field of sociophonetics to consider the implications of the reality that formant values are always estimates within a range of likely error. With this in mind, how do we calibrate the "importance", or meaningfulness, of a vowel difference? This will surely depend on the goals for an analysis. To close, we urge sociophoneticians to take seriously the inherent imprecision of LPC-based formant estimation, and the fact that choices of settings matter, as we continue as a field to refine our methods for uncovering patterns of variation and sound change in language.

Acknowledgements: The data used for this project were collected with support from grants # BCS-0518264, BCS-1123460, and BCS-1122950 from the National Science Foundation.

References

Atal, B. S. & Suzanne Hanauer. 1971. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America* 50. 637–655.

Boersma, Paul & David Weenink. 2019. *Praat: Doing phonetics by computer. Version 6.1*. [software; available from http://www.fon.hum.uva.nl/praat/].

Di Paolo, Marianna, Malcah Yaeger-Dror, & Alicia Beckford Wassink. 2011. Analyzing vowels. In Marianna Di Paolo & Malcah Yaeger-Dror (eds.), *Sociophonetics: A student's guide*, 87–106. New York: Routledge.

Dissen, Yehoshua, Jacob Goldberger, & Joseph Keshet. 2019. Formant estimation and tracking: A deep learning approach. *Journal of the Acoustical Society of America* 145. 642–653.

Duckworth, Martin, Kirsty McDougall, Gea de Jong, & Linda Shockey. 2011. Improving the consistency of formant measurement. *International Journal of Speech, Language, & Law* 18. 35–51.

Farrington, Charlie, Tyler Kendall, & Valerie Fridland. 2018. Vowel dynamics in the Southern Vowel Shift. *American Speech* 93(2). 186–222.

Harrison, Philip. 2004. *Variability of formant measurements*. MA Dissertation. York, UK: University of York.

Harrison, Philip. 2013. *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. PhD dissertation. York, UK: University of York.

Kendall, Tyler & Valerie Fridland. 2012. Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift. *Journal of Phonetics* 40(2). 289–306.

Kendall Tyler & Valerie Fridland. 2017. Regional relationships among the low vowels of U.S. English: Evidence from production and perception. *Language Variation and Change* 29(2). 245–271.

Kendall, Tyler & Erik R. Thomas. 2010. *Vowels. R: Vowel Manipulation, Normalization, and Plotting in R. Version 1.2-2*. [R software package; available from https://cran.r-project.org/web/packages/vowels/].

Kendall, Tyler & Charlotte Vaughn. 2015. Measurement variability in vowel formant estimation: A simulation experiment. In The Scottish Consortium for ICPhS 2015 (eds.), *Proceedings of the International Congress on Phonetics (ICPhS) 2015*. Glasgow: University of Glasgow.

Kewley-Port, Diane & Amy Neel. 2006. Perception of dynamic properties of speech: Peripheral and central processes. In Steven Greenberg & William A. Ainsworth (eds.), *Listening to speech: An auditory perspective*, 49–61. Mahwah, NJ: Lawrence Erlbaum.

Kewley-Port, Diane & Yijian Zheng. 1999. Vowel formant discrimination: Towards more ordinary listening conditions. *Journal of the Acoustical Society of America* 106. 2945–2958.

Labov, William, Ingrid Rosenfelder, & Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1). 30–65.

Lobanov, Boris M. 1971. Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America* 49. 606–608.

Markel, John D. & Augustine H. Gray Jr. 1976. *Linear Prediction of Speech*. Berlin: Springer.

McAuliffe, Michael, Arlie Coles, Michael Goodale, Sarah Mihuc, Michael Wagner, Jane Stuart-Smith & Morgan Sonderegger. 2019. ISCAN: A system for integrated phonetic analysis across speech corpora. In: *19th International Congress of Phonetic Sciences (ICPhS 2019)*, 1322–1326. Australia: Melbourne, 5–9 August 2019.

Morrison, Geoffrey Stewart. 2008. Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. *International Journal of Speech, Language, & Law* 15. 249–266.

Nolan, Francis & Catalin Grigoras. 2005. A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and Law* 12. 143–173.

O'Shaughnessy, Douglas. 1988. Linear predictive coding: One popular technique for analyzing certain physical signals. *IEEE Potentials* 7(1). 29–32.

Reddy, Sravana & James Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1). 15–28.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. *FAVE (Forced Alignment and Vowel Extraction) program suite. Version 1.2.2*. [software; available from https://github.com/JoFrhwld/FAVE].

Thomas, Erik R. 2011. *Sociophonetics: An introduction*. Houndmills: Palgrave.

Thomas, Erik R. & Tyler Kendall. 2007. *NORM: The vowel normalization and plotting suite. Version 1.1*. [online resource; available at http://lingtools.uoregon.edu/norm/; accessed February 2019].

Vallabha, Gautam K. & Betty Tuller. 2002. Systematic errors in the formant analysis of steady-state vowels. *Speech Communication* 38. 141–160.

Watt, Dominic. 2010. The identification of the individual through speech. In Carmen Llamas & Dominic Watt (eds.), *Language and Identities*, 76–85. Edinburgh: Edinburgh University Press.

Watt, Dominic, Anne Fabricius, & Tyler Kendall. 2011. More on vowels: Plotting and normalization. In Marianna Di Paolo & Malcah Yaeger-Dror (eds.), *Sociophonetics*: A student's guide, 107–118. New York: Routledge.

Zhang, Cuiling, Geoffrey Stewart Morrison, Felipe Ochoa, & Ewald Enzinger. 2013. Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *Journal of the Acoustical Society of America* 133. EL54–EL60.

Appendix: Full results from Simulation 1

Tables A1 and A2 provide the vowel category average medians for F1, F2, and F3 of the 4 tokens per vowel category from Simulation 1, with speakers grouped by sex (Table A1 for males, Table A2 for females). The tables also provide the average IQRs (over the 4 tokens) and the average absolute differences between the medians for each vowel token and the gold standard vowel measurement.

Inspecting the variability via average IQRs in these tables, it is evident that some speaker-vowel category pairings obtain somewhat unstable estimates. The individual speakers' IQRs range for F1 from a low of 3.4 Hz (Ryan3381's /i/) to a high of 135.5 Hz (Ann5805's /ɑ/) (mean = 37.5 Hz) and for F2 from a low of 19.7 Hz (Bryan2168's /ɪ/) to a high of 1567.1 (Paige5200's /e/) (mean = 162.2 Hz). Further, we find some support for Harrison's (2013) observation that formant estimation error is not uniform across the vowel space. High vowels tend to obtain tighter distributions for F1 while low vowels obtain the widest distributions (as visible in Figure 2 in the main text). For F2, less consistent patterns emerge. For instance, /ɑ/ and /ɔ/ obtain some of the smallest IQRs, but /æ/, also a low vowel, obtains some very large IQRs. Generally, /e/ obtains the highest IQRs for F2. And, although /ɪ/ has some of the lowest F2 IQRs for some speakers, it also has some of the highest for others (941.6 Hz for Natalie2800 and 1174.2 Hz for Paige5200). The range of variability across individual speakers and vowel classes reinforces the fact that LPC results are dependent on speaker-level and vowel token-level factors, but does not highlight specific patterns for the sources of that variability. We do find systematic patterns, however, by comparing the average IQRs across speaker sex (male: F1 = 32.4, F2 = 80.9; female: F1 = 42.7, F2 = 243.4), individual speakers (F1 = 37.5 [range = 18.0–50.7]; F2 mean = 162.2 [range = 63.2–491.1]), and vowel categories (F1 = 37.5 [range = 13.5–68.3]; F2 mean = 162.2 [range = 49.0–402.9]), showing that indeed higher formant frequencies (i.e. female voices and formants above F1) track less well with LPC and result in less stable estimates (Vallabha and Tuller 2002; Zhang et al. 2013). Speakers and vowel categories, however, yield roughly similar IQR ranges, indicating that they contribute equivalently to LPC-based estimation variability.

Table A1: Simulation 1 results for individual male speakers.

	F1			F2			F3		
	Val	**IQR**	**Dif**	**Val**	**IQR**	**Dif**	**Val**	**IQR**	**Dif**
aaron3265									
/i/	271.5	26.5	7.2	2313.4	68.7	17.2	2800.4	408.6	44.3
/ɪ/	390.0	40.9	8.1	1964.9	46.4	12.1	2598.2	125.1	13.7
/e/	416.5	37.5	8.2	2094.1	79.8	25.6	2657.4	198.1	40.5
/ɛ/	525.9	47.6	14.7	1791.7	45.8	13.5	2581.1	186.2	44.1
/æ/	645.5	29.3	5.6	1661.6	53.2	8.5	2474.9	141.2	49.1
/ɑ/	680.1	37.2	19.4	1137.8	40.4	26.0	2397.0	101.8	32.3
/ɔ/	633.9	67.3	26.1	1043.3	105.4	21.9	2407.0	385.2	29.8
/ʌ/	613.1	48.0	9.7	1297.7	50.8	14.3	2479.8	61.8	16.6
/o/	494.8	53.0	18.6	1093.9	79.1	16.2	2352.2	90.9	15.6
/u/	324.1	21.8	6.9	1585.9	109.2	23.1	2255.8	48.1	5.7
Mn		*40.9*	*12.5*		*67.9*	*17.8*		*174.7*	*29.2*
SD		*13.6*	*6.9*		*24.9*	*6.0*		*126.7*	*15.3*
Bryan2168									
/i/	286.4	6.4	0.4	2437.3	67.4	16.1	3016.0	326.2	47.4
/ɪ/	392.8	9.6	2.6	2095.6	19.7	7.0	2885.9	51.1	14.9
/e/	438.1	15.3	3.5	2191.5	154.8	15.0	2930.2	224.0	13.2
/ɛ/	568.0	16.0	4.1	1880.8	29.7	2.7	2848.5	65.2	10.8
/æ/	780.6	30.9	9.2	1711.7	145.1	10.3	2690.1	306.2	67.7
/ɑ/	701.0	26.6	6.7	1051.4	22.5	5.4	2660.3	135.6	10.8
/ɔ/	635.6	20.2	10.8	970.2	39.5	11.5	2668.0	901.9	25.8
/ʌ/	643.1	14.5	5.0	1277.4	42.3	7.1	2753.6	60.1	16.3
/o/	490.3	30.6	8.3	1101.6	77.5	9.5	2626.5	109.2	3.0
/u/	340.2	10.3	3.3	1392.5	107.5	24.9	2374.5	103.0	13.4
Mn		*18.0*	*5.4*		*70.6*	*11.0*		*228.3*	*22.3*
SD		*8.8*	*3.3*		*49.9*	*6.4*		*256.9*	*20.0*
Eric1510									
/i/	260.5	5.6	0.9	2309.4	301.0	13.1	2831.5	418.5	118.7
/ɪ/	446.2	12.4	2.7	1809.7	24.9	6.4	2682.2	71.6	6.9
/e/	474.5	10.3	3.7	1987.3	340.7	6.0	2683.3	291.3	24.0
/ɛ/	580.7	13.1	3.0	1717.0	31.3	6.5	2612.8	155.9	11.8
/æ/	709.0	17.3	5.4	1581.4	141.8	11.4	2373.9	739.4	30.5
/ɑ/	669.3	27.3	9.4	1129.2	23.2	12.3	2417.1	87.2	5.6
/ɔ/	629.5	74.7	30.7	1062.6	118.7	30.4	2471.4	407.0	16.0
/ʌ/	611.3	20.6	4.6	1327.8	47.4	6.4	2476.8	330.7	19.2
/o/	515.7	32.2	5.5	1149.0	54.4	8.7	2379.4	217.4	25.7
/u/	337.9	11.1	4.3	1362.1	151.1	27.4	2214.1	268.3	23.9
Mn		*22.5*	*7.0*		*123.5*	*12.9*		*298.7*	*28.2*
SD		*20.1*	*8.6*		*114.8*	*8.9*		*196.0*	*32.8*
Johnny4055									
/i/	245.0	7.6	2.2	2475.1	63.7	12.3	3067.0	284.5	78.4
/ɪ/	373.1	30.0	3.9	1969.6	48.6	18.6	2681.0	57.4	19.3
/e/	403.9	32.0	4.1	2163.5	115.4	15.8	2635.5	91.8	53.8
/ɛ/	563.1	44.2	20.4	1706.9	40.9	9.5	2627.6	57.9	36.0
/æ/	771.0	37.6	6.3	1565.7	64.2	12.7	2485.9	162.2	21.9
/ɑ/	657.9	124.4	31.8	998.7	102.8	25.4	2216.5	381.5	10.6
/ɔ/	634.1	63.0	30.0	962.6	53.8	17.0	2208.7	54.6	22.7
/ʌ/	607.2	55.2	14.0	1290.2	53.7	10.0	2466.4	87.9	20.3
/o/	442.5	37.3	13.1	1050.1	62.6	24.6	2465.9	101.6	13.1
/u/	266.2	9.9	3.0	1373.8	190.5	21.8	2392.8	68.3	16.1
Mn		*44.1*	*12.9*		*79.6*	*16.8*		*134.8*	*29.2*
SD		*33.1*	*11.2*		*45.6*	*5.8*		*111.5*	*21.4*

Table A1: (continued)

	F1			F2			F3		
	Val	**IQR**	**Dif**	**Val**	**IQR**	**Dif**	**Val**	**IQR**	**Dif**
Ryan3381									
/i/	257.5	3.4	1.5	2334.7	57.1	7.3	3070.3	168.2	16.4
/ɪ/	405.5	22.6	4.4	2044.0	37.9	8.5	2683.8	89.0	17.5
/e/	422.2	11.2	3.5	2175.8	102.2	19.1	2700.1	238.7	28.4
/ɛ/	529.3	17.9	5.2	1857.6	50.5	10.8	2570.9	143.4	18.5
/æ/	730.4	24.3	4.5	1743.0	47.7	8.3	2434.1	157.0	10.3
/ɑ/	742.3	134.0	21.6	1133.7	67.9	15.5	2165.3	352.3	41.2
/ɔ/	696.2	77.7	17.2	1067.5	27.3	8.9	2296.2	256.7	17.0
/ʌ/	613.3	35.3	8.2	1308.5	58.6	12.0	2496.0	199.9	37.0
/o/	467.2	25.0	8.5	1131.1	72.1	17.8	2183.0	34.0	8.1
/u/	293.2	12.5	3.9	1480.4	110.3	29.2	2245.4	57.3	6.0
Mn		*36.4*	*7.9*		*63.2*	*13.7*		*169.7*	*20.0*
SD		*39.9*	*6.5*		*26.3*	*6.8*		*97.2*	*11.9*

Mean for each vowel category for individual token medians (Val), mean interquartile ranges for individual token distributions (IQR), and mean absolute differences from gold standard values (Dif), separated by formant, for F1, F2, and F3. All values in Hz. Means and standard deviations provided for IQRs and differences.

Table A2: Simulation 1 results for individual female speakers.

	F1			F2			F3		
	Val	**IQR**	**Dif**	**Val**	**IQR**	**Dif**	**Val**	**IQR**	**Dif**
Ann5805									
/i/	364.2	11.8	2.7	2878.2	54.4	7.6	3298.2	182.0	28.2
/ɪ/	490.2	18.5	3.4	2392.7	36.5	6.2	3167.9	120.7	18.1
/e/	455.1	28.8	6.0	2619.7	88.7	10.5	3167.1	261.6	11.6
/ɛ/	730.8	28.9	10.7	2173.0	145.9	12.2	2993.3	92.7	7.8
/æ/	874.2	54.8	7.3	2088.8	147.5	60.9	2778.3	127.6	34.8
/ɑ/	928.2	135.5	15.8	1301.0	49.9	7.6	2825.5	117.7	38.4
/ɔ/	739.6	101.6	37.9	1096.4	61.4	22.0	3028.1	96.2	26.5
/ʌ/	814.6	47.0	14.2	1727.9	58.2	11.3	2983.1	201.4	36.1
/o/	547.9	25.8	7.5	1184.7	59.0	7.3	2972.5	108.5	10.6
/u/	411.1	11.5	4.8	1770.2	86.1	19.7	2942.1	39.7	12.6
Mn		*46.4*	*11.0*		*78.8*	*16.5*		*134.8*	*22.5*
SD		*41.2*	*10.4*		*39.0*	*16.5*		*63.5*	*11.7*
Jocelyn1675									
/i/	352.6	25.7	1.2	2828.4	69.1	6.4	3167.1	133.1	38.9
/ɪ/	507.6	19.6	2.4	2296.7	35.4	5.8	3142.1	67.8	13.3
/e/	445.2	8.2	0.8	2569.1	80.9	6.5	3105.2	395.2	23.8
/ɛ/	675.0	28.9	2.1	2069.4	161.4	9.8	3095.2	327.6	19.7
/æ/	861.9	44.6	7.2	1838.2	76.9	3.7	2867.2	224.3	16.2
/ɑ/	834.1	36.4	6.8	1264.9	30.3	6.6	3150.8	48.3	9.2
/ɔ/	808.3	21.9	5.8	1268.9	31.5	2.9	3154.5	31.0	7.9
/ʌ/	758.3	28.7	15.3	1714.6	44.1	19.9	2966.7	186.3	18.9
/o/	563.0	25.8	11.8	1383.5	103.6	44.4	2845.3	173.8	39.0
/u/	440.9	6.7	2.7	1954.8	233.6	43.0	2836.9	145.3	26.9
Mn		*24.7*	*5.6*		*86.7*	*14.9*		*173.3*	*21.4*
SD		*11.5*	*4.8*		*65.5*	*15.9*		*118.0*	*11.0*
Lindsey1595									
/i/	329.0	32.3	1.5	2815.5	52.4	3.7	3410.9	106.6	10.2
/ɪ/	521.2	30.9	5.6	2234.5	57.0	13.0	3103.6	69.0	9.6
/e/	414.2	32.6	9.4	2626.4	102.6	29.2	3168.4	195.7	39.8
/ɛ/	698.5	43.9	15.6	2043.7	393.7	31.8	3082.9	325.0	8.9

Table A2: (continued)

	F1			F2			F3		
	Val	**IQR**	**Dif**	**Val**	**IQR**	**Dif**	**Val**	**IQR**	**Dif**
/æ/	870.0	134.0	36.4	1835.7	414.4	27.4	2991.2	538.3	28.8
/ɑ/	812.6	52.5	14.0	1119.3	80.9	24.0	2807.6	54.5	11.3
/ɔ/	801.3	55.8	19.1	1205.1	115.0	43.6	2807.3	350.4	15.8
/ʌ/	731.0	88.8	18.6	1435.4	380.3	96.6	2964.8	641.3	4.3
/o/	527.6	29.4	10.2	1156.7	118.1	29.0	2869.3	90.5	13.6
/u/	371.3	6.4	2.9	1816.6	207.5	57.2	2717.6	111.3	28.5
Mn		*50.7*	*13.3*		*192.2*	*35.6*		*248.3*	*17.1*
SD		*36.4*	*10.2*		*147.3*	*26.0*		*208.5*	*11.4*
Natalie2800									
/i/	364.8	47.3	12.7	3082.5	119.2	25.4	3625.5	284.8	26.0
/ɪ/	485.7	30.0	22.6	2609.0	941.6	12.7	3412.7	853.3	36.8
/e/	449.9	20.7	10.0	2939.7	1396.3	28.2	3303.2	511.8	83.7
/ɛ/	749.0	37.6	15.4	2357.2	410.5	11.4	3341.1	879.7	46.4
/æ/	966.4	113.7	22.7	2210.4	354.0	30.1	3093.5	764.7	61.3
/ɑ/	808.5	80.6	41.2	1050.7	41.7	3.4	3027.3	132.2	31.2
/ɔ/	817.2	78.0	22.9	1016.3	42.3	15.5	2967.7	547.7	41.8
/ʌ/	803.7	42.0	9.7	1567.4	165.9	19.5	3139.7	863.3	40.9
/o/	548.9	30.9	13.5	1221.8	99.2	19.1	3014.0	81.8	37.2
/u/	424.0	11.1	6.9	1820.8	112.1	34.3	2912.3	70.7	23.0
Mn		*49.2*	*17.8*		*368.3*	*20.0*		*499.0*	*42.8*
SD		*31.9*	*10.1*		*452.3*	*9.6*		*335.7*	*18.0*
Paige5200									
/i/	274.2	27.9	12.2	2838.1	112.4	31.0	3355.5	396.3	364.2
/ɪ/	475.9	30.7	23.5	2165.0	1174.2	11.3	2963.1	801.3	490.2
/e/	485.7	25.8	28.0	2362.0	1567.1	27.0	2916.8	841.0	455.1
/ɛ/	651.0	42.9	14.3	1907.1	787.3	7.4	2836.9	927.2	730.8
/æ/	870.8	106.1	19.9	1733.7	456.6	26.7	2543.8	656.9	874.2
/ɑ/	781.7	28.7	10.3	1154.3	30.8	8.6	2456.2	313.8	928.2
/ɔ/	775.9	49.5	11.1	1145.0	50.8	18.1	2315.5	180.8	739.6
/ʌ/	756.0	48.5	16.1	1419.2	362.1	34.8	2515.8	799.9	814.6
/o/	580.5	29.5	16.7	1438.9	223.2	44.8	2488.1	342.3	547.9
/u/	443.8	33.9	10.3	1795.7	146.7	15.0	2719.4	204.4	411.1
Mn		*42.4*	*16.2*		*491.1*	*22.5*		*546.4*	
SD		*24.0*	*6.0*		*523.8*	*12.4*		*287.2*	