MEASUREMENT VARIABILITY IN VOWEL FORMANT ESTIMATION: A SIMULATION EXPERIMENT

Tyler Kendall & Charlotte Vaughn

University of Oregon
tsk@uoregon.edu; cvaughn@uoregon.edu

ABSTRACT

This paper considers sources of error in common vowel formant extraction techniques, investigating the extent to which the temporal point of measurement and software settings influence the formant values obtained. To do this, we report on the results of a vowel measurement simulation where, rather than extracting a single measurement for each vowel token, thousands of measurements are taken for each vowel with varied settings in jittered measurement locations (seeded by measurements from human analysts). Such an undertaking, we argue, yields important insight into the bounds of measurement error in vowel analysis.

Keywords: vowel analysis, linear predictive coding, formant settings, simulation.

1. INTRODUCTION

Formant analysis has become a central methodology in the acoustic study of vowel systems in a range of phonetic disciplines. Drawing conclusions from observed differences between formant values under varying conditions (e.g., across vowel categories, across speakers or regions, across speaking styles, or as a result of particular manipulations in the laboratory) is an important source of phonetic knowledge. Most present-day analysts acknowledge that best practices are needed in areas like statistical analysis and inter-analyst agreement for acoustic research. Furthermore, recent years have seen progress in improving vowel analysis techniques, such as through vowel normalization procedures (e.g., [1, 5, 6, 16]), and automated and semi-automated vowel alignment and extraction (e.g. [12, 15]). Despite this attention to methodological rigor, the fact remains that error can be introduced by a number of sources in acoustic vowel research, and many of these potential sources have been under-addressed.

The source of error we take up in this paper is that introduced in the measurement of vowel formants. Vowel formant measurement is in fact *estimation* and no single "correct" value exists as an essential property of an individual vowel token [17, 18]. That is, vowel measurement techniques, such as the formant tracking implemented in Praat [2], seek to obtain the most accurate estimates of formant frequencies through the use of signal processing algorithms, often using linear predictive coding (LPC) analysis [2, 17, 18]. However, it is known that there are limitations in the precision of vowel estimates as a function of LPC methods [2, 10, 11, 17, 18], and as a function of the properties of the acoustical signal being studied [3] (e.g. noise [14]).

Little research has explicitly or quantitatively studied the extent to which differences in analysts' settings and decisions matter for the outcome of an investigation (though see [4, 10, 11]). In this paper, we report on the results of a simulation experiment which sheds light on the bounds of variability obtained through formant estimation. In particular, we look at the effect of different measurement time points and different LPC settings.

2. THE SIMULATION

Our investigation is rooted in the widespread use of Praat and its LPC algorithm for formant analysis. Further, we base our study on the premise that if any analyst is given a list of specific vowel tokens from a single audio recording and asked to measure the formant frequencies of the first two formants (F1, F2), the analyst has a limited set of decisions to make: Where exactly is the vowel (i.e. what are its boundaries)? What time point(s) in the vowel should be measured? And, what LPC settings should be used? Thus, in terms of actually measuring (i.e. estimating) formant values from a given set of tokens in a given recording, all variability, unreliability, and inaccuracy arise from only two sets of parameters the analyst controls: choices involving the time point(s) measured, and choices involving the LPC algorithm and its settings. Two settings commonly manipulated by analysts are the number of LPC coefficients and the maximum formant frequency (set by Praat users in Praat's "Formant Settings…" menu). While the manipulation of other settings is possible in Praat (e.g., time step, window length, pre-emphasis), we do not consider them here. The number of LPC coefficients and maximum frequency together are generally accepted as crucial adjustments to the formant tracker and are recommended to be made on a per-

speaker and per-token basis. In terms of the temporal location of the vowel measurements we follow a common convention (e.g. [8]), and conceptualize target formant values as single measurement points for each vowel, at 1/3 the duration of the vowel.

The data for this simulation, recorded word list elicitations, come from a series of audio recordings collected for a project on regional production and perception differences in U.S. English [8]. Speakers were recorded with a Tascam digital recorder and a Shure WH30XLR head-mounted microphone in a quiet university office or home, with just the fieldworker and participant present. The original measurements, made by a team of trained analysts, and the settings used for those original measurements, are treated as seed values for the simulation. A bootstrapping algorithm was given the seed values as input, and varied the following parameters according to a normal probability distribution:

- The time point of measurement (Time)
- The number of formants (NumFs)
- The maximum frequency (MaxHz)

Time was varied around the seed value time points within a distribution of ±10% of each vowel's duration, NumFs was varied within ±1.5 formants (±3 LPC coefficients) of the seed setting, and MaxHz ranged from ±1,000 Hz around the seed. While NumFs and MaxHz can interact in their effects on the formant tracking, for simplicity we allowed them to vary independently here. We ran the bootstrap simulation for 1,000 iterations over 10 vowel categories (1 token per category) for four speakers, a male and female from the Western U.S. and from the Southern U.S.

The tokens come from the following vowel categories in American English: /i/, /ɪ/, /e/, /ɛ/, /æ/, /ɑ/, /ɔ/, /ʌ/, /o/, /u/. For each speaker, we selected the individual vowel token from the original, human-analyst measurement set (from [8]) closest to that vowel category mean. Fig. 1 displays the original vowel plot for one of the four speakers, the Western female. The plot depicts her mean for each vowel category (depicted by the IPA symbol) with ellipses indicating one standard deviation. Xs indicate the position of the individual tokens used for the simulation for each category, and standard orthography depicts the word from which the token comes (e.g. SEAT, SOUP, etc.).

Overall, the simulation can be thought of as representing 1,000 "reasonable" analysts, each picking a time point for each token 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 a few will be fairly bad choices of settings and time points.

Figure 1: Vowel plot showing the seed values for the Western female speaker.

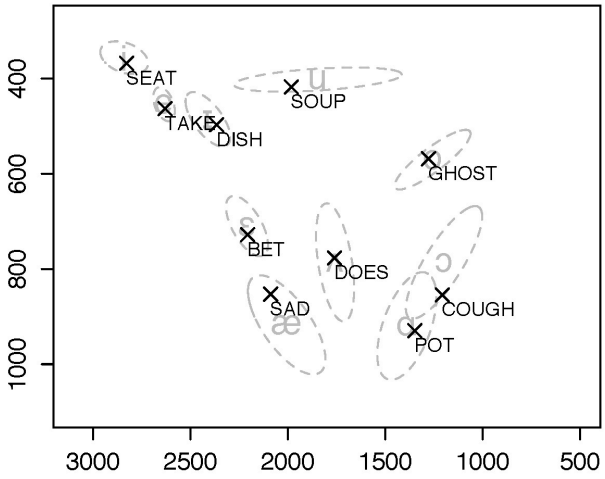
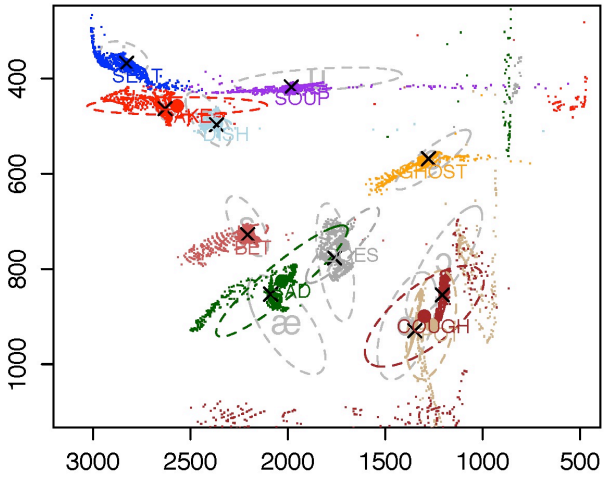


Figure 2: Vowel plot showing the results of the simulation for the Western female speaker.



3. RESULTS

The simulation results in 10,000 vowel measurements per speaker, 1,000 measurements for each of the 10 vowel tokens. Fig. 2 shows these resulting formant values overlaid on the monochromatic vowel plot of the speaker shown in Fig. 1. The other three speakers are not shown, for sake of space. Each point, colored by vowel category, represents a simulated measurement and the colored ellipses depict one standard deviation from the mean of each of the 1,000 measurements, with the means represented by larger dots.

In order to present the overall results for the four speakers, Figs. 3 & 4 display boxplots for F1 and F2, respectively, for each vowel for each speaker. The figures for F1 and F2 are displayed with the same scale (0 Hz – 3,500 Hz) for ease of comparison.

The results of the simulation could be reasonably interpreted in either of two ways. First, it can be noted that many of the distributions are fairly tight, as indicated by the small size of many of the boxes

in the boxplots. Thus, we could say that despite some noise, the simulation largely obtains coherent estimates. On the other hand, given that we derived the simulation measurements around seed values from "good" estimates in the first place, we could be struck by the amount of variability, as shown in the large number of outliers in the plots. Regardless of which view we take, our current state of knowledge does not provide us a best-practice way to determine the boundary between what we might interpret as "good" estimates and "bad" estimates, a problem to which we return in §4.

Figure 3: Boxplots showing F1 values for each vowel and speaker.

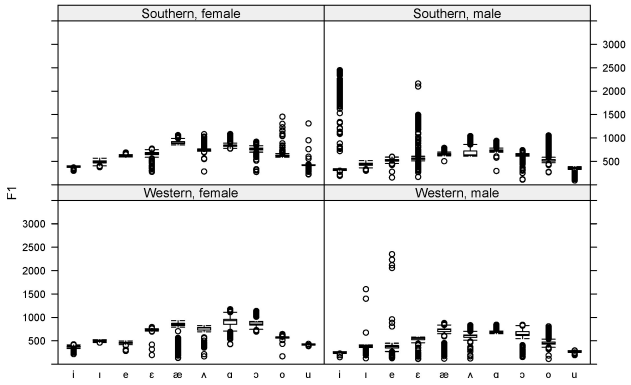
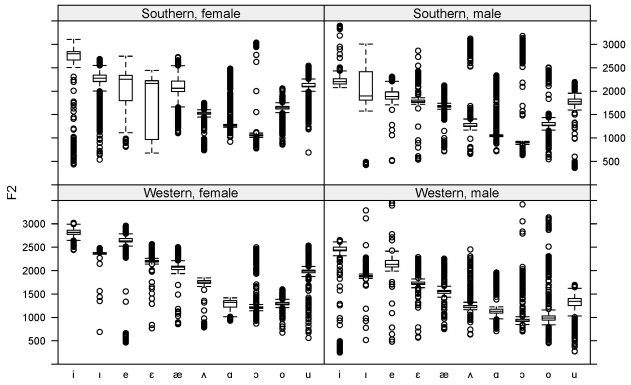


Figure 4: Boxplots showing F2 values for each vowel and speaker.



3.1. /u/ fronting: A case study

To examine the effect of measurement parameters on formant estimation more closely, we turn to consider one vowel in detail, the /u/ vowel in SOUP for the Western female. /u/ is well known to be fronting across varieties of English around the world [7, 9, 12] and, thus, the F2 positions of a speaker's /u/ productions are often of interest in sociophonetic studies. Figs. 2-4 show that F2 is variable across the simulation for all speakers; the LPC settings used and time points measured in the vowel impact the values obtained. Here, we explore which specific parameters account for this variability in the resultant measurements of /u/ F2.

Figure 5: F2 of /u/ by Time, NumFs, and MaxHz for the Western female speaker.

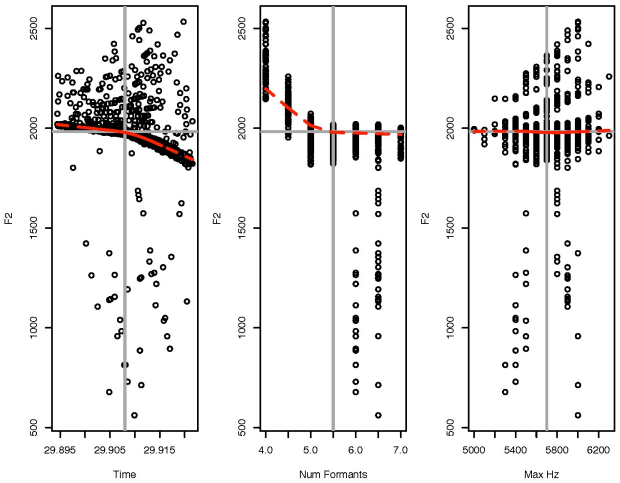


Fig. 5 displays the simulation-generated F2 values for the SOUP token from the Western female speaker as a function of the Time (left), NumFs (center), and MaxHz (right). The gray lines depict the seed values in each panel and the dashed red line depicts the output of a lowess smoother. It is clear that no one parameter accounts for all of the variability. The estimated F2 values of /u/ appear to be largely a function of both the time point measured and the number of formants used in the LPC analyses. Specifically, /u/ backs more with later measurement points. This is a sensible result, since /u/ is generally characterized in English varieties by a diphthongal, back-gliding trajectory. Also, low numbers of formants in the LPC settings result in higher F2 values. This too makes sense given that fewer coefficients in the LPC algorithm's settings will pull the tracking of F2 higher, towards F3. MaxHz here appears to be less influential in impacting the output of the measurements (likely, the simulation did not set MaxHz values low enough in this case to have a large impact on F2 estimates), although the largest outlier measurements probably result from an injudicious combination of NumFs and MaxHz.

Most important here is the observation that many of the "incorrect" values for /u/ F2 look like perfectly valid values but would confound an analysis of back vowel fronting. That is, an analyst using an earlier time point or a lower number of formants than another analyst would conclude that the speaker is engaging in more /u/ fronting. Thus, slightly different analytical decisions in measurement parameters yield different but reasonable looking results, impacting the degree of /u/ fronting "discovered" for this speaker.

4. DISCUSSION

As discussed in §3, we can conceptually group the simulation output into two broad categories. We obtain some measurements that are clearly bad, resulting from poor choices of parameters. Many of these "erroneous" values would presumably catch the eye of human analysts and would be corrected or removed from the dataset. We can also identify measurements, however, that fall within the bounds of what could be considered acceptable, even good, measurements by a human analyst. The issue is: How do we define the cut off between acceptable and unacceptable tokens?

In this simulation, we began by starting with "reasonable" values generated by trained human analysts and then created sets of similar measurement parameters based on those seed values. It may seem circular to assess the success of the simulation based on those original values. However, our goal is not just to assess how well the simulation recreates the seeded values. Rather, the simulation gives us a means to examine the boundary between "good" and "bad" measurements.

One way to do this is to compare the central tendencies of the simulation results against the original seed values. Since many extreme outliers fall into the "erroneous" category and would likely be avoidable by trained human analysts (e.g. by not making obviously inappropriate choices about formant settings and by remeasuring clearly bad formant values), we first trim, for all speakers, the simulation's extreme measurements for each vowel token, those that are more than 2 standard deviations from the mean for that token (this removes between 18-160 measurements per token, M = 81.9). Then, we calculate the absolute difference between the median for the remaining simulation measurements and the original seed value.

Table 1: Absolute Hz value differences between median of simulation results (trimmed at 2 SDs) and the original seed values.

	W, fem.		W, male		S, fem.		S, male	
V.	F1	F2	F1	F2	F1	F2	F1	F2
/i/	2.6	10.6	5.8	7.1	0.3	0.4	0.2	4.5
/ɪ/	0.7	0.1	4.8	7.2	0.8	14.7	14.0	138.0
/e/	3.8	7.8	8.2	11.2	6.6	31.4	9.7	9.8
/ɛ/	5.2	8.1	20.8	12.4	18.8	14.1	8.1	5.1
/æ/	5.3	13.2	0.9	16.6	18.5	62.4	2.2	7.0
/ʌ/	0.3	2.5	19.5	16.1	3.9	6.9	4.9	6.2
/ɑ/	3.8	24.7	0.0	5.7	0.3	7.2	6.3	3.3
/ɔ/	3.5	2.3	22.8	6.9	1.2	9.5	1.1	2.7
/o/	1.7	11.0	9.9	8.4	3.7	2.4	3.6	1.3
/u/	2.3	6.1	1.0	10.6	0.8	3.6	1.6	37.1
M	*2.9*	*8.6*	*9.4*	*10.2*	*5.5*	*15.3*	*5.2*	*21.5*
SD	*1.7*	*7.0*	*8.7*	*3.9*	*7.2*	*18.8*	*4.4*	*42.2*

These results, for F1 and F2 for each speaker's vowels, are displayed in Table 1. For most vowels, we obtain F1 values within about 6 Hz of the seed values and F2 values within about 14 Hz (based on the mean of speaker means), but we also find some cases with much larger differences, as much as 138 Hz for the F2 of the Southern male for /ɪ/. Thus, even when only including settings similar to those that would be used by sensible human analysts, we find variability in the resultant formant values. Notably, some of this variability is of similar magnitude to vowel differences reported as significantly distinguishing between groups or conditions in the research literature.

5. CONCLUSION

The simulation presented here is meant as a first attempt at better understanding and quantifying variability, inaccuracy, and error in vowel formant measurements. Its results suggest that analysts should be cautious in interpreting small Hz differences as meaningful, whether in comparing between studies, in comparing between individual subjects in the same study, or even in comparing between individual vowel tokens. Specifically, the average differences between our simulated "analysts" and original seeds suggest, conservatively, that we should not interpret F1 differences less than 6 Hz and F2 differences less than 14 Hz as meaningful.

The variability observed in the simulation could be assessed in a variety of ways. In this paper we opted to compare the simulation results to a set of "gold standard" values (the seed values) produced by trained human analysts. Alternatively, simulations like this could be used to determine not what are the most *correct* formant values, but what are the most *probable* formant values. We suggest that this is a promising direction given that (1) formant measurement is always estimation and therefore no single value will necessarily best represent an estimate and (2) such a method could help advance automated techniques. While our simulation was based on a set of seed values from human analysts, a similar version could be built to bootstrap the entire distribution of possible values. We leave this for future work. In closing, we note that simulations of analyst behavior provide a means to make measurement variability in vowel formant estimation more tractable.

6. ACKNOWLEDGMENTS

The data used for this project were collected with support from grants # BCS-0518264, BCS-1123460,

and BCS-1122950 from the National Science Foundation.

7. REFERENCES

[1] Adank, P., Smits, R., van Hout, R. 2004. A comparison of vowel normalization procedures for language variation research. *JASA* 116, 3099-3107.

[2] Boersma, P., Weenink, D. 2013. Praat: Doing phonetics by computer. http://praat.org/

[3] De Decker, P., Nycz, J. 2013. Technology of conducting sociolinguistic interviews. In: Mallinson, C., Childs, B. Van Herk, G. (eds.), *Data Collection in Sociolinguistics*. NY: Routledge, 118-126.

[4] Duckworth, M., McDougall, K., de Jong, G., Shockey, L. 2011. Improving the consistency of formant measurement. *International J of Speech, Lang., & Law* 18, 35-51.

[5] Fabricius, A.H., Watt, D., Johnson, D.E. 2009. Comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Lang. Var. & Change* 21, 413-435.

[6] Flynn, N. 2011. Comparing vowel formant normalisation procedures. *York Working Papers in Ling.* 11, 1-28.

[7] Fought, C. 1999. A majority sound change in a minority community: /u/-fronting in Chicano English. *J. Socioling.* 3, 5-23.

[8] Fridland, V., Kendall, T. 2012. Effect of regional vowel differences on vowel perception and production: Evidence from U.S. vowel shifts. *Lingua* 122, 779-793.

[9] Harrington, J., Kleber, F., Reubold, U. 2008. Compensation for coarticulation, /u/-fronting, and sound change in Standard Southern British: An acoustic and perceptual study. *JASA* 123, 2825-2835.

[10] Harrison, P. 2004. Variability of formant measurements. MA Dissertation. York, UK: University of York.

[11] Harrison, P. 2007. Formant measurement errors: Preliminary results from synthetic speech. 2007 IAFPA Annual Conference. Plymouth, UK.

[12] Labov, W., Ash, S., Boberg, C. 2006. *Atlas of North American English*. Berlin: De Gruyter.

[13] Labov, W., Rosenfelder, I., Fruehwald, J. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Lang*. 89, 30-65.

[14] Rathcke, T., Stuart-Smith, J. 2014. On the impact of noise on vowel formant measures. Methods in Dialectology XV Conference. Groningen, NL.

[15] Rosenfelder, I., Fruehwald, J., Evanini, K., Yuan, J. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.

[16] Thomas, E.R., Kendall, T. 2007. NORM: The vowel normalization and plotting suite. http://slaap.lib.ncsu.edu/tools/norm/

[17] Vallabha, G., Tuller, B. 2002. Systematic errors in the formant analysis of steady-state vowels. *Speech Comm.* 38, 141-160.

[18] Weenink, D. 2015. *Speech Signal Processing with Praat.* http://www.fon.hum.uva.nl/david/sspbook/sspbook.pdf