Statistics of vowel formants: Normality of distributions across the syllable

D. H. Whalen

CUNY Graduate Center
Haskins Laboratories
Yale University
Back to Basics

• The statistics in this talk are very basic.
• I find them interesting because we have assumed certain aspects of them rather uncritically for decades.
• Further, the data that we base our statistics on need to be of the right granularity.
• Measurements that give accurate central tendencies may not give accurate individual values.
• More and more analyses are assessing individual values, so a look at accuracy is needed.
The Issue of Normality

• Speech, like other biological systems, is quite variable.

• The statistics we use to analyze speech typically assume normality in the distributions of this variability.

• It is justified for speech?

• We really have not had enough evidence to tell.
Measuring Normality

• There are statistical tests of normality.
  – Graphical (e.g. QQ plots).
  – Numerical (4 moments).
  – Normality test (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Lilliefors, etc.)

• Such tests are useful, but they do not indicate whether a single distribution is involved, only whether the distribution itself is Gaussian.
Number of Distributions

• One way of testing the number of distributions is with Gaussian mixture models.
Still, Why Do We Care?

• If the center frequency of a formant is the target, we would expect a single distribution, whether it is normal or not.

• If reaching the edge of a range of acceptable values is all that the speaker is attempting to do, we would expect two distributions, at either end of the acceptable range.

• The next slide shows model data for F1 values with a mean of 300 Hz.
F1 distribution (mean = 300 Hz)
Moments are Nearly Identical

- A model of F1 data fitting those distributions was created.
- The mean (300) was identical (yay!).
- The sd was somewhat larger for the sum of the edges (29 normal, 41 edges).
- Skewness was 0 for both.
- Kurtosis slightly smaller for edges (-0.32 normal, -0.68, edges).
  - The non-zero value for the normal distribution is probably due to rounding errors and/or small sample size.
Gaussian Mixture Model

• A Gaussian Mixture Model (GMM) was applied to this data.
• Both AIC and BIC indicate that a 2 kernel model provided a better fit.
  – AIC: 1 kernel: 10,251; 2: 10,225.
  – BIC: 1 kernel: 10,261; 2: 10,250.
• For 100 runs of the 2 kernel, 98 had 2 kernel solutions, while 2 had 1 kernel solutions.
  – There was also an asymmetry in the 2 kernel distributions.
Three Models of Vowel Targets

• Central target, single distribution.
  – This is the (generally unstated) position in phonetics.

• Edge targets.

• Flat distribution within a range.
  – This is most consistent with Keating’s “window” model of coarticulation (1990).
  – “As will be seen below, this window is not a mean value with a range around that mean, or any other representation of a basic value and variation around that value. It is an undifferentiated range representing the contextual variability of a feature value.” (p. 455).

• As we saw in a previous slide, edges give the same undifferentiated values as predicted by the window model, so it seems unlikely we will be able to differentiate them.
Good News on Central Tendencies

• Multiple repetitions help with estimating the central tendencies of vowels.
• Even with relatively uncontrolled data, having at least 100 repetitions gives good results.
• Our results for formant analysis in Yoloxóchitl Mixtec (ISO 639 code xty) were promising.
  – 261 words in isolation x 6 reps x 10 speakers = 15,660 tokens
Monte Carlo Simulations of Sample Size

VIII. Results: how much data is needed to approximate the mean?
Focus on F1, Smaller Samples
What about the Distribution?

- Although the results are excellent for the central tendencies, it is clear that the distributions are more subject to measurement error.
- Nonetheless, I forged ahead with an experiment.
  - The scientific equivalent of “It is easier to ask for forgiveness than permission”?
Experiment: Within-Speaker Variability

• As mentioned, every speaker is variable in production.
• But is this variability normally distributed and from a single distribution?
• We don’t really have the data, since, if we measure 20 tokens, we feel like we’ve done a lot.
• But you can’t tell whether 20 tokens are normally distributed or not, much less whether they come from a single distribution.
Design

• To see typical variability, we need hundreds of repetitions.
• But saying the same thing over and over results in massive variability.
  – E.g., the “bucket” experiment, Kello et al. (2008).
• Paying people to be consistent is not very successful and focuses undue attention on it.
Design, cntd.

• I selected four forms to be repeated: “heed,” “geek,” “owed/ode” and “dote.”
  – The homonyms were selected for other reasons.

• There were 100 repetitions of each form per day, with a planned 500 repetitions.
  – Due to an unbelievable sequence of events, I only have 200 so far.

• They were randomized with 50 filler words, each occurring twice, once in the first half, once in the second.

• Acoustic measurements only.
Predictions

- Relatively steady-state vowels ("heed," "owed/ode") will be normally distributed.
- Highly coarticulated vowels ("geek," "dote") will be skewed toward the consonantal values.
Are These Measurements Accurate?

• Formant measurements are influenced by the nearest harmonic.
  – Klatt (1986); Vallabha & Tuller (2002).
• (Of course, we are really interested in the resonances, not the formants, but I will still call them formants.)
• Even recent versions of formant trackers have this problem.
  – Shadle, Nam and Whalen (2016).
Our Accuracy

• It seems that our results are not quite as subject to this problem.

• The next figure shows F1 for “heed” plotted against the F0 for that token.
  – All points from normalized time slice 3-9 were used.
  – F0 is also coded by color.

• The solid lines are where the formant value would be if the closest harmonic was used.
Fairly Immune to the F0 Problem

• The values do not line up with the nearest harmonic to any great extent.
• There may be a bit of an attraction at the higher F1 values.
• But overall, our data seem pretty good.
/i/ in "heed"
Is F0 Really a Problem?

• Yes!
• The next slide shows some speakers from the XRMB database.
• F1 is plotted against F0, with the harmonic lines marked as in the previous slide.
• For some speakers, there is not much influence, but for others, it’s huge.
Analysis

• All tokens of the four forms (five words) were time normalized into 11 points.

• The formants were taken from Praat’s LPC analysis.
  – Shadle, Nam & Whalen (2016) had found the Alku algorithm somewhat superior, but this advantage was outweighed by Praat’s rejection of outliers in its formant tracking.

• A smoothing spline ANOVA was used for examining the trajectories.
SSANOVA plot, /i/
SSANOVA plot, /o/

Group: owed_ode_dote_F2

Group: owed_ode_dote_F1

Word
- dote
- ode
- owed
Formant (Kernel Density Estimates)
Results

• Distributions are normal for all utterances.
  – The slight discrepancy for “dote” is largely because of a small change in rate between the two days of recording.

• Skewness and kurtosis measures for “geek” for the 11 normalized time points were generally small and varied between positive and negative across the time points.

• It seems likely that the time normalization itself is partly responsible for this inconsistency.
  – I don’t see a way around this at this time.
Results, cont’d

• The coarticulatory context had large effects on the trajectories, as expected, but it did not seem to have any effect on the distributions.

• From the trajectories, it is clear that taking a mid-vowel measurement alone does not fully describe the results.

• The dynamic movement of the formants should also be taken into account.

• This is not a new observation, but I myself do not have a satisfactory way of handling full trajectories.
Gaussian Mixture Model

• The next slide shows an F1/F2 plot of the GMM three time points (normalized points 2, 4 and 6 out of the 11 analyzed) of “heed” for 1, 2, and 3 kernel models.

• The GMM’s BIC found that a single kernel was the best for points 4 and 6, but 2 kernels for point 2.
  – For all 11: 1, 2, 10 and 11 were > 1; 3-9, only 1.

• The AIC, on the other hand, pointed to 2 kernels or more.
Concern about Edges

• The formant values at the beginning and ending of the words seem more variable than expected.
• It seems likely that this is due to measurement error and not to any intrinsic difference in the distributions.
• Unfortunately, verifying this would at a minimum require hand correction of each token, and possibly a different kind of formant analysis, such as reassigned spectrum analysis (Fulop 2010).
Conclusions

• Vowel formant variability seems to be normally distributed.
  – More speakers are needed.

• The BIC indicates that the values come from a single distribution.

• The AIC indicates that the values come from two distributions.

• The dynamic nature of even the “steady-state” vowels needs to be better accommodated.
Next Steps

• Determine how much the F0/formant interaction matters.
  – Could have many synthesized tokens, to compare input (=resonance) values to measured (=formant).

• Find better low dimensional descriptions of formant trajectories.

• Analyze more data!
Thank you!

• Special thanks to Dr. Wei-Rong Chen for help with MATLAB and R.
• Work supported by NIH grant DC-002717 to Haskins Laboratories.
References


References, cont’d


