The perception of formant-frequency range is affected by veridical and judged fundamental frequency

Santiago Barreda & Terrance M. Nearey

Introduction

The range of formant frequencies (FFs) produced by a speaker will be most strongly determined by that speaker's vocal tract length [1]. In a previous experiment, we trained listeners to report that acoustic characteristic associated with the average FFs produced by a speaker, which we refer to as FF-scaling [2]. We found that f0 affects listener's reporting of FF-scaling and that f0 and FF-scaling errors are negatively correlated. However, that experiment used a small number of voices and a relatively coarse, discrete response space, making it difficult to investigate the reporting of FF-scaling in a detailed manner.

In this experiment, listeners were asked to report FF-scaling using a large number of voices arranged on a (quasi-) continuous response space in order to investigate:

1) The accuracy with which listeners can report FF-scaling.
2) The possible correlation in f0 and FF-scaling errors.
3) The role of f0 information in FF-scaling estimation.

Methods

Participants: 34 listeners (27 native English speakers).

Stimuli: The stimuli were made up of the vowels /i æ/ of 4000 'voices' which differed from each other in their FFs and/or their f0s in 40 FF levels and 100 f0 levels. The FFs for the lowest FF level are given in Table I. FF levels increased by 1.2% for each level relative to the previous one. F0 levels ranged from 100 to 300 Hz in equal logarithmic steps.

Procedure: Stimuli were arranged on a 900 x 700 pixel response board as in Figure 1. FF levels were spaced 20 pixels apart, while f0 steps were 6 pixels apart. There was a 60 pixel buffer on the horizontal ends and a 53 pixel buffer on the vertical ends of the board to reduce possible truncation effects caused by sudden limits on the response space.

Listeners were played a voice and given a number of guesses to indicate the location. The voice was selected using a random uniform draw from among all stimulus voices. Each time a listener guessed, the voice associated with the location of their guess was played to them. After their allotted number of guesses, the location of the stimulus voice was displayed on the board.

Listeners participated in three blocks organized by number of guesses. In the first block they were allowed 3 guesses, 2 guesses in the second block and a single guess in the final block.

Vowel	F1	F2	F3	F4
/ i /	275	2114	2711	3500
/ æ /	705	1473	2281	3500

Table I. *Formant frequencies (in Hz) for the stimulus vowels representing voices at the lowest FF range level.*

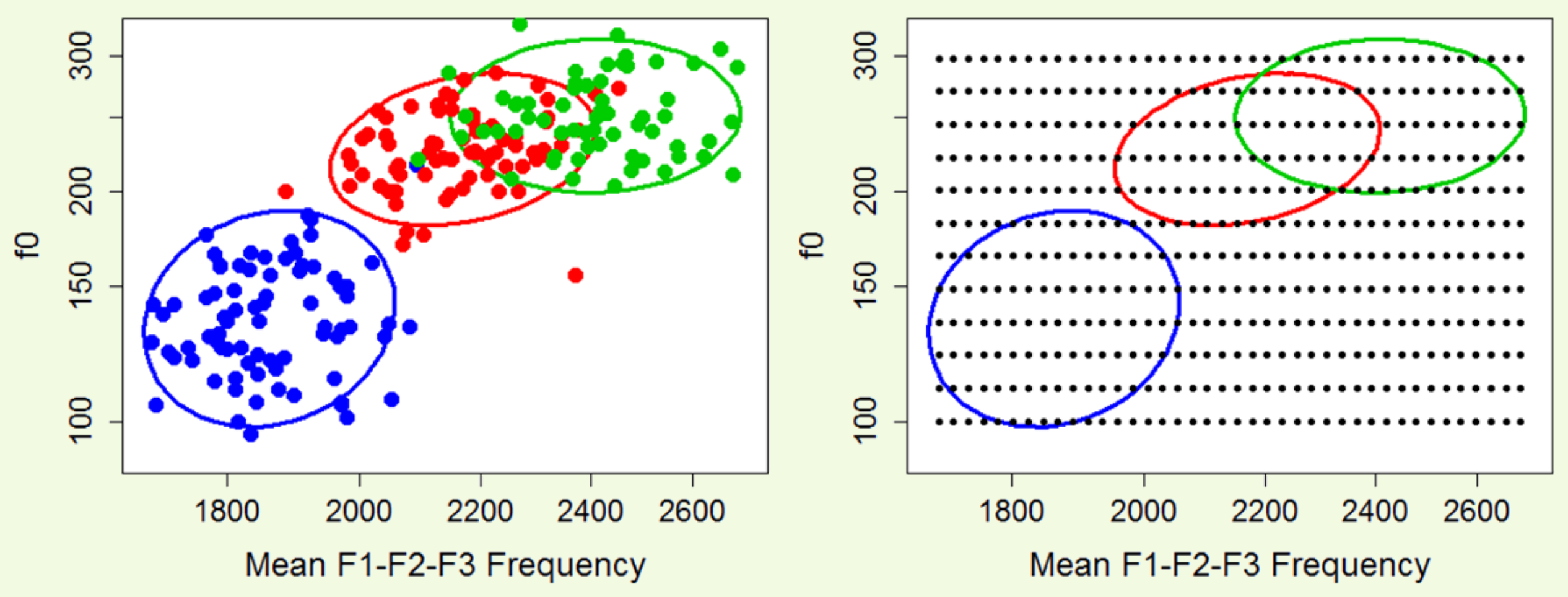


Figure 1 *– (a) The x-axis indicates the mean of the first three formant frequencies for productions of /i/. Ellipses enclose two standard deviations of the distribution of real voices from 215 speakers: adult males (blue), adult females (red), and children (green) [3,4]. The points indicate the locations of the voices of individual speakers. (b) The locations of stimulus voices are indicated by the filled points. To maintain the legibility of the figure, only every 9th f0 level is indicated.*

Results

FF-scaling and f0 reporting accuracy.

- The magnitude of the average absolute FF-scaling error across listeners was 8.3% (min = 5.4%, max = 12.1%).
- The magnitude of the average absolute f0 error across listeners was 15.6% (min = 10.9%, max = 27.3%).

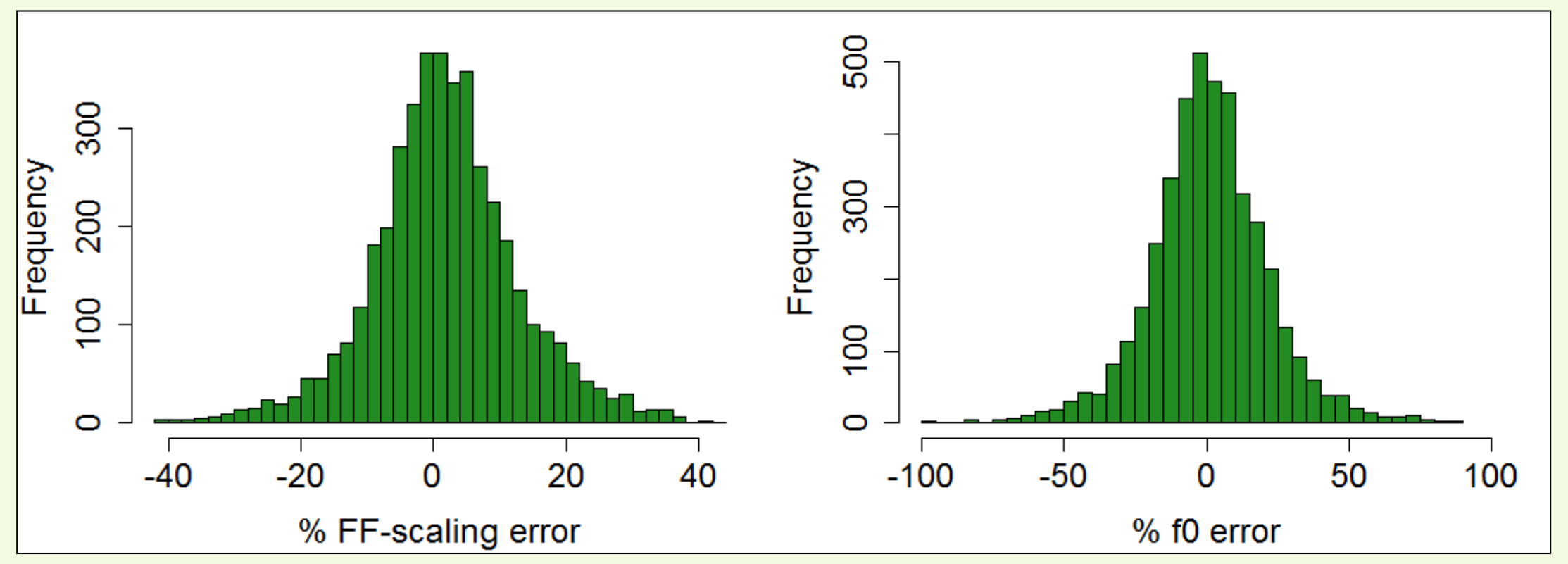


Figure 2. *FF-scaling and f0 reporting errors, pooled across all listeners. Percentage error refers to stimulus values in Hertz.*

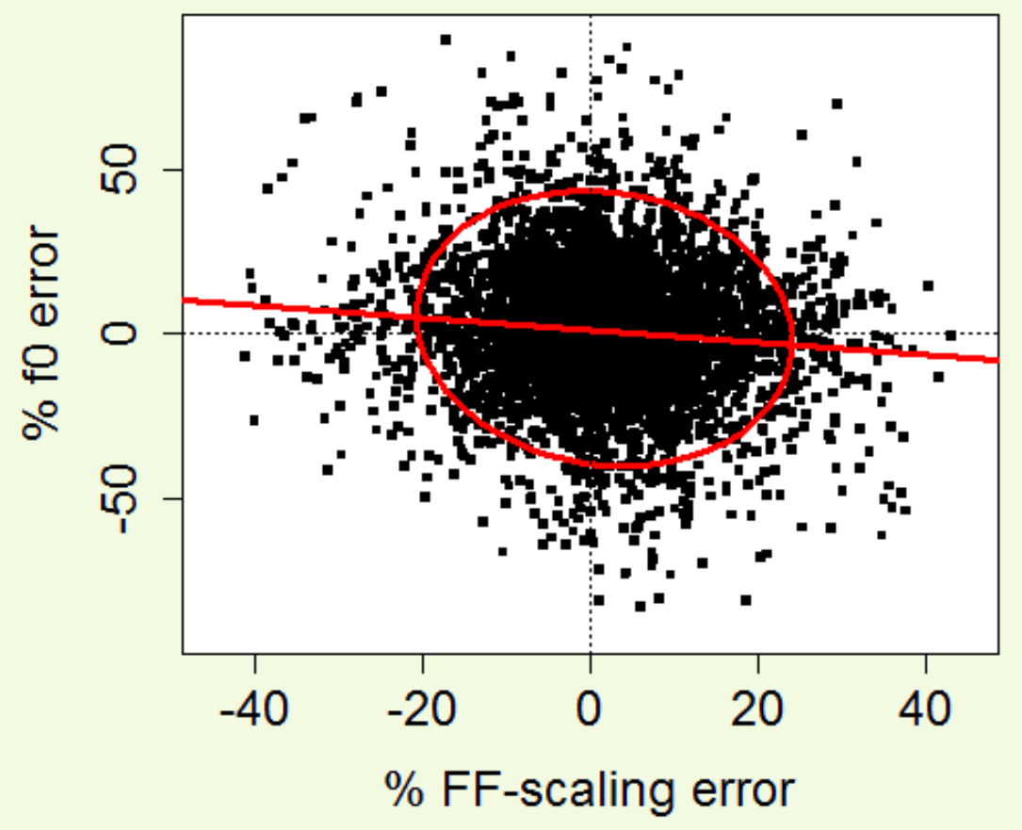
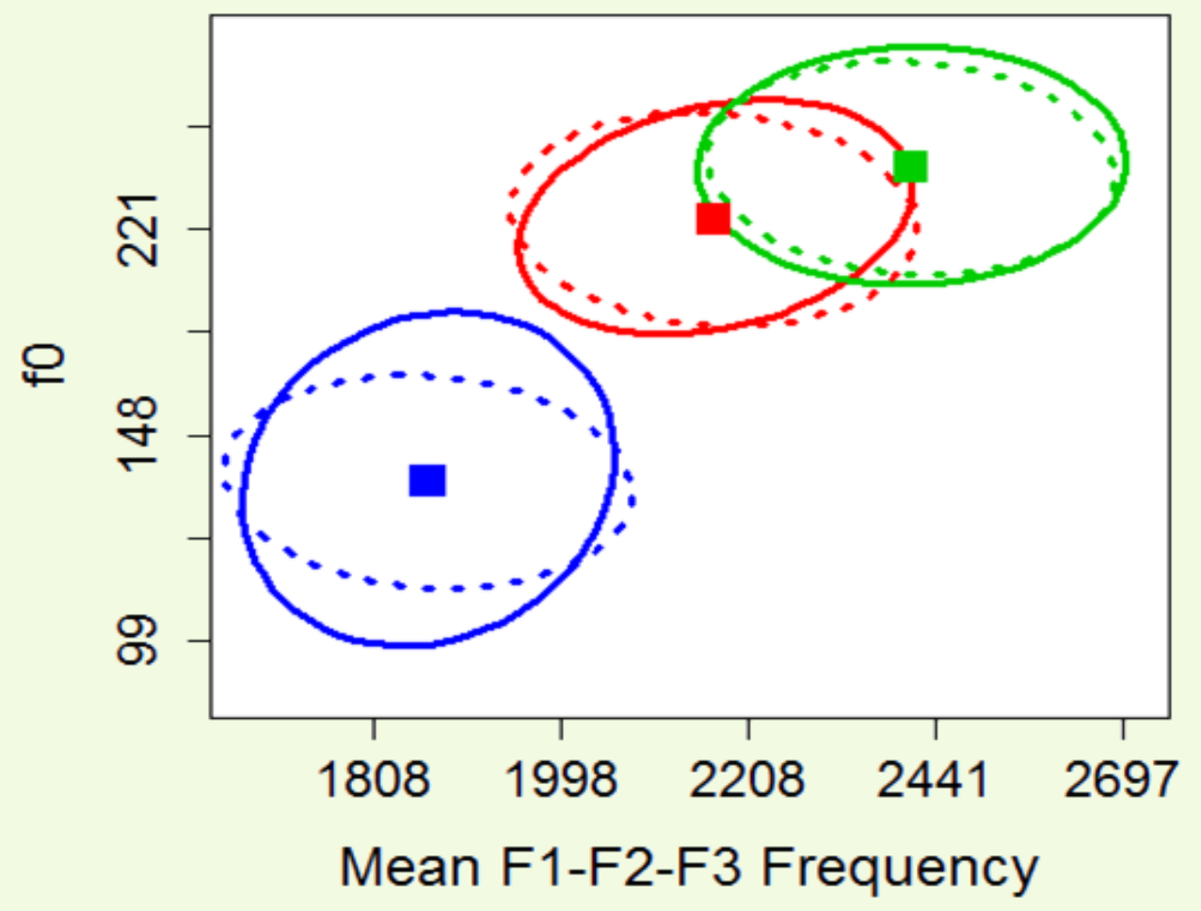


Figure 3. *Each point indicates the location of an individual error, pooled across all listeners. The ellipse encloses 2 standard deviations of pooled errors. The red line indicates the line of best fit relating FF-scaling errors and f0 errors, pooled across all participants*

Figure 4. *The dotted lines indicate a single standard deviation of the error ellipse (from Figure 3), centered about the squares of the same color. These are compared to ellipses indicating the distribution of real voices (as in Figure 1), shown in solid lines.*



FF-scaling and f0 errors.

- The mean, within-listener correlation coefficient between f0 and FF-scaling errors was -0.09 [t(33) = -3.4, p = 0.0017].
- This means that errors were slightly negatively correlated, despite the positive marginal correlation between these properties across all speakers.
- Figure 4 compares the distribution of errors to the voices of a range of speakers.

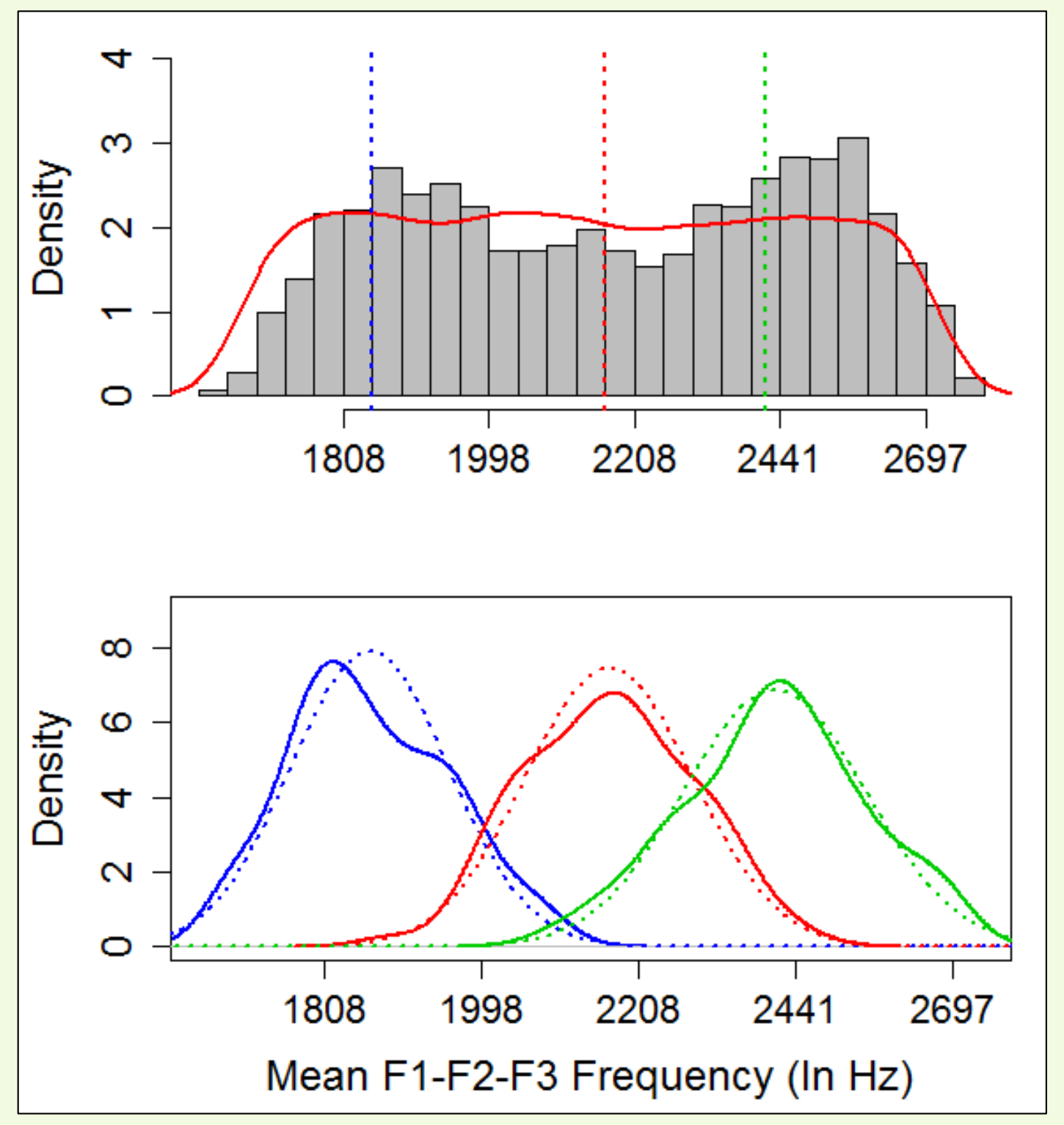
Information used in FF-scaling estimation.

- A random coefficients regressions analysis indicated that that stimulus FF-scaling [t(33) = 25, p < 0.0001], stimulus f0 [t(33) = 0.002] and f0 error [t(33) = -5.1, p < 0.0001] all have a significant effect on judged FF-scaling.
- However, a model fit to the pooled data across all listeners revealed that stimulus FF-scaling explains 44.6% of the variance in judged FF-scaling, while judged f0 and f0 error explain only 0.52% and 0.89% respectively.

Use of prior knowledge regarding the distribution of FF-scaling?

- There has been some debate about how much listeners know regarding the distributional properties of f0 and FF-scaling [5,6].
- The density of *p(reported FF-scaling | actual FF-scaling)* can be estimated from Figure 2, and is approximately normal.
- If guesses were solely based on this information, with no influence for prior knowledge, we would see a pattern like the red line in Figure 5a, which estimates the density of the sample of stimulus voices presented to listeners.
- In contrast, the distribution of responses had 2 (or 3) modes, corresponding roughly to the average locations of the voices of men women and children.
- This suggests that the modes in the histogram of Figure 5a may arise from the influence of prior knowledge about the distribution of FF-scaling between different kinds of speakers.

Figure 5. *(a) Gaussian kernel density estimates of FF-scaling stimuli actually presented to listeners (red) compared to a histogram of FF-scaling responses obtained from listeners. Vertical lines indicate the locations of the means of the normal distributions from the bottom panel. (b) Gaussian kernel density estimates of the distribution of FF-scaling from two large data sets [3,4], organized by speaker type: men (blue), women (red) and children (green). The dotted lines indicate normal distributions with parameters set based on the observed data.*



References

[1] Fant, G. (1960). Acoustic Theory of Speech Production. The Hague: Mouton. pp.107-138.
[2] Barreda, S. and Nearey, T. (2013). Training listeners to report the acoustic correlate of vocal tract length using synthetic voices. Journal of the Acoustical Society of America 133(2): 1065-1077.
[3] Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.
[4] Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97: 3099-3111.
[5] Assmann P.F., Nearey T.M., and Dembling S. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. Proceedings of the Ninth International Conference on Spoken Language Processing, pp. 889-892, Pittsburgh, PA, September 17-21, 2006.
[6] Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (2006). Vowel normalisation: Time-domain processing of the internal dynamics of speech. in Dynamics of Speech Production and Perception, edited by P. Divenyi. Amsterdam: IOS Press. pp. 153-170.

Contact Information

Santiago Barreda (sbarreda@ualberta.ca)

Terrance M. Nearey (t.nearey@ualberta.ca)

Department of Linguistics, 4-32 Assiniboia Hall, University of Alberta, Edmonton, Alberta, Canada T6G 2E7

Conclusion

- Listeners can identify voice FF-scaling with good accuracy .
- These estimates are strongly influenced by stimulus FF and weakly influenced by stimulus f0.
- FF-scaling and f0 reporting errors were slightly negatively correlated.
  - This may reflect listener's guessing strategies (i.e., 'working backwards' from apparent speaker characteristics.).
- There is some evidence that listeners are using knowledge of the distribution of FF-scaling to make their judgments.