Measuring the informativity of F3 for rounded and unrounded high-front vowels in Central Swedish

Anna Persson[1], T. Florian Jaeger[2]
[1] *Swedish Language and Multilingualism, Stockholm University, Sweden*
[2] *Brain and Cognitive Sciences, University of Rochester, USA*
anna.persson@su.se, fjaeger@ur.rochester.edu

Abstract

The Swedish vowel space is densely populated, especially in the high-front part of the space. One hypothesis holds that third formant information (F3) helps distinguish rounded and unrounded high-front vowels. The present study uses a simple model of speech perception to investigate the predicted consequences of including F3 for the perception of the high-front vowels. Ideal observer models were trained on vowel production data under different assumptions about the cue space (F1-F2 vs. F1-F2-F3) and evaluated on how well they predict the category intended by the talker. Results indicate that F3-inclusion facilitates recognition accuracy.

Introduction

The Central Swedish vowel space is densely crowded in the first two formants (F1-F2). It contains a total of 21 vowel categories that differ in quantity—assumed to be primarily cued by vowel duration—and quality—assumed to be primarily cued by formants. Category overlap in F1-F2 space is particularly pronounced among the high-front vowels, highlighted in Figure 1 (Persson, 2024). This has prompted research on what additional acoustic cues might allow listeners to distinguish between these vowels. One long-standing hypothesis holds that the third formant (F3) correlates with lip-rounding, allowing listeners to distinguish rounded and unrounded vowels, such as [iː]-[yː] (e.g., Fant, Henningsson & Stålhammar, 1969; Fujimura, 1967; Kuronen, 2000; for a review see e.g., Rosner & Pickering, 1994).

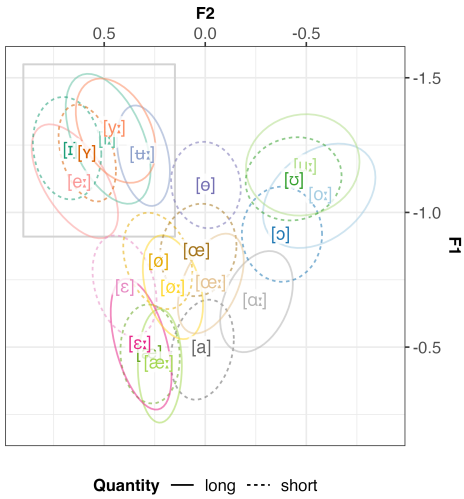

Figure 1. Summary of vowel productions from the *SwehVd* database described in more detail below. High-front vowels that constitute the focus of the present study are highlighted. Adapted for the present purpose from Persson (2024). Ellipses show 95% probability mass of bivariate Gaussians fit to the data.

This question has been approached in a number of ways. One approach has been to qualitative compare the realization of F3 in rounded and unrounded vowels, typically on a small number of recordings. For example, Kuronen (2000) compared the realization of [iː] and [yː] by four talkers, and found that the two vowels tended to differ in their F3 (see also Fant et al., 1969). Another approach has been to measure the effects of F3 on listeners' perception. For example, Fujimura (1967) exposed Swedish listeners to synthesized [iː]-[yː]-[ɯː]

continua, while varying F3. Based on listeners' responses, Fujimura argued that F3 affected listeners' perception. This particular approach has, however, been criticized for confounding multiple cues (e.g., Rosner & Pickering, 1994): while F1 and F4 were kept constant across the synthesized tokens, F2 and F3 covaried inversely. This makes it impossible to conclude that F3 was the cause for changes in listeners' perception.

A third approach—the one we pursue here—is to use computational models of speech perception to assess the *predicted* perceptual consequences of F3-inclusion. For example, Johnson & Sjerps (2021) compared the performance of support vector machines in categorizing US English vowels, either with only as F1 and F2 as input, or while also including F3. This yielded improved recognition accuracy when F3 was included. Another recent study has applied a similar approach to the recognition of Central Swedish vowels. In Persson & Jaeger (2023), we compared fifteen different normalization accounts for the perception of Central Swedish vowels. Critically, we did so under different assumptions about what cues Swedish listeners rely on in categorizing vowels: just F1 and F2 or also additional cues. The results indicated that the inclusion of F3 resulted in equivalent or improvements in recognition accuracy for all fifteen normalization accounts, compared to models based on F1 and F2 alone. This result held across all 21 vowels of Central Swedish.

Here, we follow-up on this work but focus specifically on the high-front vowels—i.e., the vowels for which F3 has been hypothesized to be particularly important. Following Persson & Jaeger (2023), we use a general model of speech perception based on Bayesian inference, ideal observers (e.g., Nearey & Hogan, 1986; for review, see Xie, Jaeger & Kurumada, 2023). We investigate to what extent a model trained on F1-F2-F3 predicts higher recognition accuracy for the high-front vowels, than a model trained on F1-F2. If F3 indeed carries information about high-front vowels, we would expect the model that includes F3 to achieve higher recognition accuracy relative to a model based on only F1 and F2.

Materials and methods

Materials

The materials are a subset of the *SwehVd* database of Central Swedish *h-VOWEL-d* words, recorded in 2020-2024 at Stockholm University (described in Persson & Jaeger, 2023). SwehVd contains recordings and annotations of 44 (24 female) talkers of Central Swedish, born and raised in the Greater Stockholm area, of 18-44 years of age at the time of recording (mean age = 30, SD = 6.82).

All talkers were recorded reading 10 repetitions of all 21 Central Swedish vowels. The database contains F1-F2-F3 measurements for each talker at five different time-points of the vowel (at 20, 35, 50, 65 and 80% into the vowel), as well as vowel duration and mean f_0 across the entire vowel segment.

For the present study, we focus on the cluster of high-front vowels [iː], [ɪ], [yː], [ʏ], [eː], and [ʉː] (see Figure 1). We included all productions of these vowels from all talkers in the database, excluding outliers and mis-pronounced vowels (using the same criteria as described in Persson, 2024).

To obtain a reliable estimate of the formant values at the steady state of the vowel, we obtained the geometric mean across the three mid-points of the vowel (at 35, 50 and 65% into the vowel).

To adjust for inter-talker variability in formant realization, as caused by differences in talkers' vocal tract physiology, formants were normalized using Nearey's uniform scaling account (Nearey, 1978). Uniform scaling was used because it has been found to

provide equally good of better fits against human perception than more complex normalization accounts (Barreda, 2021; Persson, Barreda, & Jaeger, 2024).

Bayesian ideal observers

Ideal observers have been found to provide a good fit against human speech perception (e.g., Norris & McQueen, 2008; Kleinschmidt & Jaeger, 2015; Kronrod, Coppess & Feldman, 2016). Ideal observers describe optimal use of available information, based on previous experience. For categorization, ideal observers describe the posterior probability of a given category as dependent on both the prior probability of the category in the current context, *p(category)*, and the likelihood of the acoustic input under the hypothesis that it pertains to the category, *p(cues|category)*:

$$p(category|cues) = \frac{p(cues|\mu,\Sigma) \times p(category)}{\Sigma_c p(cues|\mu_c,\Sigma_c) \times p(category_c)} \quad (1)$$

Here, we follow the implementation of ideal observers adopted in Persson & Jaeger (2023), including the simplifying assumptions of e.g., uniform priors and multivariate Gaussian category representations. For the 6 vowels included here, *p(category)* = .167. The cue likelihood, *p(cues|category)*, is assumed to be a multivariate Gaussian distribution, defined by the category mean (μ) and the variance-covariance matrix (Σ).

Five-fold cross-validation

To avoid over-fitting to the sample, a five-fold cross-validation approach was adopted to obtain 5 separate estimates of model predictions for each combination of cues: F1-F2 or F1-F2-F3. Following Persson & Jaeger (2023), the data was split by talker and category into five separate folds. For each fold, an ideal observer was fit on four of the folds, and subsequently tested on the fifth held-out fold (using the R package MVBELIEFUPDATR, Jaeger, 2024). This resulted in 2 (cue spaces) * 5 (folds) = 10 ideal observer models.

Results and discussion

Recognition accuracy depending on F3-inclusion

Averaging over the six high-front vowels, F3 improves recognition accuracy of the ideal observer from 55% to 69%. This change was statistically significant, as confirmed by a logistic regression predicting accurate (1) vs. inaccurate (0) recognition as a function of the phonetic space (F1-F2-F3 vs. F1-F2; $\hat{\beta}$ = .129, p < .0001). This result suggest that F3 carries helpful information for the categorization of high-front vowels.

To further assess how the inclusion of F3 affects the recognition of individual vowels, Figure 2 summarizes the recognition accuracy for all six high-front vowels.

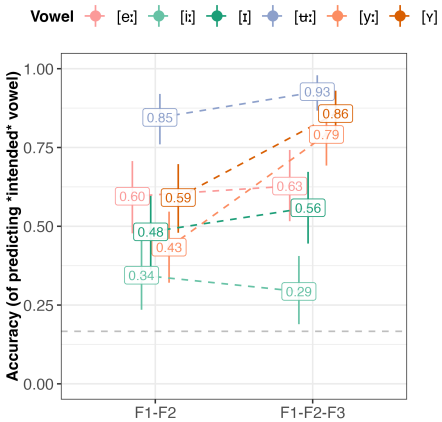

Figure 2. Vowel-specific recognition accuracy of ideal observer in F1-F2 and F1-F2-F3 space. Point ranges indicate the average mean accuracy and average 95% bootstrapped confidence intervals across the five folds. Chance level is indicated by gray line.

Two general observations can be made from Figure 2. First, and perhaps unsurprisingly, some vowels have lower recognition accuracy than others, at least under the simplifying assumption of uniform category priors. Presumably, this is

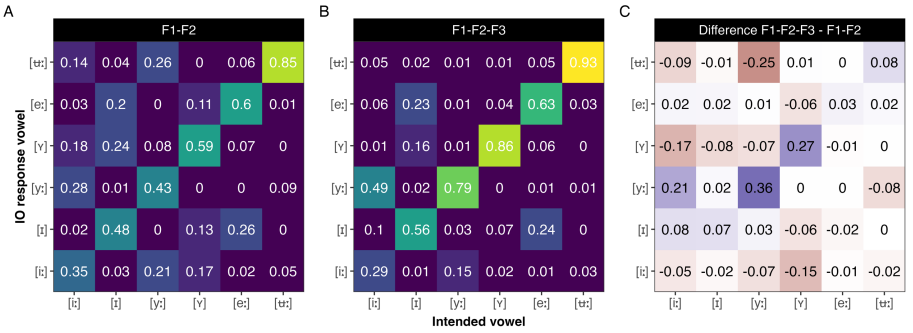

Figure 3. Confusion matrices for ideal observers trained in F1-F2 space (**Panel A)** and F1-F2-F3 space **(Panel B**). Columns show the intended vowel; rows show the recognized vowel. Probabilities sum to 1 in each column, indicating the distribution of responses for each intended vowel (averaged across the five folds). The difference matrix (**Panel C**) illustrates the differences between the Panels A and B. More **purple** indicates an increase in posterior probability for the former (F1-F3) over the latter (F1-F2) model, more **red** indicates an advantage for the latter over the former.

a consequence of their location in the acoustic-phonetic space relative to neighboring vowels. The more overlap, the lower the accuracy by which vowels are predicted to be recognized. For instance, as suggested by Figure 1, [i:] has many close competitors and partly overlaps all other categories considered, and subsequently has the lowest predicted recognition accuracy. This highlights that the recognition of vowels is dependent on their relative rather than absolute position in the phonetic space (e.g., Peterson & Barney, 1952; Kuhl, 1991; Polka & Bohn, 2003).

Second, the overall improvement in recognition accuracy when including F3 seems to be driven by a subset of the vowels. In particular, the rounded long and short [y:] and [ʏ] seem to benefit the most from the inclusion of F3, while the predicted recognition accuracy of [i:] numerically *de*creased in the F1-F2-F3 model. This was confirmed by fitting separate logistic regression models for each vowel. This revealed significant improvements for [y:], [ʏ], and [ʉ:] (all *ps* < .0001). For [i:], [ɪ] and [e:], recognition accuracy was statistically indistinguishable between the F1-F2 and F1-F2-F3 models (all *ps* > .10).

To better understand *how* the inclusions of F3 improved recognition accuracy, we investigated changes in category confusability.

Category confusability

Figure 3 summarizes the category-to-category confusability for the F1-F2 and F1-F2-F3 model.

Panel A of Figure 3 confirms that the low recognition accuracy of [i:], [ɪ], and [y:] in F1-F2 space is primarily due to confusions with their neighboring categories. Panels B and C show that inclusion of F3 decreases category confusability for most vowels. The increase is particularly noticeable for the rounded long and short [y:] and [ʏ], for which F3 seems to carry helpful information.

Figure 3 further suggests that the numerical decrease in the recognition accuracy of [i:] in the F1-F2-F3 model is primarily driven by the increased probability of confusing [i:] with [y:]: when F3 is included, there is a 21% increase in ideal observer responses for [y:] for when the intended category is [i:] (Figure 3, Panel C). One possible explanation for this is that F3 does *not* add helpful information about this particular contrast, and that additional measurement noise associated

with F3 causes the increased [iː]-[yː] confusability.

As discussed in Persson (2024), the centralization of [iː] and [yː] in the *SwehVd* database might suggest an ongoing merger of the two categories in Central Swedish. This merger might be caused by a relaxation of lip-rounding in the long [yː], or indications of talkers producing the two vowels as damped versions. Either of these two explanations might possibly account for the predicted confusability of the two vowels, and the lack of informativity carried by F3 for the contrast.

Conclusions

We assessed the informativity of F3 for high-front vowel distinctions in Central Swedish using ideal observer models of speech perception. This has allowed us to assess whether—and how—the inclusion of F3 affects the recognition accuracy that can be achieved for the high-front vowels. To ultimately establish the importance of F3-inclusion, however, it will be necessary to assess how well the types of models we evaluated here predict listeners' *actual* vowel perception.

Ideal observers offer a comparatively simple computational model that enables researchers to estimate the consequences of different perceptual systems. The use of these models has become substantially easier with the development of freely available R libraries like the one we used here (MVBELIEFUPDATR), which provides functions for fitting, plotting, and evaluating ideal observers and exemplar models.

Acknowledgements

We thank audiences at the *LMU Institute for Phonetics and Speech Processing* in October 2023, at *FiNo* 2024, and at Department of Swedish language and multilingualism at Stockholm University for feedback on earlier presentations of this work. We are particularly grateful to Maryann Tan for collaboration in preparation of the *SwehVd* database.

References

Barreda, S. (2021). Perceptual validation of vowel normalization methods for variationist research. *Language Variation and Change, 33(1),* 27-53. doi: 10.1017/S0954394521000016

Fant, G., Henningsson, G. & Stålhammar, U. (1969). Formant frequencies of Swedish vowels. *STL-QPSR 10(4),* 26-31

Fujimura, O. (1967). On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants. *Language and Speech 10(3),* 181-193. doi:10.1177/002383096701000304

Jaeger, T. F. (2024). MVBeliefUpdatr: Fitting, Summarizing, and Visualizing of Multivariate Gaussian Ideal Observers and Adaptors. R package version 0.0.1.0006, https://github.com/hlplab/MVBeliefUpdatr

Johnson, K. & Sjerps, M. J. (2021). Speaker normalization in speech perception. In Pardo, J. S., Nygaard, L. C., Remez, R.E. & Pisoni, D. B. (eds)., *The handbook of speech perception.* John Wiley & Sons, Inc. doi:10.1002/9781119184096.ch6

Kleinschmidt, D. & Jager, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148-203. doi.org/10.1037/a0038695

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified model of categorical effects in consonant and vowel perception. *Psychological Bulletin and Review, XX*, 1681-1712. doi: 10.3758/s13423-016-1049-y

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics 50*, 93-107. doi:10.3758/BF03212211

Kuronen, M. (2000). *Vokaluttalets akustik i sverigesvenska, finlandssvenska och finska.* Jyväskylä: University of Jyväskylä

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. Indiana

Nearey, T. M. & Hogan, J. (1986). Phonological contrast in experimental phonetics: Relating distributions of measurements in production data to perceptual categorization curves. In: *Experimental Phonology*, eds. Ohala, J. J. & Jaeger, J., New York: Academic Press

Norris, D. & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review 115*, 357-395. doi: 10.1037/0033-295X.115.2.357

Persson, A. (2024). The acoustic characteristics of Swedish vowels. [*manuscript*]

Persson, A., Barreda, S. & Jaeger, T. F. (2024). Comparing accounts of formant normalization against US English listeners' vowel perception. *[manuscript]*

Persson, A. & Jaeger, T. F. (2023). Evaluating normalization accounts against the dense vowel space of Central Swedish. *Frontiers in Psychology, 14,* 01-21. doi:10.3389/fpsyg.2023.1165742

Peterson, G. E. & Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America 24,* 175-184

Polka, L. & Bohn, O-S. (2003). Asymmetries in vowel perception. *Speech Communication 41,* 221-231

Rosner, B. S. & Pickering, J. B. (1994). *Vowel Perception and Production.* Oxford: Oxford Psychology Series 23

Xie, X., Jaeger, T. F. & Kurumada, C. (2023), What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review. *Cortex 166,* 377-424.