Comparison of several proposed perceptual representations of vowel spectra

Terrance M. Nearey† **and Michael Kiefte**‡

† University of Alberta, Canada

‡ Dalhousie University, Canada

E-mail: t.nearey@ualberta.ca, mkiefte@dal.ca

ABSTRACT

We report some results of modeling the categorization of a three-formant synthetic vowel continuum by speakers of English and of Finnish. We focus here on assessing the relative merits of initial representations based on a standard 3-dimensional formant-frequency space compared to those based on 1) a 2-dimensional F1 by F2-prime space, and on 2) Hermansky's PLP5. PLP is a 5-dimensional representation of spectral shape that manifests an integration of closely spaced formants in a manner suggestive of F2-prime. Linear logistic models suggest that PLP5 modestly outperforms 3 formants, which in turn substantially outperforms F1 by F2-prime. However, if quadratic rather than linear methods are allowed, the 3 formant space provides a better account of listeners' response patterns than do other models of comparable complexity. This result is confirmed using modern model selection techniques.

1. INTRODUCTION

Two of the most persistent issues in vowel perception are 1) whole-spectrum versus formant-like coding of stimulus information and 2) presence or absence of large scale integration (e.g., 3.5 Bark 'centers of gravity') of spectral energy in the F2 to F4 range, resulting in phenomena related to F2-prime. We are engaged in an ongoing study of perceptual categorization of a large vowel continuum to address aspects of these issues. Our modeling results so far are consistent with a simple formant-based representation that preserves F1, F2 and F3 related information as separable quantities.

2. THE PERCEPTUAL EXPERIMENT

A synthetic vowel continuum of 972 stimuli was constructed, spanning an F1 x F2 x F3 space satisfying the following criteria: F1 ranged from 250 to 760 Hz in 10 steps, F2 from 750 to 2260 Hz in 15 steps and F3 from 1350 to 3080 Hz in 12 steps. Step sizes for all three formants were all approximately equal to 0.5 Bark. All combinations of these steps were generated subject to the following constraints: adjacent formant separations (F3-F1 and F2-F1) must be at least 350 Hz and the sum of F1 + F2 must not exceed 2650 Hz. Those constraints lead to a reasonable approximation of the feasible three-formant space of a typical adult male. The vowels were synthesized digitally at 10 Hz using the CSRE3 synthesis software in cascade mode. Vowels were 230 ms in duration with a linearly falling F0 contour from 125 to 100 Hz.

Responses were collected from 14 speakers of Canadian English at the University of Alberta and from 14 speakers of Finnish at the University of Turku. Using specially designed software appropriate to each language, listeners selected from an inventory of 10 vowels in English [i ɪ e ɛ æ ɒ o ʊ u ɚ] or from an inventory of 8 vowels in Finnish [i e y ø æ ɑ o u].

3. INITIAL MODELING RESULTS

Our first attempts at modeling fitted linear logistic regression equations to the pooled responses of each listener group. We investigated four parametric representations of the stimulus space. 1) *F123,* a 3-dimensional formant frequency representation using the nominal formant targets used in synthesis [F1, F2, F3], transformed to Bark. 2) *F1F2p*, a 2-dimensional F1 by F2-prime space. 3) Hermansky's 5-dimensional *PLP5* [2]. 4) *F12*, a simple two-dimensional formant representation with no information about F3. The formula of Bladon and Fant [3] was used to calculate F2-prime and subsequently F1 and F2-prime were transformed to Bark. Hermansky's 'perceptual linear prediction' method calculates LPC cepstral coefficients from an approximation of a Bark-sone scaled amplitude spectrum. We calculated PLP5 from analysis of mid-vowel sections of the stimuli, using an implementation of algorithms described in [2]. Hermansky emphasizes that this representation can encode no more than two complex poles, resulting in a kind of two-formant approximation of the spectrum that bears some affinities to large-scale (3.5 Bark) integration bands and to F2-prime [2].

The four stimulus representations were used in (polytomous) linear logistic regression analyses of the pooled responses [4]. Two statistics will be reported representing the overall goodness of fit of the models. *G2* represents the likelihood ratio chi-square that is optimized in the fitting process and *error rms* represents the standard deviaition of errors in percentage points. In general, the

ISBN 1-876346-48-5 © 2003 UAB

lower the scores, the better the fit of the model. Table 1 shows the results. Model size is the number of parameters used by the model. It is equal to the product (*D* +1) (*V* - 1), where *D* is the number of dimensions of the stimulus representation, and *V* is the number of vowels in the language.

Model	*G2*	*Error (rms)*	*Model size*	*D*
	English			
1. F123	20151.9	5.93	40	3
2. PLP5	16110.8	5.19	60	5
3. F1F2p	46304.1	10.68	30	2
4. F12	44297.0	10.63	30	2
	Finnish			
1. F123	8206.1	4.75	28	3
2. PLP5	6462.56	4.08	42	5
3. F1F2p	17127.9	8.10	21	2
4. F12	14662.2	7.57	21	2

Table 1. Summary of initial modeling in two languages.

Three points are worth noting in Table 1. *First,* the 2-dimensional *F1F2p* representation performs very similarly (and actually slightly worse) than F1 and F2 alone. *Second,* these two representations fare much worse than the others. Error rms values for models 1 and 4 are at least 80% larger in English and about 60% larger in Finnish than for models 2 and 3. *Third,* the larger the model, the better the fit. The 5-dimensional PLP model fits slightly better than the 3-dimensional F123 model. All other things being equal, more free parameters generally lead to better optimized fits, and we run the familiar risk of model overfitting [6, 7].

4. MODEL COMPARISON USING AIC

A number of methods have been sugested for comparing models 'more fairly' across model sizes. We will use two main varieties here. The first is Akaike's information criterion or AIC. It is defined as the log likelihood of a model *plus* (i.e. penalized by) two times the number of fitted parameters. (We substitute G2 which differs from other log-likelihood measures only by a constant which depends on the data but not on the model). For *individual* (but not pooled) data, it seems reasonable to assume that responses to each stimulus follow multinomial distributions. We adapt here methods based on work of Maddox, Molis and Diehl [5] in comparing models via AIC for models fitted to individual listeners' data.

The rows of Table 2 summarize the mean AIC (across 14 subjects per language) of 15 different models that were investigated. The first four models are the same as those in Table 1. (Others are discussed in section 5.) The ordering for the AIC in Table 2 is similar to that of G2 in Table 1. For both English and Finnish, PLP5 fits best, followed by F123, while F1F2p and F12 models fit considerably less well. The order of the goodness of fit is reversed between the F1F2p versus the F12 models for the English speakers.

The columns labeled 'Rank' show the mean rank of the models among all 15 tested. This was calculated by ranking all 15 models separately for each listener, then calculating the mean ranking for the model across listeners. In general the rankings were fairly stable across listeners, with average SD about 0.7 and the maximum SD about 2.0. Not surprisingly, the relative ordering of the ranks within the 15 models is the same for models 1 to 4.

		English		*Finnish*	
Model	*D*	*AIC*	*Rank*	*AIC*	*Rank*
1. F123	3	3265.52	13.0	1947.40	13.0
2. PLP5	5	2975.84	9.4	1753.25	8.7
3. F1F2p	2	5287.32	14.4	2794.89	14.9
4. F12	2	5352.51	14.6	2503.23	14.1
5. QF123	9	2358.33	1.0	1450.44	1.1
6. F123S12	5	2759.46	6.0	1694.02	6.7
7. F123S13	5	2931.38	8.4	1692.99	6.9
8. F123S23	5	2863.55	7.4	1764.53	8.6
9. F123S1	4	3025.87	10.4	1792.94	9.6
10. F123S2	4	2977.83	9.5	1853.12	11.0
11. F123S3	4	3168.94	11.9	1849.12	11.1
12. [1]+[2]	8	2673.14	5.1	1627.78	5.4
13. [5]-X13	8	2409.11	2.6	1486.56	2.8
14. [5]-X23	8	2403.02	2.5	1470.82	2.1
15. [5]-X12	8	2520.46	3.9	1547.65	3.9

Table 2. Comparison of 15 models on the AIC.

5. MODEL COMPARISON USING CROSS-VALIDATION

Although AIC is a 'fairer' criterion than raw G2, it is still somewhat biased toward the selection of larger models. [5, 6]. An alternate approach to model selection is cross-validation. The method adopted here is a 'leave-out-one subject' cross-validation, referred to in [8] as round-robin cross-validation. Within each language, for each of the 14 subjects in turn, a model is trained on the 13 other subjects. The model parameters are used to provide predicted response probabilities for the subject who was withheld from training. The G2 measure reported in Table 3 is the mean of the 14 cross-validated G2 measures for each subject. The rank is calculated analogously to the calculation of the ranks of the AIC measure described in section 4. (The ranks are not as stable as in the case of the AIC. The SDs of ranks average 2.4 for English and 2.6 for Finnish.)

Confining attention to the first four rows, we see that the ranking of the models is the same as for AIC. There are two key observations. First, F1F2p representation is not very good. Including F3 as a separate variable clearly helps predictions. Second, PLP5 does not seem likely to be advantaged over F123 solely by its larger model size, judging by either the AIC or cross-validation criterion. A cautionary note is in order here, however. It is also known that leave-one-out cross-validation methods are biased toward selecting larger models [6].

		English		*Finnish*	
Model	*D*	*G2*	*Rank*	*G2*	*Rank*
1. F123	3	6954.31	10.9	3849.99	10.9
2. PLP5	5	6737.37	9.8	3722.61	6.7
3. F1F2p	2	8415.36	14.0	4477.77	14.5
4. F12	2	8582.51	14.7	4277.53	14.3
5. QF123	9	6434.78	2.8	3693.27	3.4
6. F123S12	5	6599.64	6.7	3796.85	8.6
7. F123S13	5	6741.43	8.7	3779.24	7.3
8. F123S23	5	6716.86	7.2	3762.99	7.2
9. F123S1	4	6750.80	9.4	3824.85	9.5
10. F123S2	4	6744.31	7.9	3812.72	10.5
11. F123S3	4	6942.89	10.4	3804.43	9.0
12. [1]+[2]	8	6549.23	5.6	3706.42	5.1
13. [5]-X13	8	6462.88	3.6	3686.18	3.7
14. [5]-X23	8	6444.59	3.4	3692.38	4.0
15. [5]-X12	8	6954.31	10.9	3735.88	5.4

Table 3. Leave-out-one-subject cross-validation results.

6. NONLINEAR MODELS IN F1 F2 F3 SPACE

Granting that PLP5 is better than F123, the question arises why this might be so. It is true that PLP5 can model only two complex poles (so at most two formant-like peak frequencies could be derived). However, PLP5 coefficients are 5-dimensional, not 2-dimensional. If we project PLP5 onto a 3-dimensional subspace consisting of the largest three principal components of the covariance matrix of the stimuli, we find that the results are not as good as for three formants. For the reduced-space PLP analysis, English G2 is 23125.9 (rms 6.90) compared to 20151.9 in Table 1. for the F123 analysis.. For Finnish, G2 is 9961.43 (rms = 5.71), compared to 8206.1 for F123.

Furthermore, reasonably good approximations to the first three formants may be obtained from cepstral coefficients derived from standard LPC analysis [9]. Regression analysis based of the current stimuli reveal that linear functions of the PLP5 coefficients predict Bark-scaled formants very accurately. (R^2 values of .99, .95 and .90 respectively for predictions of F1, F2 and F3.) Thus, there exists a 3-dimensional linear subspace of the 5-dimensional PLP representation that is very good approximation of the F123 space.

Since spectra of the original stimuli are functions of the 3 formant control parameters used to generate the stimuli, PLP is actually a 5-dimensional nonlinear expansion of the 3-dimensional control space. Nonlinearities arise both in the conversion of formant frequencies to a waveform and in the subsequent operations described in [2] for calculating PLP5 coefficients.

Such a nonlinear expansion can be viewed from the perspective of artificial neural networks. One way of understanding the operation of the popular multilayer perceptron is to view its output layer as a linear classifier that sits atop hidden layers which serve generate a series of (typically) nonlinear functions of the input variables. The linear partition of the (typically higher-dimensional) transformed space projects to non-linear boundaries of the original space. Rather than training hidden layers, Pao [10] advocates the use of *functional link nets*. These are single-layer perceptrons, for which stimulus representations have been enriched with explicit pre-defined nonlinear functions of the original stimulus space. Linear logistic regression applied to PLP5 would qualify as functional link net applied to the 3-formant control space.

Alternately, we can supplement F123 representation with quadratic functions of the original stimul, i.e. with squares and cross products of the formant measures. A complete quadratic expansion of the formant space leads to 9 stimulus terms, 3 original formants plus 3 squared formants plus three pairwise cross products, F1 x F2, F2 x F3 and F1 x F3. This 9-dimensional representation is shown as QF123 (model 5) in Tables 2 and 3. We see that it performs substantially better than any other of the models discussed so far both AIC and cross-validation measures.

Since both AIC and (leave-one-out) cross-validation are known to lead to selection of overly large models, it seems safest to compare PLP5 to quadratic formant models of comparable dimension. We have done this in models 6 through 8 in Tables 2 and 3. These models include F1, F2, and F3 supplemented by the squares of two formants, whose numbers are indicated following the letter 'S'. For example F123S12 uses 5 dimensions F1, F2, F3, $F1^2$ and $F2^2$. For the AIC measures in Table 2, the 5-dimensional formant representations in rows 6 through 8 outperform PLP5 (model 2) in English and Finnish. The story for cross-validation in Table 3 is mixed. In English, the formant models are better, while PLP5 wins for Finnish.

Rows 9 through 11 of Tables 2 and 3 assess formant representations enhanced by only one squared formant ($F1^2$, $F2^3$ or $F3^2$, respectively). PLP5 beats all three models on AIC for English and on both AIC and cross validated G2 for Finnish. However, for English, models 9 and 10 (with $F1^2$ and $F2^2$ added respectively) beat PLP5 in Table 3. Overall, restricted quadratic formant models of the same dimension as PLP5 are about as good, while smaller ones are slightly inferior. We also explored several other larger models. Model 12 represents the combination of F123 and FPLP parameters. Models 13 through 15 are versions of the full quadratic model with one cross-product term eliminated. Thus Model 13, labeled as '[5]-X13' represents model 5, QF123, with the F1 x F3 cross-product variable removed. We see that all such 8 dimensional models perform better than model 12 on both AIC and cross-validated G2 for both English and Finnish. This suggests that quadratic expansion is a more efficient way to supplement the three formant measures than the complex nonlinear PLP5 method.

But why should we contemplate quadratic terms in a classification model in the formant space? Is this just an admission that a formants are inadequate? Not necessarily. A linear logistic analysis of a three formant space is capable only of generating linear (planar) boundaries between any pair of categories. Planar boundaries are optimal classification of natural stimuli only if certain strict conditions are met by the distributions of measurements. This includes data that follow multivariate normal distributions with distinct means for each vowel, but with an identical covariance pattern for all vowels. But what if (co)variances are not homogeneous? Then quadratic boundaries are required for optimal classification. In fact, data from the classic Peterson and Barney study available through [11] show fairly strong evidence of heterogeneity. Box's M test for equality of covariance matrices shows highly significant differences ($p<.001$) across vowels for three formant (Bark-scaled) data from the male speakers.

7. FUTURE PLANS

We plan to expand this research in two directions. First, we hope to apply the computationally intensive bootstrap methods developed by Shao [see 6] on some of these models. We have done some preliminary work with some of the models discussed here using leave-out-*k* subject cross-validation, where *k* subjects are used for training and *n-k* for testing. Shao's work suggests that *k* needs to be a substantial proportion (e.g., 2/3) of *n* for this to lead to unbiased model selection. However, it is suggested in [7] that such a high proportion may lead to bias toward models that are too small unless *n* is fairly large. Our work so far on models 1, 2, 5 and 11 reveal suggest the same ordering of results as AIC. We are also exploring whether we can discover an explicit 2-dimensional representation that fares anywhere near as well as the baseline 3-dimensional F123 model. So far we have not found one and we strongly suspect that none exists.

8. CONCLUSIONS

Our conclusions thus far are four. *First*, explicit F1 by F2-prime models do not account well for vowel classification in a three-formant continuum. We strongly expect that suggestions like those in [12] will not pan out. *Second,* a 3-dimensional formant space provides a remarkably good approximation to listeners' behavior in these two languages, even with only linear logistic methods. *Third,* more improvement results from using quadratic expansion of the 3-dimensional space than is afforded by PLP5. *Fourth*, nonetheless, PLP5 represents a very creditable method of representing the stimulus space for these experiments. We certainly plan to keep it in our toolbox, especially for application to problems where automatic formant extraction may prove infeasible.

ACKNOWLEDGEMENTS

This research was supported by SSHRC. We would like to thank Ulla Fischer and Olli Aaltonen and his colleagues at the University of Turku for their assistance in carrying out the Finnish experiments.

REFERENCES

[1] Rossner , B. and J. Pickering.Vowel perception and production., Oxford: Oxford University Press, 1994.

[2] Hermansky, H. Perceptual linear predictive (PLP analysis of speech. *J. Acoust. Soc. Amer., 87*(4), pp. 1738-1752, 1990.

[3] Bladon, A.. and G. Fant. A two-formant model and the cardinal vowels. *Speech Transmission Laboratory Quarterly Progress Report*, vol. 1, pp. 1-8. 1978

[4] Nearey, T. The segment as a unit of speech perception. *Journal of Phonetics,* vol. 18, pp. 347-373, 1990.

[5] Maddox, W., M. Molis and R. Diehl. Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception and Psychophysics, 64*(4), pp. 584-597, 2002

[6] Shao, J., & Tu, D. *The jackknife and the bootstrap*. Berlin: Springer Verlag, 1995.

[7] Davison, A. and D. Hinkley, *Bootstrap methods and their application*. Cambridge: Cambridge University Press, 1997.

[8] Watson, C. and J. Harrington, J. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *J. Acoust. Soc. Am.,* vol. 106, pp. 458-468, 1999.

[9] Broad, D. and F. Clermont, Formant estimation by linear transformation of the LPC cepstrum. *J. Acoust. Soc. Am.*, vol. 86, pp. 2013-2017, 1989

[10]Pao, Y.-H. *Adaptive pattern recognition and neural networks*. Reading MA: Addison-Wesley, 1989.

[11]Carnegie-Mellon University. Peterson and Barney: Vowel formant frequency database. http://www–2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/speech/database/pb/, 1995.

[12]Bladon, A. Two formant models of vowel perception: shortcomings and enhancements. *Speech Communication*, vol. 2, 305-313, 1983.