A Neural Network Approach to the Dimensionality of the Perceptual Vowel Space

Terrance M. Nearey[1] and Michael Kiefte[2]
1) Dept. of Linguistics, University of Alberta, Edmonton, AB, Canada, T6G 2E7, t.nearey@ualberta.ca
2) School of Human Comm. Disorders, Dalhousie University, Halifax, NS, Canada, B3H 1R2, mkiefte@dal.ca

1. INTRODUCTION

The question of the dimensionality of the perceptual vowel space for monophthongs has a long history (see [1] for a review). Although three or more formants are typically used to synthesize acceptable vowels, under some circumstances a smaller number of spectral prominences (one or two) can produce acceptable vowel quality. The main question we address is: Can a two-dimensional perceptual space, corresponding roughly to F1 and F2-prime adequately represent the perceptual properties of vowels? F2-prime is believed by some to result from large-scale perceptual integration (e.g., 3.5 Bark 'centers of gravity') of spectral energy in the F2 to F4 range [2]. Others are skeptical of this notion; see [1]. We sketch below a novel method to examine the degree to which a two- or three-dimensional space can accurately represent listeners' perception of a large three-formant vowel continuum. Our results suggest that no two-dimensional representation adequately accounts for listeners' behavior.

In prior work [3], we modeled the categorization of a large (972 stimuli) F1-F2-F3 continuum by 14 speakers of English and 14 speakers of Finnish. Briefly, the stimuli filled a feasible F1-F2-F3 space of an adult speaker in steps of 0.5 Bark on each formant. English speakers responded to each stimulus with one of 11 possible vowel choices, including the rhotic vowel as in the word *her*, which is characterized by a low F3. Finnish speakers responded with one of 8 possible vowel choices, representing the full inventory of short monophthongs in Finnish. Details of the stimuli and procedures are given in [3]. Our prior modeling [3] used several fixed representations of the stimuli. Notably, a two-dimensional F1 by F2-prime representation performed markedly worse than F1-F2-F3 in predicting listeners' categorization of the stimuli via logistic regression. Although the explicit F1 by F2-prime representation of Bladon and Fant [2] is clearly inferior to the three-formant representation, the more general question remains whether some other two-dimensional space is adequate.

2. METHOD

The key to our analysis is a neural network architecture that is capable of implementing an optimal, non-parametric, two-dimensional representation of the stimulus space. This model is rooted in an input-layer with a saturated 'dummy-variable' coding of the 972-point stimulus space, i.e. on presentation of stimulus k, the activation of input node *k* is set to 1.0 and activations of all 971 other nodes are set to 0.0. This input layer fans in to a two-node hidden layer. From this two-dimensional bottleneck, it then 'fans out' again to a group of 11 vowel response output nodes for English and a group of 8 vowel output nodes for Finnish. Each group of output nodes is provided with a softmax transfer function (see [3] for references). This effectively implements a polytomous logistic regression of the response patterns (pooled over listeners) of each language on a common set of two derived stimulus variables. A sketch of the model with a two-dimensional hidden layer is shown in Figure 1.

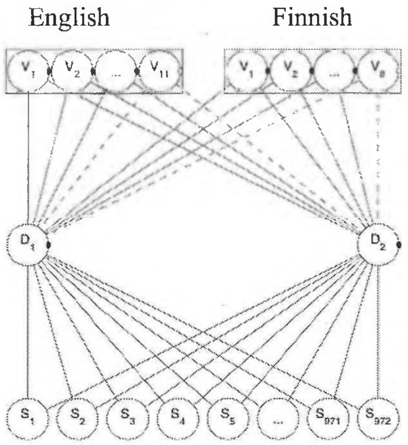

Fig. 1. Sketch of two-dimensional bottleneck neural net. The bottom input layer represents the stimuli. The middle hidden layer represents the two-dimensional reduced space. The top output layer represents listeners' responses.

The weights from the input layer to each hidden unit can be viewed as scores on a single dimension and the trained weights for each stimulus represent the analog of factor scores in multidimensional scaling space of that stimulus. Simulations of the two-dimensional structure were run using random initial weights. To aid convergence, the initial input-to-hidden layer weights were a mixture of 25% random normal deviates and 75% standardized (to zero mean and unit variance) values of F1 to unit D_1 or 75% standardized values of Bladon and Fant's F2-prime to unit

D_2. (Frequencies were transformed to Bark before standardization). Because of the high dimensionality of the problem and the possibility of 'stalling', 200 different random initializations were run. We report the best results of the 200 starts below. A similar set of analyses was run with a three-dimensional hidden layer. Here, convergence properties were somewhat better, so complete random initialization was used with 200 starts.

3. RESULTS

Initial results summarized suggest that two-dimensional bottleneck does not provide a very good account of listeners' categorization, while a three-dimensional hidden layer works quite well.

Table 1. Comparison of goodness of fit of four models. See text.

Model	*rms%*	*Error*	N_p
I. F12p	9.3%	434.4	51
II. Opt2D	8.3%	338.5	1995
III. F123	5.3%	204.5	68
IV. Opt3D	4.3%	167.5	2984

Table 1 presents a comparison of four models. Model I, *F12p* represents a simple softmax model with two inputs F1 and F2-prime. The architecture of the model is equivalent to the top two layers of Figure 1, with F1 and F2-prime applied to the two input nodes. The column labeled *rms%* shows the rms error of predictions compared to observations when responses are measured in percent. The column labeled *Error* is the error criterion that was optimized in the network fitting process. (This error was calculated on normalized proportions of responses, rather than raw response counts. If counts had been used, as in the modeling in [3], this number would have been approximately 69.5 times the value shown.) The column labeled N_p is the number of (non-redundant) free parameters required to fit the model. Although some of these numbers are quite large, it should be noted that there are 16,524 degrees of freedom in the response data. Models I through IV thus exhaust about 0.3, 12, 0.4 and 18 percent of the available degrees of freedom

Model II represents the optimal two-dimensional solution corresponding to the architecture of Figure 1. Comparing models II and III, we see that there is only a modest gain of about one percentage point in rms error with an additional 1944 degrees of freedom. Model II represents a baseline three-dimensional solution. It is similar to model I except that there are three input nodes, corresponding to the synthesis control parameters F1, F2 and F3. We see that there is about a four-percentage point reduction in rms and roughly a halving of the softmax error compared to model I. This large gain obtains with only 17 more free parameters. Model IV corresponds to an optimal three-dimensional solution. This model is like that of Figure 1, except that a third unit is added to the hidden layer. It is the largest model and fits the best of all.

4. DISCUSSION

The fact that the optimal two-dimensional model II is superior to the much smaller two-dimensional model I is not surprising because of the large increase in the flexibility of the model. Similar remarks apply to model IV and model III. Deciding whether the enormous increase in degrees of freedom is justified for the gains observed constitutes a difficult problem in model selection, especially since the problem represents a case of repeated measures categorical data. However, the comparison between models II and III is a much simpler matter. Despite the vastly greater number of degrees of freedom available to model II (which enables an optimal non-parametric, two-dimensional mapping of the stimuli), the two-dimensional bottleneck of the hidden layer apparently makes it impossible to provide a good fit to the data. The fact that model II provides a substantially poorer fit to the responses that the default three-dimensional (F1, F2 and F3) representation of model III leaves little doubt that the dimensionality of the perceptual space underlying vowel perception is greater than two and that no modification of the F1 by F2-prime space can adequately serve as a basis for modeling listeners' categorization.

REFERENCES

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

[2] 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.

[3] Nearey, T. and M. Kiefte. Comparison of several proposed perceptual representations of vowel spectra. *Proc. 15th Int. Cong. Phonetic Sciences,* vol. 1, pp. 1005-1008. 2003.

ACKNOWLEDGEMENTS

This work was supported by SSHRC. Thanks to U. Fischer, O. Aaltonen and their colleagues at the University of Turku, Finland for their help in the collection of the Finnish data.