Formant transitions as partly distinctive invariant properties in the identification of voiced stops[1]

T. M. Nearey and S. E. Shammass
Department of Linguistics, University of Alberta
Edmonton, Canada T6G 2E7

Abstract:

The F2 trajectories for /b, d, g/ in /CVd/ syllables are often summarized by the initial F2 frequency (F2i) and that of the "steady-state vowel" (F2v). Trajectories were measured for 660 Canadian English /CVd/ syllables (3 stops x 11 vowels x 10 speakers x 2 repetitions). Plots for each stop (vowels pooled) indicated a strong linear relationship between F2i and F2v. A regression line fitted to each plot represents an invariant relational property of the corresponding consonant. F2 trajectories are not sufficient to specify the stops uniquely, since the lines for the three consonants intersect (indicating category overlap). However, the slopes and intercepts for the three consonants are distinct and thus represent *partly distinctive invariant properties* or *partial invariants*. Similar patterns obtain for F3. Use of partial invariants of F2/F3 trajectories in a classification algorithm (based on minimum distance from category regression lines) result in an identification rate of over 70%, which compares favorably with a number of other statistical classification schemes. Possible extensions of this approach and relationships to aspects of perception are discussed.

Sommaire:

On représente souvent les trajectoires de F2 pour /b,d,g/ en syllabes du type /CV/ par la fréquence initiale de F2 (F2i) et sa fréquence dans l'état stable de la voyelle (F2v). Les graphiques obtenus pour chaque occlusive (avec regroupement de voyelles) à partir des mesures effectuées sur 660 syllabes du type /CVd/ en anglais canadien (3 occlusives x 11 voyelles x 10 locuteurs x 2 répétitions) révèlent sans équivoque un rapport linéaire entre F2i et F2v. Pour chaque graphique, la droite de régression représente une propriété relative invariante de la consonne, bien que les trajectoires ne soient pas suffisantes pour décrire les occlusives de façon non-ambigüe, puisque les droites de régression des trois consonnes se coupent (ce qui indique un chevauchement des catégories). Cependant, les pentes de ces droites et leurs points de rencontre avec les axes de coordonnées ont des valeurs distinctes et par conséquent représentent des *propriétés invariantes partiellement distinctives* ou *invariants partiels.* On note des résultats semblables pour F3. L'utilisation d'invariants partiels pour les trajectoires de F2/F3 dans une classification algorithmique (basée sur la distance entre chaque point et les droites de régression des trois catégories) aboutit à un taux d'identification de plus de 70%, résultat qui s'avère au moins aussi bon que ceux obtenus par plusieurs autres procédés statistiques de classification. L'article se termine par une discussion des ramifications possibles de cette approche et de ses rapports avec des problèmes de perception.

Introduction

The purpose of this study is twofold: 1) To show that vowel-dependent variation in the onsets of F2 and F3 transitions in stop+vowel syllables is systematic. 2) To show how this systematic variation can be exploited in a pattern recognition model for place in voiced stops. Variation in the onsets of CV transitions as a function of both the

consonants and the vowels involved has been well documented (e.g., Fant 1973, see also Shammass 1985 for an extensive review). Preliminary examination of plots of formant transition data from the literature indicated that strong linear relationships existed between the onsets and steady states for voiced stop+vowel syllables. The present study was undertaken to clarify the nature of these relationships and to attempt to exploit them in a consonant recognition scheme.

Experiment

Subjects

Subjects were 10 (5 male and 5 female) phonetically trained speakers of Canadian English and were all graduate students or faculty members in linguistics at the University of Alberta.

Materials and Methods

Speakers were provided with a randomized list of phonetically transcribed syllables which they were asked to read. Two repetitions of each of 33 /CVd/ 's (with C ranging over /b, d, g/ and V ranging over /ɪ, ɪ, e, ɛ, æ, ʌ, ɔ, o, ɷ, u, ɚ /, were collected from each of the speakers. The 660 tokens were digitized at 16 kHz and analyzed as follows. A 16 ms Hamming window was advanced in 5 ms frames over the first 80 ms following stop release. Each frame underwent an autocorrelation LPC-based spectral analysis. A lag window (rectangular in the frequency domain with a bandwidth of 50 Hz; see Tohkura, Itakura and Hashimoto 1978) was applied to the autocorrelation coefficients prior to the estimate of the inverse filter. A 20 coefficient analysis was used for all male speakers and 16 to 18 coefficients were used for females. Printouts of estimated formant frequencies and amplitudes (using a method similar to that described by Christensen, Strong and Palmer 1976, involving the second derivative of the smoothed log magnitude spectrum derived from the LPC analysis) were examined and four measurements were derived manually: 1) F2v, the frequency of "steady-state vowel" F2 at 60 ms following stop release; 2) F2i , the "initial" frequency of F2, taken as early as possible after stop release, subject to continuity of the track with F2v; 3) F3v and 4) F3i, analogous measures for F3. The measurement points for a typical stimulus are illustrated[2] in Figure 1. For some of the female voices, the initial estimates proved difficult to track. For each of these voices, an *ad hoc* adjustment of the number of coefficients was made on a few syllables until usable results were obtained. Data for these speakers was then re-analysed with the 16 ms windows which were advanced in 2.5 ms frames.

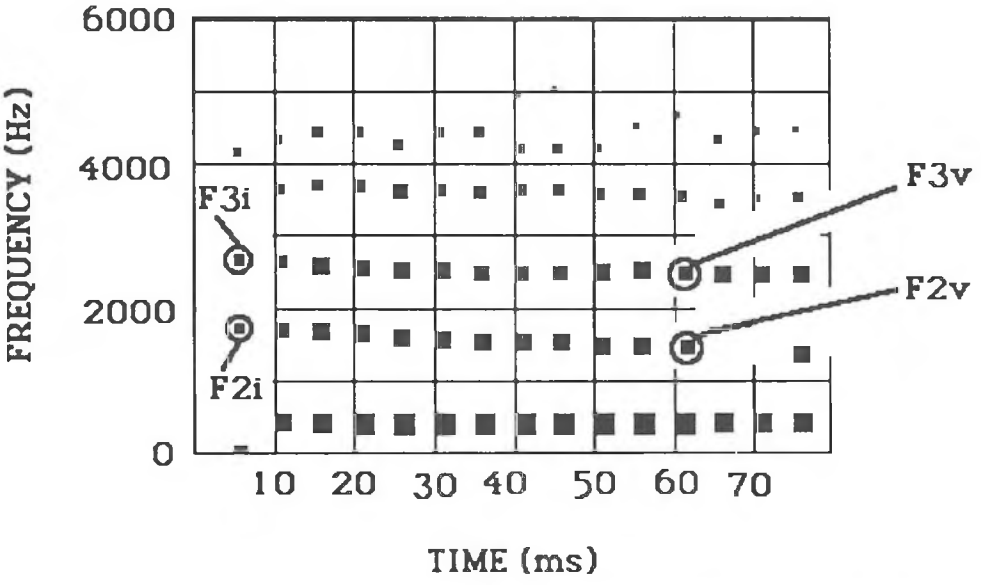

Figure 1. Schematic spectrogram showing measurement points for initial (F2i, F3i) and "steady-state" (F2v, F3v) formant frequencies.

ANALYSIS

Graphical analysis

Scatterplots of initial F2 as a function of F2 of the following vowel nucleus (F2i X F2v) confirmed that a strong linear relationship existed for each of the three consonants considered separately. These plots are shown in the left half of Figure 2. All three plots indicate a strong positive correlation between F2i and F2v. However, /d/ shows less "tuning" of F2 onset (F2i) with changes in F2v, consistent with a "/d/-locus" near 1800 Hz (Delattre et al. 1955; Fant 1973). On the other hand, /b/ and especially /g/ are more strongly vowel-dependent. Note that /b/-onsets generally occur at or below the diagonal (F2i=F2v), while /g/-onsets lie slightly above it . Similar patterns exist for the three consonants for F3. (See right half of Figure 2.) However, the differences in the distributions of the three consonants in F3 is less striking.

Regression analysis

The left half Figure 2 also displays the results of least-squares regressions[3] of F2i on F2v for each of the consonant categories considered separately. A similar analysis of F3i and F3v is presented in the right half. The regression coefficients coefficients reported in Figure 2 may be used as the basis of a simple minimum distance classification procedure as described below. It should be noted that the relationships described here are similar in many ways to those exploited by Klatt for formant frequency transition calculation in speech synthesis by rule (see Allen, Hunnicut and Klatt 1987: 111-116).[4]

Classification results

Minimum distance classification

After using the data as a training set for the regression lines, each spoken syllable was re-classified as a member of /b/, /d/ or /g/ on the basis of its distance to each of the corresponding regression lines. More precisely, each token is mapped into the F2i X F2v plane, where its vertical distance, D2c, to the regression lines for each consonant (c) is calculated. A similar set of distances, D3c, is calculated in the F3i X F3v plane. The decision was based on the combined distance measure $D^2c = [(D2c)^2 + (D3c)^2]$. A token is classed as the consonant for which D^2c is minimum.

This minimum distance classification rule results in a correct partition rate of 73.9%, when the training data are re-classified. A cross-validation approach, in which the data from an arbitrary subsample of five of the speakers were used as the training set while the remaining 5 were used as the test set, actually yielded slightly higher correct classification rate for the test data, 76.1%.

Alternative parametric classification methods

The present analysis shows that there is considerable information available in formant frequency transitions for the identification of stop consonants. It should be noted that the analysis used *acoustic context only* and did not require prior phonetic categorization of the following vowel. Kewley-Port (1982) investigated linear discriminant analyses of /b/, /d/ and /g/ based on F2-F3 transition measurements of syllables spoken by a *single* speaker. She found that automatic classification of place features for stop consonants was quite high (97%) when linear discriminant analyses were carried out separately for individual vowel contexts. However , it should be born in mind that only 5 repetitions for each vowel token were involved. Shammass (1985) reports separate vowel-wise linear discriminant analyses for the present data and finds an overall correct identification rate of 81% and ranging from 72% for / ɚ / to 90% for /ʌ/. Shammass's classification results involved 20 points per vowel (2 repetitions by each of 10 speakers). Furthermore her results were based on the so called U-method (or jackknife, Gray and Schucany 1972) of classification which reduces bias in classification scores for small samples. Classifications results reported by Kewley-Port were considerably lower (68% correct) when a single linear discriminant analysis (pooling over vowels) was conducted. A single linear discriminant analysis of the present multi-speaker data yielded an identification rate of 66% (compared to about 74% for the regression method described above).

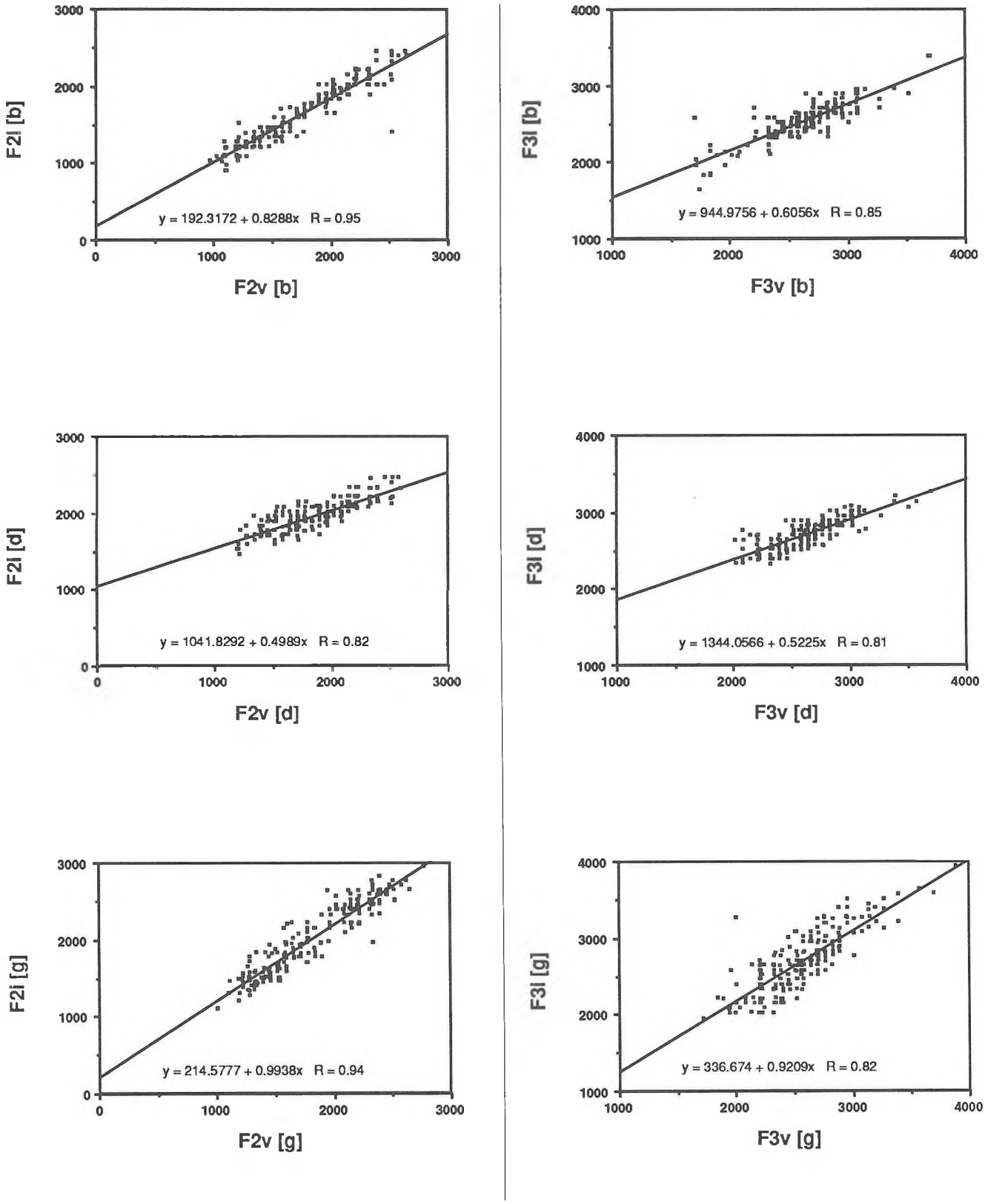

Figure 2. Left panel: Scatterplot of the frequencies of initial F2 (F2i) and F3 of the vowel at 60 ms from consonant release (F2v). Right panel: analagous plots for F3i and F3v. (Top to bottom in each panel: data for /b/, /d/, /g/). Least-squares regression lines, their coefficients and Pearson's r are shown for each analysis.

Linear discriminant analysis is based on the assumption that the samples used in the determination of the classification rule are drawn from normal distributions that differ only in their mean vectors (i.e., the means for F2i, F2v, F3i, F3v may be distinct for each of /b/, /d/, /g/) but that the groups have a common covariance matrix. The scatter plots and regression analysis indicate that this latter assumption is likely not correct, since, e.g. the regression of F2i on F2v has a substantially more positive slope for /g/ than for /d/. (Indeed, the differences in the orientations of these distributions seem more salient than their overall locations.[5]) A formal test of the equality of covariance matrices provides strong evidence for rejection of the common covariance assumption[6] (Box's M = 249.0, approximate $F(6,120)=41.3$, $p<.0001$). Bayesian classification schemes using separate estimates of the covariance matrices for each group are more appropriate than linear discriminant analysis in such cases. These procedures, sometimes referred to as quadratic discrimination (Lachenbruch 1975: 20-23), may be performed either in the full space of the original measurements or in a reduced dimensional space based on a prior linear discriminant analysis (Tatsuoka 1971: 232 -233). Classification of the present data using separate covariance estimates in a reduced (2-dimensional) space was investigated by Shammass (1985), resulting in 72% correct identification. A full 4 -dimension quadratic discriminant analysis (equivalent to a maximum likelihood classification scheme) actually showed a slightly lower overall classification rate of 71%. These rates are similar to that of the minimal distance regression rule. Confusion matrices for the regression classification, linear and quadratic discriminant analysis are given in Table 1. The similarity of the error patterns for the regression and quadratic methods may indicate they are exploiting essentially the same properties of the distributions.[7]

TABLE I: Consonant-wise identification rates (in percent) for selected classification procedures:

Linear Discriminant Analysis

Actual Group	Predicted Group		
	/b/	/d/	/g/
/b/	89.5	9.1	1.4
/d/	16.4	47.7	35.9
/g/	15.0	25.0	60.0

Quadratic Discriminant Analysis (2-Dimensional)

Actual Group	Predicted Group		
	/b/	/d/	/g/
/b/	84.1	10.0	5.9
/d/	10.5	71.4	18.2
/g/	11.8	26.4	61.8

Minimum distance Regression Classification (all data pooled)

Actual Group	Predicted Group		
	/b/	/d/	/g/
/b/	89.5	6.4	4.1
/d/	11.8	69.1	19.2
/g/	15.0	21.8	63.2

Discussion

The regression lines for each consonant may be regarded as representing invariant relational properties for each consonant. These invariants are not sufficient individually to separate the groups in all cases (superposition of Figures 2, 3 and 4 shows considerable category overlap). But, since the lines for the three categories are not identical, the properties may be considered partially distinctive.[8]

Other factors such as the shape of onset spectra and VOT are important cues in analytic recognition of stops (Blumstein and Stevens 1979; Edwards 1981; Searle et al. 1979; Kewley-Port 1982) as well as in speech perception (Blumstein and Stevens 1980). Walley and Carrell (1983) show that formant frequency information can, in certain cases, override other cues. Shammass (1985) confirms that both spectral shape information and formant frequencies play a role in listeners' perception of synthetic stops. She compares a regression classification of synthetic stimuli (similar to that presented here) to listeners' judgments. Predictions are good only in the case of front vowels. Further research is needed to clarify the nature of the interaction of cues in listeners' perception and to apply these results to automatic speech recognition. We hope that, given a suitable parametric representation of spectral shape information, methods similar to those applied here and in Shammass (1985) will result in improved recognition performance.

Notes

1. Portions of this research were presented as Poster Session D15, 111th Meeting Acoustical Society of America Cleveland, Ohio 13 May 1986 and stem from work performed in conjunction with the doctoral dissertation of the second author.

2. The size of the marks in Figure 1 is linearly related to the formant amplitude estimates (in dB). This figure is included for illustrative purposes only. The actual measurements were extracted manually from numerical printouts.

3. As noted by an anonymous reviewer, ordinary least squares regression is strictly justifiable only when the independent variables are error free. Daniel and Wood suggest: "As a rule of thumb, least squares analysis can be used safely if the variance of x is less than a tenth of the average scatter of the x's about their mean (1971:32)." Estimates of "pure error" (Draper and Smith 1981) from repetitions of the same syllable by the same speaker range between 53 and 152 Hz in the present data for all measures, while the standard deviations of F2v and F3v about their means range from 279 to 440 Hz. All but one of the regression lines in Figure 2 exceed Daniel and Wood's rule of thumb. The one exception, F3v of /d/ falls just short of this criterion, with a ratio of .106, or about 6% greater than their "safe" ratio. The reviewer suggested that principle components analysis might be preferable to regression. Then, presumably, the eigenvector associated with the largest eigenvalue would replace the regression line and a suitable distance metric would have to be applied (e.g. perpendicular distance to the eigenvector). We agree with the reveiwer that this is unlikely to make much difference in the present case. It might be more important in other cases where both variables showed large "pure error" variation compared to their ranges.

4. However, Klatt found that, for velars and alveolars, more than one regression line per consonant was required for accurate modeling of onset/target relationships, depending on whether the following vowel was back rounded, back unrounded or front. The present data do not seem to warrant such an approach (but see notes 5 and 8).

5. The mean vectors (F2i,F3i,F2v,F3v) are: (1601, 2442, 1699, 2637) for /b/; (1969,2727,1858, 2648) for /d/; and (2029,2702,1826, 2568) for /g/. /d/ and /g/ are very close , deviating by no more than 60 Hz on all measures. /b/ shows more markedly lower values (about 150 to 370 Hz) on all measures except F3v.

6. Separate histograms of F2i, F2v, F3i and F3v also indicate that the assumption of simple multinormal distribution about the mean is suspect. Several of these histograms showed tendencies toward bimodality, with one peak for front vowels and another for back. More detailed modeling of

these probability distributions in classification schemes is planned. It should be noted that more complex distributions could be used in classification schemes without explicit reference to categorical knowledge of the vowel categories in the classification phase itself.

7. This note is a response to some interesting comments from an anonymous reviewer. The regression classification technique would, in effect, constitute a maximum-likelihood classification procedure under the following conditions: 1) all errors of measurement are in the dependent variables; 2) error variances for F2i and F3i are equivalent for all the vowels; 3) residuals from F2 and F3 analyses are uncorrelated; 4) overall location of the distributions along the F2v and F3v axes are independent of the consonant (no systematic co-articulation effects of the consonants on the steady states). There is evidence that some of these conditions are at least moderately violated by this data. Condition 1 has been commented on in note 3. Regarding condition 2, standard errors for both F2 and F3 regressions for /b/ and /d/ ranged from 107 to 122 Hz. However standard errors for the /g/ regressions were 152 Hz for F2 and 214 Hz for F3. Experimentation with a modified regression analysis that weighted distances (in inverse proportion to the error variances) yielded highly similar identification rates to the unweighted method. Regarding condition 3, correlation analysis of the residuals of F2 and F3 analyses showed no significant relationship for /b/, but both /d/ and /g/ showed significant positive correlations: R=.299 for /d/ and R= .156 for /g/, accounting for about 9% and 2% of the residual variance respectively. Regarding condition 4, see note 5. Quadratic discriminant analysis (QDA) can accommodate all the above violations of assumptions of a maximum likelihood regression model. The fact that QDA does not show improvement over the regression model may be because 1) the violations involved are relatively mild; or 2) the violations occur in "directions" that do little harm; or 3) there are violations of additional assumptions of QDA itself (see note 6).

8. Shammass (1985) presents evidence that slightly different regression lines may characterize female versus male data for the same consonants in both F2 and F3. While this raises interesting questions related to speaker-normalization, exploratory investigation indicated that the differences are relatively unimportant for this data set. In particular, classification based on separate regression lines for males and females leads to only modest improvement in classification scores over the pooled male and female data reported below. Nonetheless, we believe the issue merits further study in larger data sets.

References:

Allen, J., M. S. Hunnicutt and D. Klatt. **(1987).** *From Text to Speech: The MITalk System.* Cambridge University Press: Cambridge.

Blumstein, S.E. and K.N. Stevens. **(1979).** Acoustic invariance in speech production:evidence from measurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Am. **66**, 1001-1017.

Christensen, R., W. Strong, and E. Palmer. **(1976).** A comparison of three methods of extracting resonance information from predictor-coefficient encoded speech. IEEE Trans. Acoust. Speech and Signal Processing ASSP **24**, 8-14.

Daniel, C. and F. Wood. **(1971).** *Fitting Equations to Data.* Wiley: New York.

Delattre, P.C., A.M. Liberman, and F.S. Cooper. **(1955).** Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. **27**, 769-773.

Draper, N. and H. Smith. **(1981).** *Applied Regression Analysis,* Second Edition. Wiley: New York.

Edwards, T. **(1981)**. Multiple feature analysis of intervocalic English stops. J. Acoust. Soc. Am. **69,** 535-547.

Fant, G. **(1973)**. Stops in CV syllables. In G. Fant (ed.) *Speech Sounds and Features.* MIT Press: Cambridge MA. 110-139.

Gray, H. and W. Schucany. **(1972)**. *The Generalized Jackknife Statistic.* New York: Marcel Dekker.

Kewley-Port, D. **(1982)**. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. J. Acoust. Soc. Am. **72,** 379-389.

Lachenbruch, P. A. **(1975)**. *Discriminant Analysis.* Haffner Press: New York.

Searle, C., J.Z. Jacobson and S.G. Rayment. **(1979)**. Stop consonant discrimination based on human audition. J. Acoust. Soc. Am. **65**, 799-809.

Shammass, S.E. **(1985)**. Formant transitions, spectral shape and vowel context in the perception of voiced stops. Unpublished Ph. D. dissertation. University of Alberta.

Tohkura, Y., F. Itakura, and S. Gasgunitim. **(1978)**. Spectral smoothing trechnique in PARCOR speech analysis-synthesis. IEEE Trans. Acoust. Speech and Signal Processing ASSP **26:**587-596.

Walley, A.C. and T.D. Carrell. **(1983)**. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. J. Acoust. Soc. Am. **73**. 1011-1022.