Morrison, G. S., & Nearey, T. M. (2007). Testing theories of vowel inherent spectral change. *Journal of the Acoustical Society of America, 122*, EL15–EL22. DOI: 10.1121/1.2739111.
http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=JASMAN00012200000100EL15000001&idtype=cvips

Testing theories of vowel inherent spectral change[a)]

Geoffrey Stewart Morrison[b)] and Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, Alberta, T6G 2E7, Canada
gsm2@bu.edu *t.nearey@ualberta.ca*

Abstract: Three competing accounts of vowel inherent spectral change (VISC) in English all agree on the importance of initial formant frequencies; however, they disagree about the nature of the perceptually relevant aspects of formant change. The onset + offset hypothesis claims that the final formant values themselves matter. The onset + slope hypothesis claims that only the rate of change counts. The onset + direction hypothesis claims that only the general direction of change in formant frequencies is important. A synthetic-vowel perception experiment was designed to differentiate among the three. Results provide support for the superiority of the onset + offset hypothesis.

PACS numbers: 43.71.-k, 43.71.An, 43.71.Es
Date Received: * **Date Accepted:** *

1. Introduction

Traditionally, the English vowel system is often said to comprise true diphthongs, /aɪ, aʊ, ɔɪ/, phonetic diphthongs, /e, o/ (frequently transcribed as /eɪ, oʊ/), and monophthongs, e.g., /i, ɪ, ɛ, æ/. However, for many North American dialects it has been reported that several nominal monophthongs show significant *vowel inherent spectral change* (VISC), and VISC appears to be important for perception [Andruski and Nearey, 1992; Assmann and Katz, 2005; Assmann, Nearey, and Hogan, 1982; Hillenbrand, Clark, and Nearey, 2001]. Figure 1 provides examples of mean F1–F2 formant trajectories for natural productions of Western Canadian English /e/, /ɪ/, and /ɛ/ in the same context as used in the present study (data from Morrison, 2006).

There are three main accounts of the perceptually relevant aspects of VISC [Gottfried, Miller, and Meyer, 1993; Nearey and Assmann, 1986]. All three hypotheses agree that the initial formant frequencies are perceptually relevant for vowel identification, but disagree on what additional cues are relevant. The *onset + offset* hypothesis states that the relevant perceptual cues are the formant values at the end of the vowel. The *onset + slope* hypothesis states that the relevant cue is the rate of change of formants over time. The *onset + direction* hypothesis states that the only relevant factor is the direction of formant movement in an F1–F2 (or similar) space. To elucidate the difference among the three hypotheses: In the direction hypothesis the rate of change in time (hereafter speed) of formant movement and the formant values achieved at the end of the trajectory are irrelevant. In the slope hypothesis the direction and speed of formant movement are relevant, but the formant values achieved at the end of the trajectory are irrelevant. If the glide portion of one vowel is longer than the glide portion of another vowel, then they could both have the same formant slopes but different final formant values. In the offset hypothesis the formant values achieved at the end of the trajectory are relevant (i.e., direction and magnitude of formant movement are relevant), but the speed of movement towards those values is irrelevant. If the glide portion of one vowel is longer than the glide portion of another vowel, then they could both have the same final formant values but different formant slopes.

Nearey and Assmann (1986), and Gottfried, Miller, and Meyer (1993) used pattern

[a)]Portions of this paper was presented at the 150th Meeting of the ASA, Minneapolis, November 2005.
[b)]Present address: Department of Cognitive & Neural Systems, Boston University, 677 Beacon Street, Boston, Massachusetts 02215.

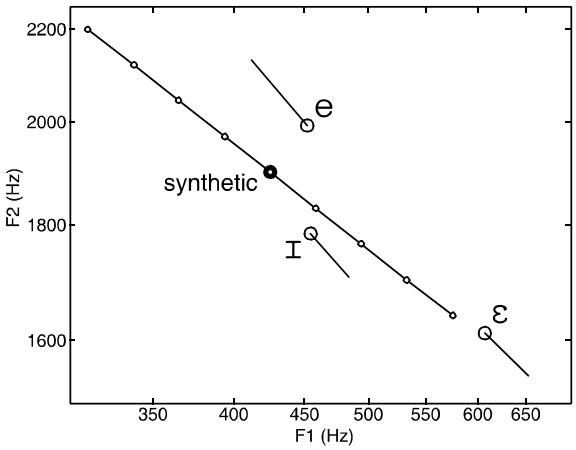

Fig. 1. First and second formant properties of natural and synthetic vowels. The comets labeled /e/, /ɪ/, and /ɛ/ represent the mean onsets and offsets of these vowels produced in isolated-word /bVpə/ context by seven male speakers of Western Canadian English (ten replications per speaker). Comet heads and tails represent the formant values at 25% and 75% of the duration of the vowels respectively. (Geometric mean vowel durations for /e/, /ɪ/, and /ɛ/ were 114, 69, and 82 ms respectively.) For the synthetic stimuli, the large black dot represents the initial formant values and the small white dots represent the nine sets of final formant values.

recognition models to test alternate parameterizations consistent with the three VISC hypotheses. In both studies, although there was a slight advantage for the offset model, all three parameterizations performed well in terms of correct identification (and, in the earlier study, correlation with listeners' response patterns).[1] In contrast to the former studies, the stimuli in the present experiment are explicitly designed to differentiate among the three hypotheses. Specifically, the hypotheses are directly compared in a perceptual experiment using a synthetic /e/–/ɪ/–/ɛ/ continuum which has a fixed onset, but which varies in offset, duration, and slope.

2. Method

2.1 Stimuli

Stimuli were synthesized in a male-speaker range using an implementation of the Klatt cascade formant synthesizer [Klatt and Klatt, 1990]. The stimuli were /bVpə/, consisting of synthetic /bVpˀ/ plus a natural recording of the final /pə/ syllable from the burst onward.[2] Care was taken (adjusting glottal slope and breathiness parameters) to match the voice quality of the synthetic speech to the natural portion of the stimuli. To limit the stimulus space, pilot studies were conducted to find a single initial set of F1–F2 values from which it was possible to obtain /e/, /ɪ/, and /ɛ/ percepts by changing only the final formant values and duration. The initial F1 and F2 values were 425 and 1900 Hz respectively (6.052 and 7.550 log Hz), see Figure 1.

Three types of vowels were synthesized: Formant trajectories were either *straight* (in log Hertz) from the beginning to the end of the vowel (Figure 2a,c,e); or they were *elbowed* with an initial steady state for the first 25% of the vowel duration, followed by a glide (Figure 2b,d,f). Elbowed diphthongs thus had the same initial and final formant values as straight diphthongs, but the slope of the glide portion was 33% steeper. The third vowel type was *flat*, having no formant movement during the vowel.

Each stimulus had one of three directions of movement. The flat stimuli had zero movement. The other two directions were opposite in the log F1–F2 space (see Figure 1), stimuli either had converging VISC (F1 rose and F2 fell, see Figure 2c–f), or diverging VISC (F1 fell and F2 rose, see Figure 2a–b).[3] Non-flat stimuli had one of four magnitudes of VISC (multiples of ±0.0756 log Hertz for F1 and ∓0.0367 log Hertz for F2). Because natural productions of the three vowels examined here differ in duration as well as spectral properties, stimuli also ranged over

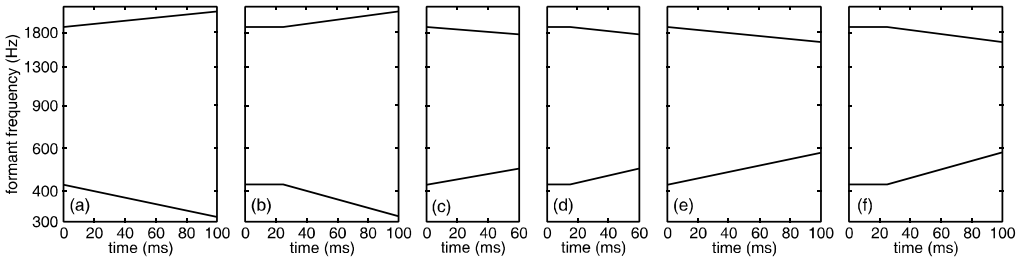

Fig. 2. Examples of formant trajectories for synthetic stimuli (excluding consonant transitions). The examples are straight (a, c, e) and elbowed (b, d, f) versions of the stimuli which, in the perception experiment, received the greatest number of /e/ (a, b), /ɪ/ (c, d), and /ɛ/ (e, f) responses. The percentages of responses (pooled across listeners) for each category were: (a) 85 /e/, 3 /ɪ/, 12 /ɛ/. (b) 79 /e/, 4 /ɪ/, 17 /ɛ/. (c) 21 /e/, 60 /ɪ/, 20 /ɛ/. (d) 22 /e/, 53 /ɪ/, 26 /ɛ/. (e) 23 /e/, 35 /ɪ/, 41 /ɛ/. (f) 25 /e/, 31 /ɪ/, 44 /ɛ/. Audio files Mm. 1 through 6 correspond to Figures 2a through 2f.

Mm. 1. Audio file (a) (8.03kB). This file is of type "wav".
Mm. 2. Audio file (b) (8.03kB). This file is of type "wav".
Mm. 3. Audio file (c) (7.25kB). This file is of type "wav".
Mm. 4. Audio file (d) (7.25kB). This file is of type "wav".
Mm. 5. Audio file (e) (8.03kB). This file is of type "wav".
Mm. 6. Audio file (f) (8.03kB). This file is of type "wav".

three duration values: 60, 80, and 100 ms (excluding consonant transitions). This resulted in a total of 51 stimuli: Eight straight, eight elbowed, and one flat, all multiplied by three durations.

2.2 Listeners

Listeners were 23 undergraduate-student volunteers. They were monolingual English speakers who had grown up in western Canada, and reported no hearing deficits.

2.3 Procedures

Listeners were tested one at a time in a sound booth. The stimuli were played at a comfortable volume (approximately 65 dB SPL) via a Roland Edirol UA-30 sound card and a calibrated Mackie HR824 studio monitor. In each trial, the listeners heard a stimulus, then saw three buttons on a computer screen labeled *baypa*, *bippa*, and *beppa* representing /e/, /ɪ/, and /ɛ/ respectively (prior to the experiment, listeners were trained on the orthography to vowel-category relationship). They indicated their response via a mouse click. A new stimulus was automatically presented 500 ms after a response to the previous stimulus was provided. The whole set of stimuli were presented eight times in randomized blocks, resulting in a total of 408 trials per listener.

3. Results and discussion

Direct statistical tests of the three rival hypotheses are not possible because such tests require a strict nesting of terms, and the hypotheses are not related to each other in this manner. An indirect strategy analogous to partial correlation analysis was therefore adopted. The procedure is as follows: Models of two rival hypotheses (e.g., A and B) are fitted to the data, then a third larger model (e.g., C) including all the terms of the first two models is fitted. Thus, both smaller models are properly nested within the third. If the larger model fits significantly better than A but not significantly better than B, then it is reasonable to infer support for hypothesis B. Roughly speaking, B includes essentially all the relevant information in the larger hypothesis, while adding the extra terms from A adds little explanatory information.

3.1 Offset vs slope

In order to compare the adequacy of the offset versus the slope hypotheses, *Vector smoothed Generalized Additive logistic regression Models* (VGAM) [Yee and Wild, 1996] were fitted to the perceptual response data (the raw counts of responses for each category given to each

stimulus), and compared for goodness-of-fit. Three models were fitted to each of the listeners' data sets. In all models, duration (three values) was entered in milliseconds and fitted via a (saturated) quadratic polynomial. Duration parameters were included to control for the effect of duration which is not of interest in the present paper. Other parameters entered into the models related directly to formant movement. These were fitted via smoothing splines with two effective degrees of freedom.[4] Since all three models include initial formant specifications, the differences among them can be summarized by their characterization of spectral change:

(*A*) *Offset Model*: $\Delta F1$, the change of F1 in log Hertz from the beginning to the end of the vowel.

(*B*) *Slope Model*: $\Delta F1/\Delta t$, the slope of F1 in log Hertz per second over the glide portion of the vowel (from the beginning or elbow of the vowel, for straight and elbowed stimuli respectively, to the end of the vowel).

(*C* = *A*+*B*) *Combined Model*: A model containing both $\Delta F1$ and $\Delta F1/\Delta t$. This allows for both the offset and the slope to have effects on perception.

Parameterizations A and B are equivalent to those of Assmann and Nearey (1986). Since F2 in

Table 1. Likelihood ratio tests for combined model (*C*) versus offset model (*A*) and slope model (*B*). ΔG^2 is decrease in deviance statistic (lack of fit measure) in the larger versus smaller model. Δedf is the change in effective degrees of freedom (see note 4). Reported *p* values are nominal significance levels for individual tests referred to a χ^2 distributions with Δedf degrees of freedom.

	Models compared					
	C vs *B* (added terms $\Delta F1$)			*C* vs *A* (added terms $\Delta F1/\Delta t$)		
Listener	ΔG^2	Δedf	*p*	ΔG^2	Δedf	*p*
1	13.53	3.42	.005[a]	1.95	3.47	.666
2	3.77	3.93	.427	5.38	3.97	.247
3	11.17	3.94	.024[a]	4.49	3.81	.318
4	13.09	4.03	.011[a]	12.38	3.83	.013[a]
5	3.39	3.49	.414	2.56	3.51	.556
6	5.35	3.68	.217	3.91	3.72	.376
7	4.68	3.90	.308	6.46	3.77	.148
8	8.62	3.78	.062	2.98	3.82	.533
9	3.01	3.71	.510	3.00	3.75	.517
10	4.98	4.04	.295	3.89	3.98	.418
11	13.05	3.57	.008[a]	11.54	3.75	.017[a]
12	30.00	4.00	.000[a]	12.33	3.68	.012[a]
13	8.17	3.99	.085	5.51	3.87	.225
14	13.86	3.63	.006[a]	2.98	3.73	.518
15	4.73	3.91	.305	1.52	3.90	.812
16	11.76	3.95	.019[a]	2.44	3.90	.640
17	4.14	3.93	.377	2.97	3.86	.540
18	9.47	3.94	.048[a]	2.17	3.76	.670
19	14.10	3.99	.007[a]	3.40	3.91	.478
20	4.45	3.77	.318	4.08	3.70	.352
21	7.49	3.91	.107	1.25	3.80	.850
22	12.86	3.76	.010[a]	1.61	3.50	.740
23	7.68	3.66	.085	5.62	3.77	.206

[a]significant at a nominal α level of .05

the synthetic stimuli in the present study was perfectly correlated with F1, it was only necessary to enter one set of formant values into the models.

The model-fitting procedure minimizes the deviance statistic (G^2). Table 1 gives results of ΔG^2 likelihood ratio tests comparing the differences in the goodness-of-fit between the larger model (model C) and each of the two smaller models (models A and B). Models fitted to data from 10 of the 23 listeners had a significant (nominal $p<.05$) improvement in goodness-of-fit when the offset parameters were added to a model already containing slope parameters. In contrast, when slope parameters were added to a model already containing offset parameters, there was a significant improvement in goodness-of-fit for data from only three listeners, and these three data sets had also had a significant improvement of fit when the offset parameters were added to the slope model. The results therefore provide greater support for the onset + offset hypothesis than for the onset + slope hypothesis.

3.2 Offset vs direction

If the direction hypothesis is correct, then there should be no difference in listeners' responses to stimuli which have the same VISC direction in the F1–F2 plane but different final formant values. The direction versus offset hypotheses were tested in the same manner as the slope versus offset hypotheses. The formant-movement parameters in the models were:

(*A*) *Offset Model*: Δ*F1* (as above).

Table 2. Likelihood ratio tests for combined model (*E*) versus offset model (*A*) and direction model (*D*). (See caption of Table 1 for explanation of abbreviations.)

	Models compared					
	E vs *D* (added terms Δ*F1*)			*E* vs *A* (added terms *direction coding*)		
Listener	ΔG^2	Δ*edf*	*p*	ΔG^2	Δ*edf*	*p*
1	26.97	3.57	.000[a]	6.64	3.98	.155
2	6.16	3.93	.181	5.14	3.99	.273
3	13.45	3.73	.007[a]	3.93	3.83	.390
4	11.93	3.85	.016[a]	4.89	4.01	.300
5	2.44	3.48	.573	5.05	4.00	.282
6	13.07	3.68	.008[a]	3.19	4.01	.529
7	5.38	3.91	.240	9.44	3.99	.051
8	6.47	3.73	.144	0.69	4.00	.952
9	33.78	3.66	.000[a]	4.66	3.94	.316
10	13.73	4.03	.008[a]	7.10	3.99	.130
11	35.36	3.70	.000[a]	3.16	4.00	.531
12	16.98	3.71	.001[a]	3.05	4.03	.554
13	3.38	3.95	.488	6.28	3.99	.178
14	18.99	3.69	.001[a]	1.58	4.03	.815
15	13.66	3.85	.007[a]	0.57	4.00	.966
16	14.30	3.86	.006[a]	5.37	4.00	.251
17	2.27	3.91	.672	1.93	4.00	.750
18	19.77	3.82	.000[a]	5.90	4.00	.207
19	12.19	3.93	.015[a]	5.18	3.99	.268
20	1.15	3.80	.868	11.67	4.02	.020[a]
21	16.33	3.83	.002[a]	2.66	3.99	.616
22	26.23	3.64	.000[a]	6.25	3.81	.165
23	13.83	3.78	.007[a]	5.51	4.03	.242

[a]significant at a nominal α level of .05

(*D*) *Direction Model*: The three directions of formant movement in the stimuli, diverging, flat, and converging, were entered as three discrete interval levels, −1, 0, and +1.[5]

(*E* = *A*+*D*) *Combined Model*: A model containing both ΔF1 and direction parameters.

Results of the ΔG^2 likelihood ratio tests are given in Table 2. Models fitted to data from 16 of the 23 listeners had a significant (nominal $p<.05$) improvement in goodness-of-fit when the offset parameters were added to a model already containing direction parameters. In contrast, when the direction parameters were added to a model already containing offset parameters, there was a significant improvement in goodness-of-fit for data from only one listener. The results therefore provide much greater support for the offset hypothesis than for the direction hypothesis.

3.3 Remaining issues

There was considerable inter-listener variation in response patterns. Also, a number of listeners gave a preponderance of /e/ responses, even for stimuli with converging VISC. Note that inter-listener variation and bias do not invalidate the results of the analyses. All that is required for the reasonable application of the modeling procedure is that the response probabilities change noticeably under at least some of the stimulus manipulations at issue. For each listener there was a significant improvement in goodness-of-fit when a model including only duration information was compared with the offset model (model A), this indicates that the requirement was met with respect to formant movement.

All hypotheses agree on the importance of initial formant values, and the onset point in the experiment may have given a strong cue for /e/ perception. Examination of Figure 1 indicates that the ratio of initial F1 and F2 values were closest to the mean of natural /e/ productions (the Morrison, 2006, data was collected after the present study was conducted). It is also possible that no single onset point will allow for clear percepts of all three vowels by all listeners, and additional experiments using multiple onset points may be warranted.

Furthermore, it is conceivable that a perceptual mechanism registering rate of spectral change may have operational limits (e.g., temporal smearing) that produce differences in 'effective slope' that are less than the nominal 33% differences between corresponding straight and elbowed stimuli. Thus the distinction between slope and offset hypotheses should probably also be assessed in contexts with longer mean vowel durations (e.g. isolated CVCs), where absolute physical differences between straight and elbowed stimuli would be larger.

Finally, there is perhaps a weak indication that a more complex characterizations of formant trajectories should be investigated (See note 1 for some possible parameterizations). The combined models *C* and *E* contain information not available in their constituents (*A*+*B* and *A*+*D*, respectively). Three of 23 listeners for model *C* and one for model *E* showed significant improvement over both the relevant simpler models. Since at the nominal .05 α level, one might expect to find only about two false positives (Type I errors), it would seem judicious to examine more complex trajectory models in a larger and presumably more powerful experiment.

4. Conclusion

The weight of evidence presented here indicates that the onset + offset hypothesis is superior to the onset + slope and onset + direction hypotheses.

Acknowledgments

This research was supported by the Social Sciences and Humanities Research Council of Canada. Thanks to Peter F. Assmann, Michael Kiefte, the Associate Editor and two anonymous reviewers for comments on earlier drafts of this paper.

Notes

1. More complex curve-fitting parameterizations of VISC have been tested [Zahorian and Jagharghi, 1993]; however,

Hillenbrand, Clark, and Nearey (2001) failed to find an improvement over the onset + offset model for models using polynomials and discrete-cosine transforms. But see discussion in section 3.3.

2. This was one of a series of several experiments designed to examine English and Spanish listeners' perception of VISC. The /bVpə/ pattern results in possible but non-existent words in both languages. In some experiments the /bVpə/ words were embedded in natural-speech carrier sentences in the same voice as the natural-speech /pə/ portion of the stimuli.

3. The stimulus identification rates reported in the caption of Figure 2 are based on Western Canadian English listeners' responses, and the reader may categorize these stimuli differently. Also, the reader may have difficulty distinguishing the straight and elbowed versions of each pair of example stimuli (although this was not the case for the younger of the two authors). This might be expected if slope is not a perceptually relevant aspect of VISC.

4. The fitting procedure implements smoothing splines via a maximum penalized likelihood method. The analyst can request a target degrees-of-freedom value, and the software will seek a smoothness penalty that comes close to that value. See Hastie and Tibshirani (1990); Hastie, Tibshirani, and Friedman (2001 ch. 9); and Yee and Wild (1996) for details. Two requested degrees of freedom in the representation of a stimulus property results in (approximately) an added four degrees of freedom in the fitted models because the smoothing splines are estimated independently for each of two non-redundant response contrast terms for the three-way vowel distinction. Similar results were obtained with linear stimulus terms; however, for most listeners there was a significant improvement in goodness-of-fit when offset and slope models were fitted using smoothing splines with two effective degrees of freedom. This modest increase in the complexity of the models maintained all the essential differences among the three main hypotheses examined, and provided some added flexibility to allow for factors such as possible floor and ceiling effects in the stimulus variables in question.

5. Direction in the stimuli is restricted to one dimension (see Figure 1), and the three values −1, 0, and +1 exhaust all possible one-dimensional directions. A discrete-level coding system was therefore adopted, with direction dummy coded using two parameters (i.e., [1 0], [0 0], and [0 1] for −1, 0, and +1 respectively). The smallest non-zero ΔF2 magnitudes in the stimuli (68 Hz) just exceeded the largest mean threshold for F2 movement detection (66 Hz) reported by Kewley-Port and Goodman (2005). The stimuli in the present study also had F1 movement (minimum 31 Hz), and the combined effect of F1 and F2 movement was therefore expected to make the difference between zero-magnitude and the smallest non-zero-magnitude stimuli detectable (results of Monte Carlo permutation tests were consistent with this expectation).

References

Andruski, J. E., and Nearey, T. M. (**1992**). "On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables," J. Acoust. Soc. Am. **91**, 390–410.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**, 975–989.

Assmann, P. F. and Katz, W. F. (**2005**). "Synthesis fidelity and time-varying spectral change in vowels," J. Acoust. Soc. Am. **117**, 886–895.

Gottfried, M., Miller, J. D., and Meyer, D. J. (**1993**). "Three approaches to the classification of American English diphthongs," J. Phonetics **21**, 205–229.

Hastie, T., & Tibshirani, R. (**1990**). *Generalized Additive Models* (Chapman and Hall, London).

Hastie, T., Tibshirani, R., & Friedman, J. (**2001**). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).

Hillenbrand, J. M., Clark, M. J., and Nearey, T. N. (2001). "Effect of consonant environment on vowel formant patterns," J. Acoust. Soc. Am. **109**, 748–763.

Kewley-Port, D., & Goodman, S. G. (**2005**). "Thresholds for second formant transitions in front vowels," J. Acoust. Soc. Am. **118**, 3252–3560.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Morrison, G. S. (**2006**). *L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach,* PhD diss., U. Alberta.

Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of vowel inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.

Zahorian, S., and Jagharghi, A. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–1982.

Yee, T. W., & Wild, C. J. (**1996**). "Vector generalized additive models," *J. Royal Stat. Soc. Series B (Methodological)* **58**, 481–493.