A COMPARATIVE STUDY OF EMPIRICAL FORMULAE FOR ESTIMATING VOWEL-FORMANT BANDWIDTHS

Mehrdad Khodai-Joopari & Frantz Clermont

School of Computer Science,
University of New South Wales (ADFA Campus),
Canberra, ACT 2600, Australia

ABSTRACT: The problem of estimating formant bandwidths (B1, B2, B3, etc.) for spoken vowels, under closed-glottis conditions, has in previous studies been approached in two major ways: one accommodates properties of the vocal tract, and the other is driven by curve fitting of measured data. This paper describes quantitative results obtained from a comparative investigation of selected formulae, which are representative of these two approaches. The accuracy of each formula is assessed using bandwidths measured by Fujimura & Lindqvist (1971) for 16 Swedish vowels. It is concluded that an empirical formulation, which is based on curve fitting and does not take into account lossy properties of the vocal tract or phonetic characteristics of the vowels, is likely to yield questionable B2 and B3 estimates. These potential inaccuracies therefore cast some doubts on the perceptual and the acoustic-articulatory robustness of bandwidth estimates obtained using curve-fitting approaches that are oblivious to intrinsic properties of the vowels or the vocal tract.

INTRODUCTION

It is well known that although formant frequencies of vocalic sounds can be estimated fairly accurately, formant bandwidths are notoriously difficult to measure from natural speech. Fortunately, the mechanism of human auditory system is such that it can compensate for some of these inaccuracies and therefore formant bandwidths are usually not considered to provide primary phonetic cues. For example, Carlson et al. (1979) showed that differences in formant bandwidths have rather little effect on listeners' judgments of vowel-like speech sounds.

By contrast, formant bandwidths have been shown to provide important cues for certain phonetic distinctions (e.g., vowel nasality) and, when discarded (Hermansky & Javkin, 1987), some otherwise available phonetic information might be lost. Perhaps less known is the related finding that certain formant bandwidths contribute to the auditory-phonetic integrity of dynamically changing sounds. This new perspective arises from perceptual evaluations of Dutch diphthongs synthesised from dynamically changing formant-bandwidths, which have led Peeters (1991; pp. 107-108) to argue that *"Normally it is taken for granted that, under normal conditions, bandwidths do not influence the phonetic quality of steady-state vowels…. We noticed, however, a strong influence of bandwidth values on our synthetic dynamic sounds which manifested itself in two ways: (1) as a factor of overall sound-quality, (2) as a factor that could split up a diphthong, especially in the case of B3"*.

Formant bandwidths have also been shown to play a determining role in recovering vocal-tract area functions from speech acoustics. For example, Rice & Ohman (1976) and Fant (1980) illustrated the possibility of producing infinitely different area functions from only one set of formant frequencies. The differences in bandwidth patterns produced in each case, suggested that the predictive power of bandwidths might be a solution to avoid ambiguities of this kind. In this vein, Mokhtari & Clermont (2000) derived a new functional form, which uniquely relates bandwidth-related components to odd-indexed coefficients of the Fourier sine series of the logarithmic area function. However, the use of formant bandwidths for alleviating the many-to-one mapping of area functions is still fraught with appreciable difficulty, since they are not reliably measured from natural speech.

Admittedly, several techniques for estimating bandwidths from formant frequencies have been proposed over the years, but their relative performance and reliability are not completely understood. Therefore, the focus of this study is to investigate such methods both theoretically and experimentally and find out their advantages, disadvantages, and limitations in yielding useful and accurate approximations.

MATERIALS AND METHODS

Reference Bandwidth Data

The formant-patterns (or F-patterns = [F1, F2, F3]) and bandwidth-patterns (or B-patterns = [B1, B2, B3]) used as reference data for this study are the sets of values (averaged over two male speakers' data), which are listed in Table I-C-II of Fant (1972, p. 48) but were measured by Fujimura and Lindqvist (1971) in their preceding study of 16 Swedish vowels. It is worth noting that, in their 1971 study, Fujimura & Lindqvist used an analysis-by-synthesis procedure to improve Fant's (1962) method of recording the frequency response of the vocal tract, thereby providing more reliable estimates of the vowel bandwidths.

Empirical Formulae

The above studies prompted other researchers to develop several mathematical formulations for estimating closed-glottis bandwidths (see Table 1). The formulae presented therein were selected for the following reasons: 1) they all represent closed-glottis bandwidths; 2) they were derived from measurements reported in Fant's (1962) and Fujimura & Lindqvist's (1971) studies on Swedish vowels, which employed similar measurement techniques and included female data; 3) two of the formulae accommodate female bandwidths. It is relevant to note that the resonances of the vocal tract are different in open- and closed-glottis conditions. Therefore, bandwidths for natural speech are expected to be somewhat larger because of the partially open glottis during phonation (Fujimura & Lindqvist 1971, p. 549). Below we briefly discuss these formulae and their respective motivation.

Table 1: Selected formulae for estimating vowel bandwidths from the resonances of the vocal tract under closed-glottis condition.

1. Fant-1972	$B1 = 15(500/F_1)^2 + 20(F_1/500)^{1/2} + 5(F_1/500)^2$ $B2 = 22 + 16(F_1/500)^2 + 12000\ /(F_3 - F_2)$ $B3 = 25(F_1/500)^2 + 4(F_2/500)^2 + 10F_3/(F_{4a} - F_3)$
2. Hawks & Miller-1995	$F_b = S \times (k + (x_1 \times F_c) + (x_2 \times F_c^2) + (x_3 \times F_c^3) + (x_4 \times F_c^4) + (x_5 \times F_c^5))$ $\begin{cases} k = 165.327516, \\ x_1 = -6.73636734\, e^{-01}, \\ x_2 = 1.80874446\, e^{-03}, \\ x_3 = -4.52201682\, e^{-06}, \\ x_4 = 7.49514000\, e^{-09}, \\ x_5 = -4.70219241\, e^{-12}, \end{cases}$ *for* $F_c < 500\,Hz$ $\begin{cases} k = 15.38146139, \\ x_1 = 8.10159009\, e^{-02}, \\ x_2 = -9.79728215\, e^{-05}, \\ x_3 = 5.28725064\, e^{-08}, \\ x_4 = -1.07099364\, e^{-11}, \\ x_5 = 7.91528509\, e^{-16}, \end{cases}$ *for* $F_c > 500\,Hz$ $S = 1 + 0.25\left(\frac{F0 - 132}{88}\right)$, *where average* $F_0 = \begin{cases} 132 & for\ men \\ 220 & for\ wemen \end{cases}$
3. TMF-1963	$Bn = 50\left(1 + F_n^2/(6 \times 10^6)\right)$

1. **Fant-1972:** derived a set of empirical formulae for estimating the first three formant-bandwidths (B1, B2, and B3), which embody relations between formant frequencies (F1, F2, F3, and averaged values of F4) up to 4KHz, and account explicitly and separately for each source of energy loss in the vocal tract. The factors $(500/F1)^2$, $(F1/500)^{1/2}$ and $(F1/500)^2$, respectively, account for wall vibration, cavity surface and radiation losses; the term (F3–F2) in B2 describes the radiation loss of a constricted tract with converging F2 and F3. The term F4a in B3 accounts for the subject's averaged formant frequency of F4, for which Fant suggested a value of 3400Hz. The B1 and B3 formulae can be extended to adult females by increasing the coefficient of the first term in B1 from 15 to 20, and by setting F4a to 3700Hz, while the B2 formula is applicable to both genders. A disadvantage of Fant's formulae is that they require knowledge of four formant frequencies and give estimates only up to B3.

2. **Hawks & Miller-1995 (HM-95):** derived a single formula based on regression analysis of Fujimura & Lindqvist's measured data. Two different sets of coefficients are needed for frequencies below and above 500Hz. This formula can be applied to both genders by altering the scalar S shown in Table 1, and is claimed by the authors to accommodate variations in "vocal tract size" (p. 1343).

3. **Tappert, Martony, and Fant (TMF)-1963**: used Fant's (1962) measured data to derive a formula, which approximates bandwidths by a parabolic function. This formula is completely independent of the F-pattern and indeed the simplest of those selected here for comparison.

EVALUATION PROCEDURES

There are two stages in our evaluation of each of the formulations described above. The first stage consists of a global evaluation, for which we adopt the same approach as that used by Fant (1972) and Hawks & Miller (1995). Bandwidth errors are expressed in dB [$20\log_{10}(B_m/B_p)$], where B_m and B_p are the measured and predicted bandwidths, respectively. The first stage also includes computations of: a) correlations between measured and predicted bandwidths using each formula; b) percentages of estimated bandwidths within acceptable Difference Limens (DL%) of the measured bandwidths using Flanagan's (1957) principle of JND's for formant bandwidths. The second stage consists of a detailed evaluation of each formulation, which is guided by articulatory-phonetic characteristics of vowels.

GLOBAL EVALUATION

The results of this evaluation are presented in Table 2. First, it can be seen that the TMF-63 formula has the lowest performance especially in the B1 range, which we seek to elucidate below.

Table 2: Absolute Mean Errors (dB), Correlation (ρ), and (DL%) between measured and predicted bandwidths for 16 Swedish vowels (Fujimura & Lindqvist, 1971) and estimated bandwidths (Fant-72, HM-95, and TMF-63).

		B1			B2			B3		
		Fant-72	**HM-95**	**TMF-63**	**Fant-72**	**HM-95**	**TMF-63**	**Fant-72**	**HM-95**	**TMF-63**
|Error| (dB)	**All**	1.25	1.12	2.55	1.29	1.95	3.21	2.06	3.23	3.72
ρ	**All**	0.89	0.90	0.46	0.78	0.16	0.27	0.77	0.44	0.45
DL%	**All**	75.0	75.0	37.5	100	87.5	50	75.0	56.25	56.25

A closer inspection of the TMF-63 formula indeed reveals that, in the F1 region, it tends to approximate bandwidths by a monotonically increasing function of the formant frequency (Fig. 1a). However, an important property reported by Fujimura and Lindqvist (1971, p. 548) is that bandwidths have higher values for lower frequencies of the first formant, which is due to wall vibration losses and is accounted for in Fant's (1972) formula by the F1-related term $(500/F1)^2$. To verify the apparent void in the TMF-63 formula, the latter was modified to include this F1-related term for frequencies below 500Hz. The effect of this modification is illustrated in Fig. 1b. It is clear from Fig. 1 and Table 2 that the TMF-63 formula is not only inadequate for estimating B1, but it also yields the worst B2 and B3 estimates by comparison with Fant-72's and HM-95's formulae. In consequence, the TMF-63 formula is no further considered.

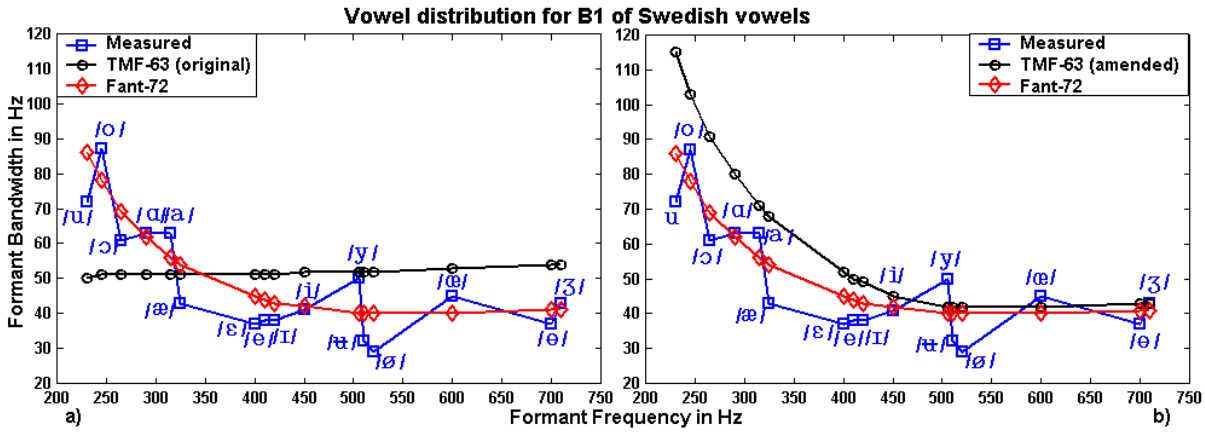

Figure 1: *Left graph*: TMF-63 original formulation. *Right graph*: TMF-63 formulation modified to account for sources of energy loss in the low frequencies.

Although the other formulae (Fant-72 & HM-95) have excelled in predicting B1, the former has outperformed the latter in the B2 and B3 regions. This is not surprising as these two formulae embody completely different approaches - Fant's is articulatorily and phonetically motivated, while Hawks & Miller's depends simply on the properties of fitted curves, and, although their contrastive performance is evident in Table 2, the consequences of choosing one over the other necessitate a more detailed evaluation, which takes into account the phonetic characteristics of the vowels.

PHONETICALLY MOTIVATED EVALUATION

In Fig. 2a, we have plotted the overall pattern of measured bandwidths versus formant frequencies of Swedish vowels (given as two-male speakers' averages in Table I-C-II of Fant (1972)) and their corresponding estimates from Fant-72 and HM-95 formulae. A first glance through Fig. 2a shows that the HM-95 formula estimates bandwidth values along a nonlinear curve, which does fit through the clouds of measured bandwidth values and of those obtained from Fant-72 formula, thus suggesting that it might be a solution to an F-pattern independent formula. Although Fig. 2a provides an overall impression of the strength of this formula in B1 and of its weaknesses in B2 and B3, it still does not give much information in phonetic terms. Therefore, in order to bring out the relations between specific categories of vowels and bandwidth patterns in more detail, we have decomposed Fig. 2a into Figs. 2b, 2c and 2d and kept the order of vowels in the same minimal phonetic steps as those adopted by Fant. We have also decomposed the Absolute Mean Errors from Table 2 on the basis of two broad phonetic categories (front & back), whose respective contributions are given in Table 3.

Table 3: Absolute Mean Errors (dB) - decomposed for Front & Back vowels - between measured and predicted bandwidths for 16 Swedish vowels (Fujimura & Lindqvist, 1971) and estimated bandwidths (Fant-72, HM-95)

		B1		**B2**		**B3**	
		Fant-72	**HM-95**	**Fant-72**	**HM-95**	**Fant-72**	**HM-95**
|Error| (dB)	**Front**	1.15	0.90	1.24	2.36	2.37	3.17
	Back	1.13	1.26	1.08	1.09	1.25	3.91

Inspection of the vowel sequence charts in Figs. 2b, 2c and 2d, and values from Table 3 for the B1, B2 and B3 regions, uncovers further results, which are not readily observed in the overall picture depicted in Fig. 2a, but which shed more light on the curve-fitting approach underlying the HM-95 formula.

1) Results from Tables 2 and 3, and Fig. 2b for B1, show that the separate set of coefficients built into the HM-95 formula for frequencies below 500 Hz guarantee a relatively high level of accuracy at low frequencies. One could argue that this separate treatment of the B1 range is equivalent to accounting for wall vibration losses at low frequencies.

2) By contrast, observations of B2 and B3 and the decomposed values of overall errors (see Table 3 and Figs. 2c and 2d), show relatively low levels of accuracy, and weak correlations between measured B2 and B3, and their estimated values. Even significant discrepancies are observed in some cases; for example, one can observe that in the B2 region, the HM-95 formula yields higher values for closed-front vowels. This is clearly a contradiction to what was reported for measured bandwidths by Fujimura & Lindqvist (1971, p. 548), *"There are some correlations of the bandwidth values with some particular articulatory features. The second formant bandwidths appear to be generally higher for more-open vowels than for closer vowels. This is clear in the case of front vowels".*

3) Further elucidation of the B2 profile obtained from HM-95 and shown in Fig. 2c prompted the calculation of correlations between formant frequencies and bandwidths for both measured and estimated bandwidths. The results indicate a strong correlation of 0.944 between measured F2 and HM-95's B2, while the correlations between measured F2 and B2, and those between measured F2 and Fant-72's B2 are only 0.315 and 0.339, respectively. It is these weak correlations between B2 and F2, which are more aligned with Fujimura & Lindqvist's (1971, p. 548) empirical finding that B2 has less dependency on formant frequencies than B1 does. Observations concerning B3 reveal a significant increase in magnitude of the error for both back- and front-vowels with absolute mean

values of 3.91 and 3.17 dB respectively (see Table 3), and a significant reduction of percentage values within the acceptable Difference Limen (see Table 2).

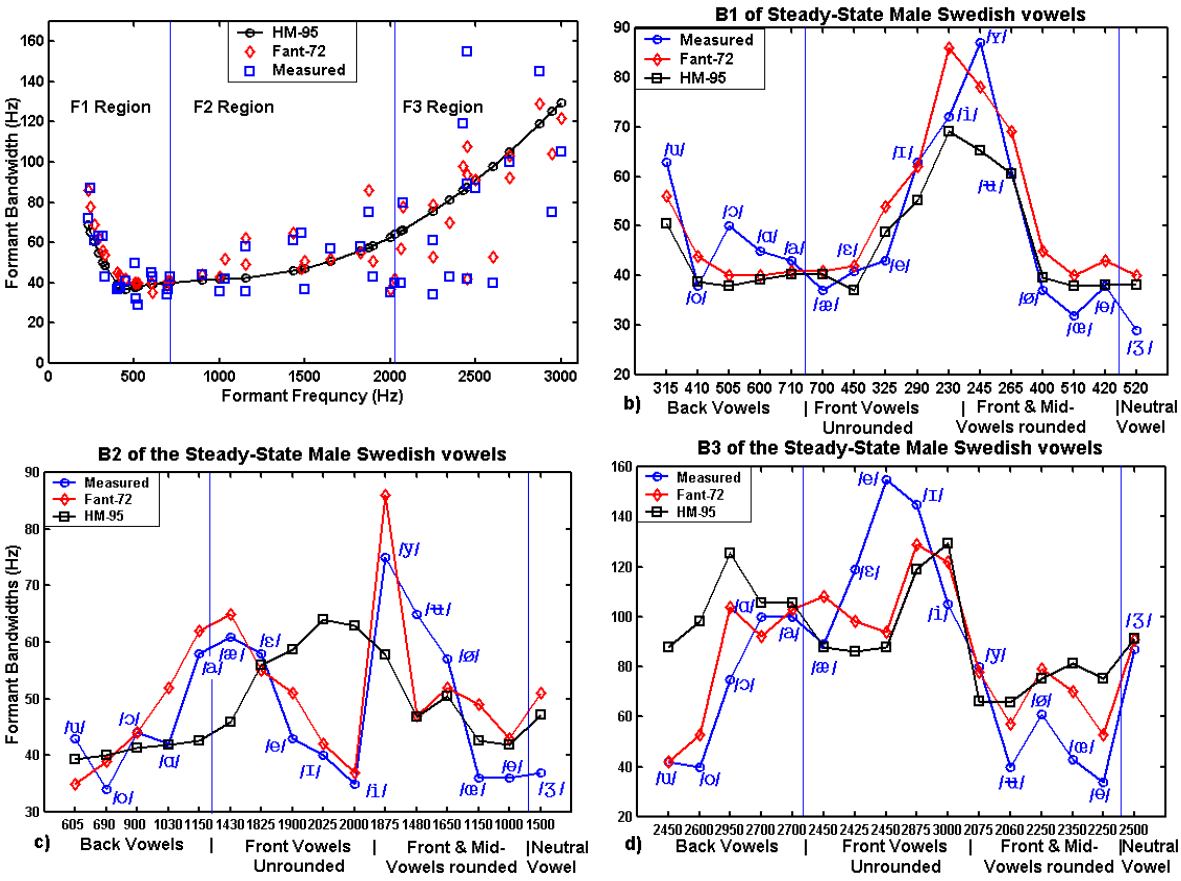

Figure 2: a) Formant frequencies versus bandwidths (Fujimura & Lindqvist's measured, Fant-72- and HM-95 formulae) for 16 Swedish male vowels. b), c), and d) represent B1, B2 and B3 respectively, versus vowels in minimal phonetic steps.

We can now confidently argue that the HM-95 formula's low performance in the high-frequency range results from using the same set of coefficients for all frequencies above 500Hz and from discarding any relations between frequencies and corresponding bandwidths in this region. To obtain a good approximation of bandwidth values, therefore, it is necessary to somewhat constraint the inter-formant dependency relation of bandwidths, and to take into account different sources of energy loss corresponding to different frequencies and different articulatory positions of the vocal tract. According to Fant (1960, p. 121), *'the second formant is to a large extent tuned by the back cavity and tongue constriction."* Consequently, if Hawks & Miller had enforced this characteristic of the second-formant frequency, their formula could have provided more accurate B2 and B3 estimates. However, enforcing such constraints would have conflicted with their original proposal of an F-pattern independent model.

CONCLUDING DISCUSSION

We have found that, in order to obtain phonetically meaningful estimates of measured bandwidths in the high-frequency range, certain factors are needed to distinguish between front and back vowels, which are not automatically accounted for in a curve-fitting approach. It was also found that an F-pattern dependent formula is necessary to account for sources of energy loss in the vocal tract. In support of these criteria, our experimental results show that Fant's 1972 elaborate formulae give excellent approximations to measured (closed-glottis) bandwidths.

By contrast, Hawks & Miller's formula, which is based solely on curve fitting and does not distinguish between different sources of energy loss at higher frequencies, fared poorly against the same

measured data, particularly in B2 and B3. These findings are at odds with Hawks & Miller's claim regarding the accuracy of their formula (1995, p. 1344), "*Overall, the comparison suggests that the F-pattern independent, bandwidth estimation procedure can provide reasonably accurate estimates of bandwidth measurements from natural speech and should serve well for some types of speech synthesis"*. While this may still be valid to some degree, the authors provide neither the statistical evidence nor do they outline the synthesis approaches that would support the viability of their formula.

Admittedly, the accuracy of lower- and upper-formant bandwidths may not be an important perceptual factor in some cases, but it is worth re-stating Peeters' (1991) finding that higher-formant bandwidths, especially B3, seem to play a percept-conditioning role in dynamic sounds. Since Peeters used pre-fabricated estimates for bandwidths, it would be useful to determine the perceptual robustness of the range of estimates yielded by Fant's and Hawks & Miller's formulae.

Formant bandwidths are no less important in the acoustic-articulatory domain, where, for example, they have been shown to play a constraining role in the recovery of unique area functions from the formants. In this vein, Fant (1980, p. 85) suggests that for estimating vocal-tract area functions, "A *more efficient set of acoustic parameters would be F1, F2, F3, and B3,"* and *"… B3 is a good correlate of degree of lip opening and also mouth opening."* Should B3 prove to be a critical parameter for securing unique or plausible area functions, then the use of a curve-fitting approach *à la* Hawks & Miller might be questionable since it produces B2 and B3 estimates that are quite different from measured bandwidths. Further work is therefore warranted to determine the extent to which accuracy and selection of lower and upper-formant bandwidths are critical to the search for more realistic area functions of the vocal tract.

REFERENCES

Carlson, R. Granstrom, B. & Klatt. D, (1979) *Vowel perception: The relative perceptual salience of selected acoustic manipulations.* Speech Transmission Laboratory Quarterly Progress and Status Report, 3-4, 73-83.

Fant, G. (1960) *Acoustic theory of speech production*. Mouton, The Hague, The Netherlands.

Fant, G. (1962) *Formant bandwidth data,* Speech Transmission Laboratory Quarterly Progress and Status Report, 1, 1-2

Fant, G., (1972) *Vocal tract wall effects, losses, and resonance bandwidths,* Speech Transmission Laboratory Quarterly Progress and Status Report, 2-3, 28-52.

Fant, G. (1980) *The relations between area functions and the acoustic signal,* Phonetica, 37, 55-86.

Flanagan, J.L. (1957) *Difference limens for formant amplitude.* Journal of Speech and Hearing disorders, 22, 205-212.

Fujimura, O. & Lindqvist, J. (1971) *Sweep-Tone Measurements of Vocal-Tract Characteristics,* The Journal of the Acoustical Society of America, 49, 541-558.

Hawks, J.W. & Miller, J.D. (1995) *A Formant bandwidth estimation procedure for vowel synthesis.* Journal of Acoustical Society of America, 97(2), 1343-1344.

Hermansky, H. & Javkin, H.R. (1987) *Evaluation of ASR front-ends using synthetic speech.* Research Reports. Speech Technology Laboratory Division of Panasonic Technologies, Inc., 1, 11-1, 11-13

Mokhtari, P. & Clermont, F. (2000) *New perspectives on Linear-Prediction modeling of the vocal-tract: Uniqueness, formant-dependence and shape parameterization*, Proceedings of the 8th Australian International Conference on Speech, Science and Technology, 478-483.

Peeters, W.J.M. (1991) *Diphthong Dynamics: A cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and* German, PhD thesis, The University of Utrecht, The Netherlands.

Rice, L. & Ohman, S. (1976) *On the relationship between formant bandwidths and vocal tract shape features.* UCLA Working Papers in Phonetics 31, 27-31.

Tappert, C.C., Martony, J. & Fant, G. (1963) *Spectrum envelopes for synthetic vowels.* Speech Transmission Laboratory Quarterly Progress and Status Report, 3, 2-6.