SYSTEMIC COMPARISON OF SPOKEN AND SUNG VOWELS IN FORMANT-FREQUENCY SPACE

Frantz Clermont
School of Computer Science,
University of New South Wales (ADFA Campus),
Canberra, ACT 2601, Australia
frantz@cs.adfa.edu.au

ABSTRACT: A new approach is introduced for uncovering dominant characteristics of steady-state vowels produced in sung phonation. The approach is motivated by the articulatory-phonetic notion of a vowel space and uses the correlate space of formant frequencies as a systemic tool. Two spatial representations (F1-F2, F1-F2-F3) are examined using formant frequencies measured from Australian English vowel-nuclei of /hVd/ monosyllables, which were spoken at an estimated fundamental-frequency (F0) of 80-Hz and sung at a nominal F0 of 110-Hz by a semi-professional *bass* singer. An asymmetric retraction of the sung relative to the spoken polygon is clearly evident in F1-F2 space, where the unrounded front vowels are shifted towards the apparently less susceptible back vowels. In F1-F2-F3 space, spoken and sung vowels cluster tightly about distinct quadratic surfaces, which facilitate a more complete interpretation of the retraction in F1-F2 space. The emergent perspective is that spatial representations of spoken and sung vowel-formants provide useful pathways for interpreting acoustic-articulatory strategies in singing.

INTRODUCTION

The underlying theme of this study is partially echoed in one of several familiar *dicta* from the historic Italian School of vocalism (Miller, 1996: Ch. 15) – "*si canta come si parla*" (one sings as one speaks), which is at the core of some of the main issues related to the pedagogy and the science of the singing voice. On the one hand, it seems quite reasonable to presume that, since the same vocal tract can accommodate speaking and singing, both acts must obey the same laws of resonance-tube acoustics and be subject to the same articulatory constraints. As far as the physical sound instrument is concerned, therefore, it can be argued that one indeed sings as one speaks. On the other hand, the taught or the perceived act of singing does "not simply reduce to sustained speech spun over wide-ranging pitch fluctuations" (Miller, 1996: p. 51); rather it involves careful gestures aimed at maintaining phonetic integrity through proper shapes of the vocal tract. This tension between the anatomically unchanging vocal tract and the necessity of altering its behaviour for effective and correct singing, does prompt the question of whether the singer uses a different articulatory strategy from that for speaking, or whether s/he adopts certain invariant articulatory postures that are learnt for speaking.

A full elucidation of this question demands, for a range of singing voices, a whole body of acoustic, articulatory, physiological and perceptual data, none of which are currently available in comprehensive form nor are they trivial to measure because of either the time-consuming, immature or invasive procedures involved. This notwithstanding, one accessible and important port of entry is linked with the fact that, for relatively low phonation frequencies, it is possible to reliably measure the low-ordered formant-frequencies directly from the acoustic signal, and to interpret them according to vocal-tract length and certain articulatory movements. It is, therefore, possible and useful to examine formants' behaviours in sung phonation of sounds like the vowels, about which a great deal is already known in spoken phonation. However, our point of departure from previous studies with similar aims, such as those reported by Sundberg (1970; 1974), is that we propose to exploit the articulatory-phonetic notion of a vowel space for cross-examination of spoken and sung vowels, thereby seeking to expose dominant characteristics that may not be evident or discernible on a per-formant or a per-vowel basis.

Our overall contention then is that a systemic analysis of the vowel formant space is necessary in order to gain a more holistic perspective on acoustic-articulatory differences between speaking and singing. Furthermore, the proposed analysis of the formant-frequency structure of sung versus spoken vowels has apparently not been attempted for Australian English and, consequently, this study contributes fresh results to the assembly of knowledge on the acoustics of the singing voice.

VOWEL DATASET & ACOUSTIC PARAMETERISATION

Subject, Script and Recording Procedures

The subject engaged for, but not told about, the purpose of this study is an adult-male, semi-professional singer, who is a native speaker of Australian English. He has several years of training in Western classical singing and has won regional Australian championships in the *bass* voice category.

The word list used for recording consists of 11 /hVd/ monosyllables ("heed", "hid", "head", "had", "hard", "hudd", "hod", "hoard", "hood", "who'd" and "herd"), which were presented on separate flash cards at intervals of approximately 4 seconds. In a quiet and non-reverberant room, the subject used a head-mounted microphone and adopted a standing posture for all recordings. He was asked first to speak 5 randomised tokens of each syllable at his habitual speaking rate (estimated at 80 Hz on average). Following a pause of about 15 minutes, a 110-Hz tone was played, several times in succession, to facilitate vocal preparation for the sung recordings. In the same random order used for the spoken recordings, the subject sang the /hVd/-syllables at a fundamental-frequency found on average to be within 10 Hz of the tone played. These procedures were performed once in September 2001 and once in November 2001, but only the September data have thus far been analysed. For both datasets, the analogue signals were sampled at 11,025 Hz and quantised to 8 bits.

Formant-Frequency Measurement

Prior to any acoustic processing, the vowel-nucleus' steady-state section was isolated from each of the 110 (55 spoken and 55 sung) syllables recorded in September 2001. This segmentation stage consisted of identifying spectrographic sections through which the darkest bands of energy appeared to be stationary, concurrently with an auditory validation of the corresponding waveform intervals.

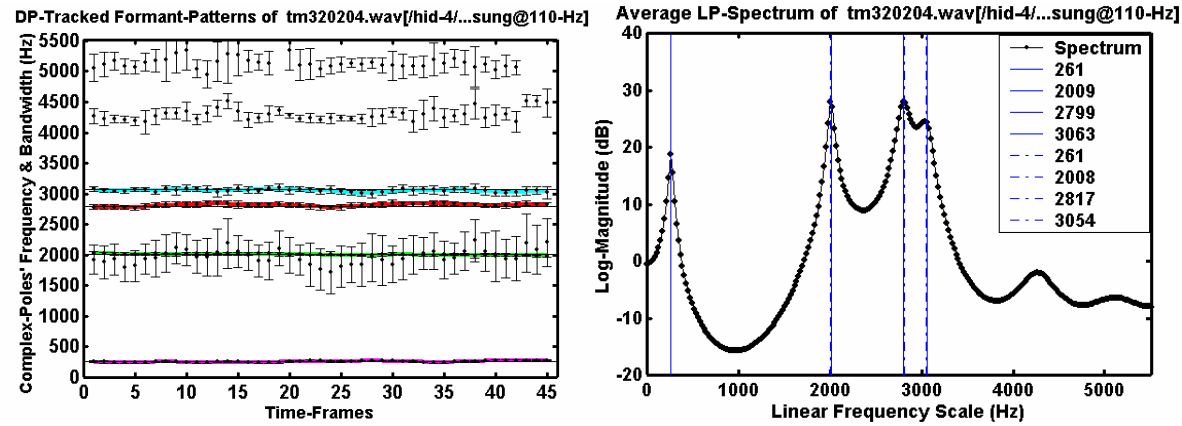

Figure 1: Formant tracks and average LP-spectrum of steady-state section of sung /hid/

The next stage consisted of linear-prediction (LP) analyses through Hanning-windowed frames of 30-msec duration sequenced every 10 msecs. For 10% of the spoken data, the LP-order had to be increased to 16 from a default value of 14, and to 20 for 20% of the sung data, in order to enhance the F3 and F4 regions. For each steady-state nucleus, then, the LP-analyses yielded a set of frame-by-frame LP-poles, among which the 4 lowest formant-frequencies ([F1, F2, F3, F4] = F-pattern) were selected using a tracking method (Clermont, 1992) based on analysis-by-synthesis (AbS) and dynamic programming (DP). DP-tracked, F-patterns are illustrated in the left graph of Figure 1 and seen to pass through the narrow-bandwidth poles. This is achieved by the AbS, which identifies the poles corresponding to prominent spectral peaks, jointly with the DP optimisation, which secures temporal continuity. The singer's *vibrato* manifests itself as a superimposed undulation, which poses a methodological problem in determining representative F-patterns. However, the steady-state nature of the data to be ultimately analysed justifies retaining frame-averaged F-patterns (illustrated by dashed vertical lines at right graph), which were further checked for consistency (*inter alia*: White, 1999) with the F-patterns corresponding to the frame-averaged spectra (illustrated by solid vertical lines at right graph). For the purpose of this study, only F1, F2 and F3 are subsequently examined.

MEASUREMENT CONSISTENCY & INTER-TOKEN VARIABILITY

Measured F-patterns are expected to exhibit a certain degree of variability induced not only by the measurement method employed, but also by one's inability to reproduce sounds that are spoken and presumably sung, in *exactly* the same way. It is therefore relevant to first note that the mean differences between frame-averaged and averaged-spectra's F-patterns range from 0.34 to 4.7 Hz for the spoken vowels and from 0.07 to 3.7 Hz for the sung vowels. This indicates that the F-patterns measured are indeed very similar across the steady-state frames and that the formant-tracking method used is internally consistent. It is then justified to turn to inter-token dispersions (ITD) for any intrinsic regularity among the averaged, steady-state F-patterns. The ITDs given in Table 1 are quite small for F1 of both spoken and sung vowels, but larger for F2 and F3 of sung vowels. These contrasts could be said to reflect a strong consistency in mandibular positions for both phonation types, but more variability in the subject's lingual positions during singing. The per-formant ITDs increase expectedly from F1 to F3, and their respective ranges ([13-16, 29-68, 51-69]-Hz) lie within difference-limens (Flanagan, 1955) for human perception. In sum, there appear to be no gross measurement errors or unexpected irregularities that could cast doubts on our F-patterns' viability and hence discourage further analyses.

Table 1. Inter-token dispersions (ITD) (= standard deviations in Hz) for averaged, steady-state F-patterns.

BROAD VOWEL CATEGORY	SPOKEN			SUNG		
	F1	F2	F3	F1	F2	F3
FRONT	13	43	60	13	68	69
BACK	16	29	51	15	33	61
ALL	15	36	55	14	51	65

SEQUENCE CHARTS OF SPOKEN & SUNG VOWELS

Sequence charts like those shown in Fig. 2 (Australian English (AE) at left and Swedish (SW) at right) have a long history but, more importantly, they embody a compelling structure within which variations in F-patterns are readily observed vowel by vowel along the abscissa. For example, a glance through both charts immediately reveals that, between the spoken and sung vowels under consideration, there are subtle differences in F1 across all vowels, whereas the high-F2 vowels display the most striking contrasts with lower F2 and lower F3 values in sung phonation. A brief cross-examination of the AE and SW charts also highlights spoken-sung contrasts that appear to be quite similar in F1 and F2, but relatively less so in F3. Beyond the componential observations that are clearly afforded by the vowel-sequenced chart on a per-formant basis, it is difficult to progress towards a systemic perspective without appealing to inter-dependencies among formant frequencies in the entire vowel space. Consequently, it becomes necessary to consider co-varying representations of the F1, F2 and F3 dimensions in order to be able to examine, *in toto*, both the high-F2 or front vowels targeted above and the seemingly less susceptible, low-F2 or back vowels.

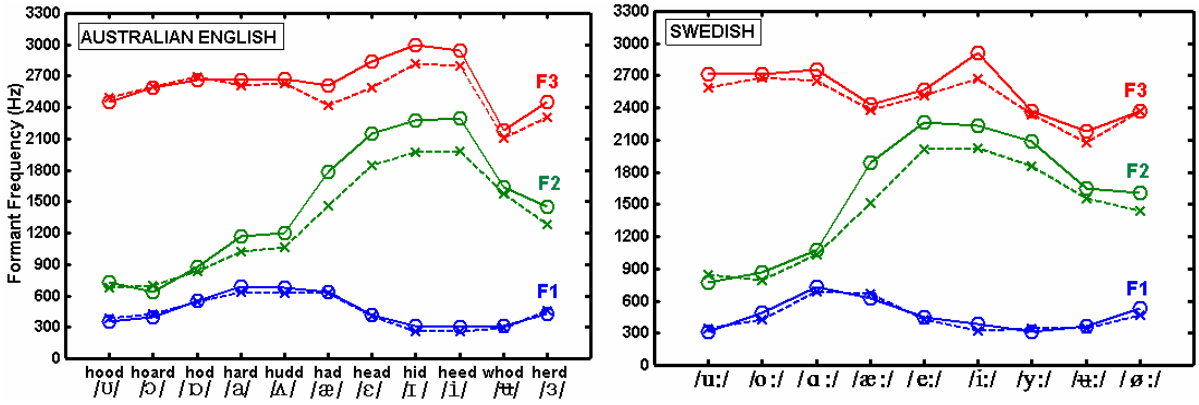

Figure 2: *Left graph*: F-patterns (5-token averages) of Australian English (AE) vowel-nuclei spoken and sung in /hVd/ context by 1 Australian bass singer. *Right graph* (captured digitally from Sundberg's (1970: p. 30) Fig. 1: singer-B4): F-patterns (3-token averages) of Swedish (SW) vowel-nuclei spoken and sung in /rV/ context by 1 Swedish bass singer. Symbols: [circles➔spoken; crosses➔sung at nominal F0=110 Hz.].

F1-F2 SPACE OF SPOKEN & SUNG VOWELS

As resonance frequencies of the vocal tract, F1 and F2 are interpretable in terms of two major articulatory dimensions (mandibular and lingual, respectively) in vowel production. In addition, the ensemble of (F1,F2)-coordinates tends to specify operating limits (Potter & Steinberg, 1950), within which maximal separation is expected among steady-state vowels from the same speaker or presumably the same singer. For these reasons, the F1-F2 space should prove useful for gaining a systemic impression of articulatory-phonetic contrasts between spoken and sung, steady-state vowels.

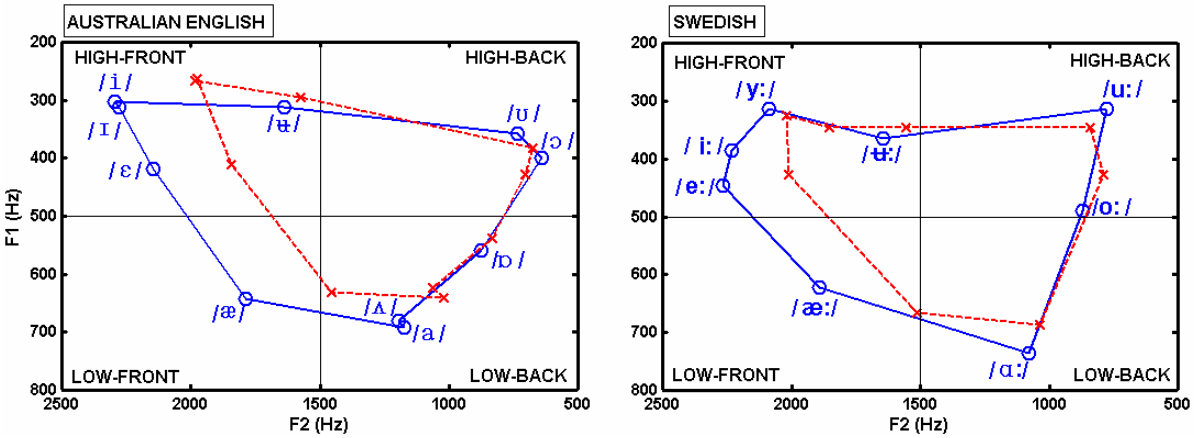

Figure 3. *Left Graph*: Australian English vowel polygon. *Right Graph*: Swedish vowel polygon. Symbols: [circles➔spoken; crosses➔sung at nominal F0=110 Hz; standard phonemic labels are assigned to the spoken vowels to enhance visual contrasts with the sung counterparts, i.e., no explicit auditory validation is implied.]

A global inspection of Fig. 3 (at left) reveals an asymmetric retraction of the sung relative to the spoken polygon, resulting from a major shift of the front vowels towards the back vowels, except for /ʉ/ presumed to be rounded in spoken phonation. The same contrasts apply to Swedish (Fig. 3 at right), including also a minor shift of the homologue, front vowel /ʉ:/. The sung polygon is retracted and, *a fortiori*, reduced in area with respect to the spoken polygon. The unrounded, high- and low-front vowels contribute the significant proportion of the reduction, while the high- and low-back vowels seem less susceptible to changes brought upon by the bass singing voice. The global impression is this – nearly all front vowels of 2 different vowel systems sung by 2 respective bass singers have been produced with different articulatory gestures from those used by the same singers for speaking these vowels. What are these gestures likely to have been and what could they imply?

A closer inspection of the polygons from an acoustic-articulatory point of view helps shed some light on this question. It is clear that the significant drops in F2, which indicate a lengthening of the vocal tract, have caused the major shift of the unrounded front vowels. A deeper interpretation arises from the fact that the vocal tract will increase in length as a result of larynx depression and/or lip rounding/protrusion. One could thus argue with Sundberg (op. cit.) that both singers must have lowered their larynx for singing the unrounded front vowels. Since these vowels' F3 are also lowered (Fig. 2), more so for the Australian singer, one could surmise that both singers have yielded to the concomitancy of lip rounding with larynx movement (Perkell, 1969). However, a tighter perspective seems to emerge if the small shift of the rounded front vowels is put in context with the major shift of the unrounded front vowels from spoken to sung phonation – namely that larynx height and labial re-adjustments are desirable if not necessary as articulatory settings (Laver, 1980; Nolan, 1983) for (bass) singing, albeit with differing degrees that will depend on the vowel's susceptibility to such settings. Accordingly, for the back vowels, which are presumably spoken with a relatively low tongue position (Perkell, op. cit.) and thus less likely to require major adjustments of vocal-tract length in (bass) singing, no significant drop might be expected in F2, which is indeed seen in both singers' data. By contrast, mandibular re-adjustments are likely to occur, which are to some degree manifest in the F1 data from both singers' back vowels.

The evidence based on F1-F2 data for 2 bass singers of 2 different vowel systems tends to lean against the *dictum* from the early Italian School of vocalism. Is this view upheld in F1-F2-F3 space?

F1-F2-F3 SPACE OF SPOKEN & SUNG VOWELS

The striking phenomenon in F1-F2 space is the retraction of the sung polygon towards the back-vowel side of the spoken polygon. However, for both AE and SW, the retraction does not alter the fact that the sung polygon is still basically contained within the spoken polygon. Are the two geometrically related phenomena reconcilable with a distinct articulatory strategy in singing?

In an attempt to answer this question, we turned to a higher-dimensional representation that relates F3 to F1 and F2. This approach is not new (Broad & Wakita, 1977: bi-planar formulation; Kasuya & Yoshizawa, 1992: quadratic formulation) but it has apparently not been applied to a sung-vowel space. The quadratic formulation was adopted because it requires no *a priori* partitioning of the vowel space, and two modelling steps were followed: (i) to evaluate the goodness of 2 quadratic surfaces separately fitted to spoken (Fig.4: left graph) and sung (Fig. 4: right graph) vowels; and (ii) to evaluate the goodness of 1 surface fitted to the pooled spoken and sung vowels.

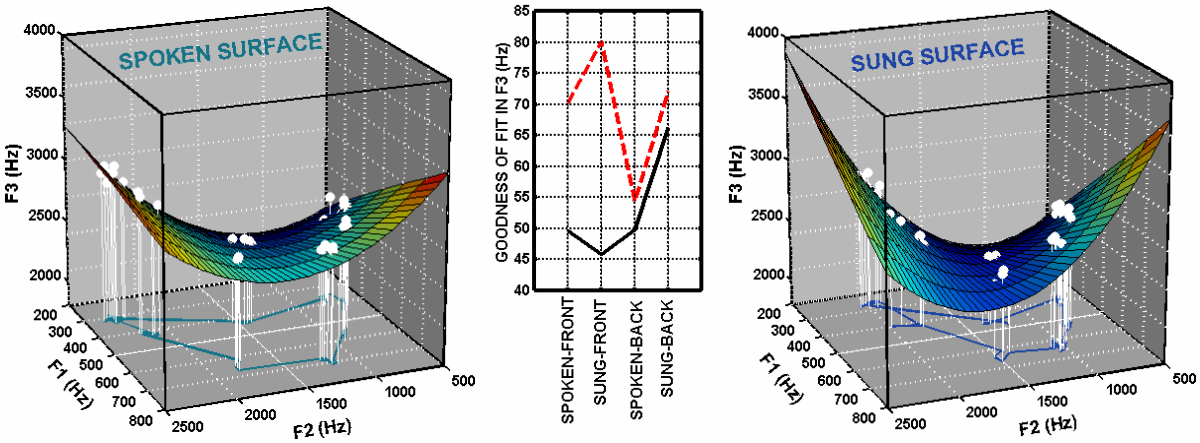

Figure 4. 3-D representation of steady-state F1, F2 and F3 (5 tokens) for Australian English vowels (/ʉ/ excluded) spoken (*left*) & sung (*right*) by 1 Australian bass singer: [while balls represent individual tokens projected onto F1-F2 space; quadratic surfaces (see Table 2) are fitted through the white balls]. *Middle graph*: root-mean-squared (r.m.s.) errors in fitting back- & front-vowels F3 [solid line➔2 separate surfaces (spoken at left; sung at right); dashed line➔1 surface to all spoken & sung vowels pooled.]

The results obtained in step (i) are encouraging. When 2 quadratic surfaces are separately fitted to spoken and sung vowels, the r.m.s values for F3 (Table 2 & solid line in Fig. 4's middle graph) are within or close to the ITDs for F3 (Table 1), which indicates a tight fit. Surface saddles are parallel to spoken and sung back-vowels, which appear to anchor both spaces. Sung back-vowels have a slightly reduced F1-range but are otherwise parallel to the spoken ones. Sung front-vowels are displaced towards sung back-vowels owing to lowered F2 and F3, which together depress the sung surface.

Table 2. Spoken & sung surface parameters, and goodness of fit (r.m.s. in Hz)

PHONATION	$F_3 = \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_1^2 + \alpha_4 F_2^2 + \alpha_5 F_1 F_2$						Goodness of Fit (Hz)	
	α_0	α_1	α_2	α_3	α_4	α_5	BACK	FRONT
SPOKEN	2261	2.32	-8.14e-001	-1.68e-004	4.99e-004	-1.04e-003	49.88	49.68
SUNG	2186	3.20	-1.11	2.82e-004	7.79e-004	-2.08e-003	66.30	45.79

The results obtained in step (ii) are enlightening. When a single quadratic surface is fitted to the pooled spoken and sung vowels, a clear dichotomy is observed (dashed line in Fig. 4's middle graph): (a) the goodness of fit to the back-vowels' F3 remains nearly unchanged from that obtained with separate quadratics; (b) in sharp contrast, the goodness of fit to the front vowels worsens by a factor close to 2.

Collectively, results (i) and (ii) provide complementary confirmation of a genuine dichotomy in the way in which the sung-vowel space differs from the spoken surface. This also strengthens the plausibility of the evidence, uncovered in F1-F2 space, for a distinct strategy in singing particularly the front vowels that are unrounded in spoken phonation.

CONCLUSION

We have introduced in this paper a comparative approach for uncovering dominant behaviours of the human vocal tract in singing and speaking phonation. The approach imports the articulatory-phonetic notion of a vowel space, and it appeals to inter-dependencies among the 3 lowest formant frequencies as a systemic tool. This is our point of departure from previous studies, where approaches have tended to be more componential.

The bass singing voice has been the focus of this study, for which formant frequencies were measured from steady-state nuclei of spoken and sung vowels in Australian English. An asymmetric retraction of the sung relative to the spoken polygon is clearly evident in F1-F2 space, which is caused by a major shift of the unrounded front vowels towards the less susceptible back vowels. In F1-F2-F3 space, spoken and sung vowels cluster tightly about distinct quadratic surfaces, which provide complementary confirmation of a genuine dichotomy between sung front- and spoken front-vowels. The systemic perspective arising from these results points to certain articulatory settings in singing, which do not seem to indicate a rigid adherence to a speaking strategy. Further work will be necessary to substantiate this claim in the articulatory domain, where vocal-tract shapes inferred from the formants are expected to provide more explanatory clues to the phenomena uncovered here.

The dimensionality of the data used for this study is small and, therefore, no general claim can be made about the acoustics of the bass singing voice. However, it should be recalled that the spoken-sung contrasts in our Australian English (F1, F2) data were found to be comparable to those manifest in Sundberg's (1970) Swedish data for the same voice category. This lends credence to our formant measurements, but it also raises the question of a plausible homogeneity within voice categories. Parametric surfaces such as those proposed here could be exploited to investigate this question.

ACKNOWLEDGEMENTS

I thank Mr Tom Millhouse for agreeing to donate his spoken and sung vowels. I express my sincere gratitude to Dr David Broad, Dr Philip Rose, Mr Michael Morse and Mr Mehrdad K.-Joopari for their encouragement and helpful comments.

REFERENCES

Broad, D.J. and Wakita, H. (1977). *Piecewise-planar representation of vowel formant frequencies,* Journal of the Acoustical Society of America 62, 1467-1473.

Clermont, F. (1992). *Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching,* Proceedings of the 4th Australian International Conference on Speech Science and Technology, Brisbane, Australia, 48-53.

Flanagan, J.L. (1955). *A difference limen for vowel formant frequency,* Journal of the Acoustical Society of America 27, 613-617.

Kasuya, H. and Yoshizawa, S. (1992). *Geometric representation of speaker individualities in formant space and its application to speech synthesis*, Proceedings of the 14th International Congress on Acoustics, Beijing, China, paper G3-10.

Laver, J. (1980). *The phonetic description of voice quality,* Cambridge University Press.

Miller, R. (1996). *On the art of singing,* Oxford University Press.

Nolan, F. (1983). *The phonetic bases of speaker recognition,* Cambridge University Press.

Perkell, J.S. (1969). *Physiology of speech production,* M.I.T Press.

Potter, R.K. and Steinberg, J.C. (1950). *Toward the specification of speech,* Journal of the Acoustical Society of America 22, 807-820.

Sundberg, J. (1970). *Formant structure and articulation of spoken and sung vowels,* Folia Phoniatrica 22, 28-48.

Sundberg, J. (1974). *Articulatory interpretation of the "singing formant",* Journal of the Acoustical Society of America 55, 838-844.

White, P. (1999). *Formant frequency analysis of children's spoken and sung vowels using sweeping fundamental frequency production*, Journal of Voice 13, 570-582.