Systemic Relationship between Bass Singers' Spoken and Sung Vowel-Formant Spaces

Frantz Clermont
School of Computer Science,
University of New South Wales (ADFA campus),
Canberra, ACT 2600, Australia
E-mail: frantz@cs.adfa.edu.au

ABSTRACT

A systemic approach is described for uncovering characteristics of sung with respect to spoken vowels obtained from two bass singers – Australian English and Swedish. The approach imports the notion of articulatory settings to elucidate spoken-to-sung transformations that are basically linear in the space of the two lowest, formant frequencies. This study departs from previous ones, which have tended to focus on individual components and, instead, focuses on how the singer's vowel system is affected.

1. INTRODUCTION

We describe in this paper a systemic approach for elucidating acoustical consequences of certain articulatory settings (Laver, 1980; Nolan, 1983), which are applied by two bass singers (Australian and Swedish) in their respective transition from spoken to sung vowels. The systemic concept is conveyed in Fig. 1, where spoken and sung vowels are represented as individual components of two vowel systems, which are related to each other via a certain transformation in formant-frequency space. This is our point of departure from previous studies (*inter alia*: Sundberg, 1970; 1974), where approaches have tended to focus on individual vowels and their individual formants. Here we show that there are deeper insights to be gained by looking at the vowel system as a whole, and particularly insights into the changes in articulatory settings, which underlie the necessary modification of speaking for effective singing gestures (Miller, 1996).

The two lowest formant-frequencies (F1 and F2) will be our systemic tool for comparing spoken with sung vowels, as the former are broadly interpretable in terms of two major articulatory dimensions (mandibular and lingual, respectively) in vowel production. Our first step will be to seek systemic regularities between spoken and sung vowels as they are manifest in the two singers' F1-F2 spaces. Our second step will be to determine a parametric form of the spoken-to-sung transformation in F1-F2 space as a means of comparing objectively the two singers' transitions from speaking to singing. Our third step will be to confirm the dominant setting for that transition, in the light of Lindblom and Sundberg's (1971) formant data on Swedish vowels simulated with normal and lowered larynx positions.

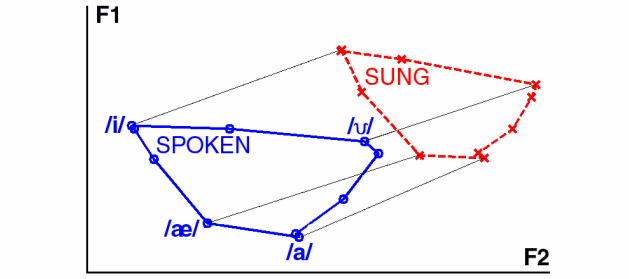

Figure 1: Systemic concept underlying the spoken-to-sung vowel transformation in formant-frequency space.

2. SPOKEN & SUNG VOWEL DATA

Two of the three formant datasets employed for this work were imported from previous studies of Swedish vowels, for which only a brief outline is given below. The dataset used for comparing spoken with sung vowels in Australian English is new and, therefore, our recording and formant-estimation procedures are more detailed.

Swedish (SW) Vowels

Average F-patterns for SW vowel-nuclei are shown in Fig. 3(b), which were measured (Sundberg, 1970) from /rV/-syllables spoken and sung by a SW bass singer. Fig. 3(c) gives F-patterns for SW vowels, which were simulated by Lindblom and Sundberg (*op. cit.*) with (i) a normal larynx position obtained from a SW-speaker's X-ray of [ɑ]; (ii) and a lowered position fixed at 1 cm below normal.

Australian English (AE) Vowels

Our subject is an adult-male, semi-professional singer, and a native speaker of Australian English. He has several years of training in Western classical singing, and won regional Australian championships in the *bass* voice category. He spoke 5 randomised tokens of 11 monosyllables ("heed", "hid", "head", "had", "hard", "hudd", "hod", "hoard", "hood", "who'd" and "herd") at his habitual speaking rate (F0=80 Hz on average). Queued with a 110-Hz tone, our subject sang the same tokens at an F0 close to the tone played. The analogue signals were sampled at 11,025 Hz and quantised to 8 bits. The steady-state section of every vowel nucleus was then isolated spectrographically and auditorily, prior to formant-frequency measurements.

ISBN 1-876346-48-5

Formant-Frequency Measurement

Formants were measured using linear-prediction (LP) analyses through (Hanning) windowed frames of 30-msec duration, by steps of 10 msecs. For 10% of the spoken data, the LP-order was increased to 16 from a default value of 14, and to 20 for 20% of the sung data, in order to enhance the upper formant regions. For each steady-state section, the LP-analyses yielded a set of frame-by-frame poles, among which F1, F2, F3, and F4 were estimated using a method (Clermont, 1992) based on cepstral analysis-by-synthesis (AbS) and dynamic programming (DP).

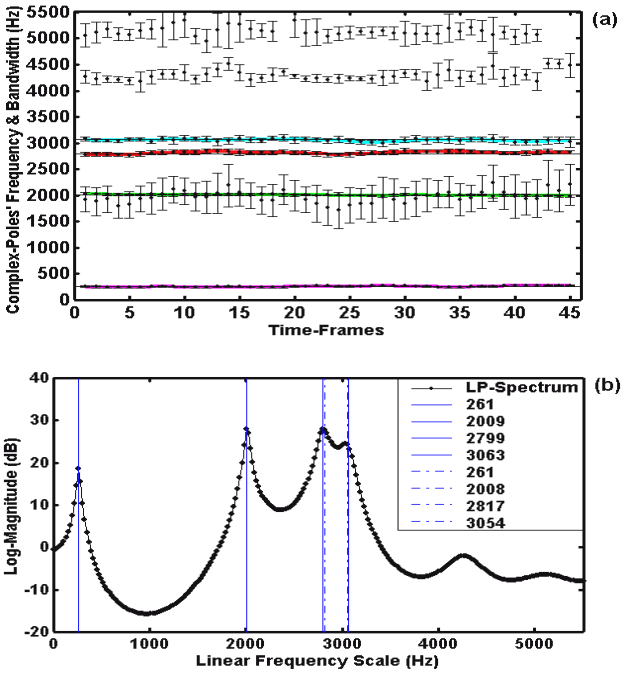

Figure 2: (a) Pole-gram & superimposed DP-tracks for AE /i/ sung at 110 Hz; (b) Frame-averaged LP-spectrum for (a).

DP-tracked, F-patterns are illustrated in Fig. 2(a), and can be seen to pass through the narrow-bandwidth poles. This is achieved with the AbS, which selects the poles corresponding to prominent spectral peaks, jointly with the DP optimisation, which secures temporal continuity.

Inter-Frame and Inter-Token Variability

The singer's vibrato appears as a superimposed undulation, which could pose a methodological problem in determining representative F-patterns. To minimise any vibrato effects, F-patterns were averaged (dashed lines in Fig. 2(b)) across the steady-state frames of each nucleus and, to check for consistency (*inter alia*: White, 1999), the former were compared with F-patterns (solid lines in Fig. 2(b)) obtained from the frame-averaged LP-spectra using the same method of formant estimation. The mean differences calculated between the two sets of 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. In view of these small intra-nucleus differences, the frame-averaged F-patterns were averaged over the 5 tokens and finally retained for AE (see Fig. 3(a)).

It is well known that measured F-patterns exhibit some variability caused not only by the measurement method used, but also by one's inability to replicate sounds that are spoken and presumably sung, in exactly the same way. Consequently, the inter-token dispersions (ITDs) about the average F-patterns retained for AE were used as a tool for gauging intrinsic variability.

Table 1. Inter-token dispersions (=standard deviations in Hz) of frame-averaged, F-patterns of AE vowels (bass subject).

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

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. One might speculate that these contrasts reflect a strong consistency in mandibular positions in spoken and sung phonations, but more variability in our subject's lingual positions during singing. The ITDs increase expectedly from F1 to F3, and their respective ranges lie within difference-limens (Flanagan, 1955) for human perception.

In sum, there appear to be no gross measurement errors or unusual variability that should discourage further analyses.

3. VOWEL SEQUENCE CHARTS

It is instructive to begin our comparative study of spoken (SPV) and sung (SUV) vowels, and of vowels simulated with normal (NLV) and lowered larynx (LLV), by using the sequence charts shown in Fig. 3. Sundberg (1970) has pioneered this type of comparative analysis, which provides a good overview of vowel-by-vowel variations in F-patterns.

For example, only subtle differences in F1 are visible in all three charts, thus suggesting that, from a one-dimensional perspective, the transition from speaking to singing by bass subjects or the transition from normal to lowered larynx in the simulation has minor effects on F1 across all vowels. The high-F2, SUVs and LLVs display an unmistakably lowered F2 by comparison with their SPV and NLV counterparts. The spoken-to-sung lowering of F3 is relatively less marked across most vowels, with some notable singer-dependent variability; NLVs and LLVs appear to be quite close in F3 except for [u] and [ɑ].

The consistent contrasts along the F2 dimension indicate at least a certain similarity in the way in which both bass singers modified their speaking positions of the tongue body. The further similarity in F2-contrasts between the bass singers' and the lowered-larynx data tends to support previous findings (*inter alia*: Sundberg, *op. cit.*) on larynx-height adjustments in singing voices. To progress beyond these component-by-component observations, we turn to the F1-F2 plane in search of a systemic perspective.

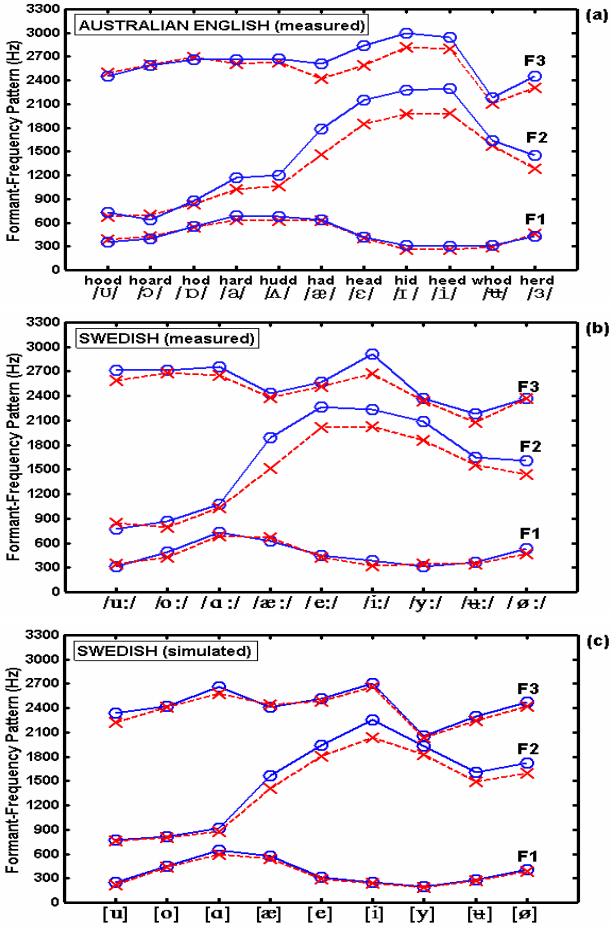

Figure 3: Vowel Sequence Charts. {[3(a): AE (measured: 5-token averages); [3(b)]: SW (measured: 3-token averages captured digitally from Sundberg's (1970: p.30) Fig. 1: singer B4); [3(c)]: SW (simulated: Lindblom & Sundberg (1971: p.1176)}. {Plot Symbols: [3(a,b)]: o(blue)=SPV, x(red)=SUV; [3(c)]: o(blue)=NLV, x(red)=LLV}.

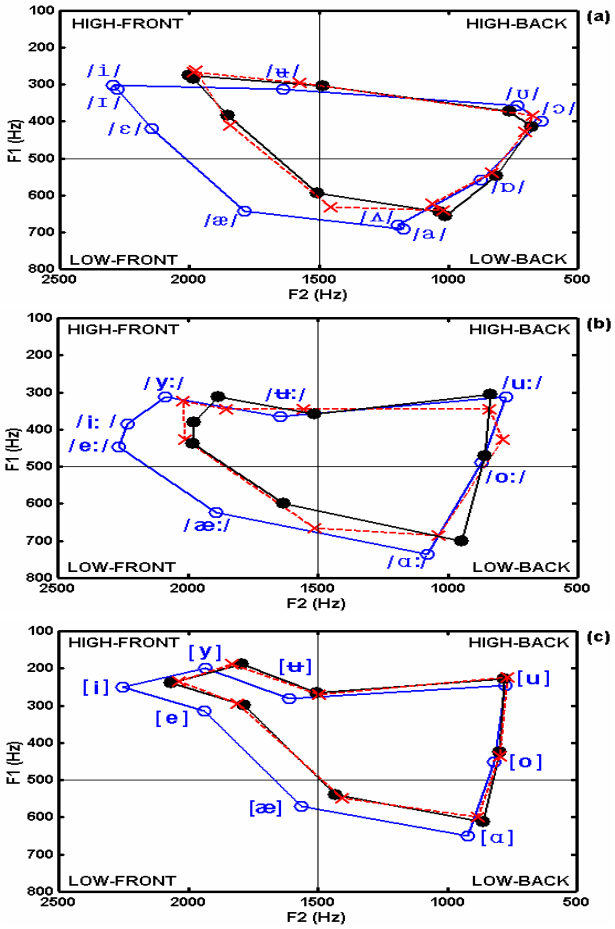

Figure 4: Planar distributions of F2 versus F1 are plotted with the same symbols shown in Fig. 3 for SPVs and SUVs; the linear regression fits (Table 2: SPV➔SUV, NLV➔LLV) are overlaid with filled circles (black). The (500,1500)-Hz lines (idealised uniform tube) are visual delimiters for the 4 broad vowel categories. {Vowel symbols: phonemic for AE (measured: [4(a)]), and for SW (measured: [4(b)]) as per Sundberg (op. cit.), but phonetic for SW (simulated: [4(c)]) as per Lindblom & Sundberg (op.cit.)}.

4. VOWEL SYSTEMS IN F1-F2 SPACE

As resonance frequencies of the vocal tract, F1 and F2 afford articulatory interpretations of vowel production in terms of jaw (high/low) and tongue (back/front) positions. In addition, the ensemble of (F1, F2)-coordinates tends to specify operating limits (Potter and Steinberg, 1950), within which maximal separation can be expected among steady-state vowels from the same speaker or presumably the same singer. Thus, the F1-F2 space should prove useful for gaining a systemic impression of articulatory-phonetic contrasts between spoken and sung vowels.

Indeed, if we re-plot the F1- and F2-patterns from Fig. 3 onto a classic vowel diagram, we obtain the more global view shown in Fig. 4. For example, both Figs 4(a) and 4(b) immediately reveal an asymmetric shift of the SUV relative to the SPV polygon, consistently for AE and SW. A similar shift from the NLV to the LLV polygon appears in Fig. 4(c). The unrounded, high- and low-front vowels largely contribute to the shifts, while high- and low-back vowels seem to be less susceptible to the changes brought upon by the bass singing voice. The geometric consequence is clearly an asymmetric compression, which appears far greater along the F2 than along the F1 axis, and relatively more pronounced from SPVs to SUVs than from NLVs to LLVs.

The overall impression is this – the spoken-to-sung transformation appears to be similar for both vowel systems (AE and SW), thus implying that both singers employed similar articulatory settings to achieve the transformations. What are these settings likely to have been?

A preliminary answer to this question emerges from the visibly similar transformation from NLVs to LLVs, which could indicate that larynx lowering and the resulting lengthening of the vocal tract constitute a dominant setting for the bass singing voice. In the next section, the plausibility of this proposition is evaluated in the light of a more quantitative characterisation and comparison of the systemic relationships depicted above.

ISBN 1-876346-48-5

5. SYSTEMIC TRANSFORMATIONS

What is the nature of the transformation underlying the systemic shifts noted above? In addressing this question, our objectives are to find a mathematical form that provides a numerically plausible mapping between the spoken and sung polygons, and then to return to the notion of dominant setting advanced earlier.

The main geometric properties of the spoken-to-sung transformation are clear – a major compression along F2 of most front SPVs and a minor one along F1 of all SPVs, the result of which appears as a nearly parallel translation of the front-vowel towards the back-vowel side of the spoken polygons. The same consistency applies to NLVs and LLVs. The transformation is expected therefore to be nearly linear, and expressible in the parametric forms given in Table 2.

Indeed, the linear-regression fits shown in Fig. 4 are quite good, with fitting errors for F1 (rms1) and F2 (rms2) also given in Table 2, which are close to or within ITDs for AE, relatively tolerable for SW (measured) and remarkably low for SW (simulated). Thus, numerically speaking, the linear transformations seem viable. The intercept-related parameters (row 1 of Table 2) are data-dependent and less consequential as far as the core of the transformation is concerned. Instead, it is the core matrices in row 2 of Table 2, which are brought to bear on the following discussion.

Table 2. Parameters (*a*'s and *b*'s) and errors (rms in Hz) of $\tilde{F}1 = a0 + a1*F1 + a2*F2$ & $\tilde{F}2 = b0 + b1*F1 + b2*F2$, where $F1$ and $F2$ are either the measured values for SPVs or those simulated for NLVs, while $\tilde{F}1$ and $\tilde{F}2$ are the fitted values that are compared with either those measured for SUVs or those simulated for LLVs, respectively.

	AE SPV➔SUV	**SW** SPV➔SUV	**SW** NLV➔LLV
$[a0 \quad b0]$	[77 295]	[10 313]	[−8 137]
$\begin{bmatrix} a1 & b1 \\ a2 & b2 \end{bmatrix}$	$\begin{bmatrix} 0.89 & -0.29 \\ -0.03 & 0.78 \end{bmatrix}$	$\begin{bmatrix} 0.93 & -0.31 \\ 0.005 & 0.80 \end{bmatrix}$	$\begin{bmatrix} 0.95 & -0.11 \\ 0.004 & 0.87 \end{bmatrix}$
$[rms1 \quad rms2]$	[20 45]	[39 63]	[7 22]

First, the diagonal elements of all core matrices confirm the small F1 and large F2 compressions noted earlier. Second, these matrices are strikingly similar for the 2 bass singers despite their different vowel systems and the 30-year gap in the data. It is thus likely that both singers employed similar articulatory settings. Third, the NLV➔LLV core matrix is quite similar to the SPV➔SUV ones, but numerically closer to the SPV➔SUV one for SW as might be expected.

In sum, the above analyses lend support to the proposition that larynx lowering dominates the spoken-to-sung transformation by our bass singers. The off-diagonal differences between the SPV➔SUV and NLV➔LLV matrices are not all negligible, and could be due to the lack of control of larynx height or concomitant lip rounding/protrusion (Perkell, 1969) in the simulated data.

6. CONCLUSIONS

We have introduced a systemic approach for uncovering characteristics of sung with respect to spoken vowels. The approach imports the notion of articulatory settings to describe systemic shifts in F1-F2 space.

The spoken-to-sung shifts for two different vowel systems (Australian English and Swedish) are found to be essentially linear for two respective bass singers. A similarly linear relationship derived from simulated lowered-larynx data, lends support to our proposition that this setting is dominant in the bass singing voices studied.

Although no general claims are currently possible owing to our small number of tokens and limitation to bass voices, the consistency in systemic contrasts observed here suggests that the approach should be useful for further elucidating the relation between spoken and sung vowels.

REFERENCES

[1] F. Clermont, "Formant contour parameterisation of vocalic sounds by temporally constrained spectral matching", Proc. 4th Australian Intl Conf. on Speech Sci. & Tech., Brisbane, pp. 48-53, 1992.

[2] J.L. Flanagan, "A difference limen for vowel formant frequency", J. Acoust. Soc. Am. 27, pp. 613-617, 1955.

[3] J. Laver, *The phonetic description of voice quality*, Cambridge University Press, 1980.

[4] B. Lindblom and J. Sundberg, "Acoustical consequences of lip, jaw, and larynx movements", J. Acoust. Soc. Am. 4, pp. 1166-1179, 1971.

[5] R. Miller, *On the art of singing*, Oxford University Press, 1996.

[6] F. Nolan, *The phonetic bases of speaker recognition*, Cambridge University Press, 1983.

[7] J.S. Perkell, *Physiology of speech production*, M.I.T Press, 1969.

[8] R.K. Potter and J.C. Steinberg, "Toward the specification of speech", J. Acoust. Soc. Am. 22, pp. 807-820, 1950.

[9] J. Sundberg, "Formant structure and articulation of spoken and sung vowels", Folia Phoniatrica 22, pp. 28-48, 1970.

[10] J. Sundberg, "Articulatory interpretation of the "singing formant"", J. Acoust. Soc. Am. 55, pp. 838-844, 1974.

[11] P. White, "Formant frequency analysis of children's spoken and sung vowels using sweeping fundamental frequency production", J. Voice 13, pp. 570-582, 1999.