ACOUSTIC DESCRIPTION OF A SOPRANO'S VOWELS BASED ON PERCEPTUAL LINEAR PREDICTION

Thomas Millhouse [1] *& Frantz Clermont* [2]

[1] Sydney Conservatorium of Music, University of Sydney, Australia
[2] JP French Associates and University of York, United Kingdom
[1] thomas.millhouse@bigpond.com [2] akustikfonetiks@yahoo.com.au

ABSTRACT

A perceptually-motivated model known as Perceptual Linear Prediction (PLP, [6]) is employed to parameterise and to interpret the cardinal vowels sung by a professional soprano at pitches ranging from 220 to 880 Hz. The PLP model yields perceptual formants (F_1' and F_2'), which encode the low and high-spectral regions, respectively. These formants are found to be tractable and robust, thereby facilitating a more complete description of the sung-vowel space.

1. INTRODUCTION

A major problem inherent to the acoustic analysis of sung vowels is the lack of a parameterisation method that can resolve phonetic and timbreal information whilst maintaining robustness for rising pitch. Traditional formant analysis has been the primary focus for some time due to the information that formant frequencies readily carry about vocal-tract shapes, phonetic distinctiveness and speaker specificity. However, traditional formant techniques cannot account for a complete characterisation of sung vowels across all singing voice types. Previous works clearly reflect this limitation.

1.1. Background

Acoustic formant analysis of sung vowels has been successful for low-pitched voice registers. The wide spacing of harmonics in high-pitched singing however, makes the evaluation of acoustic formant frequencies problematic and unreliable. The use of spectrographic parameterisation and componential description of the acoustic formant structure (i.e., formant by formant, and vowel by vowel), was pioneered by [11], whilst [3] proposed a systemic approach to sung vowels in the phonetic space spanned by the three lowest formants. The results of these studies provided valuable phonetic and timbreal information about the sung vowel but were still limited in regard to high-pitched voices.

In an attempt to overcome the problem inherent to high-pitched singing, there have been a number of perceptually-motivated acoustic studies of the singing voice. Principal Component Analysis of 1/3 octave filter bank outputs was utilised by [1] to study the differences between spoken and sung vowels. This dimensionality-reduction approach yielded a spatial representation of sung vowels, which affords discrimination in a phonetic-like space spanned by the two major dimensions. However, the approach is dependent on the availability of a statistically-significant sample of sung vowels, required to define each vowel as a function of its displacement from other vowels, making cross speaker or single vowel comparison problematic.

A more promising approach known as Perceptual Linear Prediction (PLP) has arisen from [6]. It affords the possibility of extracting spectral features automatically from the acoustic signal, which are related to formant frequencies while being perceptually-motivated. The PLP model initially developed for spoken sounds was exploited only recently in [8], a study of sung vowels. The results reported therein provide evidence of the interpretive power of PLP-derived formants for spoken as well as for sung vowels.

1.2. Objectives and outlines of this study

The work reported in this paper seeks to extend the work of [8] by looking at the behaviour of PLP-derived formants for vowels sung by a soprano through her full range of pitches.

The body of the paper consists of three major sections. In Section 2 the sung-vowel material and the PLP procedure are outlined, together with a brief evaluation of the PLP-derived formants. Section 3 gives a componential description of each of the sung vowels, while Section 4 provides a systemic perspective of the sung-vowel space. Section 5 summarises our findings.

2. METHODS AND MATERIALS

2.1. Soprano Subject and Vowel Data

The subject enrolled for this study is a native speaker of Australian English and a professional operatic soprano with a taxonomy rating of 3.1b, meaning a *'nationally recognised opera singer employed in principal roles'* according to the taxonomy scale defined in [2]. She was recorded in the studios of the State Opera of South Australia, using the MBox professional sound equipment and a B&K head microphone.

Three cardinal vowels in /hVd/ context were recorded for analysis. Our subject first spoke 5 randomised tokens of the Australian English monosyllables "heed", "hard" and "hood" at her habitual speaking rate of ≈190 Hz. Then, queued with a series of pulsed sine-wave tones (220, 275, 330, 440, 550, 660, 880 Hz), she sang 3 tokens of the syllables at about the same pitch as the tone. The analogue signals were sampled at 11,025 Hz and quantised to 8 bits. The sampled data was manually isolated and stored into files containing individual spoken and sung syllables.

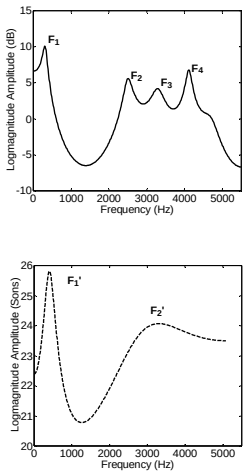
2.2. Perceptual Formant Parameterisation

The algorithmic procedure (PLP, [6]) was used for parameterising the acoustic signal. It is outlined at right of Fig. 1, and uses psychoacoustic transforms to produce an auditory spectrum illustrated in dotted line at left of Fig. 1. This smooth spectral representation is based on the autoregressive coefficients of the PLP polynomial (typically of order 5), and reflects the auditory integration of two or more formant peaks into a single peak. The example shown at left of Fig. 1 for a sung front-vowel shows two broad peaks that are referred to as the perceptual formants F_1' and F_2'.

These formants can be automatically extracted by selecting the broad peaks of the auditory spectrum. However, the assignment of perceptual peaks requires special attention when two peaks merge together ($F_1' \approx F_2'$). To validate such cases often related to back vowels, the peak-selection process is aided by manual interpretation of corresponding spectrograms.

The automatic and manual steps outlined above were followed to estimate F_1' and F_2' through the entire spoken and sung syllables, for every 30-msec frame advancing at a rate of 10 msecs.

Figure 1: Procedural outline of Perceptual Linear Prediction (PLP) according to [6].



Acoustic representation
↓
Critical-Band Analysis
↓
Equal-Loudness Pre-emphasis
↓
Intensity-Loudness Conversion
↓
Linear Prediction
↓
Auditory Representation

Using the resulting perceptual formant-tracks, the 7 steadiest frames of each vowel nucleus were identified visually. The 7-frame averaged values for F_1' and F_2' were finally retained for the descriptive analyses given in Sections 3 and 4. The "goodness" of our F_1' and F_2' estimates is first assessed by way of inter-frame/token consistency.

2.3. Inter-Frame and Inter-Token Variability

Table 1 lists inter-token and inter-frame dispersions, which are interpreted as baseline measures of reliability for our F_1' and F_2' estimates. Note that such measures have not been previously reported for sung vowels.

On the whole, inter-frame dispersions are smaller or near their inter-token counterparts. This finding lends support to our expectation that, through the steady-state range of the vowel, F_1' and F_2' should vary relatively less than from token to token. The dispersion values are also well within the difference limens for human perception of vowel sounds [5]. The inter-frame and inter-token dispersions examined below, lead us to believe that there are no gross measurement errors that could cast doubts on our F_1' and F_2' estimates and, hence, discourage further analyses.

Table 1: Dispersions across tokens and across frames (in parentheses) computed as standard deviations (Hz).

	SPOKEN		SUNG	
	F`1	F`2	F`1	F`2
/heed/	16 (8)	40 (67)	9 (9)	43 (43)
/hard/	38 (19)		19 (14)	41 (25)
/hood/	7 (11)		9 (10)	22 (15)

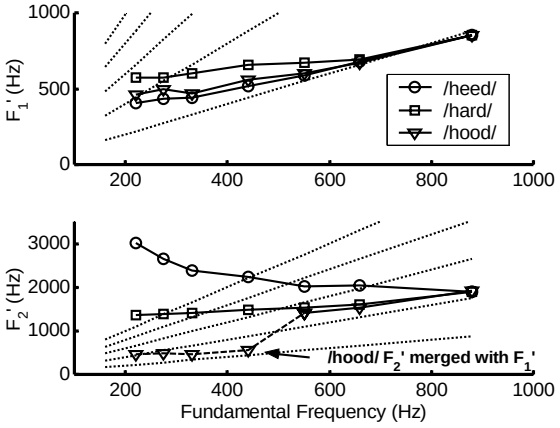
3. COMPONENTIAL DESCRIPTION

As a first step, therefore, we will proceed with a deeper interpretation of our data in the form of componential (formant by formant) charts as shown in Fig. 2. Such charts embody a compelling structure within which variations in F_1' and F_2' are readily observed horizontally as functions of increasing pitch, and obliquely as functions of per-pitch harmonics.

Reported previously in [4] the number of derived perceptual formants for spoken language is dependant on the pharyngeal vowel placement. For front vowels a definitive F_1' and F_2' are normally derived for each vowel, however for spoken back vowels only a single F_1' is observed with its centre frequency dependant on the average value of it's acoustically derived formants particularly F_1 and F_2. Normally it is expected that for the back vowel "hood", the single F_1' will be quite low typically 500 Hz where as for the back vowel "hard", the single F_1' can extend as high as 1300 Hz in female subjects.

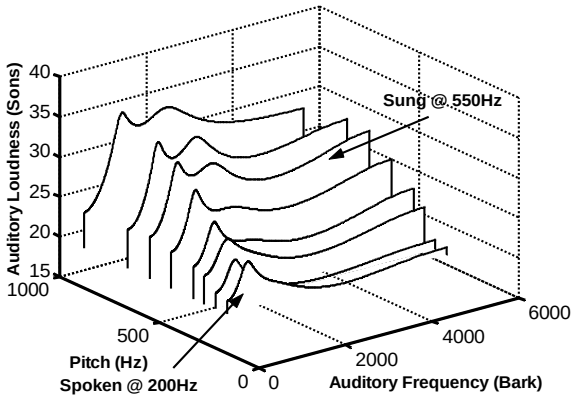
Returning now to our sung data, we see that for all three vowels, F_1' is aligned with the fundamental frequency as pitch increases, whereas F_2' exhibits three distinct movements prior to convergence. The F_2' in "heed" gradually drops towards 2 kHz, while the F_2' in "hard" starts lower than 2 kHz and rises very slowly towards 2 kHz. The F_2' in "hood" is merged with F_1' (as expected from [4]) thus appears to remain constant until pitch reaches 550Hz. At 550Hz a second F_2' is resolved with a converging pattern for F_2' in "hood" similar to that for F_2' in "hard".

Figure 2: Componential charts of perceptual formants (F_1' at top and F_2' at bottom). Per-pitch harmonics are represented as oblique dotted lines.



As reported previously in [8], the major differences between speech and singing in the auditory domain include an increase in energy in the upper spectral region and a detectable F_2' for sung vowels. However, not all sung vowels have a pronounced F_2' for all pitches. This is highlighted in the auditory spectra given in Fig. 3 for the sung back-vowel in "hood". While there is an increase in spectral energy around the higher-spectral regions, a second perceptual formant is not resolved until the pitch value extends beyond 550 Hz.

Figure 3: PLP spectra for the vowel in /hood/: spoken; and sung as a function of rising pitch. Resolution of F_1' and F_2' occurs beyond a pitch value of 550 Hz.



From the charts given in Fig. 2, note that all three vowels are seen to possess distinctive perceptual formants as far as a pitch value of 550 Hz but, beyond this pitch frequency, the sung-vowel spectra become quite homogenous. This phenomenon has been observed in perceptual intelligibility studies. For example [9] reported that sung-vowel intelligibility drops below 50% for pitches above 440 Hz. Thus, our PLP results tend to be aligned with those obtained from the point of view of perceptual intelligibility.

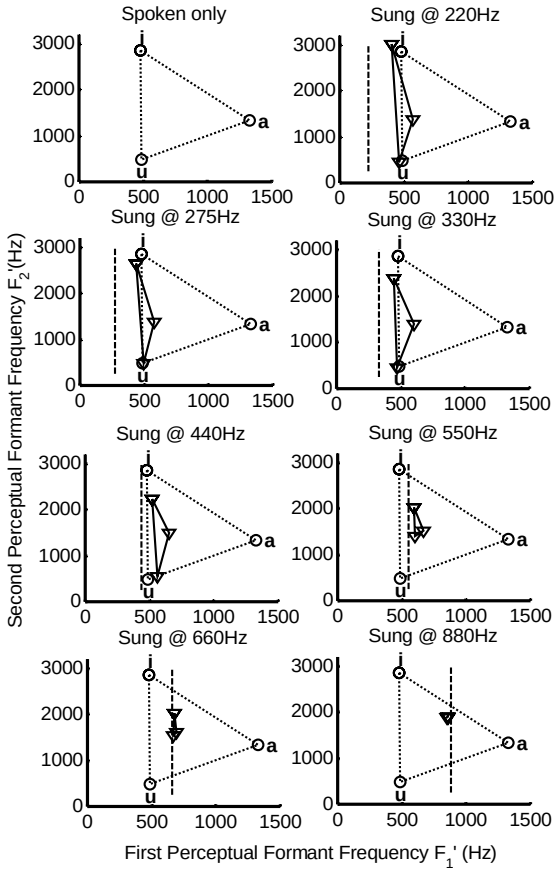
4. SYSTEMIC DESCRIPTION

Beyond the componential observations that are afforded by the componential charts shown above, it will be instructive to progress towards a systemic description by considering the space spanned by F_1' and F_2'. The co-variation of these formants should yield a more cohesive perspective on the auditory formant space of sung vowels as a function of rising pitch.

Figure 4 displays a sequence of F_1'- F_2' planes, which highlight the transformation of the sung-vowel triangle from lower to higher pitch values.

The sung vowels are bounded by the increasing fundamental frequency as well as the $F_1' = F_2'$ line, which forms the lower base of the spoken triangular space. Initially, the sung vowels migrate towards the lower F_1' side of the triangle bounded by the vowels in "heed" and "hood". The sung-vowel triangle is continually compressed towards the vowel in "hard" as pitch rises. An analogous description of this phenomenon has been given in [10] an acoustic-articulatory study of the acoustic vowel space as a function of rising fundamental frequency.

Figure 4: F_1'-F_2' planes showing the spoken-vowel space represented by circles, and the superimposed sung-vowel space (triangles). The pitch frequency is drawn vertically as a dotted black line.



The final resting position for our soprano's sung vowels is clearly notable in the F_1'-F_2' plane. All three vowels converge to a final value for F_1' equal to the fundamental and a final value for F_2' close to 2 kHz. Could this neutral-like vowel position be the optimum auditory and articulatory setting for the production of a soprano's highest pitched notes?

5. CONCLUSIONS

A core theme running throughout this paper is the ability of the PLP procedure to elucidate perceptual formant spaces for a soprano's sung vowels at high pitched frequency. The formant parameterisation afforded by PLP is a significant advance over previous attempts rooted in traditional linear prediction. The PLP model has also facilitated both componential and systemic perspectives on the soprano's sung vowels, which become more and more homogenous as a function of rising pitch. The merging and clustering of F_1' and F_2' appear to reflect a neutral-like vowel position at high pitches, which will need to be validated in the articulatory domain. Future work will involve a more complete acoustic-auditory study of sung vowels by enrolling additional subjects to cover other voice types. A key question for future work will be to describe the acoustic formant region known as the singer's formant in the PLP derived perceptual domain, as it relates so strongly with the pedagogical description of good voice quality.

6. REFERENCES

[1] Bloothooft, G. & Plomp, R. 1985. Spectral analysis of sung vowels. II. The effect of fundamental frequency on vowel spectra. *J. Acoust Soc Am* 77, 1580-1588.

[2] Bunch, M. & Chapman, J. 2000. Taxonomy of singer's used in Scientific Research. *J. Voice.* 14(3), 363-369.

[3] Clermont, F. 2002. Systemic comparison of spoken and sung vowels in formant-frequency space. *Proc. 9th Australian International Conference on Speech Science and Technology.* SST02 124-129.

[4] Fant & Risberg 1962. Auditory matching of vowels with two formant synthetic sounds. *STL-QPRS* 4, 7-11.

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

[6] Hermansky, H. 1990. Perceptual linear prediction. *J. Acoust. Soc. Am.* 87(4), 1738-1752

[7] Johansonn, C. Sundberg, J. & Wilbrand, H. 1982. X-Ray study of articulation and formant frequencies in two female singer's. *STL-QPSR* 4, 117-134.

[8] Millhouse, T. J. & Clermont, F. 2006. Perceptual characterization of the singer's formant region: A preliminary study. *Proc. 11th Australian / New Zealand International Conference on Speech Science and Technology.* SST06 253-258

[9] Scotto di Carlo, N. & Germain, A. 1985. A perceptual study of the influence of pitch on the intelligibility of sung vowels. *Phonetica* 42, 188-197

[10] Story, B. 2003. Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics. *Proc. Stockholm Music Acoust. Conf.* SMAC03, 435-438

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