INEXACTNESS AND ROBUSTNESS IN CEPSTRA-TO-FORMANT TRANSFORMATION
OF SPOKEN AND SUNG VOWELS

Thomas J. Millhouse
Sydney Conservatorium,
University of Sydney,
Sydney Australia
Thomas.Millhouse@defence.gov.au

Frantz Clermont
Dept. of Computer Science, Maths & Science,
The American University of Paris,
Paris, France
frantz.clermont@ac.aup.fr

ABSTRACT

Accurate measurement of formant frequencies is important in many studies of speech production and perception. The mapping of formant frequencies from cepstra coefficients by Broad and Clermont (1989) is an important evolution in formant estimation techniques from previous studies. This method also holds significant potential in estimating formant frequencies of sung vowels due to its improved robustness over traditional formant tracking techniques. The employment of this method in the estimation of formant frequencies from a dataset of spoken and sung phonation has led to some findings, not only on the inexactness of the model, but also on the resonant nature of the sung and spoken vowels themselves.

1. Introduction

A persistent irony in measuring the difference in formant structure between sung and spoken vowels is that the interesting peculiarities of sung vowels are inherently difficult to measure using traditional formant tracking techniques. Formant estimation for high-pitched phonations is a well-defined obstacle for present day automatic formant estimation algorithms. Traditional methods of formant estimation, such as spectrogram analysis and automatic linear prediction (LP) are widely accepted techniques for formant analysis of low-pitched vowels. Errors occur within these techniques when applied to vowels of a higher fundamental frequency as quantified by [2]. A wide displacement of the harmonics within the frequency spectrum is the principal cause of formant estimation error in higher pitched vowels. The clustering of upper formants and a pronounced vibrato also contribute to a breakdown of traditional formant tracking algorithms when analysing sung vowels. Seeking to overcome this and other difficulties in formant estimation of sung vowels, studies have appeared in the literature that have used non-traditional formant estimation techniques, but are faced by yet an additional problem in that no clear set of resultant formant values are available to verifying these tracking systems.

Sundberg [3] used a servox artificial voice source to generate a low frequency source on a soprano’s throat, whilst the soprano was singing at high pitch. This study generated acceptable results and demonstrated that the acoustical theory of speech production still applies to high-pitched soprano singing. Bloothooft and Plomp [4] approached this problem from a perceptual perspective and used a statistical method known as Principal Component Analysis to observe the behaviour of sung formants as a function of fundamental frequency. Their results provided good agreement with the results in [3] however reported formant values as a function of new statistical dimensions based on perception, rather than specific frequencies. Titze [5] used an analysis-by-synthesis method to estimate complete spectral envelopes, given limitations for formant frequencies, and glottal source parameters. This method proved effective at high pitches based on the Tenor voice type, however is reported as being computationally expensive. Titze’s technique uses an analysis-by-synthesis approach with parameter optimization to obtain a best-fit formant spectrum over the harmonic data. This sort of procedure would prevent it from being effective as a tool for real time feedback of formant frequencies.

There are two clear indications from the evolution of previous studies. Initially the parameterisation selection for the sung vowel should be chosen so that for a common vowel, a minimum amount of spectral distortion occurs with pitch modification. Secondly that formant estimation in sung vowels is better approximated when a technique that uses the whole frequency spectrum is used to estimate the individual formant frequencies. The complexity and time consuming method of Titze’s analysis-by-synthesis has left the path open for a less complex but significantly robust method to make the next evolutionary leap.

Broad and Clermont (1989; cited henceforth as BC89 [1]), presented a method of formant estimation via linear mapping from the cepstra. This technique is not as accurate as LP but is more robust. It will always deliver an approximate value for a formant frequency given a frame of cepstra and its use of the cepstra has the potential to deliver greater reliability at higher pitches. Of most important note is its evolution from the study of [4] in that the process uses a full spectral approach to predict specific formant frequencies rather than a specific frequency peak picking technique.

The aim of this paper is to assess the capability of BC89 cepstra-to-formant procedure in its application to vowels sung by a Bass-Baritone voice at frequencies where formant information can be reliably determined. It will also assess the performance of the cepstra-to-formant model on sung vowels by comparison with a spoken data set, together with a first attempt at elucidating the resonant nature of the sung vowel

2. Materials and Methods

2.1 The Materials

To demonstrate the effectiveness of the BC89 model, we used a complete set of Australian English vowels spoken and sung by the same male subject. He is a part-time chorus artist with a professional Australian Opera Company, and a native speaker of Australian English. He has had several years of training in Western classical opera singing, has won regional Australian Vocal Championships, and regularly performs as a
professional chorus artist and in minor operatic roles. Initially he spoke five randomized tokens of 10 monosyllables in /hVd/
context, at his habitual speaking rate (F0=80Hz on average). Then queued with a 110Hz tone our subject sang the same
tokens at approximately the same pitch as the queued tone. The analogue signals were sampled at 11,025 Hz and quantized
to eight bits.

2.2 The Method

The BC89 cepstra-to-formant procedure essentially consists of two stages. Initially a database of formant frequencies and their corresponding cepstra is created in order to obtain a series of regression coefficients \( a_i \). These coefficients are calculated using the Multiple Linear Regression (MLR) statistical technique (see [6] for further
details). These regression coefficients can then be used to predict the formant frequencies of any new cepstra \( c_i \), based on
estimates of the form

\[
F = F = a_0 + \sum_{i=1}^{M} a_i c_i
\]

where \( M \) = 14 cepstral coefficients. Initial formant trackings on our data were carried out using the analysis-by-synthesis procedure outlined in [7]. For each frame of measured formant information, linear predictive cepstra coefficients (LPCC) were also calculated and used as the basis for the cepstra-to-formant mapping procedure. When more than 14 cepstra coefficients were required to define the first four formants, only the low-indexed 14 coefficients were retained in the model.

2.3 Test Procedure

To test the effectiveness of the BC89 model on the dataset outlined above, we used a procedure known as ‘piggy backing’
whereby the first four tokens were used to predict the fifth. This process created five sets of results from which the average
of each was used to give the final figures of the root mean
squared (rms) error and the Pearson correlation coefficient of
the LPC derived formants compared with the formants of each
spoken token. Comparison of the absolute glottal
reflection coefficient for both spoken and sung vowels (see
Figure I) indicates that the sung vowels have a consistently
higher glottal reflection coefficient, confirming narrower
bandwidths and thus indicating a more resonant vocal tract.

Additionally on this issue, the glottal reflection coefficient which is directly proportional to the mean resonance
bandwidths has been reported by [10] and implemented by [8]
to compare the mean bandwidth of the poles of the sung with
those of spoken spectra. Comparison of the absolute glottal
reflection coefficient for both spoken and sung vowels (see
Figure I) indicates that the sung vowels have a consistently
higher glottal reflection coefficient, confirming narrower
bandwidths and thus indicating a more resonant vocal tract.

The results and explanations given above already suggest
the reduction of formant frequency bandwidths for the
sung vowels improves performance in sung formant-frequency
estimation from a database of sung cepstra data.

4. Discussion

The question raised above lends itself to an explanation based on the spectral differences between sung and spoken
vowels. The hypothesis is that the cepstral nature of the sung vowel presents itself as more supporting to formant estimation
via MLR. The more detailed explanation is what characteristics of the sung cepstra make it more so.

4.1 Formant bandwidth narrowing

The sung database might deliver better results based on the
difference in average formant bandwidths between the sung
and spoken spectrums. No information regarding the bandwidths of sung formants has been apparently published,
with the exception of [8]. Using a band-selective glottal reflection coefficient, these authors demonstrated that the sung
formant bandwidths of the upper formants were in fact narrower than their spoken counterparts. The reasoning why
the singer may narrow their bandwidths could stem from a need to increase the amplitude of the upper formants, or it could be
surmised that the narrowing of the bandwidths occurs as a mechanism for improving the intelligibility of the vowels as
indicated in spoken language [9].

Alongside the improved perceptual intelligibility, [2] were
able to demonstrate that the automatic prediction power of the
LPC is more accurate when dealing with narrower bandwidths.
An analysis of the LPC derived formant bandwidths although
unreliable yields that the sung bandwidths are on average considerably lower than their spoken counterparts (Table II).

| Table I: The rms errors and Pearson correlation coefficients between known and predicted formant frequencies in applying Eq (1) with \( M = 14 \) cepstral coefficients. |
|---------------------------------|---------|---------|---------|---------|
|                                  | \( F_1 \) (Hz) | \( F_2 \) (Hz) | \( F_3 \) (Hz) | \( F_4 \) (Hz) |
| Spoken data: 5 tokens x 10 vowels | 0.9908     | 0.9932     | 0.9583     | 0.9249     |
| r                               | 16.17      | 55.46      | 53.56      | 59.27      |
| Rms error                       | 38.69      | 33.30      | 35.86      |
| Sung data: 5 tokens x 10 vowels  | 0.9970     | 0.9954     | 0.9772     | 0.9322     |
| r                               | 8.92       | 38.69      | 33.30      | 35.86      |

The correlations and rms between the predicted formants and the LPC derived formants are given in Table I. The rms
values and Pearson correlation coefficients indicate that a clear relationship exists between the cepstra and their derived
formant frequencies for both spoken and sung data. The sung
data however gives results that are overall better than those
obtained using the spoken counterpart.

Why does the model appear to have an increased
performance in the prediction of sung formants from a sung
database, over its spoken counterpart?

| Table II: LPC Derived Spoken and Sung Bandwidths |
|---------------------------------|---------|---------|---------|---------|
|                                  | \( B_1 \) (Hz) | \( B_2 \) (Hz) | \( B_3 \) (Hz) | \( B_4 \) (Hz) |
| Spoken data                     | 69.91    | 124.63   | 192.43   | 240.56   |
| Mean                            | 41.53    | 60.31    | 81.99    | 234.30   |
| STD                             | 38.86    | 44.63    | 87.93    | 94.27    |

Additionally, on this issue, the glottal reflection coefficient
which is directly proportional to the mean resonance
bandwidths has been reported by [10] and implemented by [8]
to compare the mean bandwidth of the poles of the sung with
those of spoken spectra. Comparison of the absolute glottal
reflection coefficient for both spoken and sung vowels (see
Figure I) indicates that the sung vowels have a consistently
higher glottal reflection coefficient, confirming narrower
bandwidths and thus indicating a more resonant vocal tract.

The results and explanations given above already suggest
the reduction of formant frequency bandwidths for the
sung vowels improves performance in sung formant-frequency
estimation from a database of sung cepstra data.
Analysis of their formant frequencies indicates that the sung vowels and higher SF values than their counterparts. The analysis having higher than average SF values for spoken vowels /heed/ and /hid/ are the only ambiguities in this on average SF values for spoken vowels /heed/ and /hid/ are the only ambiguities in this analysis having higher than average SF values for spoken vowels and higher SF values than the sung counterparts.

Analysis of their formant frequencies indicates that the front vowels have the greatest displacement in F2 and F3 when comparing spoken and sung vowels [7]. In addition, the F2 in the spoken front vowel is highest for the vowels /heed/ and /hid/ and its close proximity to F3 would increase the SF factor. Analysis of the mean difference in F3-F2 reveals that it is considerably lower in the two spoken vowels /heed/ and /hid/ than for their other spoken counterparts, which account for the higher SF value in these vowels.

4.3 Resonant nature of the sung vowel

One failing point in evaluating formant bandwidths by trend analysis of the glottal reflection coefficient or SF is that neither technique is explicit about the influence of spurious poles resulting from the LP derived cepstra. One way of separating the influence of the spurious poles is to recast the LP derived cepstra as the sum of resonant cepstra and non-resonant cepstra as displayed in Figure III. The resonant cepstrum is derived from the LP detected formant frequencies and bandwidths, and the non-resonant cepstrum here consists of the spurious poles and any untracked higher formants. Such that:

\[ c_i = c_i(\text{resonant}) + c_i(\text{nonresonant}) \]  

where,

\[ c_i(\text{resonant}) = \frac{2}{\pi} \sum_{n=1}^{4} \exp\left(-\frac{inB_i}{F_s}\right) \cos\left(\frac{2inF_i}{F_s}\right) \]  

with,  

\[ F_i = \text{Sampling Frequency}. \]

A further difficulty of this parameterisation is that the non-resonant cepstra consists of both the spurious poles and any untracked higher formants. The measure will be zero for a flat spectrum. Observations on the influence of the separate components of the resonant and non-resonant cepstra can now be made.

The influence or contribution of the separate components of the resonant and non-resonant cepstra can provide information on the resonant nature of the vowel itself. A specific definition of vowel resonance with respect to resonant and non-resonant cepstra is yet to be quantified, but it could be tentatively defined as the contribution of the resonant cepstra to the complete cepstra within the bandwidth of interest. This contribution of the resonant cepstra can be measured by observing the spectral difference between the resonant cepstra and the complete cepstra made over a distinct bandwidth of frequency resolution. Conversely, observations on the distance between the non-resonant cepstra and a perfectly flat spectrum will present a quantitative measure of the contribution of the spurious poles to the complete cepstra.

Figures IV and V were calculated by measuring the differences in vowel resonant and non-resonant cepstra from either the full cepstra or a perfectly flat spectrum respectively. The difference values for spoken and sung vowels are given on a vowel-by-vowel basis and presented as an average over all vowels for the spoken and sung phonations.
By observing the distance measure of the resonant cepstra from the full cepstra over the limited frequency band of 0 – F4, we obtain a measure of the contribution of the resonant cepstra to the complete cepstra. Figure IV clearly demonstrates that the resonant cepstra is much closer to the complete cepstra, in sung vowels than in their spoken equivalents. This would imply that the sung resonant cepstra is a greater contributor to the complete cepstra than its spoken counterpart is. Likewise, Figure V indicates that the sung non-resonant cepstra is much closer to a perfectly flat spectrum than its spoken counterpart. As the distance measure is limited to the frequency of F4, this distance measure indicates that the contribution of the spurious poles to the complete cepstra up to F4 is much less for sung vowels than for their spoken counterparts.

The increased contribution of the resonant cepstra and decrease in non-resonant cepstra would imply that for this single subject, his sung vowels are quantitatively more resonant than his spoken counterparts. This final acoustical discussion strikes similarities with singing pedagogical knowledge and everyday observations in that the sung vowel is indeed more resonant than its spoken equivalent.

5. Conclusions

This paper has discussed the application of the BC89 model to predicting formant frequencies for a single subject who spoke and sung the Australia English vowels. The model is more accurate in predicting the four formant frequencies of the sung vowels than those of the spoken vowels.

We have identified the sung vowel spectra as being more resonant than their spoken counterparts in this single subject pilot study. The increase in sung spectral resonance is in part due to the reduction in spurious poles at the lower end of the spectrum. This phenomenon remains to be confirmed across a range of singing voices, and to be explained in articulatory and/or physiological terms. To enable this experiment it will be necessary to develop a robust method for estimating the formants of higher pitched vowels to enable the higher voice types to be analysed with this described method. To this end the accurate estimation of higher pitched vowels still remains a pivotal research question in the science of the singing voice.

6. References


