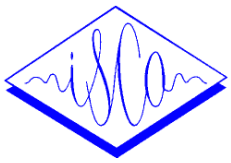

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# FAST FORMANT ESTIMATION OF CHILDREN'S SPEECH

*A. A. Wrench[1], J. M. M. Watson[2], D.S. Soutar[3], A.G.Robertson[4] and J. Laver[1]*

1 Centre for Speech Technology Research, University of Edinburgh, South Bridge, Edinburgh, EH1 1HN, Scotland.
2 Department of Speech Pathology and Therapy, Queen Margaret College, Edinburgh, EH12 8TS.
3 Plastic Surgery Unit, Canniesburn Hospital, Bearsden, Glasgow.
4 Beatson Oncology Centre, Western Infirmary, Glasgow.

## ABSTRACT

Constrained Multiple Centroid Analysis is a method of formant estimation already used successfully in the analysis of adult speech. In this paper we examine the performance of this approach when applied to the speech of children. The accuracy of the estimated formant frequency is measured for a range of excitation frequencies. In order to gauge absolute performance, an acoustic analogue of the vocal tract is used. Speed and accuracy of the Constrained Multiple Centroid Analysis are examined in a comparative study with a Linear Prediction based formant estimation algorithm. Results show that Multiple Centroid Analysis provides less bias towards the nearest harmonic than the established method based on Linear Prediction. Furthermore, computational efficiency makes it an attractive alternative to Linear Prediction.

## 1. INTRODUCTION

At ICSLP '92 [1] we described an acoustic phonetic analysis that used the Multiple Centroid Analysis algorithm to estimate the segmental speech quality of patients who had undergone oral surgery. The analysis was based on extraction of acoustic features, key among those being the frequency, energy and bandwidth of the speech formants. At EUROSPEECH '93 [2] we demonstrated a Personal Computer implementation of the analysis which as well as assessing segmental speech quality, provided real-time visual feedback to aid speech rehabilitation. This system is currently used in a clinic by patients who have undergone treatment for intra-oral cancers.

It is intended that the same system may be expanded for use by cleft palate children; to monitor speech progress and help in their speech therapy. To this end we have investigated the performance of the formant estimation algorithm when applied to children.

In this paper, the performance of this algorithm is tested for accuracy when applied to high pitched speech and is compared with a Linear Prediction based formant estimation algorithm.

## 2. BACKGROUND

Estimation of the frequency of formants is most commonly performed in two stages. Firstly, for each frame of speech, a set of spectral peaks and bandwidths are estimated. Then, to ensure that each formant is assigned to the appropriate peak, a tracking algorithm is employed. Typically, the formant frequencies are selected from candidates proposed by solving for the roots of the linear predictor polynomial or by peak picking a spectrum derived from the polynomial. The formant frequencies are computed at regular intervals from the speech waveform before dynamic programming is applied. In the dynamic algorithm, the local cost of all possible assignments of formants to the complex roots (or picked peaks) is made with respect to frequency and bandwidth. The cost of each of these assignments given the assignment of the previous frame is then calculated and the minimum cost assignment selected using a modified Viterbi algorithm. [3]

The need for a tracking algorithm can be attributed to the over specification of the order of the Linear prediction analysis which permits one or two extra pole pairs over the number required to map directly to the expected formants. These extra pole pairs are necessary to compensate for spectral tilt and losses along the vocal tract but may also produce extra peaks in the estimated spectrum which do not correspond to formants.

The method of Multiple Centroid analysis, in contrast, does not adhere to the all-pole model and does not require extra free variables to compensate for the vagaries of this model.

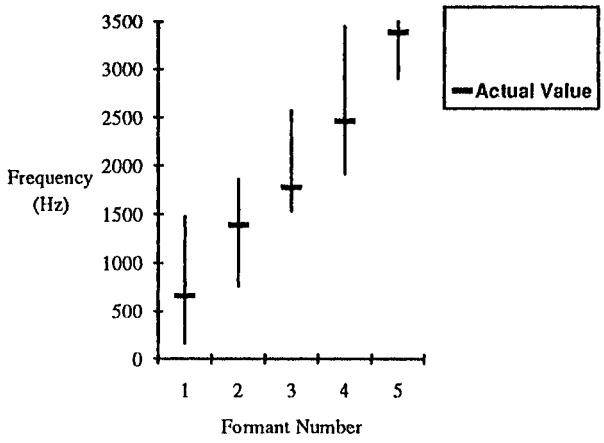


Figure 1. Shows the formant range limits set for this experiment. These limits change with age and sex and in this case are adjusted to the formant structure of the acoustic tube. Note that the ranges overlap.

Consequently the number of centroids can be chosen to map one to one to the formants and no tracking algorithm is required.

## 3. MULTIPLE CENTROID ANALYSIS

Multiple Centroid Analysis is an umbrella description covering any technique which evaluates more than one centroid (or centre of gravity) from a single multi-modal distribution. Crowe first proposed this type of analysis and applied it to formant estimation in 1987 [4]. He formulates the efficiency of a

multiple centroid description by partitioning the distribution. An estimated centroid of a specified partition of the distribution bounded by n=k1 and n=k2 is defined as a squared error given by :

$$\sum_{n=k_1}^{k_2}(n-k)^2 P(n)$$

The minimum squared error is found for each partition and the sum of these minimum errors is calculated for each possible combination of partitions. The partitioning of the distribution is assumed to be nonoverlapping and the number of partitions defines the number of centroids to be determined. This then, is a global least squares solution to the estimation of a specified number of centroids of a multimodal distribution.

## 4. CONSTRAINED LEAST SQUARES

The global least squares method is computationally expensive, requiring $2N^2$ operations for four partitions (where N = number of bins in the discrete spectrum). When applying centroid analysis to the power spectral distribution of sonorant speech we can, from our knowledge of the physical limitations of the human vocal tract and the associated articulators, define limits to the range of frequencies associated with each formant. These limits can be interpreted as constraints on where the boundaries of each partition may lie when estimating the centroids. By reducing the possible combinations of partitions in this way, a considerable saving in computation can be made with no loss of information. We can say that there is no loss of information because it is known *a priori* that formants cannot lie outside the partition limits. Looking at this another way: If a global search were to produce a minimum error for a partitioned solution outside these limits we could say that this was erroneous, perhaps due to the presence of a higher formant within the analysis bandwidth or prominent harmonics. Such constraints therefore make the estimation of speech formants more robust.

## 5. COMPARISON OF FORMANT ESTIMATION ACCURACY

As a method of evaluating the Constrained Least Squares method of Multiple Centroid Analysis, the formant frequency estimation accuracy is compared with that of the Linear Prediction approach as the periodic excitation of the test signal source is varied from 100 to 500Hz. Care must be taken in the choice of test signal. Real speech cannot be used because the formant structure is not known and can only be estimated. A synthesised signal based on serial cascaded filters provides a perfect all-pole response. Such a signal is likely to distort comparative performance results in favour of Linear Predictive Analysis since this approach assumes such a model as a basis for signal estimation. To avoid any possible bias in the comparative test we have employed an analogue test signal generated by acoustic pulses sent down a length of plastic tubing and recorded at the exit using a condenser microphone. While broadly conforming to the all-pole model this physical arrangement should exhibit losses along the tube and coupling with the source as is the case with the vocal tract. The resonant characteristic of the tube was established by exciting the tube with a swept pure tone and noting the frequencies of maximum amplitude.

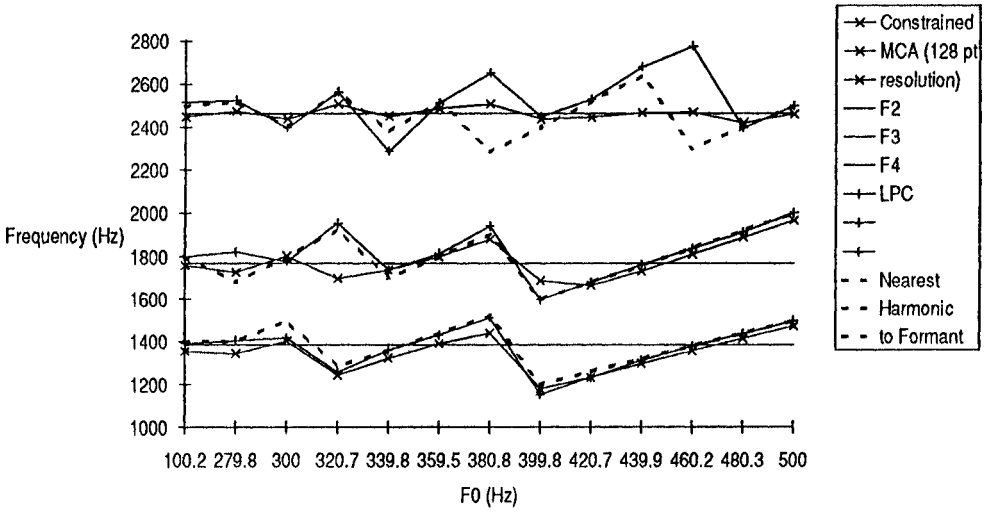


Figure 2. Formant frequency estimates of the acoustic tube excited by pulse trains of different frequencies using MCA and LPC with reference to the nearest harmonic indicating a higher degree of accuracy by the MCA. The standard deviation of these estimates never exceeded 36Hz and averaged 7Hz.

Acoustic recordings were made using a miniature condenser microphone onto digital audio tape and then transferred at 16kHz to computer where it was downsampled to 7khz so as to leave 5 formants in the analysis band. The variable frequency excitation was provided by a rectangular pulse generator through an audio amplifier with the pulse width set to 10 micro seconds. Two second recordings were made at each of the frequencies shown on the x-axis of figures 2,3 and 4. Each of these recordings was then estimated at 14ms intervals using frames of 36ms duration (for a 256pt transform) or less ( for smaller transforms) and the resulting set of resonant frequency estimates were averaged.

The results graphed in figure 2. show the harmonic speech structure produces considerable bias in the resonant frequency estimates from both methods. As might be expected this increases as the excitation frequency increases and the harmonics become more sparse. The dotted line indicates the frequency of the nearest harmonic to each formant and it is

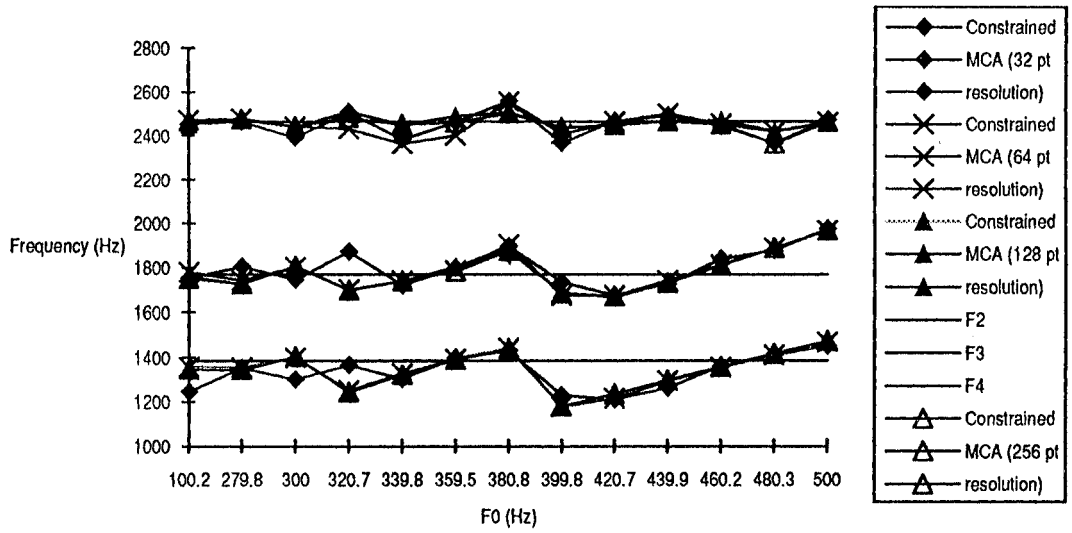


Figure 3. Demonstrates little change in MCA performance when spectral resolution is varied from 256 point FFT(27.3Hz) to a 32 point FFT (218.8).

apparent that although both methods are strongly biased towards the nearest harmonic the Constrained Centroid Analysis is affected to a lesser degree particularly for F4.

## 6. COMPUTATIONAL COMPARISON

The computational complexity of MCA is determined by the resolution of the Fourier transform. The computational complexity of the transform is proportional to NlogN (where N is the number of points in the transform). The computation involved in the MCA is proportional to (L+P)M operations where L, M and P are the ranges of the 1st, 2nd and 3rd internal boundaries respectively for the four partitions. The ranges are measured in FFT bins and so become fewer for smaller FFTs.

Formant analysis of a given 2 second waveform on a Sun sparcstation takes 12.5 seconds of user time when using the LPC method; 150.7 secs for the unconstrained least squares MCA method; 5.8 secs for constrained least squares MCA method using a 256 point transform; 1.2 secs using a 128pt transform; 0.7secs using a 64pt transform and 0.4secs using a 32pt transform. A 32pt transform corresponds to a 219Hz resolution in the frequency domain. The relative performance of MCA when different transform sizes are applied is shown in figure 3. Very little degradation in performance is apparent right down to a resolution of 219Hz. Further reduction to a resolution of 438Hz, however, produces catastrophic degradation.

Figure 4 demonstrates the robustness that can be gained by applying appropriate constraints. When unconstrained, the boundary between F1 and F2 falls below the lower limit set at 1421Hz.

A number of adjustments could be made to the Linear Prediction algorithm which might potentially improve it's performance such

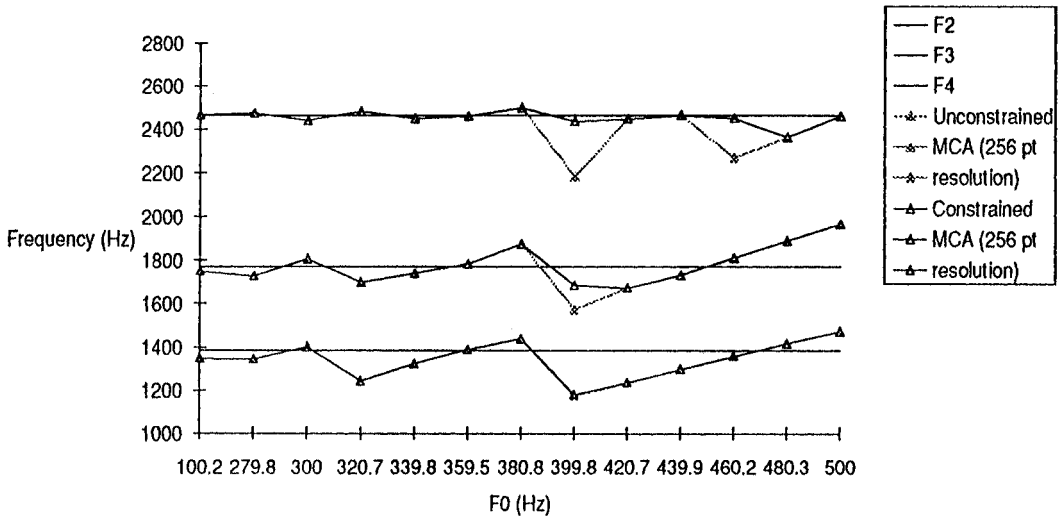


Figure 4. The imposition of constraints on partition boundaries results in reduced errors in the mean formant frequency estimates when compared to Unconstrained Least Squares Multiple Centroid analysis of the acoustic tube.

as using pitch synchronous analysis or applying another method of solution rather than autocorrelation or varying the pre-emphasis factor. However, in practice these changes significantly increase computation and can result in algorithm instability. The root solving technique of Snell and Milinazzo[5] would appear to be efficient but it is not clear how it compares in speed to standard peak picking methods. The implementation used here is that which is provided with the ESPS package by Entropic representing a practical standard.

## 7. CONCLUSIONS

The likely reason for inaccuracy in F2 and F3 estimates compared to F4 particularly at high F0 is the aliasing of energy from both F2 and F3 onto the same harmonic. We can see that F2 and F3 are less than 1 harmonic apart and that the formant energy from both has spread into the same harmonic, thus making it impossible to resolve the two and the estimation algorithms can do no better than select the harmonics. In contrast, F4 is more than one harmonic from the adjacent formants and so even though the energy has spread to the widely spaced harmonic structure it is still possible to attribute the centroid of the energy to the frequency of F4. It is in this situation that MCA wins over LPC analysis. Fortunately, it is the case that in children's speech the formants are more widely separated than in adult speech. As a result there are often more than 2 harmonics between formants (Figure5).

10kHz

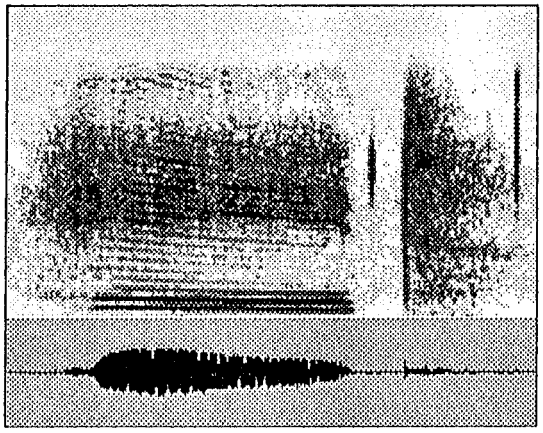

Figure 5. "Heed" spoken by 4 year old. MCA estimates F1=506Hz, F2=3025Hz, and F3=4050Hz

This study has examined several aspects of the performance of Constrained Centroid Analysis; namely its accuracy, robustness and computational efficiency. A controlled static study has indicated good performance in each of these areas but excludes important characteristics of real speech such as the dynamic nature of speech for which the tracking algorithm is required and the comparative resolving power of the two techniques is not tested., From subjective observation, the dynamic tracking performance and resolving power of Centroid Analysis appear comparable with LPC but these aspects remain to be examined in detail in future objective studies.

## 8. SUMMARY

Multiple Centroid Analysis performs favourably on high pitched data showing improvements in speed and accuracy over the conventional Linear prediction tracking method. The algorithm integrates knowledge about the range of formant positions into the spectral resonance estimation rather than applying it in a post-processing tracking algorithm thus allowing savings in computation. Furthermore, the estimation process does not incorporate an all-pole spectral model. This gives it an advantage when used to estimate speech produced by a high pitched voice or speech with a nasal or fricative quality since in these cases the model is inaccurate. The *constrained* least squares method of multiple centroid analysis can provide a substantial improvement in computational efficiency over unconstrained analysis (a factor of 26 in the example shown) and is faster than LPC based analysis (by up to 31 times in the examples shown). The imposition of a limited range of frequencies for each formant makes the analysis more robust and the accuracy of the constrained least squares algorithm degrades gracefully as the resolution is decreased, permitting a simple trade-off between computation and accuracy. Most importantly from the point of view of analysing children's speech, constrained least squares multiple centroid analysis is shown to be more tolerant to sparse harmonic structure than LPC analysis.

## ACKNOWLEDGEMENTS

The authors thank Dr M. Campbell for the use of laboratory facilities to generate the acoustic data for this work and Mr Raymond Parks for setting up the experiment. The helpful contributions of Dr Colin Watson and Alan Sharp are also gratefully acknowledged. Work leading to this paper was funded by the British Cancer Research Campaign.

## REFERENCES

[1] Wrench A. A. et al, "Objective Speech Quality Assessment in Patients with Intra-oral Cancers: Voiceless Fricatives", Proc. ICSLP 92, Banff, Vol. 2., pp 1071- 1074, 1992.

[2] Wrench A. A. et al, "A Speech Therapy Workstation for the Assessment of Segmental Quality: Voiceless Fricatives", Proc. Eurospeech 93, Berlin, Vol. 1, pp 219-222, 1993.

[3] McCandless S. S., "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-22, pp. 135-141, 1974

[4] A. Crowe and M.A. Jack. "Globally optimising formant tracker using generalised centroids" *Electronics Letters*, Vol. 23, No.19, pp 1019-1020, 1987.

[5] Snell, R. C. And Milinazzo, F., "Formant Location From LPC Analysis Data", IEEE Trans. Speech and Audio Pocessing, Vol. 1,No. 2, pp 129-134, April 1993