

## A NOVEL SPECTRUM ESTIMATION ALGORITHM OFFERING ENHANCED FEATURE RESOLUTION IN LPC SPECTRA

G. Duncan<sup>†</sup> and M.A. Jack<sup>†</sup>

### ABSTRACT

A novel formant estimation technique based on a pole-focusing mechanism applied to linear predictive coding (LPC) spectra is presented. The pole focusing technique is shown to provide superior feature resolution compared to standard LPC analysis. In its application to formant estimation from the speech signal, the technique is found to be the antithesis of standard LPC analysis, requiring no preemphasis, and demanding use of very large model orders. The technique has the additional benefit of obviating any application of model-order determining criteria through its concurrent use of within-frame increasing model order and off-axis spectral estimation.

### INTRODUCTION

A phonetic-feature-based approach to speech modelling demands explicit representation of those articulatory and acoustic features of the waveform known to be perceptually important to the listener. Notwithstanding the performance of the knowledge-based components of the speech recognition system, the acoustic front-end processor is at liberty to employ various types of signal transformation as best befits computational requirements. Nonetheless, one of the most widely used digital signal processing techniques in either category is linear predictive coding (LPC) analysis, based on modelling the speech production process (ref.1). The essence of the technique is to model the vocal tract, from the glottis to the lips, as a set of short, concatenated acoustic tubes whose overall filter transfer function is given by a digital all-pole model. The most important property of this application of autoregressive signal analysis to the composite speech signal is its ability to deconvolve the filtering action of the vocal tract from the glottal excitation waveform. This provides a convenient spectral representation of the vocal tract frequency response, being devoid of any harmonic line components due to periodicity of the waveform within the analysis window.

The pole focusing technique presented here (ref.2) employs autocorrelation-based LPC methods, but is the antithesis of standard LPC analysis in that very large model orders are used in the modelling process, but are coupled with the use of off-axis spectral estimation. That is, spectral estimation along search paths which lie well within the unit of the digital z-transformation plane. Additionally, the pole focusing technique does not require the application of preemphasis. Here, the steps required in the provision of formant estimates using the pole focusing method are compared and contrasted with those required by standard LPC processing. Finally, a comparative example is given which illustrates the superior feature resolving properties of the pole focusing technique.

### FORMANT ESTIMATION USING THE POLE FOCUSING TECHNIQUE

The amplitude value of any given sampled data point on the speech waveform,  $s_n$ , can

---

<sup>†</sup> Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, U.K.

be predicted approximately by a weighted linear sum of several, say  $p$ , past data points, that is:

$$s_n - \hat{s}_n = \sum_{k=0}^p a_k s_{n-k} = \epsilon_n ; a_0 = 1 ; \epsilon_n^2 \rightarrow 0 \quad (1)$$

where  $\hat{s}_n$  is the predicted value of the actual data point. The coefficients themselves can be found by applying the condition of least mean squares, which is a consequence of the applied constraint that, during the short time period of 25ms or so ( $= N$  sampled data points), over which the vocal tract is considered to have a fixed configuration, the total error power,  $\sum \epsilon_n^2 = \sigma^2$ , associated with the choice of prediction error coefficients must be minimal.

The key to understanding the nature of the new pole focusing technique proposed in this paper, however, is best found by an interpretation of the properties of linear prediction in the transfer function and frequency domains. Firstly, given that  $a_0 = 1$ , then in equation (1), the sequence of prediction error coefficients can in fact be viewed as a finite impulse response prediction error filter (p.e.f.) which is digitally convolved with the input speech signal to produce the prediction error, or residual,  $\epsilon_n$ . That is, in the frequency domain:

$$\frac{S(nF)}{\frac{1}{A(nF)}} = E(nF) \approx \sigma^2 \quad (2)$$

where  $F$  is the frequency resolution given an  $N$ -point data window,  $F = 1/(NT_s)$ . That is, LPC-based spectral analysis attempts to minimise the difference between the actual signal spectrum,  $S(nF)$ , and its estimate,  $1/A(nF)$ . The residual signal is thus assumed to be a Gaussian random variable with zero mean and variance  $\sigma^2$ , and  $1/A(nF)$  is an estimate of the vocal tract frequency response. Since the impulse response of the prediction error filter is simply the coefficient sequence  $a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}$ , then  $A(nF)$ , the inverse vocal tract frequency response, can readily be calculated by employing the usual discrete Fourier transform along the  $z$ -plane unit circle, with  $z = e^{j2\pi m/M}$ , where  $M$ , the length of the DFT, can be set to any desired value to give any arbitrarily-required spectral resolution,  $F = 1/(MT_s)$ . Of course, there is no reason why the  $z$ -plane unit circle need be chosen as the sole search path for calculation of the DFT. Choosing a search path in the  $z$ -plane such that  $|z| < 1$ , that is,  $z = r e^{j2\pi m/M}$ ,  $0 \leq r \leq 1$  will enhance the effect of p.e.f. zeroes, or vocal tract poles, on the inverse vocal tract frequency response. Note that spectral "amplitude" in these "off-axis" spectra however, now has little meaning in respect of providing a direct physical interpretation of spectral features, other than indicating when the off-axis search path approaches the location of a vocal tract pole.

If a formant peak is manifest in several off-axis spectra in any given analysis frame, then its centre frequency can be expected to move in a deterministic manner as the DFT search path approaches and recedes from the pole positions characterising the formant peak. This effect can best be understood by examining the properties of the damping factor,  $\zeta$ , related to the position of poles in the analog Laplace transform  $s$ -plane.  $\zeta$  itself is a measure of the closeness of any pole pair to the  $j\omega$  axis, and hence a measure of the Q-factor of the associated resonance peak in the system frequency response, although evidently this will also depend on the locations of other transfer function components in a multi-stage filter system. It is a well-known result that the equation governing the relation between the undamped natural frequency of oscillation,  $f_o$ , and the resonance centre frequency,  $f_m$ , of

the frequency response of a simple two-pole system, viewed from the  $j\omega$ -axis is:

$$\frac{f_m}{f_o} = \sqrt{1 - 2\zeta^2} \quad (3)$$

In the digital domain, using a generalised DFT search path with constant radius,  $r_s$ , then the relationship between damping factor (and hence bandwidth) of any spectral peak and radial distance,  $r_d$ , between any z-plane pole position and the search path,  $r_d = r_p - r_s$ , say, is given by:

$$\ln(1 - |(r_p - r_s)|) = \ln(1 - |r_d|) = -\zeta\omega_o T_s \quad (4)$$

and substituting for  $\zeta$  in equ.(3) gives:

$$\frac{f_m}{f_o} = \sqrt{1 - 2 \frac{(\ln(1 - |r_d|))^2}{\omega_o^2 T_s^2}} \quad (5)$$

In the pole focusing technique, the method adopted for provision of formant centre frequency data has been to employ a weighted averaging operation to the collected set of peaks extracted from all off-axis spectra. Firstly, the peak with the highest Q-factor is located and assumed to be an estimate of the location of a vocal tract pole in the z-plane,  $p_a$ , say, with undamped natural frequency  $f_{ao}$ . From a knowledge of the maximum DFT radial distance, from the assumed pole to the most distant DFT search path possible, a minimum value for  $f_{am}/f_{ao}$  is calculated from equation (5). This specifies the frequency range over which peak values in the list are to be averaged to yield a value for formant candidate  $F_a$ . All peaks lying between  $f_{ao} \pm f_{am}$  are averaged and weighted according to their Q-factor. Thus, a value for formant frequency is given by:

$$F_a = \frac{\sum_{u=0}^U f_{au} Q_{au} / Q_{ao}}{\sum_{u=0}^U Q_{au} / Q_{ao}} \quad (6)$$

where  $U$  is some unspecified number dependent upon  $f_{am}$  and  $r_{dmax}$ .  $Q_{ao}$  is the Q-factor of the pivotal peak at  $f_{ao}$ . Peaks within the averaging range are deleted as future candidates for vocal tract poles, but may still be used in the averaging process of other such poles. Note that the averaging range extends to  $f_{ao} + f_{am}$ . There is no theoretical requirement to include this upper band, but it is found to be necessary as a consequence of choosing to implement an increasing model order with decreasing radius,  $r_s$ . The peak in the list with the next-highest Q-factor, excluding those previously included in the averaging operation above, is chosen as the pivot for the next pass of the averaging process, and this operation is continued until no more pivotal peaks are available. Finally, the extracted formant candidates  $F_a$ ,  $F_b$ , etc. are arranged in frequency order, together with associated averaging weights and radius at which the pivotal peak was found, which is taken to be the approximate radial position of vocal tract pole.

Standard LPC analysis employing preemphasis and a  $p=18$ th-order model is compared against pole focusing performance in formant extraction from the phrase "allow your rule"

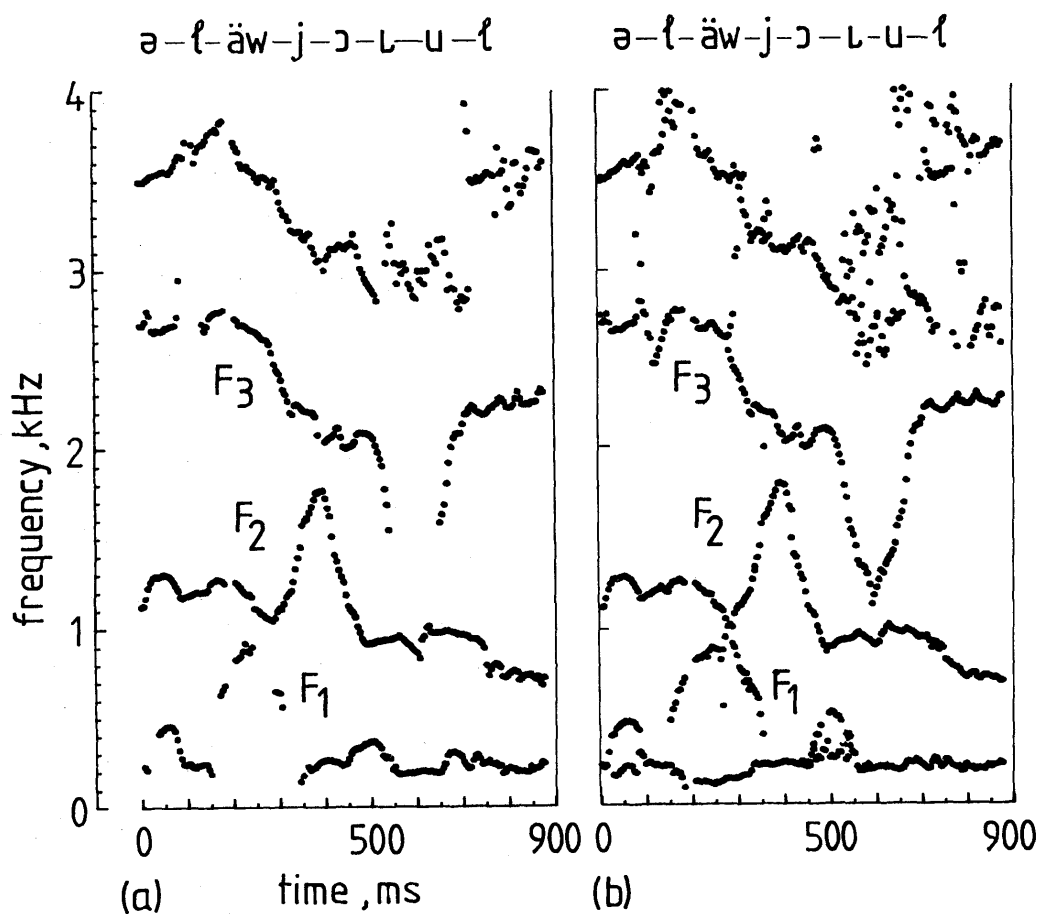
consisting of purely voiced speech spoken by a phonetically-trained male talker. In figure (1a) employing standard LPC analysis with a fixed model order, whilst the formant values correspond well to those expected, there are several regions with a low yield of formant values; specifically, during the /ow/ of "allow", there is a low yield of values for  $F_1$ . Similarly, during the transition from /oo/ to /r/ in "your",  $F_3$  is undetectable below 1.4kHz as it approaches  $F_2$ . Using an increasing model order and off-axis spectral estimation, however, figure (1b) demonstrates that there is an excellent yield using pole focusing for formants  $F_1$ ,  $F_2$  and  $F_3$ .

#### ACKNOWLEDGEMENT

This work has been supported by a U.K. Science and Engineering Research Council grant.

#### REFERENCES

1. B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America* 50(2 (part 2)) pp. 637 - 655 (1971).
2. G. Duncan and M. A. Jack, "An improved algorithm based on pole enhancement for estimation of the vocal tract frequency response," *Electronics Letters* 22(23) pp. 1213 - 1214 (Nov. 1986).



**Fig.1** Comparison of formant estimation properties for (a) standard LPC analysis,  $p=18, \mu=0.976$ , and (b) pole focusing technique. Time-aligned IPA phonetic transcription is indicated.