SPEECH SPECTRAL SEGMENTATION FOR SPECTRAL ESTIMATION AND FORMANT MODELLING

H.S. Chhatwal, A.G. Constantinides

Department of Electrical Engineering, Imperial College of Science and Technology, Exhibition Road, London, U.K.

ABSTRACT

Among the most useful methods for the spectral estimation of speech signals are those based upon the principles of Linear Prediction (LP). Due to the particular spectral energy distributions of voiced speech sounds, it is often advantageous to develop LP models of selected portions of the signal frequency response, rather than the entire spectral range. This modification leads to the formulation of the technique known as 'Selective Linear Prediction'. However, the choice of frequencies which divide the individual regions, and the predictor order to adopt in each portion of the overall spectrum are both essentially intuitive and based on merely an approximate notion of the energy distribution of the signal under consideration.

This work describes a method of splitting the speech spectrum into several segments using the approach of Dynamic Programming in order to produce the 'best' fit between real and model spectra on the basis of a suitable error criterion. This technique is finally modified for the segmentation of a frequency response into speech formant regions and implemented for several vowel utterances.

1. INTRODUCTION

The notion of a parametric representation of the speech signal is of central importance in many areas of speech processing. One of the most successful characterizations of acoustical speech behavior is based upon the excitation of a slowly time-varying linear system (representing the combined effects of the glottal wave, vocal tract and radiation of speech energy from the lips) by a quasi-periodic pulse train for voiced speech or random noise for unvoiced utterances. In the particular class of parametric models associated with Linear Predictive Analysis, the transfer function of the linear system is represented by a finite number of poles over short time intervals during which the signal is assumed to be stationary [8],[11]. A simple LP synthesizer based upon such a scheme would, therefore, be of the form shown in Fig. 1.

The pre-eminence of linear predictive methods in a wide variety of speech processing applications stems from the existence of simple and non-iterative algorithms for the evaluation of the predictor coefficients and the ability of LP analysis to provide accurate estimates of the basic speech parameters such as pitch [1],[2], formant positions [3],[4] and vocal tract area functions [5].

In addition, LP can also be used in the area of spectral estimation as an effective means of speech spectrum smoothing. This particular attribute is a consequence of the fact that LP tends to model the envelope of the spectrum without being sensitive to the fine structure caused by the periodicity of the speech signal [6]. The predictor parameters can then be stored in speaker/speech recognition systems in order to generate the model spectrum when a comparison with the corresponding model for the unknown speech signal is required.

Alternatively, the smoothed signal frequency response can be used to enhance considerably the efficacy of a formant extraction process where the positions of the peaks (or formants) in the speech spectrum are calculated via a simple peak-picking procedure. These formants also form an efficient characterization of voiced speech sounds and can be incorporated into effective formant speech synthesizer configurations.

The generalization of the method of spectral LP to the case where a given spectral range is matched by a model frequency response was first described by Makhoul [7] as 'Selective Linear Prediction'. Two important factors which determine the accuracy of the model spectrum are (i) the break frequencies between Selective LP segments and (ii) the allocation of predictor orders between the various regions. However, the choice of values for both of these variables is usually fairly arbitrary and is generally decided on the basis of a qualitative assessment of the concentration of spectral energy within various frequency ranges of the original signal spectral response.

In this paper, an approach to the Selective LP modeling of speech spectra is presented which optimizes the fit between real and model frequency responses with respect to both the segment break frequencies and predictor orders. The technique utilised in the implementation of this optimizing process is Dynamic Programming (DP). Therefore, after an initial description in Section 2 of the underlying theory of Selective LP spectral estimation and its uses, Section 3 outlines the development of an algorithm for frequency response segmentation based upon the principles of DP. The error measure to be minimized during the optimization procedure is also introduced in the context of its applicability to the particular type of signal under consideration.

Section 4 describes another application of this technique where certain additional constraints are included in the DP algorithm. These segmentation restrictions produce breaks between regions such that a single speech formant occupies each segmented region and is, therefore, described by a single, low order, Selective LP model. Finally, the results of the application of this procedure to several vowel sounds are presented.

2. SELECTIVE LP SPECTRAL MODELING

In the time domain formulation of the autocorrelation method of LP analysis, the N windowed speech samples $\{x_n: 0 \le n \le N-1\}$ within a given frame are employed directly in the evaluation of the signal autocorrelation sequence as:

$$R_i = \sum_{n=0}^{N-1-i} x_n x_{n+i} \qquad i \ge 0 \tag{1}$$

However, for the purpose of spectral estimation, only the discrete power spectrum $\{P(w_m): 0 \le m \le N-1, \ w_m = 2\pi m/N\}$ may be available and so the autocorrelation values may be obtained from the signal frequency response by an inverse DFT operation,

$$R_i = \frac{1}{N} \sum_{m=0}^{N-1} P(w_m) e^{jiw_m} \tag{2}$$

or, since $P(w_m)$ is real and even,

$$R_i = \frac{1}{N}\left\{ P(0) + (-1)^i P(\pi) + 2\sum_{m=1}^{N/2-1} P(w_m)\cos(iw_m) \right\} \quad (3)$$

For a linear predictor of order p, the first (p+1) values $\{R_i : 0\leq i\leq p\}$ are calculated and are then used in the computation of the predictor coefficients $\{a_k : 1\leq k\leq p\}$ and model gain (G) using, for example, the Levinson-Durbin algorithm described in [8] to solve the following equations:

$$\sum_{k=1}^{p} a_k R_{i-k} = -R_i, \quad 1\leq i\leq p \qquad \& \qquad G^2 = R_0 + \sum_{k=1}^{p} a_k R_k \quad (4)$$

These parameters are finally used to evaluate the discrete model power spectrum $P(w_m)$,

$$P(w_m) = \frac{G^2}{\left|1+\sum_{k=1}^{p} a_k e^{-jkw_m}\right|^2} \; ; \quad \begin{array}{l} 0\leq m\leq N-1 \\ w_m = 2\pi m/N \end{array} \quad (5)$$

In Selective LP, the above technique is generalized to the case where a given portion $w_a\leq w_m\leq w_b$ (i.e. $a\leq m\leq b$) of the speech spectrum $P(w_m)$ is to be modeled by the LP estimate $\hat{P}'(w_m')$. To compute the parameters of $\hat{P}'(w_m')$, the required region must be mapped onto the Selective LP frequency domain such that $w_a \Longrightarrow 0$ and $w_b \Longrightarrow \pi$ as shown in Fig. 2 and then the above method can be implemented to obtain $\hat{P}'(w_m')$.

The entire procedure for the evaluation of this spectral estimate is, therefore, as follows:

(a) Define a new spectrum $P'(w_m)$ based upon the required frequency range:

$$P'(w_m) = \begin{cases} P(w_m), & w_a\leq w_m\leq w_b \\ \text{undefined otherwise} \end{cases} \quad (6)$$

(b) Transform the actual frequency values (w_m) into the corresponding frequencies (w_m') within the Selective LP domain:

$$w_m' = \pi\left(\frac{m-a}{b-a+1}\right) \qquad a \le m \le b \tag{7}$$

or

$$w_m' = \pi\left(\frac{w_m - w_a}{w_b - w_a + \frac{2\pi}{N}}\right) \qquad w_a \le w_m \le w_b \tag{8}$$

This linear mapping ensures that $w_a \Longrightarrow 0$ and $w_b \Longrightarrow \approx\pi$ as required.

(c) Let $P'(2\pi - w_m') = P'(w_m')$. This defines $P'(w_m')$ over the lower half of the unit circle and w_m' now spans the entire range $0 \le w_m' \le 2\pi$.

(d) Compute autocorrelation values $\{R_i: 0 \le i \le p\}$ from (3) with $N=2(b-a+1)$ and $P'(w_m'), w_m'$ replacing $P(w_m), w_m$.

(e) Use procedure outlined in equations (4),(5) to compute the model spectrum $\hat{P}'(w_m')$.

The reason that Selective LP is of potential value lies in the particular spectral energy distributions of voiced speech sounds which suggest the desirability of splitting the speech spectrum into several regions and treating each segment as an independent spectral estimation problem.

For example, speech is often sampled at a 20 kHz rate to deal with the whole range of voiced & unvoiced sounds. However, for voiced speech, the 0-5 kHz interval is of most interest and so a high predictor order (e.g. $p_1=12$) can be allocated to this region and a lower order (e.g. $p_2=5$) for the 5-10 kHz one whereas to fit the whole 0-10 kHz range with one LP spectrum would require a predictor of order 24-28.

In this manner, not only are the computations reduced but a lower number of poles need to be interpreted if formant analysis is involved. The linear prediction parameters can also be stored for subsequent use in speaker/speech recognition systems where the Selective LP spectrum is used as a template against which the actual power spectral density of a given utterance can be compared.

Even in the 0-5 kHz interval, it would seem worthwhile to segment this range into smaller regions since the majority of signal energy is concentrated for most voiced sounds below 2-3 kHz and so it is more important to represent this area accurately than higher frequencies.

3. DYNAMIC PROGRAMMING APPLIED TO FREQUENCY RESPONSE SEGMENTATION

In the previous section, the choice of predictor orders and break frequencies between Selective LP segments could be made totally arbitrarily. However, this work is concerned with the evaluation of these parameters such that the fit between the original speech spectrum and the model frequency response consisting of several LP regions is 'best' in relation to a suitable error criterion. Therefore, the form of error measure to be employed in this analysis must first be considered.

3.1 Choice of Error Criterion

Let Total Number of Regions $= R$

Length of Region n $= r_n$

Endpoint of Region n $= t_n = \sum_{i=1}^{n} r_i$

Order of Predictor in Region n $= p_n$

Sum of first n Predictor Orders $= u_n = \sum_{i=1}^{n} p_i$

Error in Region n $= E_n(r_n, p_n)$.

In addition, the log magnitudes of the signal and model spectra are given by $\{s_i: 1 \leq i \leq N/2\}$ and $\{l_i(r_n, p_n): 1 \leq i \leq N/2\}$ respectively. Since the differences between real and model spectra are to be minimized, a squared error measure as in (9) would seem appropriate.

$$E_n(r_n, p_n) = \sum_{i=t_n-r_n+1}^{t_n} [s_i - l_i(r_n, p_n)]^2 \tag{9}$$

However, even though errors that occur at high frequencies are less important to the effectiveness of the model as the signal energy is less, they are provided as much emphasis in this scheme as those at lower frequencies. One solution to this problem lies in the use of a 'Weighted (squared) Error Criterion' of the form shown in equation (10).

$$E_n(r_n, p_n) = \sum_{i=t_n-r_n+1}^{t_n} s_i \cdot [s_i - l_i(r_n, p_n)]^2 \tag{10}$$

It is important to note that the minimization of this error criterion will actually result in a process involving two nested optimizations. The first is inherent in the choice of the model frequency response $l_i(r_n, p_n)$ within the given region since the linear prediction process itself yields this response in such a manner as to minimize the sum over all frequencies of the ratio of the signal power to the model power spectrum [6]. The second minimization is that of the weighted error measure and is carried out over the entire speech spectrum with respect to both the region sizes (r_n) and predictor orders (p_n) in order to obtain the optimal values for both of these parameters.

One major advantage of this approach is that it takes advantage of both the local spectral matching properties of LP and the suitability of the weighted error measure over a broader spectral range. This is so because LP tends to match the signal response much more effectively in the regions of spectral peaks than the areas of low signal energy [8]. The perceptually important regions corresponding to formant peaks are, therefore, well represented.

However, the LP spectral matching process operates uniformly over the entire frequency range and results in an estimate which treats regions of both high and low signal energy in the same manner. As mentioned above, the Weighted Error Criterion has certain advantages in this respect and so it is this form of error measure which is implemented in the optimization procedure discussed in the next section and is also used in all subsequent results.

3.2 Development of Segmentation Equations

With 2 regions and given values for the predictor orders p_1, p_2, the optimal break frequency can be quite easily selected as that frequency yielding the minimum total error. This problem will, however, become intractably complex if 3 or more regions are considered - and even more so if the values of $p_1, p_2, \ldots, p_R$ are also to be chosen optimally in some prescribed manner.

It is at this point that we must resort to a more sophisticated technique than simple enumeration of all possible combinations of break frequencies and predictor orders to determine that which gives the lowest error. The approach adopted in this work is that of Dynamic Programming - an algorithm that has been used quite successfully in, for example, the time domain segmentation of speech signals [9],[10].

The concept of optimality in relation to the choice of segment endpoints has been discussed above, but the implication of the term in calculating region predictor orders is less apparent. This is because it is meaningless to discuss the 'optimal' number of predictor coefficients to employ within a single LP region since the fit of the model improves indefinitely as the order p increases [11].

However, it is possible to determine the optimal allocation of coefficients between the segments to minimize the error between real and model spectra such that the sum of the individual orders over all the regions (u_n) is a constant. It is this definition of optimality which shall be considered in this section.

Let $f_n(t_n,u_n)$= LEAST COST APPROXIMATION OF FIRST t_n SPEECH SPECTRUM SAMPLES BY FIRST n SELECTIVE LP REGIONS WITH u_n AS THE SUM OF THE FIRST n PREDICTOR ORDERS

Then, with the weighted error measure of equation (10), this cost function can be expressed as:

$$f_n(t_n,u_n) = \min_{r_n} \min_{p_n} \left| f_{n-1}(t_n-r_n,u_n-p_n) + E_n(r_n,p_n) \right| \qquad (11)$$

This equation implies the following:

As r_n is the length of the nth segment, it is known that the first t_n-r_n samples are to be represented by (n-1) LP regions with a total predictor order of u_n-p_n and $f_{n-1}(t_n-r_n,u_n-p_n)$ gives the lowest possible error for this condition. The error in the nth segment itself is then given by $E_n(r_n,p_n)$. The parameters r_n,p_n are then varied over their allowable ranges in order to find those values which minimize the total error up to and including the nth segment. This result is finally assigned to $f_n(t_n,u_n)$ and can be used iteratively to compute errors arising from segmentation into (n+1) regions.

The possible range of values for r_n,t_n,p_n,u_n at each stage must now be considered. The most stringent constraint on the variation of r_n,t_n in this case is the fact that reasonable spectral matches are only obtained if the number of samples in a segment exceeds at least double the predictor order in that segment [8]. For p_n,u_n, the corresponding limitation is that, in Selective LP, the minimum number of coefficients that may normally be assigned to any one region is usually 2 - in order to produce at least 1 complex pole in the model spectrum. Using these results, the following equations constitute a complete description of the segmentation procedure.

For 1 region:

$$f_1(t_1,u_1) = E_1(r_1,p_1) \qquad 2p_1 \le t_1 \le N/2 \quad ; \; 2 \le u_1 \le u_R \tag{12}$$

For subsequent regions:

$$f_n(t_n,u_n) = \min_{2p_n \le r_n \le t_n - 2\sum_{i=1}^{n-1} p_i} \; \min_{2 \le p_n \le u_n - 2(n-1)} \{ f_{n-1}(t_n - r_n, u_n - p_n) + E_n(r_n,p_n) \}$$

$$2\sum_{i=1}^{n} p_i \le t_n \le N/2 \quad ; \; 2n \le u_n \le u_R \tag{13}$$

Starting with equation (12) and subsequently using (13), the number of segments is increased such that n=1,2,...R, where R is the required number of regions. The final result is then given by $f_R(N/2,u_R)$ which is the minimum error when all N/2 samples are represented by R Selective LP regions with an overall predictor order u_R.

However, this process does not automatically yield explicitly the optimal breakpoints for each segment and these values must be evaluated by 'Backtracking'. The values of r_n,p_n which produce the minimum error $f_n(t_n,u_n)$ (i.e. the optimizing values r_n^*,p_n^*) must, therefore, be stored at each stage.

At the end of R stages, these results can be employed in the calculation of the optimal segment endpoints $\{t_n^*: 1 \le n \le R\}$ recursively as in (14) starting at the last sample.

Initial Condition: $t_R^* = N/2$

$$t_{n-1}^* = t_n^* - r_n^* \qquad n=R,R-1,\ldots,2 \tag{14}$$

At this point, the optimal break frequencies and predictor orders are known so that the LP frequency response consisting of R regions can be reconstructed and compared with the real speech spectrum.

3.3 Segmentation Results

The above procedure was implemented for 2 and 3 region segmentations with optimal breakpoint selection and predictor order allocation. The results are illustrated in Fig. 3 along with the associated region end frequencies and errors. As u_R=12 in both cases, a 12th order single region LP spectrum is also shown for comparison. The speech segment under investigation consists of N=256 data samples from the voiced phoneme /ɔ:/ sampled at a rate of 10 kHz.

Both plots indicate a considerably improved fit between real and model spectra over the single region LP model but a more quantitative assessment of the improvement is obtained by comparing the total weighted errors indicated in Fig. 3 with the equivalent value for the single LP region given below.

WEIGHTED ERROR FOR SINGLE 12th ORDER LP REGION = 196,775

The reductions are, therefore, over 40% on this figure while employing the same overall predictor order.

4. SPEECH SPECTRAL SEGMENTATION INTO FORMANTS

4.1 Additional Constraints for Division into Formant Regions

The general breakpoint/predictor order allocation technique is now modified to allow formant-type regions to be selected and thereby obtain estimates of speech formant frequencies and bandwidths. The changes necessary are listed below.

(a) The predictor order in each region is fixed at p=2 to produce a single resonance peak modeling each speech formant. This reduces the computational requirements considerably and also allows the 2nd order model parameters to be expressible directly in terms of the autocorrelation sequence $\{R_0, R_1, R_2\}$ as in equation (15). This eliminates the need for a general routine such as the Levinson-Durbin algorithm.

$$a_1 = \frac{R_1 R_2 - R_1 R_0}{R_0^2 - R_1^2} \quad ; \quad a_2 = \frac{R_1^2 - R_2 R_0}{R_0^2 - R_1^2} \quad ; \quad G^2 = (1 - a_2^2)\frac{R_0^2 - R_1^2}{R_0} \qquad (15)$$

(b) A minimum possible size for each region (r_{min}) is established such that all segments cover a frequency range of at least 250 Hz (approx) which allows formant-type regions to be selected but precludes the possible isolation of pitch harmonics in the original frequency response (which generally have a frequency range of around 100 Hz).

(c) From the 2nd order LP model system function H(z) of equation (16), the poles of H(z) will be at the positions indicated:

$$H(z) = \frac{G}{1+a_1 z^{-1}+a_2 z^{-2}} \quad ; \quad \text{Poles at } z, z^* = \frac{-a_1 \pm \sqrt{a_1^2-4a_2}}{2} \qquad (16)$$

$$= p \pm jq$$

Therefore, if $(a_1)^2-4a_2>0$, the Selective LP model consists of a pair of real poles and the terms "resonance frequency" and "bandwidth" lose their inherent meaning. This is obviously not the type of situation which is required in the estimation of formant positions and so this form of segmentation is prevented by setting the error for the nth region to a very large value if, for a certain t_n, r_n, the predictor coefficients are such that real poles occur. This prevents this combination of t_n, r_n leading to the minimum $f_n(t_n, u_n=2n)$.

The results of spectral segmentation with the above constraints are shown in Fig. 4 for 4 different voiced phonemes with N/2=128 frequency samples and a sampling rate of 10 kHz in all cases. A single region LP spectrum of the same overall order as the Selective LP model is also shown for comparison. These plots indicate that this form of segmentation does indeed tend to produce divisions which separate the individual formants as required.

The number of regions ranges from 4 to 6 depending on the utterance and it is the choice of the total number of segments (which also determines the overall predictor order) that is discussed in the next section.

4.2 Number of Segmentation Regions

The reduction in weighted error as the number of segments increased was found to be of the same asymptotic form as that for a single LP region with increasing predictor order (although at a much faster rate). This suggested the use of a stopping criterion similar to those employed for standard non-segmented LP modeling. One such criterion mentioned by Makhoul [11] which takes account of this asymptotic behavior was investigated.

Let E^n_{tot} = TOTAL WEIGHTED ERROR FOR n REGION MODEL

$$= f_n(t_n=N/2, u_n=2n)$$

The following threshold test is then implemented:

$$T_n = 1 - \frac{E_{tot}^{n+1}}{E_{tot}^{n}} < \delta \qquad (17)$$

This implies that T_n should fall sharply to a value below δ when the required number of regions is reached and subsequently remain steady. The variation of T_n with the number of segments (n) for the four utterances investigated earlier is shown in Fig. 5. Arrows indicate the number of regions illustrated in the results of Fig. 4. This threshold test, therefore, provides a reliable stopping criterion with any value of δ in the range $0.09 < \delta < 0.18$ yielding the required number of regions.

5. SUMMARY

A procedure has been established for the partitioning of speech spectra into Selective LP regions in order to minimize the discrepancy between real and model frequency responses with respect to a weighted error criterion. This technique uses Dynamic Programming to yield both the optimal set of breakpoints and allocation of predictor orders and was found to produce considerably improved spectral estimates of the original signal when compared with standard LP modeling.

The feasibility of spectral segmentation into formants using 2nd order regions was also established for a variety of vowel utterances. An effective stopping criterion was finally developed for the number of segments to implement in this analysis - thereby offering a solution to the difficult problem of choosing an overall order for the linear predictor.

ACKNOWLEDGEMENTS

This work is carried out in collaboration with the IBM U.K. Science Centre, Winchester, U.K. and we are grateful for their help.

REFERENCES

[1] J.D. MARKEL, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, No.5, pp.367-377, December 1972.

[2] J.N. MAKSYM, "Real-Time Pitch Extraction by Adaptive Prediction of the Speech Waveform", IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No.3, pp.149-153, June 1973.

[3] J.D. MARKEL, "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation", IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, No.2, pp.129-137, June 1972.

[4] S.S. McCANDLESS, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. ASSP-22, No.2, pp.135-141, April 1974.

[5] H. WAKITA, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No.5, pp.417-427, October 1973.

[6] J. MAKHOUL, "Spectral Linear Prediction: Properties and Applications", IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. ASSP-23, No.3, pp.283-296, June 1975.

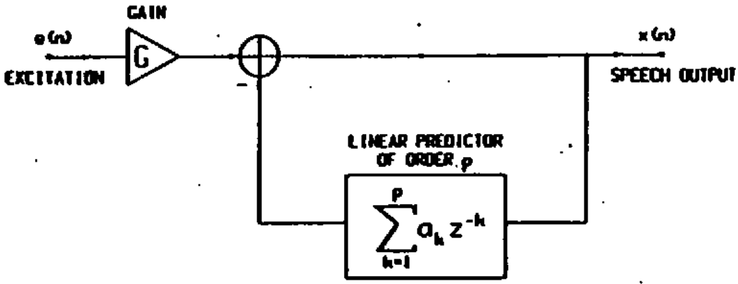
[7] J. MAKHOUL, "Selective Linear Predictive Spectral Matching", Presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, Oct.30-Nov.2 1973. (Abstract included in the Journal of the Acoustical Soc. of America, Vol. 55, No.2, February 1974, pg.397)

[8] L.R. RABINER & R.W. SCHAFER, "Digital Processing of Speech Signals", Prentice-Hall, 1978.

[9] J.S. BRIDLE & M.D. BROWN, "Connected Word Recognition using Whole Word Templates", Proc. Inst. Acoust. Autumn Conf., November 1979, pp.25-28.

[10] H. SAKOE & S. CHIBA, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on Acoustics, Speech and Signal Proc., Vol. ASSP-26, No.1, pp.43-49, February 1978.

[11] J. MAKHOUL, "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, No.4, pp.561-580, April 1975.



SYSTEM TRANSFER FUNCTION: $$H(z) = \frac{X(z)}{E(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

Fig. 1: Block Diagram of LP Model for Speech Production

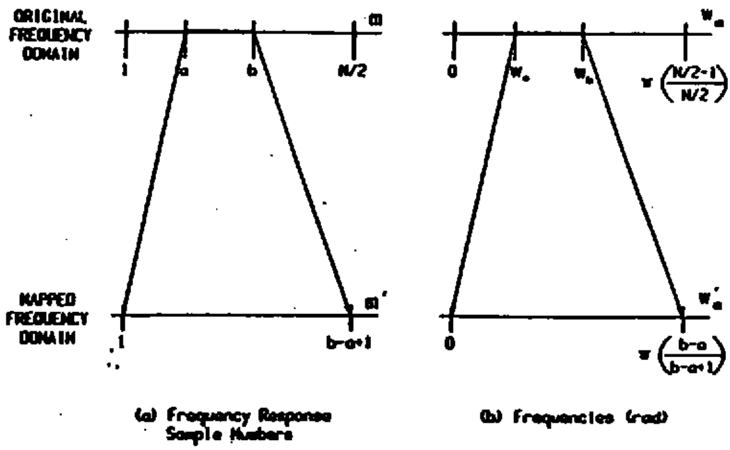
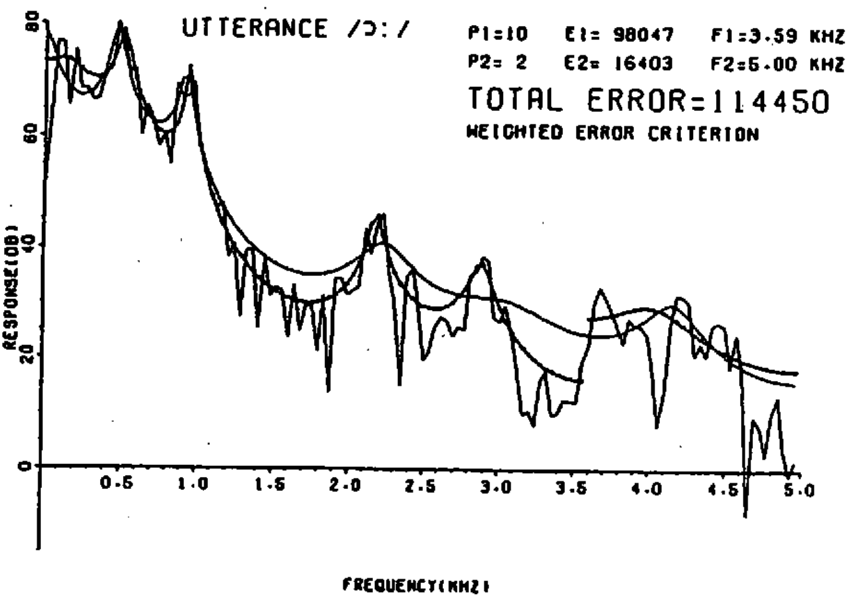
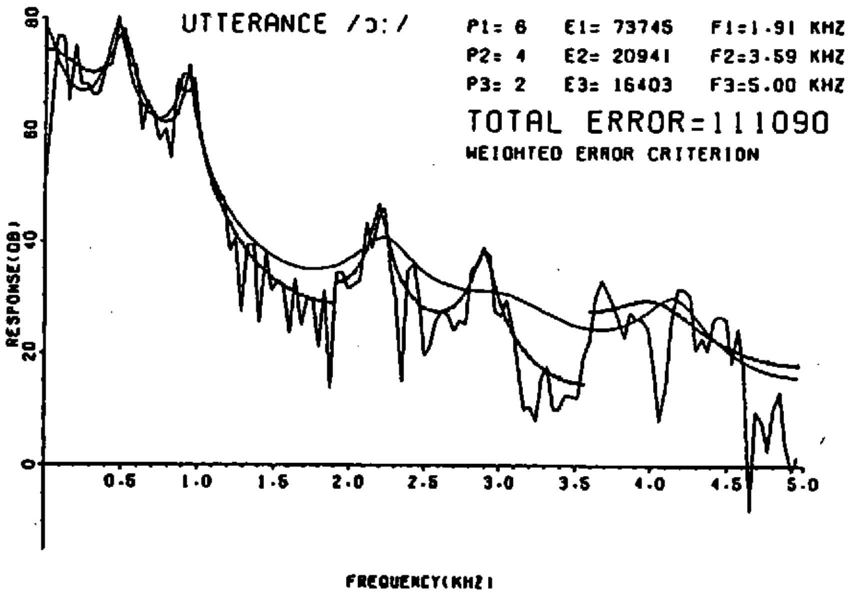


Fig. 2: Frequency Domain Transformations for Selective LP Modelling

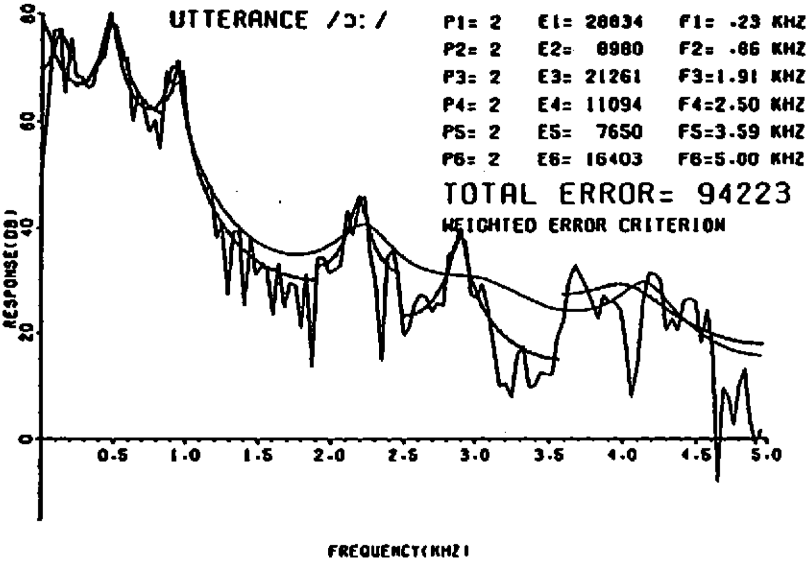




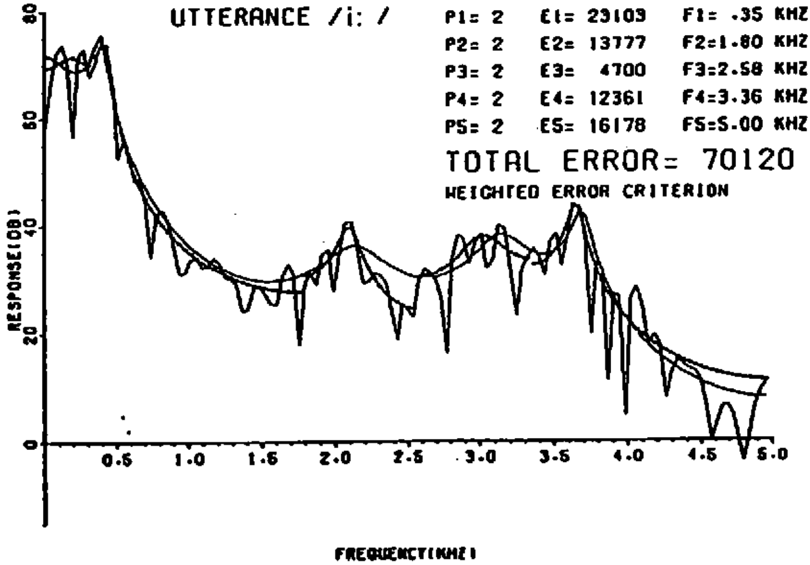
———: Standard single region LP fit to speech spectrum

———: Selective LP fit

Fig. 3: Optimal Breakpoint Selection & Predictor Order Allocation for (a)2 Region Segmentation and (b)3 Region Segmentation

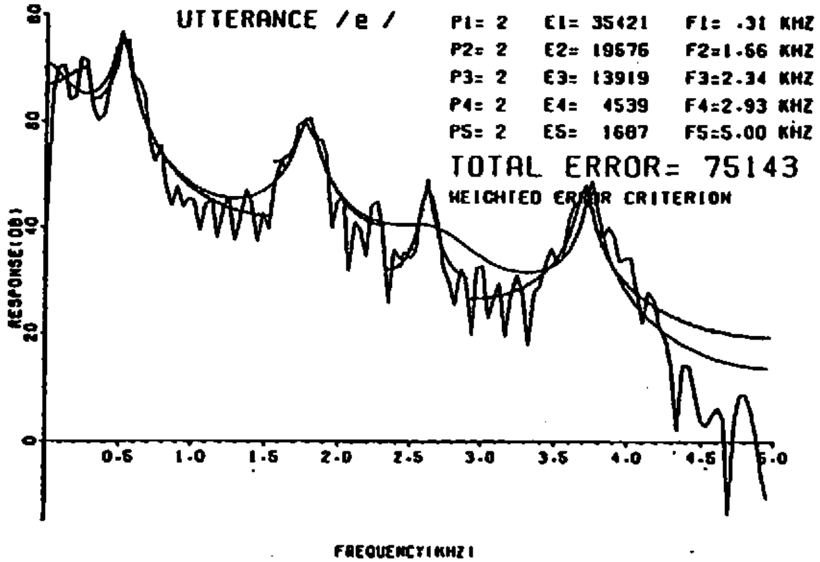


(a)

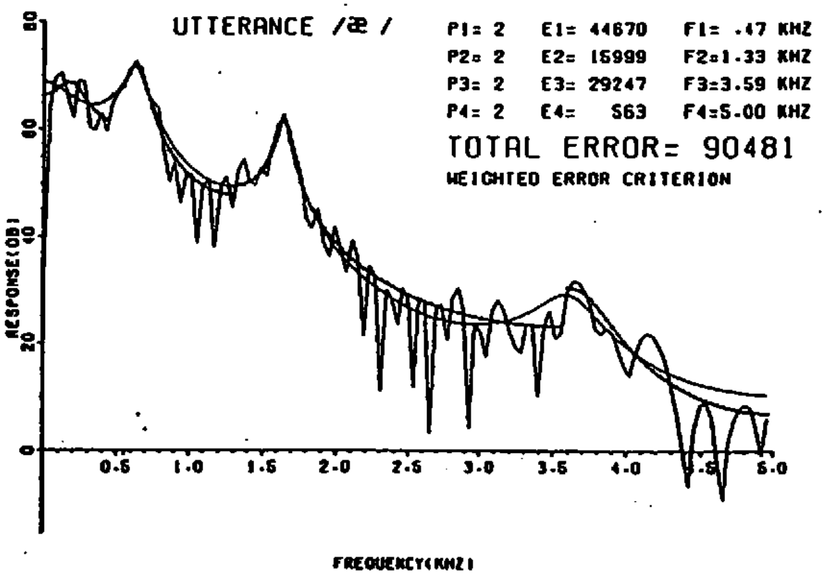


(b)

Fig. 4: Spectral Segmentation into Formants for 4 Vowel Phonemes



(c)



(d)

Fig. 4: Contd.

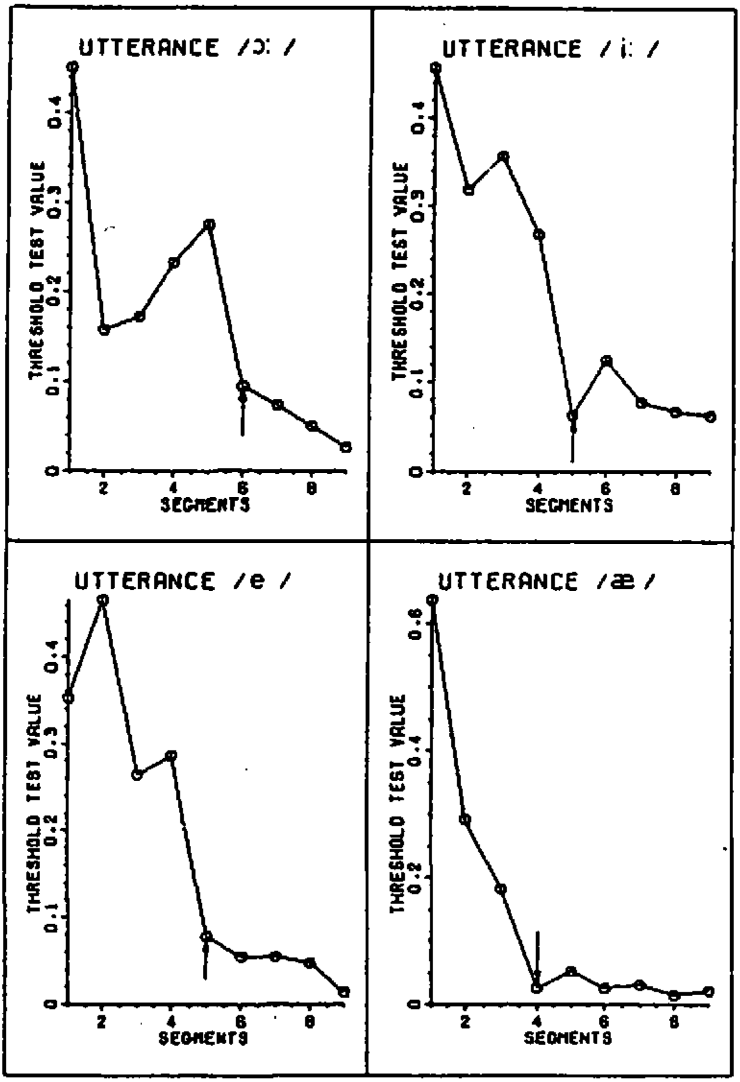


Fig. 5: Variation of Threshold Test Value with Number of Segments