Speech Communication 10 (1991) 209–221
North-Holland

Formant extraction from group delay function

Hema A. Murthy and B. Yegnanarayana
Department of Computer Science and Engineering, Indian Institute of Technology, Madras – 600 036, India

Received 29 August 1989
Revised 7 March 1991

Abstract. This paper presents an approach based on the properties of group delay functions for extracting formants from speech signals. The algorithm is similar to the cepstral smoothing approach for formant extraction using homomorphic deconvolution. The significant differences are (i) the logarithmic operation is replaced by $(\,)^r$ operation and (ii) the additive and high resolution properties of group delay functions are exploited to emphasize formant peaks. The group delay function (or the negative derivative of the Fourier transform phase) is derived for a signal which in turn is derived from the Fourier transform magnitude of the speech signal. If a suitable value of r is used, this method gives highly consistent estimates of formants compared to both the cepstral approach and the model-based linear prediction (LP) approach for smoothing the magnitude spectrum. The effects of the parameters, exponent r and window width p, on the proposed technique for formant extraction are studied.

Zusammenfassung. Dieser Beitrag stellt eine Methode zur Messung der Formantfrequenzen vor welche die Eigenschaften der Gruppenlaufzeitfunktionen ausnützt. Der Algorithmus ist der kepstralen Methode zur spektralen Abrundung ähnlich. Die zwei wichtigsten Underschiede sind (1) der Logarithmus wird durch einen $(\,)^r$ Operator ersetzt und (2) die additiven Eigenschaften und das gute Auflösungsvermögen der Gruppenlaufzeitfunktionen werden ausgenutzt um die Scheitelpunkte der Formanten hervorzuheben. Die Gruppenlaufzeitfunktionen (oder die negative Ableitung der Phase des Fourierspektrums) wird abgeleitet für ein Signal welches seinerseits von der Magnitude des Fourierspektrums des Sprachsignals abgeleitet wird. Wenn ein passender Wert für r gebraucht wird, dann ergibt die Methode Schätzwerte für die Formanten welche vergleichbar sind mit denen welche mit der kepstralen Methode oder mit der linearen Prädiktion gewonnen werden. Die Auswirkung des Exponenten r sowie der Länge des Analysefensters auf die Ergebnisse werden ebenfalls untersucht.

Résumé. Ce papier présente une technique fondée sur les propriétés des fonctions retard de groupe afin d'extraire les formants des signaux de parole. L'algorithme est semblable au lissage cepstral utilisant la déconvolution homomorphique. Les différences significatives sont les suivantes: (a) le logarithme est remplacé par un opérateur $(\,)^r$ et (b) les propriétés additive et de haute résolution des fonctions retard sont exploitées pour accentuer les crêtes des formants. La fonction retard de groupe (ou la dérivée négative de la phase de la transformée de Fourier) est dérivée pour un signal qui, à son tour, est dérivé de l'amplitude de la transformée de Fourier du signal. Si une valeur convenable de r est utilisée, cette méthode donne des estimations formantiques très cohérentes comparées à celles obtenues par la technique cepstrale ou par la prédiction linéaire. Les effets de l'exposant r et de la largeur de la fenêtre sur la technique proposée ont été étudiés.

Keywords. Fourier transform phase, spectral root cepstrum, group delay functions, formant extraction.

1. Introduction

Most speech analysis/synthesis systems are based on the acoustical theory of speech production. In this theory, the vocal tract system is approximated by a series or parallel connection of resonators. The natural resonance frequencies of the vocal tract system vary slowly and continuously with time. The resonance or formant features carry information relating to the identification of speech sounds. Detection of the formants requires that some estimate be made of either the frequency response or the transfer function of the vocal tract. The short time magnitude spectrum of a speech signal cannot be used by itself for formant extraction, as the glottal excitation and the analysis window effects manifest themselves on the spectrum of the vocal tract impulse response. Therefore standard peak picking algorithms for formant extraction cannot be applied

to the magnitude spectrum.

Linear prediction (LP) analysis (Makhoul, 1975) is a spectral modelling technique which is often employed to estimate the relevant parameters of the vocal tract. In LP analysis the vocal tract system is characterized by an all-pole model. The technique approximately deconvolves the excitation waveform and the vocal tract impulse response, thus producing a smooth spectral envelope. A simple peak-picking algorithm on the spectral envelope or a complete root solving of the resulting polynomial can be used to obtain an estimate of the locations of formants. LP analysis requires a choice of the model order which dictates the number of vocal tract poles and hence determines the shape of the estimated spectrum: too low a model order will not adequately emphasize all the formant features, while too high a model order may often generate spurious peaks. Also, it is sometimes difficult to detect a low amplitude formant adjacent to a large amplitude formant by the model-based spectrum estimation, even though the information is available in the magnitude spectrum.

Another approach which is commonly used for smoothing the magnitude spectrum is the technique based on the cepstrum (Rabiner and Schafer, 1978). In this method the inverse Fourier transform (FT) of the log magnitude function is computed. The resulting cepstrum is windowed to derive a smoothed magnitude spectrum, the peaks of which may correspond to the formants. A convolution in the time domain becomes a multiplication in the frequency domain which becomes an addition in the cepstral domain. Thus, in the cepstrum, the excitation information and the vocal tract response components become additive. But, the logarithm operation gives equal emphasis to very small and very large values. This is not desirable because even small (nearly zero) spectral magnitude values assume significance in the computation of the cepstrum, as the logarithm of a small number (≈ 0) is a large negative value.

Use of the phase spectrum (as opposed to the magnitude spectrum) generally improves the resolution of the short time spectral features (Yegnanarayana, 1978). The phase spectrum of the standard Fourier transform is difficult to interpret due to the inevitable wrapping of the phase function. The need to use an analysis window and the existence of the glottal excitation contribute significantly to the phase wrapping problem.

If the Fourier transform phase is to be processed, it should be available in an unwrapped form. The existing methods for phase unwrapping do not yield satisfactory results for all types of signals, especially when the roots of the signal z-transform are near the unit circle (Tribolet, 1977). Information about the resonances of the vocal tract are available in both the magnitude and the phase spectra. As the magnitude spectra of individual resonances are multiplied, the resolution is lost especially when the adjacent formants have significantly different amplitudes. Thus, although the information about closely spaced formants is available in the magnitude spectrum, they cannot easily be resolved.

Group delay (GD) spectra have been used to improve feature resolution available in the LP spectra (Yegnanarayana, 1978), but this technique still suffers from the need to adopt some modelling criterion. Modelling techniques may give higher resolution, but with the uncertainty of generating spurious peaks. Also, the group delay function cannot improve the frequency resolution restriction imposed by the data window. In this paper we propose a new technique based upon the spectral root group delay function for formant extraction which reduces the effect of data windows on spectral resolution. In Section 2 we discuss briefly the properties of group delay functions relevant for the detection of formant peaks. In Section 3 we discuss the proposed technique for formant extraction. We demonstrate through examples that the causal portion of the signal derived from the short time magnitude spectrum contains information about the smoothed magnitude spectrum. Performance of this technique is illustrated in Section 4 for various choices of analysis parameters and different signals (synthetic and natural speech).

2. Definitions and properties of group delay functions

Let $V(\omega)$ be the Fourier transform of the minimum phase signal $v(n)$. The following rela-

tions can be shown (Yegnanarayana et al., 1984):

$$V(\omega) = |V(\omega)|\exp(-j\theta(\omega)), \quad (1)$$

$$\ln|V(\omega)| = c(0)/2 + \sum_{n=1}^{n} c(n)\cos\omega n, \quad (2)$$

and

$$\theta(\omega) = -\sum_{n=1}^{\infty} nc(n)\sin\omega n, \quad (3)$$

where $\{c(n)\}$ are the cepstral coefficients (Rabiner and Schafer, 1978).

The group delay function which is the negative derivative of phase is related to the magnitude and phase through the cepstral coefficients $c(n)$ as follows:

$$\tau(\omega) = -\theta'(\omega) = \sum_{n=1}^{\infty} nc(n)\cos\omega n. \quad (4)$$

If $V_1(\omega)$, $V_2(\omega)$, ..., $V_p(\omega)$ are the responses corresponding to p resonators, the frequency response of the cascade of these resonators is given by

$$V(\omega) = V_1(\omega)V_2(\omega)\cdots V_p(\omega). \quad (5)$$

The group delay function of the overall system is given by the summation of the group delay functions of the individual resonators $\tau_i(\omega)$:

$$\tau(\omega) = \tau_1(\omega) + \tau_2(\omega) + \cdots + \tau_p(\omega). \quad (6)$$

It can be shown (Yegnanarayana, 1978) that each of these $\tau_i(\omega)$ has a behaviour similar to $|V_i(\omega)|^2$ around the resonance frequency and asymptotically approaches zero for ω away from the resonance frequency. Because $\tau_i(\omega)$ is proportional to the squared magnitude response and the $\tau_i(\omega)$s are additive, the resonance peaks show up much more sharply in $\tau_i(\omega)$ as compared to the magnitude response $|V(\omega)|$. Note that in $|V(\omega)|$ the influence of the response of one resonator on the other is significant due to the multiplication of $V_1(\omega)$, $V_2(\omega)$, ..., $V_p(\omega)$. Note also that while taking a logarithm on $|V(\omega)|$ does make the contribution of each term additive, the influence of one response on the other does not disappear since the logarithm of a small value is a large negative value. Figure 1 shows the $\log|V(\omega)|$ and $\tau(\omega)$ for a cascade of five resonators. The sharpness of each of the resonances and the resolving power of $\tau(\omega)$ is evident in the figure.

3. Formant extraction from group delay functions

3.1. Principle of the proposed method

For formant extraction we propose a spectral root group delay function approach which is similar to the spectral root homomorphic deconvolution (SRDS) (Lim, 1979). Our goal is to obtain good raw formant data from the speech signal. The proposed method involves deriving a signal with the characteristics of a minimum phase signal so that the phase spectrum of this signal contains the information of the magnitude spectrum (Yegnanarayana et al., 1988). The peaks of the group

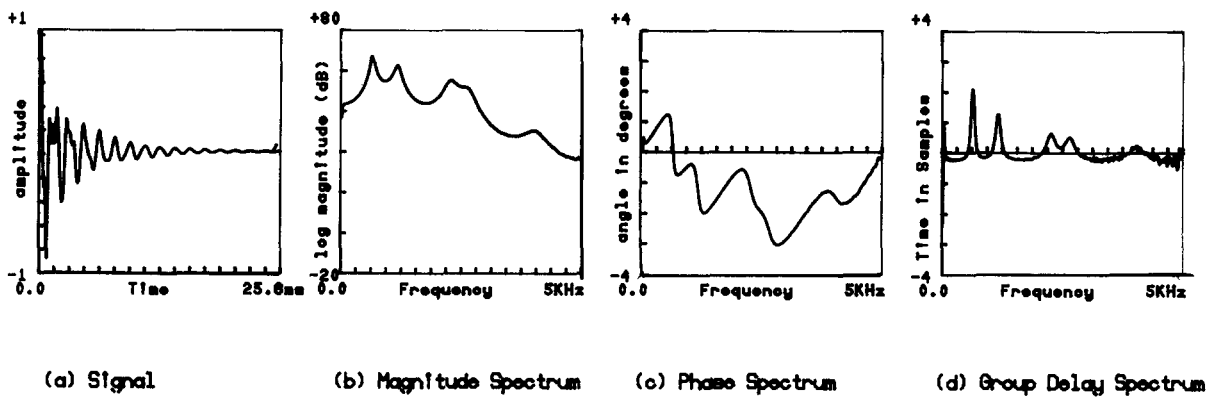


Fig. 1. Illustration of the high resolution property of group delay functions.

delay function derived from this phase function correspond to formants.

Figure 2 illustrates the new spectral root group delay function approach for formant extraction. In the figure F and F^{-1} correspond to the forward and inverse Fourier transforms, respectively. Han_w(n) is a gating function and is given by

$$\text{Han\_w}(n) = \begin{cases} 0.5 + 0.5\cos(\pi n/L), & 0 \le n \le L, \\ 0.0, & n > L, \end{cases}$$

where L is the length of the window. This technique is like the cepstral smoothing technique, except that (i) an rth power operation is used in place of the log operation and (ii) the phase group delay is computed instead of the smoothed magnitude spectrum. Figure 3(a) shows a segment of speech (25.6ms, 10kHz sampling rate). Figure 3(b–d) shows the magnitude, phase and LP spectra obtained by multiplying the speech segment with a Hamming window and then taking the Fourier transform. The inverse Fourier transform of the magnitude function gives an even sequence which is called the spectral root cepstrum. The even sequence is then truncated to include only the causal portion of it. This signal is then multiplied by half a Hann window to select the first p (corresponding to 4.2ms) samples (henceforth referred as $\tilde{x}_p(n)$). Figure 4(a) shows signal $\tilde{x}_p(n)$. It is worth noting that the magnitude and phase spectra of the original signal are unrelated, whereas the magnitude and phase spectra of signal $\tilde{x}_p(n)$ are related. Notice that peaks in the magnitude spectra of Figure 4(b) correspond to phase transitions in Figure 4(c). The differenced phase corresponds to the group delay function shown in Figure 4(d). The window size p should be taken as large as possible to obtain a good resolution of formants, but should be less than the pitch period in order to avoid fluctuations due to pitch in the magnitude and phase spectra.

Since we do not explicitly use a model in our analysis, we feel that our approach should provide a better representation of the underlying nature of the vocal tract system than the one obtained using model-based analysis. The cepstral ap-

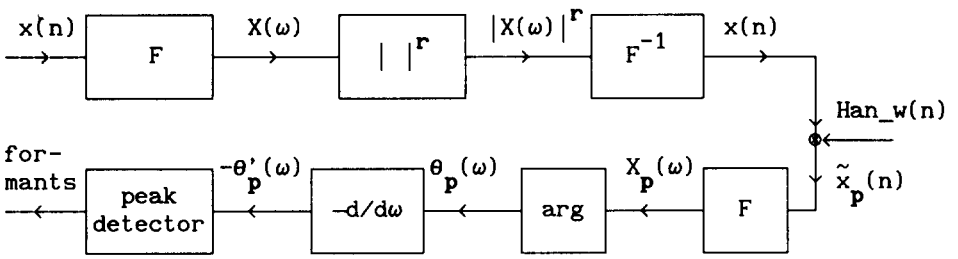


Fig. 2. Proposed method of formant extraction from spectral root group delay function.

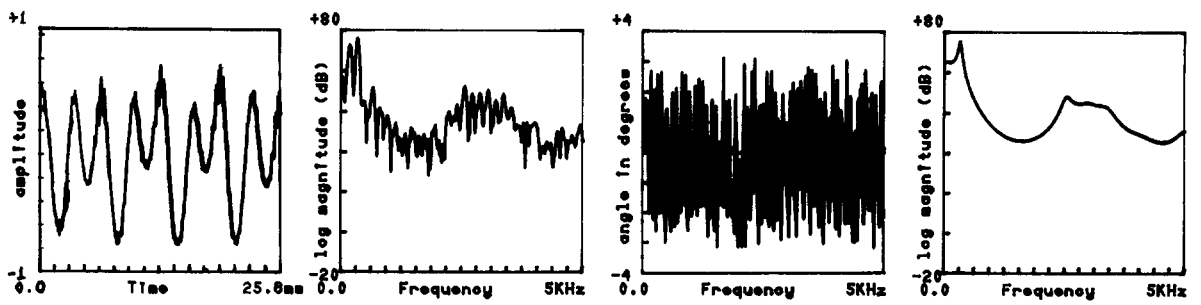


Fig. 3. A segment of speech and its spectra.

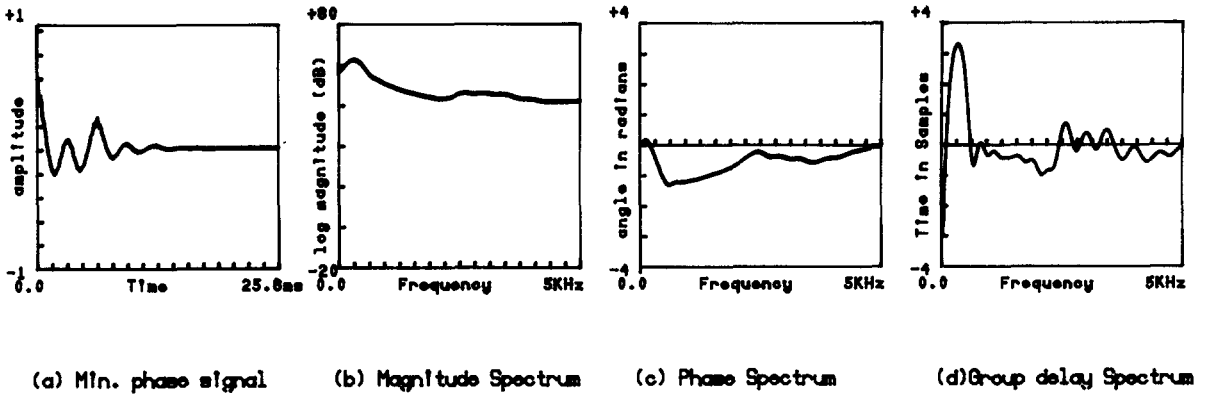


Fig. 4. Minimum phase signal and its spectra.

proach to formant extraction is not model-based either but has the disadvantage that the computation of cepstrum involves a logarithm operation. We now give some of the properties of the spectral root cepstrum.

3.2. *Properties of the spectral root cepstrum*

Let $\{x(n)\}$ be a real sequence and let $\{X(k)\}$ be its discrete Fourier transform.

1. Then $F^{-1}(|X(k)|) = \{\tilde{x}(n)\}$ is an even sequence.
2. Given that $\{x(n)\}$ is a sequence with finite support, from the Akhiezer–Krein and Fejer–Riesz theorems (Papoulis, 1977) it can be shown that

$$\begin{aligned} F^{-1}(|X(k)|^r) &= F^{-1}(|X(k)|^{0.5r}|X(k)|^{0.5r}), \\ &= F^{-1}\{Y(k)\}\{Y^c(k)\}, \\ &= \{y(n)\}*\{y(-n)\}, \end{aligned} \tag{7}$$

   where c and $*$ denote complex conjugation and convolution operations, respectively. Thus, $|X(k)|^r$ can be expressed as the Fourier transform of the autocorrelation function of some sequence $y(n)$.
3. Given that $|X(k)|^r$ is a positive even function and that $\{x(n)\}$ is a non-zero sequence, then $F^{-1}(|X(k)|^r)$ is maximum at the origin (Bracewell, 1986).
4. Minimum phase property: The above properties suggest that the truncated sequence $\tilde{x}_p(n)$ behaves like a minimum phase signal.

This is confirmed by numerically computing the roots of the z-transform of the sequence $\tilde{x}_p(n)$ for a number (≈ 50) of different frames of speech data. It was found without exception that in all the examples the roots (error $< 10^{-12}$) lie inside the unit circle. We have seen in our studies that the group delay functions derived from the magnitude and phase of the FT (Yegnanarayana et al., 1984) of $\tilde{x}_p(n)$ are identical.

From these empirical observations we conclude that the causal portion of $\tilde{x}(n)$ can be considered as a minimum phase sequence. The minimum phase condition ensures that the log magnitude and phase of such a signal are related through the Hilbert transform. Thus, the complete magnitude information is captured in the FT phase of $\tilde{x}_p(n)$. Because of the minimum phase characteristic, the cepstral coefficients and hence the weighted cepstrum can be derived recursively from $\tilde{x}_p(n)$. The group delay function can also be computed as the FT of the weighted cepstrum (Yegnanarayana et al., 1984). But in this paper we compute the group delay function through the spectrum using the discrete Fourier transform relation.

Note that use of exponent r on $|X(k)|$ does not alter the peaks in the smoothed envelope of the magnitude spectrum. Thus, the peak location information is preserved in the computation of the magnitude spectrum of the windowed spectral root cepstrum. It is generally not possible to model each of the peaks by a simple second order all-pole system (resonator), even though they

may correspond to a resonance peak in the original magnitude spectrum of $X(k)$. Therefore low order (10 to 18) linear prediction analysis of the signal $\tilde{x}_p(n)$ will not result in the desired peaks.

It is important to note that exponent r in $|X(k)|^r$ does not disturb the location of the pitch peak in $\tilde{x}(n)$. Therefore both the spectral peak locations and the value of the pitch period are not altered by the exponent factor. But the exponent factor helps to contain the significant information of the spectral envelope in a short window size p for $\tilde{x}(n)$. This helps in the choice of a p lower than that of the pitch period to avoid the influence of pitch on formant extraction.

4. Performance evaluation

In this section we demonstrate the effectiveness of the proposed group delay function approach for formant extraction as compared to the LP and cepstral approaches. In each case the raw formant data is obtained and the performance is judged by the visual inspection of the formant contours. In the group delay function approach for formant extraction there are a few parameters which decide the resolution and accuracy of the formant data that may be obtained. One of them is window size p. This is similar to the cepstral (or LP) smoothing technique in which the window size (or model order) chosen in the cepstral domain (or AR model) affects the resolution that can be achieved. In addition to window size p, exponent r also plays a significant role in the resolution that can be obtained. We now study the effects of varying these two parameters on the formant information obtained from the group delay function. To discuss the performance of our method we first consider synthetic speech data corresponding to the formant contours shown in Figure 5.

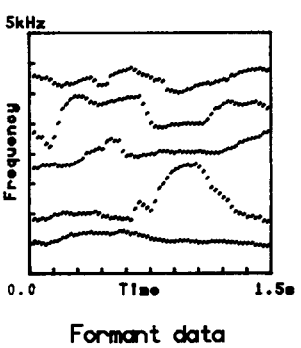


Fig. 5. Formant data for synthetic speech.

Model for the synthetic signal: The synthetic signal chosen is a voiced utterance generated by using a simplified model for speech production shown in Figure 6. The waveshape for the glottal pulse was chosen to be of the form (Rabiner and Schafer, 1978)

$$g(n) = \begin{cases} 0.5(1 - \cos(\pi n/N_1)), & 0 \le n \le N1, \\ \cos(\pi(n - N_1)/2N_2), & N_1 \le n \le N_1 + N_2, \\ 0, & \text{otherwise.} \end{cases}$$

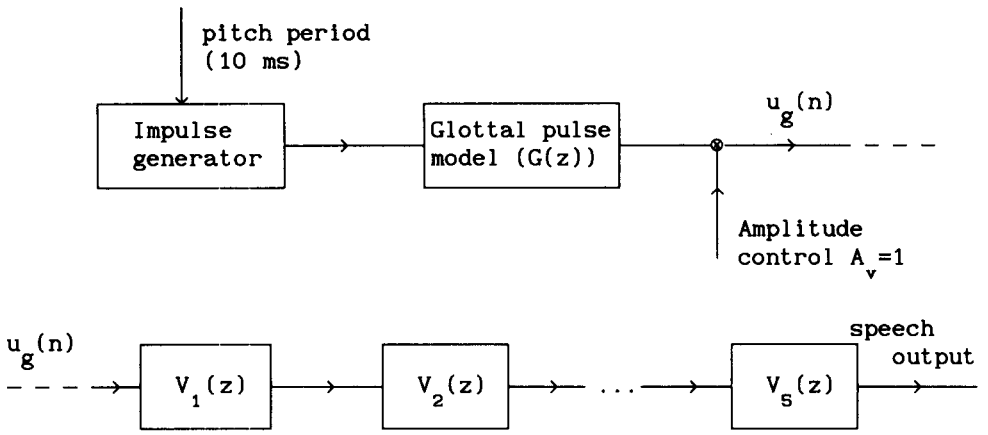
The transfer function for the vocal tract was modelled as

$$V(z) = \prod_{k=1}^{5} \frac{1 - 2e^{-\pi B_k T}\cos(2\pi F_k T) + e^{-2\pi B_k T}}{1 - 2e^{-\pi B_k T}\cos(2\pi F_k T)z^{-1} + e^{-2\pi B_k T}z^{-2}}.$$

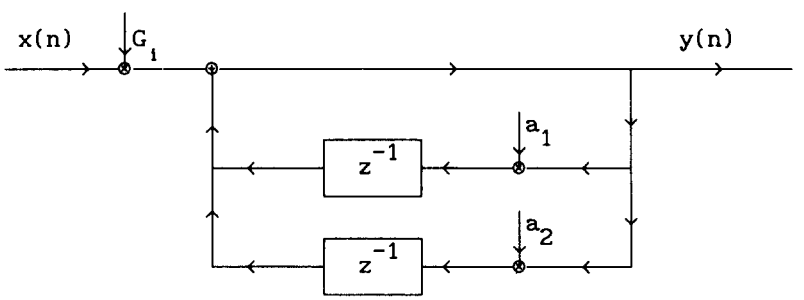
This equation describes a cascade of digital resonators that have unity gain at zero frequency. All the five formants (F_ks) vary continuously with time as defined by the format plot shown in Figure 5. T is fixed at 0.0001 sec (i.e. 10 kHz sampling rate). The formant bandwidths (B_ks) were fixed a priori at 10% of the formant frequencies.

A pitch period of 10 ms was chosen to generate the excitation signal. In the model for the glottal pulse $N_1 = 60$ and $N_2 = 10$. The formant data were designed so as to capture most of the situations encountered in practice (in the context of voiced speech), namely, proximity of formants, sudden rise in formants, sudden fall in formants.

Figure 7 shows the formant data obtained from the synthetic speech signal using the group delay (GD) approach for various window sizes (p = 5.0 ms to 8.0 ms). Figure 8 shows the formant data for the same synthetic data using LP analysis for various orders (10 to 22) and Figure 9 shows the formant data obtained using cepstrum analysis for the same window widths as used in the GD approach. Comparison of the data in Figure 7 with the raw formant data obtained from LP analysis (Figure 8) and the raw formant data obtained from cepstrum analysis (Figure 9) shows that our



where $V_i(z)$ is represented by the following filter:



$$\text{where } G_i = 1 - 2e^{-\pi B_i T}\cos(2\pi F_i T) + e^{-2\pi B_i T}$$

$$a_1 = -2e^{-\pi B_i T}\cos(2\pi F_i T)$$

$$a_2 = e^{-2\pi B_i T}$$

Fig. 6. Model for generating synthetic speech.

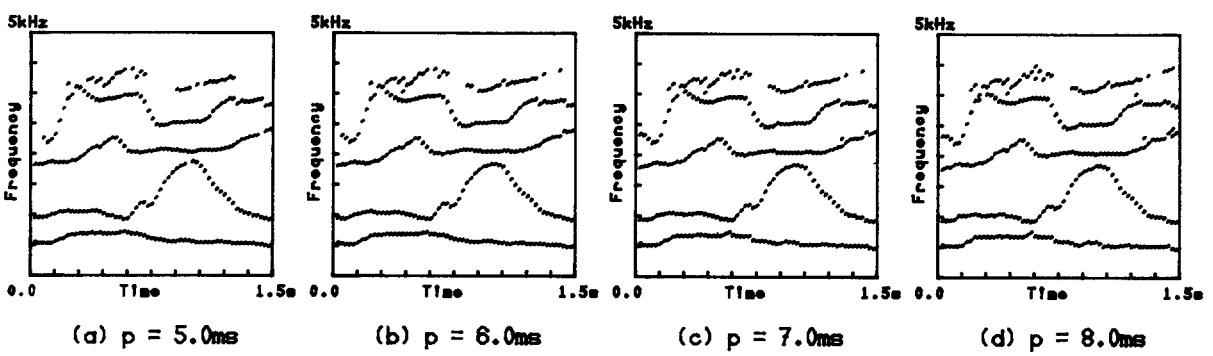


Fig. 7. Formant data obtained for synthetic speech using group delay approach. (a) p = 5.0ms, (b) p = 6.0ms, (c) p = 7.0ms and (d) p = 8.0ms.

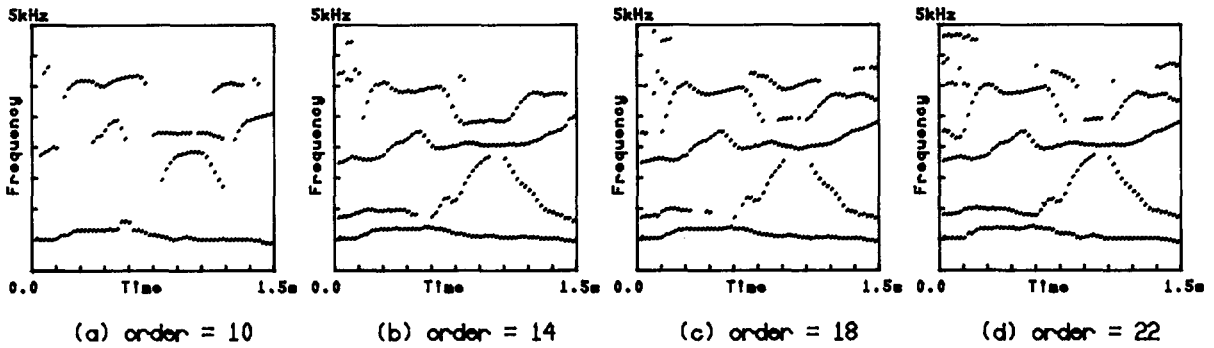


Fig. 8. Formant data obtained for synthetic speech using linear prediction analysis. (a) Order = 10, (b) order = 14, (c) order = 18 and (d) order = 22.

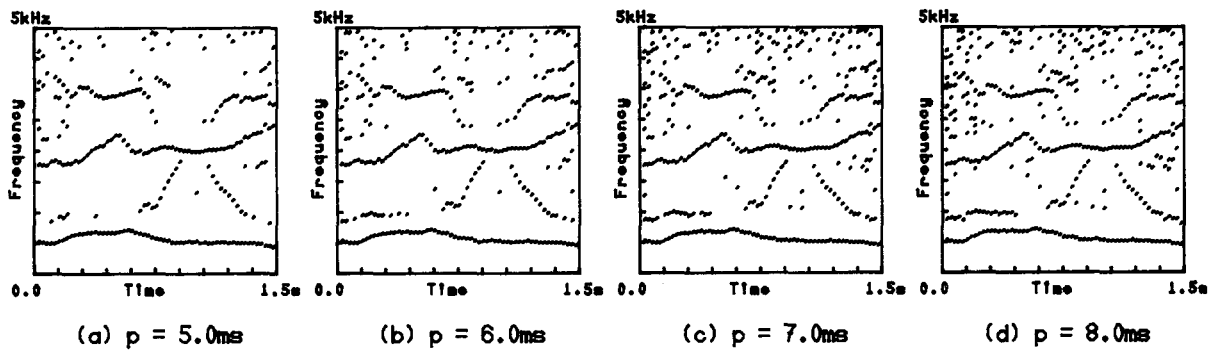


Fig. 9. Formant data obtained for synthetic speech using cepstrum analysis. (a) p = 5.0ms, (b) p = 6.0 ms, (c) p = 7.0ms and (d) p = 8.0ms.

method gives equally good but more consistent estimates of formants over a wide range of window widths for the synthetic data. It is to be noted that in the case of LP analysis (Figure 8) lower order may not bring out all the formant peaks. As the order is increased all the peaks get resolved but not many spurious peaks will be generated since the data is strictly the output of an all-pole system.

The effect of window size p on formant extraction is studied by obtaining plots of the raw formant for different values of the window in the spectral root cepstrum domain. The window size is varied uniformly from 5.0ms to 8.0ms in steps of 1.0ms, with $r = 0.5$. This is illustrated in Figure 7. Notice that an increase in window size results in an increase in the resolution of the peaks. The formant data is consistent over a sufficiently large range of window sizes (Figure 7(a–c)). But too large a window size (for example 8.0ms) causes spurious peaks to appear, as in Figure 7(d). The window size should be large enough to resolve the peaks that are close to each other but should not be too large to include the effects of pitch on formant extraction.

The choice of r (for a particular window size p) depends upon the dynamic range of the signal spectrum. The choice of r basically dictates the degree of overlap between the source and the vocal tract components in the root cepstrum domain. In the digital implementation of the SRDS algorithm the $(\,)^r$ and $(\,)^{1/r}$ operations require two phase unwrapping operations (Lim, 1979). In our approach we perform $(\,)^r$ only on the positive real function $|X(\omega)|$. Hence no phase unwrapping is necessary. Thus there is no constraint on r. We have experimentally observed that the choice of r is related to the spectral flatness. It appears log-

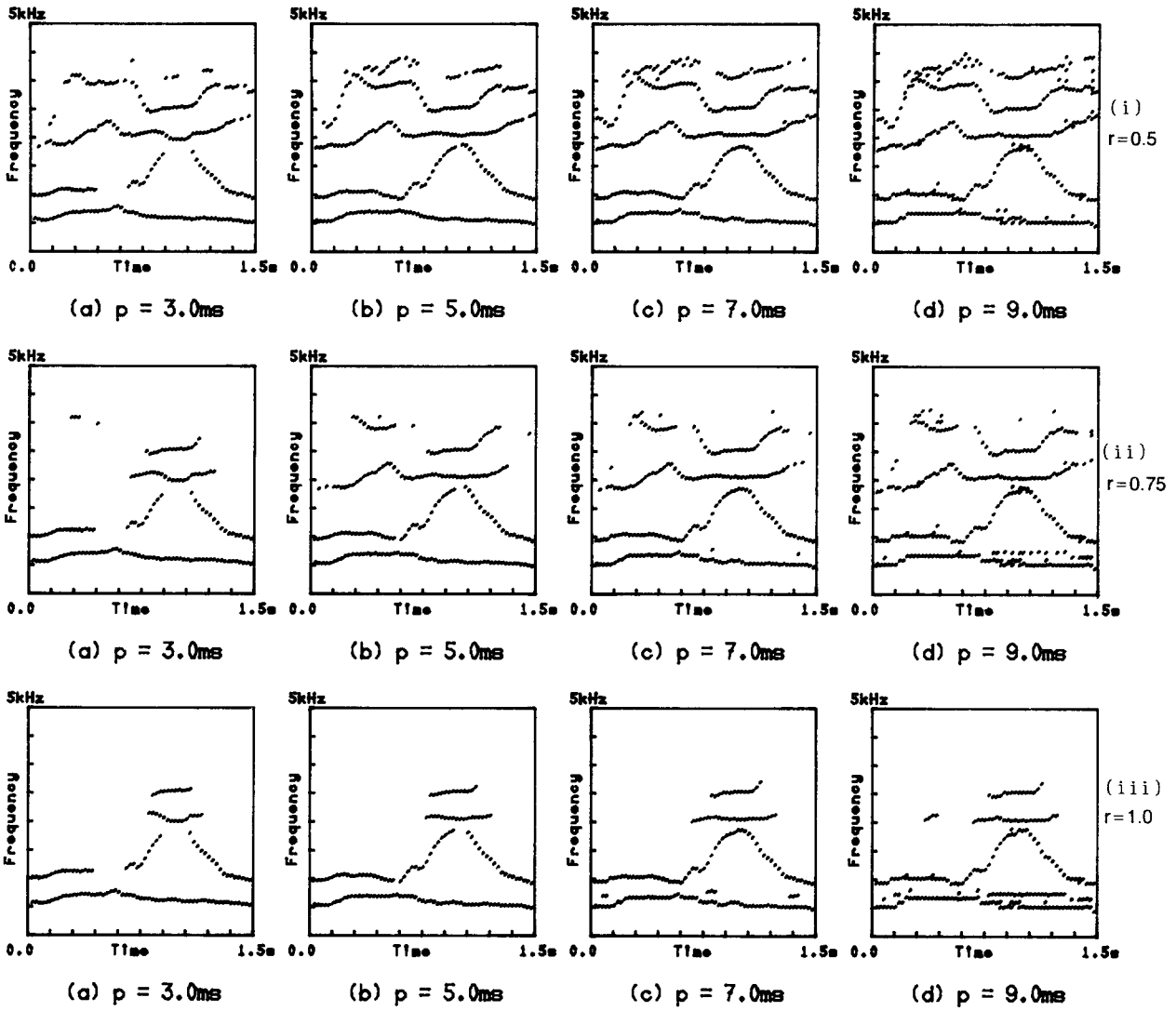


Fig. 10. Illustration of the GD formant extraction technique for various choices of p and r. (i) $r = 0.5$: (a) $p = 3.0$ms, (b) $p = 5.0$ms, (c) $p = 7.0$ms and (d) $p = 9.0$ms. (ii) $r = 0.75$: (a) $p = 3.0$ms, (b) $p = 5.0$ms, (c) $p = 7.0$ms and (d) $p = 9.0$ms. (iii) $r = 1.0$: (a) $p = 3.0$ms, (b) $p = 5.0$ms, (c) $p = 7.0$ms and (d) $p = 9.0$ms.

ical that when the dynamic range is low (as in the case of noisy speech) the peaks in the spectrum must be emphasized if they have to make a significant contribution to $\tilde{x}(n)$. This can be achieved by keeping $r > 1$. On the other hand, when the dynamic range is very large (as in the case of normal or high pitch voiced speech) the contribution by the first formant dominates the computation of $\tilde{x}(n)$. The effect of the first formant must be deemphasized. This is done by keeping $r < 1$. When $r < 1$ the vocal tract information is concentrated around the origin in $\tilde{x}(n)$ and the gating function enables a good separation of the source information from the vocal tract response. Figure 10 illustrates the performance tradeoff for various choices of r and window sizes. The effect of r can be visualized by traversing Figure 10 vertically from bottom to top along a direction corresponding to a fixed window width.

It is to be noted that parameter r and window

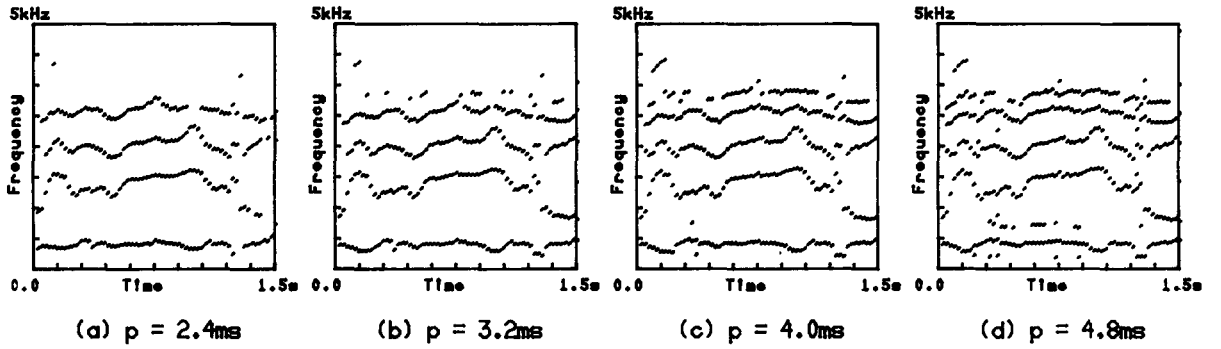


Fig. 11. Formant extraction from natural speech using GD approach. (a) p = 2.4ms, (b) p = 3.2ms, (c) p = 4.0ms and (d) p = 4.8ms.

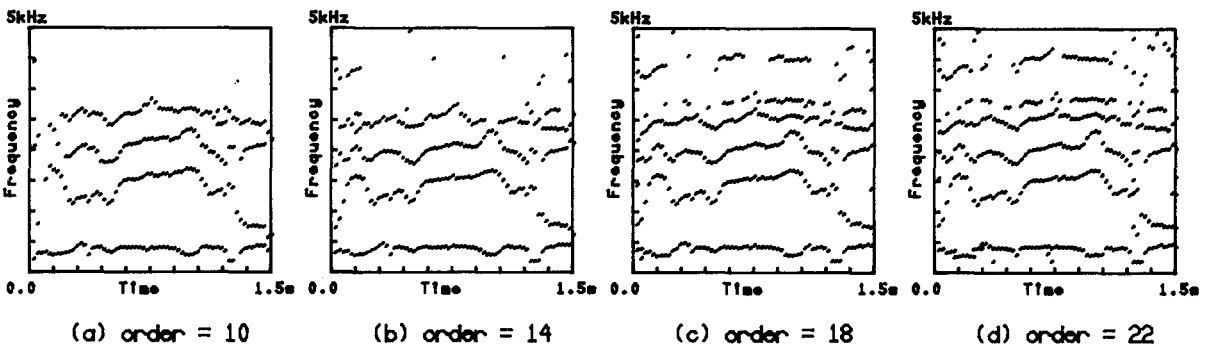


Fig. 12. Formant extraction from natural speech using LP approach. (a) order = 10, (b) order = 14, (c) order = 18 and (d) order = 22.

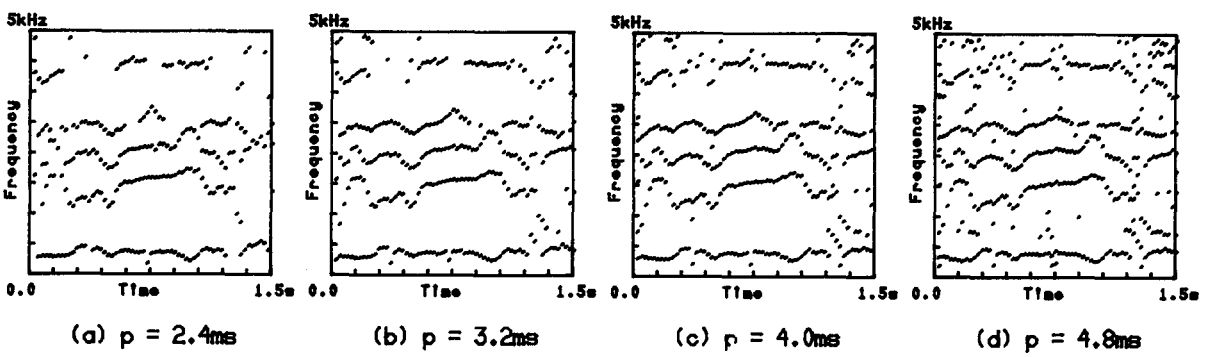


Fig. 13. Formant extraction from natural speech using cepstrum analysis. (a) p = 2.4ms, (b) p = 3.2ms, (c) p = 4.0ms and (d) p = 4.8ms.

size p are related. A smaller window size produces a poorer resolution, while a smaller r produces a better separation of source and excitation. As the window size p must be smaller than the pitch period to avoid fluctuations, r can be manipulated to obtain a good resolution of formants.

So far we have illustrated the use of this new technique for formant extraction on synthetic speech, where the synthetic speech has been modelled as the glottal excitation of a truly all-pole model.

Natural speech may not correspond to a truly

auto-regressive process of a fixed model order. We now compare the GD approach with that of LP analysis for formant extraction from natural speech. Figures 11, 12 and 13 show the formant data obtained using the GD approach, the LP approach and cepstrum analysis for the utterance "We were away a year ago", as spoken by a male speaker. The comparison confirms our earlier conclusions that the GD formant extraction technique gives more consistent formant values (for various window sizes) than that of the LP approach (for various orders) and cepstrum analysis (for various cepstral windows).

Figures 14 and 15 illustrate the formant contours for high pitched synthetic and natural speech data. Here, r is chosen to be 0.5. As long as the window size p is less than the pitch period, the proposed method works well even for high pitched speech. The synthetic speech was generated using the same procedure as indicated in Figure 6. The pitch period used for this case was 5ms. For natural speech the utterance is "We were away a year ago" as spoken by a female speaker. In Figure 15(a), in the region between 0.6–0.9s the GD method does not resolve the 2nd and 3rd formants as well as the LP method (Figure 15(b)) because the time window chosen is very small (2.4ms). The time window cannot be increased beyond 3.2ms as the average pitch period for this utterance is about 4ms. It is observed that the formants are steady for window sizes ranging from 1.6ms to 3.2ms. For the LP method in Figure 15(b) a carefully chosen order of 12 seems to be appropriate for this utterance. A lower order does not resolve the formants while a higher order generates a lot of spurious peaks. The cepstrum analysis (Figure 15(c)) generates spurious peaks especially at high frequencies, for

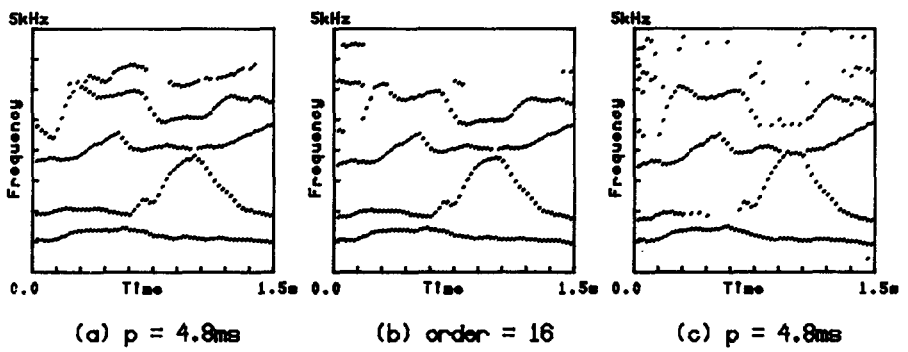


Fig. 14. Formant extraction from high pitched synthetic speech. (a) GD method (r = 0.5, p = 4.8ms), (b) LP method (order = 16) and (c) cepstrum analysis (p = 4.8ms).

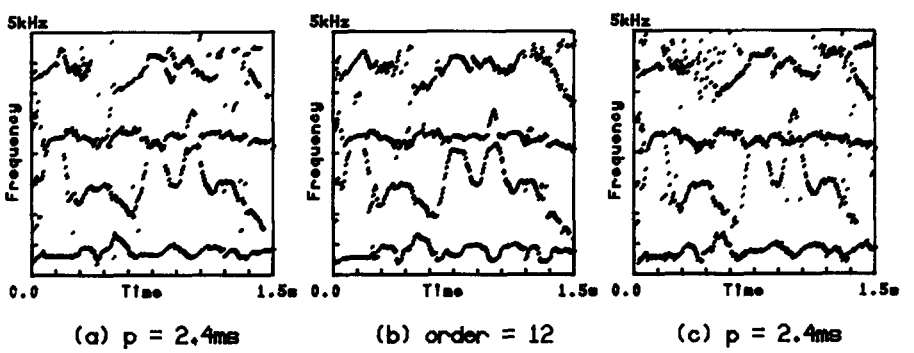


Fig. 15. Formant extraction from high pitched natural speech. (a) GD method (r = 0.5, p = 2.4ms), (b) LP method (order = 12) and (c) cepstrum analysis (p = 2.4ms). The data was collected for overlapping frames to capture the rapid fluctuations in the formant frequencies.

all window sizes.

Figure 16 shows the formant contours for an utterance in an Indian language Hindi "mai yah ca:hta: hu:n". The sentence contains segments of different categories of speech segements such as unvoiced, nasals and fricatives. For the unvoiced segments, the peak locations occur at random frequencies. For most of the voiced segments the formant frequencies are extracted well as seen from the continuity of the points.

We have also examined the performance of the proposed method for noisy speech data. Figure 17(c) shows the formant contours obtained for an utterance with an overall SNR = 10dB, using p = 3.2ms and r = 2. The variation of SNR for each frame is shown in Figure 17(b). The formant contour for the clean data is given in Figure 17(d). A comparison of Figures 17(c) and 17(d) shows that there are spurious peaks in those frames where the SNR is very low (< 0dB), while for all the other frames the formant peaks, even for the noisy data, are located at the appropriate frequencies. For some segments of the noisy data in the region 0.9s to 1.2s (in Figure 17(c)), the fourth formant is not extracted as neatly as that for the clean data. This is because for a given frame SNR is also a function of frequency. At high frequencies the SNR is usually lower than at low frequencies.

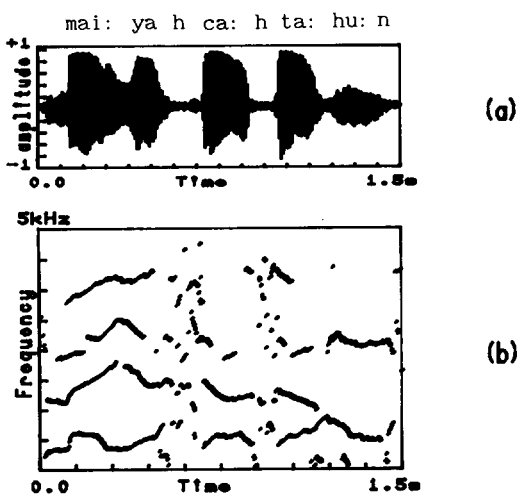


Fig. 16. Formant extraction for an utterance in an Indian language Hindi containing different categories of segments including nasals and unvoiced.

While the proposed method seems to work well for a wide variety of speech signals, computation time is significantly higher than the LP or cepstral methods. For the utterance "We were away a year ago" the computation time for the proposed method is 40sec, whereas it is 20sec for LP and 30sec for the cepstrum method.

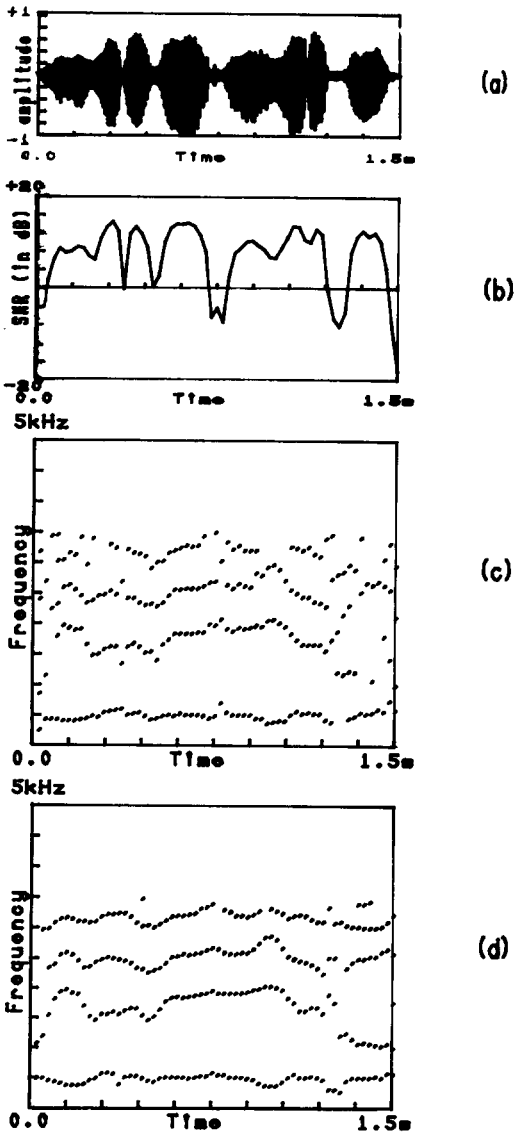


Fig. 17. Formant extraction from noisy speech. (a) Clean speech, (b) SNR as a function of time, (c) formant data for noisy speech, (d) formant data for clean speech.

4. Conclusions

We propose a new method of extracting formant information from the speech signal. We demonstrate that the additive and high resolution properties of the group delay functions can be used for extracting closely spaced and low amplitude formant information. This method for formant extraction gives a more consistent performance compared to other methods based upon smoothing the magnitude spectrum. These studies show that there is a relationship between spectral flatness and analysis parameters. We are currently attempting to establish a theoretical basis for this relation.

References

R.N. Bracewell (1986), *The Fourier Transform and its Applications* (McGraw-Hill, New York), pp. 6–23.

J.S. Lim (1979), "Spectral root homomorphic deconvolution system", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, No. 3, pp. 223–231.

J. Makhoul (1975), "Linear prediction: A tutorial review", *Proc. IEEE*, Vol. 63, No. 2, pp. 561–580.

A. Papoulis (1977), *Signal Analysis* (McGraw-Hill, New York), pp. 231–234.

L.R. Rabiner and R.W. Schafer (1978), *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ), pp. 367–370.

J.M. Tribolet (1977), "A new phase unwrapping algorithm", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-25, pp. 170–175.

B. Yegnanarayana (1978), "Formant extraction from linear prediction phase spectrum", *J. Acoust. Soc. Amer.*, Vol. 63, pp. 1638–1640.

B. Yegnanarayana, D.K. Saikia and T.R. Krishnan (1984), "Significance of group delay functions in signal reconstruction from spectral magnitude or phase", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-32, No. 3, pp. 610–623.

B. Yegnanarayana, G. Duncan and H.A. Murthy (1988), "Improving formant extraction from speech using minimum phase group delay spectra", in *Signal Processing IV: Theories and Applications*, ed. by J.L. Lacoume, A. Chehikian, N. Martin and J. Malbos (Elsevier, Amsterdam), pp. 447–450.