SOME ISSUES IN PROCESSING SHORT SEGMENTS OF SPEECH SIGNALS

A THESIS

Submitted for the award of the Degree

of

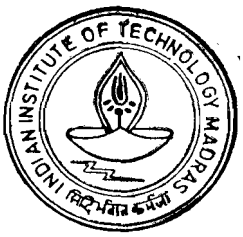
DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

by

K. V. MADHU MURTHY

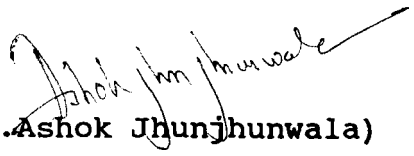


DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS-600 036
INDIA

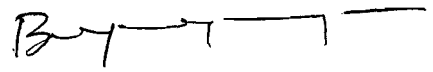
OCTOBER 1990

CERTIFICATE

This is to certify that the thesis entitled **SOME ISSUES IN PROCESSING SHORT SEGMENTS OF SPEECH SIGNALS** is the bonafide work of **Mr.K.V.Madhu** Murthy, carried out under our guidance and supervision, in the Department of Electrical **Engineering, Indian** Institute of Technology, Madras, for the award of the degree of Doctor of Philosophy in Electrical Engineering.



(Dr.Ashok Jhunjhunwala)
Professor
Department of Electrical Engineering
IIT, MADRAS



(Dr.B.Yegnanarayana)
Professor
Department of Computer Science and Engineering
IIT, MADRAS

ACKNOWLEDGMENTS

I thank my research guides Prof.B.Yegnanarayana and Prof.Ashok Jhunjhunwala for the constant encouragement and support given to me throughout the course of this research work. Prof.B.Yegnanarayana introduced me to the fascinating field of speech signal processing. I cherish many technical discussions I had with him for the past several years which helped me to get an insight into the basic issues involved in speech research. Prof.Ashok Jhunjhunwala helped me with technical and nontechnical advice whenever I was stuck up with a problem. I gratefully acknowledge the time and efforts that my research guides have spent on me.

I acknowledge the cooperation extended by all members of the Speech and Vision Laboratory, Department of Computer Science and Engineering, IIT, Madras, with whom I have been closely interacting during all stages of my research work. I especially acknowledge the many useful discussions I had with Ms.Hema A. Murthy, Mr.C.P.Mariadassou and other members of this group. These discussions helped me to understand better some of the subtle research problems.

I thank my colleagues Mr.K.G.Uday Kumar, Mr.C.Chandra Sekhar, Mr.P.Eswar and Mr.S.Rajendran for helping me in the preparation of the manuscript and the diagrams of this thesis.

I acknowledge the interesting discussions I had with Prof.V.V.Rao, Dr.K.M.M.Prabhu, Dr.G.R.Reddy, Mr.G.V.Ramana Rao, Mr.M.Prakash, Mr.R.Sundar, Mr.N.Alwar and Mr. S.R. Rajesh Kumar.

I have enjoyed the company of Mr.C.Eswara Reddy, Dr.A.T.V.Ramesh, Dr.R.Sai Prasad, Mr.T.Sudhakar Babu and Mr.P.C.Majhee with whom I used to have many technical and nontechnical interactions.

I acknowledge the encouragement and support given by Prof.T.Rangaswamy, Principal and Dr.M.M.Naidu, Head of

the Department of Computer Applications, of S.V.U.College of Engineering, Tirupati, where I am presently teaching. I also wish to thank my colleagues Ms.O.Nalini Kumari and Mr.P.Anjaneyulu for extending their cooperation and for sharing part of my academic workload for several months, to enable me to complete my thesis work faster.

I thank my parents and sister for having provided a congenial atmosphere at home to pursue my academic and research interests.

CONTENTS

Appendix

ABSTRACT

Practical signals such as speech are nonstationary in nature. Processing such signals to capture their dynamic variations requires the use of short data records for analysis. This leads to time-frequency resolution problems when standard processing algorithms such as discrete Fourier transform (DFT) are used in the analysis. The aim of our research is (i) to explore methods for analyzing signals like speech to extract the time varying parameters suitable for applications in recognition and synthesis; (ii) to study various issues connected with the time-frequency resolution problem such as type of signal representations and processing algorithms; and (iii) to propose alternate strategies to circumvent the time-frequency resolution problem while analyzing natural signals under both quasistationary and time varying assumptions.

The major contributions of this thesis are:

(1) A modified linear predication method to analyze speech under time varying formulation;

(2) A weighted linear predication method to process noisy speech under quasistationarity assumption;

(3) A study of the effectiveness of signal representation using group delay functions;

(4) Detection of the source-tract interaction in speech-like signals;

(5) A study of time-frequency resolution in composite signal decomposition.

Chapter 1

OVERVIEW OF THE THESIS

1.1. Introduction to the Problem of Digital Processing of Speech Signals

This thesis addresses issues relating to digital processing of analog signals like speech to extract useful information such as temporal and spectral features. This is useful in applications such as in low bit rate coding, speech synthesis and recognition. In particular, we consider the issue of the time-frequency resolution problem in speech during analysis of short data (less than a pitch period) to represent the effects of the source-tract interaction. We also consider the time-frequency resolution problem in the analysis of quasistationary (1 to 3 pitch periods) segments to extract formant information and pitch epochs, especially when the signal is noisy.

Given a linear system and an excitation signal, it is easy to compute the output signal using convolution. But in practice, the problem is to process the available output signal to extract the characteristics of the system and excitation, especially when they are varying with time. Even if we have some idea of the signal generating mechanism to make some valid assumptions on the model like autoregressive (AR), autoregressive and moving average (ARMA) etc., it is difficult to extract the system characteristics due to

nonstationary nature of the signal. The problem is compounded by the fact that the signal is usually processed in digital domain. Discretization and truncation are still the major sources of error in digital processing of speech-like signals. Due to this reason, it is difficult to extract important temporal and spectral features. This makes the tasks of reliable coding and quality synthesis difficult to accomplish.

In digital signal processing we are generally interested in extracting various features of the signals in time and frequency domains. For example, in the case of speech signals, time domain features of interest are pitch period, boundaries of silence/voiced/unvoiced segments, and other epochs characterizing changes either in vocal tract or excitation. Similarly, in the frequency domain, we are interested to find the pitch frequency, formant frequencies and bandwidths of the voiced sounds and the distribution of signal energy in different frequency bands for unvoiced sounds.

Since most physical signals are a result of excitation of inertial systems, the information pertaining to the system behavior within a short duration spreads over a wider time interval in the signal. Hence it is generally necessary to consider a long segment of data to process such signals for extracting information about the system as well

as the excitation source. This is possible only when the system is stationary.

But in practice, signals like speech are generated from time varying systems. It is not possible to determine the behavior of the system at every instant using only a few samples around the instant of interest. This problem of information extraction is compounded by the inevitable noise added to the signal before the signal is captured. Small segments of noisy signals do not capture the noise statistics adequately to smooth out the effects of noise. But surprisingly speech signals are perceived in the analog domain by the auditory mechanism even in the presence of noise. This suggests that there is scope for developing analysis techniques which enable effective processing of nonstationary signals like speech to extract information about both the system and the source.

1.2. Issues in Processing Speech Signals

In this thesis we explore methods for analyzing time varying signals in order to extract parameters suitable for application to speech analysis and synthesis. At first we discuss the conventional methods of signal processing and highlight the inadequacy of these methods for extracting the embedded time varying spectral information from the signal.

A crucial factor in signal processing is the issue

of representation of signal information. Digital computer provides flexibility to implement several linear and nonlinear processing techniques. For processing a signal using a digital computer, it is necessary to discretize the signal. Further, a number of processing techniques like the Fourier transform (FT) require that the signal data be processed as finite length blocks. The window effects, forced by the block processing, dominate in the case of analysis of short time signals and affect resolution in the frequency domain. Similarly, windowing in the frequency domain affects resolution in the time domain. Hence, the Fourier transform representation of signal information, though by far the most popular representation, has its own limitations.

Thus discretization and truncation operations affect the time and frequency resolutions of the signal. The severity of these limitations may be different for different signal representations. In this context, we discuss various representations and show how group delay functions (the negative derivative of the FT phase) are useful in some situations. These functions have certain desirable qualities, such as high resolution and additive properties. We show how the time and frequency resolution properties of signals, represented through the group delay functions, are affected by the discretization and truncation operations. We

then discuss the adequacy of representation of signal information through the group delay functions, and highlight the likely errors in representation. Note that group delay functions may be obtained from the Fourier transform phase and magnitude. In such a case, these functions are also subjected to the resolution limitations imposed by the block processing of data.

After establishing the dependence of time and frequency resolution properties of signals on the type of representation and processing algorithms, we plan to study these issues in extracting information from signals. In order to focus on the effects of truncation and discretization in the representations like Fourier transform and group delay functions, we have considered a signal which is a model of composite signal. A composite signal consists of a summation of basic wavelets and their scaled and shifted replicas. Model signals are chosen as they enable us to control at will the features of the basic wavelets and their arrival times. Only stationary signals are considered in this study. The primary purpose is to study the resolution obtainable in the time and frequency domains by finding out the number of basic wavelets corresponding to resonators and their arrival times (epochs). This requires identification of resonances and epochs. If the frequencies of the resonances are close, then the frequency resolution

is a problem. If the epochs are close in time domain, then the time resolution is a problem. The above problems are compounded by the presence of additive noise in the signal. In the noise free case, the effects of discretization and truncation set limits on the resolution. Even within these limits, the problem is complicated because we do not have effective methods to resolve overlapping waveforms and epochs. Hence there is a need for new techniques for resonance extraction and epoch identification from a composite signal.

The stationary composite signal model does not approximate speech signals as the vocal tract system is continuously changing with time during speech production. Hence, only short segments of speech signals are normally considered for processing. Even within a pitch period of a steady voiced sound, the system changes continuously due to loading of the vocal tract by the subglottal system. This effect is called source–tract interaction. One consequence of this effect in the speech signal is that formants may have different bandwidths at different time instants even within a pitch period. Detection of such changes requires careful analysis. We model this situation through a resonant system whose bandwidth changes within a pitch period. The effect of this change can be detected through group delay functions. We outline an inverse filter method to detect the

source-tract interaction in the case of simple model signals. However, there is no convenient signal processing method to measure this effect in more complicated signals such as natural speech.

In all the above cases the techniques developed for processing signals in which information over short intervals is of interest, are applied to model signals. We develop techniques to process real speech data also. Techniques for speech processing under nonstationary assumption are explored. A new approach called modified LP method is proposed for processing short (less than 2 msec) segments of speech data to capture variations of the vocal tract system characteristics even within a pitch period. Here we propose a technique to obtain better estimates of the autocorrelation parameters of a short data record. This technique is based on modifying the autocorrelation parameters obtained over a longer analysis frame that includes the short segment. The modification is performed using parameters extracted from the LP residual over the short segment. The modified autocorrelation parameters capture the characteristics of the speech signal over the short segment of interest.

Speech signals are processed to extract formant information and pitch epochs assuming quasistationarity over the analysis intervals (10-20 msec). We propose a new method

for processing noisy speech signals under quasistationary assumption. Here we use the principle that the effect of noise can be reduced if, in the signal processing algorithms, more weightage is given to those portions of the signal in time domain where the signal-to-noise ratio (SNR) is high. Based on the above principle, a weighted linear prediction algorithm is developed in which a weight function based on the high SNR criterion is incorporated in the standard linear prediction (LP) analysis procedure so as to obtain LP parameters that are less affected by noise. This is a noniterative technique.

To demonstrate the effects of processing speech using the techniques proposed in this thesis, we have developed a synthesis package. The following studies are made:

(1) Performance of the weighted LP method for analysis and synthesis using noisy speech for different noise levels.

(2) Comparison of the modified LP method with the standard LP method for synthesizing more natural sounding speech.

1.3. Organization of the Thesis

We have identified and investigated several issues connected with processing of short segments of signals. The

organization of the thesis is as follows:

First we introduce the issues involved in digital signal processing and the limitations associated with it. The issues studied are: (1) The interrelationships among various types of representations; (2) Types of processing algorithms; (3) Requirement of discretization and truncation operations; (4) Limitations of time-frequency resolution.

Next we introduce group delay functions as an interesting representation for signals and list their properties. We study the issue of signal representation using group delay functions with reference to the effects of discretization and truncation and conclude that the signal information can be adequately represented using group delay functions under certain conditions.

After studying the issues of signal representation, we investigate the problems imposed by discretization and truncation for extracting information from signals under stationary assumption. The main issue here is the resolution in time and frequency domains. In this context, we study the problem of composite signal decomposition using model signals. We demonstrate that better resolution in time and frequency domains can be obtained by using group delay processing for composite signal decomposition.

Next we study techniques to analyze model signals

under nonstationary assumption. Specifically, we study the problem of detecting source-tract interaction, when the bandwidths of the resonances of the vocal tract system change within a pitch period. We discuss results of our experiments with simple model signals.

We explore methods to capture the effects of the source-tract interaction from natural speech signals. Speech signals are analyzed under nonstationary formulation using a modified LP method, where a two-pass LP analysis is used to capture both global and local variations in the system. We synthesize speech using LPCs obtained from the modified LP method and study the effect of changing the width of the short data record on the quality of synthetic speech.

After studying the time-frequency resolution issues in clean speech signals, we take up analysis of noisy speech. We develop a weighted LP method where high SNR regions are emphasized to process noisy speech under stationary assumption. We experiment different weighting functions and noise levels to study the effectiveness of the method in improving the quality of synthetic speech.

At the end, we summarize the results of the investigations done in this thesis. A brief discussion on the major contributions and some comments on the scope for future work are also given. We describe a speech analysis-synthesis system based on the LP model in the Appendix.

Chapter 2

REVIEW OF ISSUES IN DIGITAL PROCESSING OF SPEECH-LIKE SIGNALS

In this chapter we briefly review the issues in digital signal processing which have given rise to the investigations presented in this thesis. First we show that digital processing of signals requires discretization and truncation operations. These operations result in loss of some signal information and this loss depends on the representation used. We introduce group delay representation and discuss its properties. We review the time-frequency resolution problem in the context of stationary composite signals where the issue of representation plays a major part. Next we consider the problem of source-tract interaction in speech. We review some analysis methods that prove the importance of source-tract interaction. Speech analysis under time varying formulation is more appropriate to capture the effects of source-tract interaction. We review important methods of speech analysis under this category. We consider the problem of noisy speech processing within the limitations of time-frequency resolution. We review some standard techniques for speech enhancement. Finally we review LP based speech analysis-synthesis methods.

With the advent of high speed digital computers,

processing of signals in the digital domain has become popular. Digital processing is concerned both with obtaining discrete representation of signals and with the theory, design and implementation of numerical procedures for processing in this representation [Oppenheim and Schafer 1989]. Digital signal processing techniques are advantageous compared to their analog counterparts due to the flexibility they offer in a variety of problem situations. In addition, the availability of integrated circuits technology to develop dedicated signal processors of increasing sophistication promise economical implementation of complex digital processing algorithms [McCanny and White 1987].

2.1. Limitations in Digital Processing of Signals

Most of the practical signals are analog in nature. In general, they contain all frequencies and may be infinite in extent. Digital processing of such signals requires operations of discretization and quantization. Once the analog data is represented adequately in digital form, the type of digital signal processing algorithms available for processing this data assumes importance. Starting from mid 1960's there has been tremendous activity in this area. In 1965, Cooley and Tukey published their famous algorithm for computation of the discrete Fourier transform (DFT) called fast Fourier transform (FFT) [Cooley and Tukey 1965]. Subsequently, a number of algorithms such as linear

prediction, harmonic decomposition, adaptive signal processing, array processing etc., [Kay and Marple 1981] were proposed for model based spectral estimation. These algorithms depend on block processing of the digital data. Block processing is also necessary when the signal is not available as a continuous stream.

There are a number of algorithms to analyze linear time invariant (LTI) systems in literature. Most of these algorithms involve solutions of linear equations which are mathematically tractable compared to systems where nonlinear equations are encountered. Thus stationary signals and systems, which can be cast in LTI form, are easier to handle. Slowly varying (nonstationary) signals such as speech are usually approximated to stationary signals over short intervals of time. To facilitate block processing, the signal has to be truncated into finite length segments. The effect of truncation is same as multiplying the signal samples with a window function.

The operations of discretization, quantization and truncation introduce signal distortion that depends on representation as well as the algorithms chosen for processing the signal. Poor sampling leads to aliasing problems and finite frame width gives rise to resolution problems. In this thesis, we confine ourselves to the issues connected with sampling and truncation and do not consider the problems arising out of quantization.

By far the most frequently used algorithm in the spectral analysis of stationary signals is the fast Fourier transform (FFT) algorithm which computes the discrete Fourier transform (DFT). The DFT relations may be graphically derived from the theory of continuous Fourier transform [Brigham 1974] as follows: Consider the continuous Fourier transform pair of an example function h(t) as illustrated in Fig. 2.1a. To discretize h(t), we multiply it by a sampling function $\Delta_o(t)$ with a period T as illustrated in Fig. 2.1b. The sampled function $h(t)\Delta_o(t)$ and its Fourier transform are shown in Fig. 2.1c. It can be observed that the operation of sampling or discretization results in aliasing in the spectrum. If the function h(t) is not band-limited, that is, $H(f) \neq 0$ for some $|f| > f_c$ $(=1/2T)$, then sampling will introduce aliasing. T has to be made smaller to reduce this error,

The Fourier transform (FT) pair in Fig.2.1c is not suitable for machine computation because an infinity of the samples of h(t) is considered, Hence it is necessary to truncate $h(t)\Delta_o(t)$ so that only a finite number of points are considered. The multiplication of the sampled function $h(t)\Delta_o(t)$ with a truncation function of width T_o (as shown in Fig. 2.1d) results in a finite duration signal shown in Fig. 2.1e. The effect of this operation in the frequency domain is convolution of the aliased Fourier transform of Fig. 2.1c with the FT of the truncation function shown in

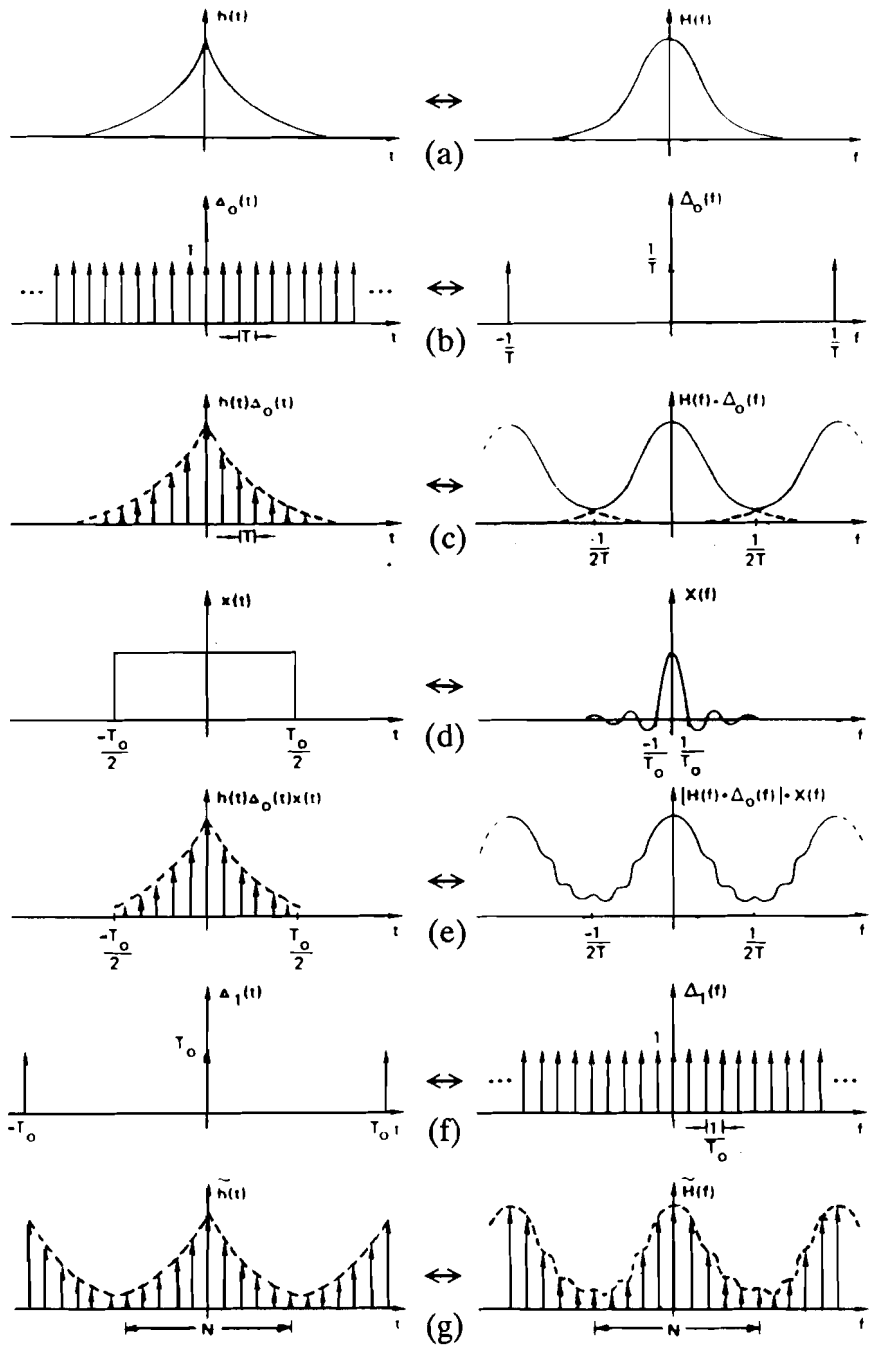


Fig. 2.1. Graphical development of the discrete Fourier transform (reproduced from [Brigham 19741).

Fig. 2.1d. The resulting FT has a ripple in it as shown in Fig. 2.1e. This effect has been amplified in the illustration for emphasis. This ripple causes spectral energy leakage into adjacent frequency bands. It can be shown that the strength of this ripple is inversely proportional to the width of the truncation function in the time domain.

The modified transform pair of Fig. 2.1e is still not suitable to handle digital signals as the frequency function is still continuous. To discretize the frequency function, we multiply it by a frequency sampling function with sampling interval $1/T_o$ shown in Fig. 2.1f. (Here we note that the sampling interval is to be chosen either equal to or less than $1/T_o$ to avoid aliasing of the block data in the time domain.)

The discrete FT pair of Fig. 2.1g is acceptable for the purpose of digital computation since both the time and frequency domain representations are discrete and finite length. But this is achieved at the cost of resolution in both time and frequency domains.

2.2. Issues of Representation: Group Delay Functions

Signal information can be represented in various domains. These are time domain, Fourier transform, z-transform and cepstrum [Oppenheim and Schafer 1989]. Each of these representations offers certain advantages in terms

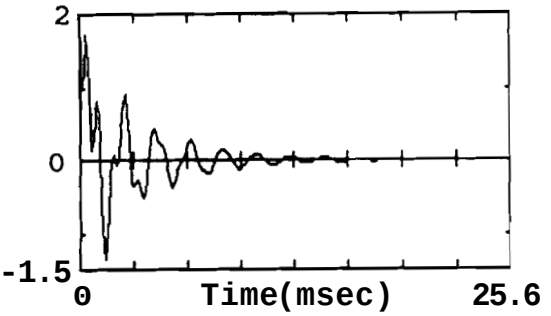
of processing of the signals for analysis and synthesis. However, they have their limitations too. We describe below some of the important representations, their advantages and limitations. These are also illustrated in Fig. 2.2.

2.2.1. Description of various representations

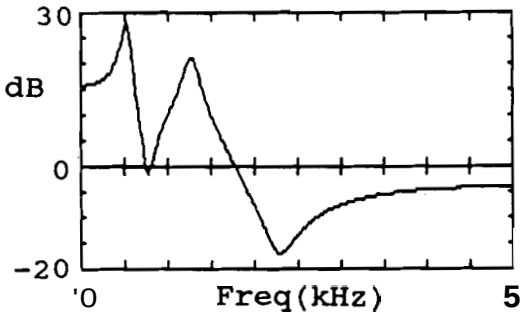
Time domain: Continuous time domain signal representation is of limited practical significance, as only analog processing is possible. Band-limited signals can be represented by the signal samples obtained at discrete intervals of time. The major advantage in this representation is that digital processing is possible. A number of analysis algorithms to find quantities such as short time energy profile, zero crossings, epochs, correlations are available in this domain. But this representation is not suitable if it is required to process signal information in different frequency bands.

Fourier transform domain: We obtain the signal in discrete Fourier transform domain by applying discrete Fourier transform to a windowed discrete time signal. Here the signal information is represented as a summation of a finite sequence of scaled complex frequency components. The accuracy of this representation is limited by the operations of discretization and truncation in time and frequency domains.

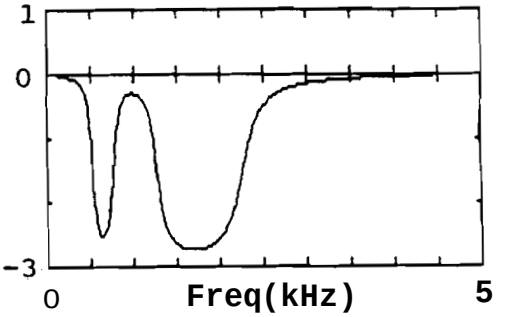
e-transform domain: In general, a discrete signal may be



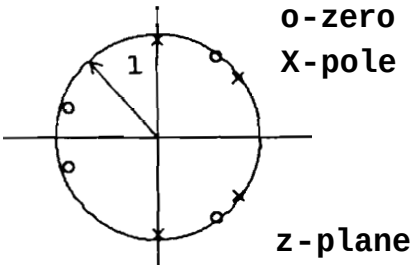
(a)



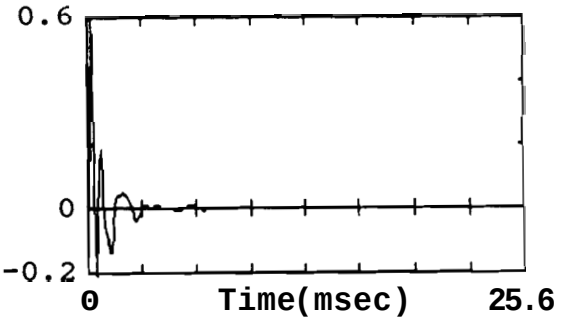
(b)



(c)



(d)



(e)

Fig. 2.2. Signal representation in various domains: (a) Time domain; (b) FT magnitude; (c) FT phase; (d) z-transform; (e) cepstrum.

represented in terms of its z-transform, as a polynomial in powers of z. The roots of this polynomial characterize the signal information.

Cepstrum: The inverse Fourier transform of log-magnitude spectrum is called cepstrum. The complex cepstrum is computed as the inverse Fourier transform of logarithm of the complex Fourier transform of the original signal. The advantage in the representation of the signal information in cepstral domain is that it is possible to easily separate the slowly varying and rapidly varying spectral features of the signal. The limitation is that when the roots of the signal are close to the unit circle in the z-plane, aliasing takes place in the cepstrum [Childers, Skinner and Kemeriat 1977]. Computation of complex cepstrum involves phase unwrapping in the Fourier transform domain for which an effective algorithm is not yet available.

We have listed some commonly used signal representations with their advantages and limitations. Recently considerable interest is shown in the FT phase function. New signal processing algorithms are proposed to exploit the desirable properties of FT phase for various signal processing applications [Hayes, Lim, and Oppenheim 1980] [Fathima 1986] [Yegnanarayana, Fathima and Hema A. Murthy 1987] [Yegnanarayana, Madhu Murthy and Hema A. Murthy 1987] [Hema A. Murthy et al. 1988] [Hema A. Murthy, Madhu Murthy and Yegnanarayana 1989A] . Now we introduce

another representation of signals in the form of group delay functions [Yegnanarayana 1984], [Yegnanarayana, Madhu Murthy and Hema A. Murthy 1988]. The group delay functions are closely related to the FT phase and inherit some of its interesting properties.

2.2.2. Representation of signals through group delay functions

In this section we briefly review the definitions and some important properties of the group delay functions. Algorithms for computing the group delay functions, as well as algorithms for deriving the signal from the group delay functions are given in [Yegnanarayana, Saikia and Krishnan 19841.

For a discrete-time signal $\{x(n)\}$, we define the group delay functions as follows: Let

$$X(\omega) = |X(\omega)|\, e^{j\theta(\omega)} \qquad (2.1)$$

be the Fourier transform of the signal $\{x(n)\}$. Then the group delay function is defined as the negative derivative of the unwrapped phase function. That is

$$\tau_p(\omega) = -d\theta_u(\omega)/d\omega \qquad (2.2)$$

where $\theta_u(\omega)$ is the phase function in unwrapped form. We call this the group delay function derived from the FT phase. Similarly we define a group delay function $\tau_m(\omega)$ derived from the FT magnitude function $|X(\omega)|$. It can be shown that $\tau_m(\omega)$ is the negative derivative of the phase of the unique

minimum phase equivalent signal derived from $|X(\omega)|$.

We give a summary of the important properties of $\tau_p(\omega)$ and $\tau_m(\omega)$ which were discussed in detail in [Yegnanarayana, Saikia and Krishnan 1984].

(i) For a minimum phase signal $\tau_p(\omega) = \tau_m(\omega)$

(ii) For a maximum phase signal $\tau_p(\omega) = -\tau_m(\omega)$

(iii) For a mixed phase signal $|\tau_p(u)| \neq |\tau_m(u)|$

(iv) Additive property: Convolution of signals in the time domain is reflected as summation of their respective group delay functions in the frequency domain, as shown in Fig. 2.3.

(v) High resolution Property: The resonances or anti-resonances (due to complex conjugate pairs of poles or zeros) of a signal are better resolved in the group delay function than in the spectral magnitude function [Yegnanarayana 1978]. Further, the signal information is confined to the narrow regions around the pole 'or zero locations, as shown in Fig. 2.4.

We elaborate below what we mean by high resolution property of the group delay functions. As DFT is used in the computation of the group delay functions they also suffer from the effects of the finite window size. We note that in the FT magnitude function, the contributions due to individual resonances are multiplicative in nature. Hence a

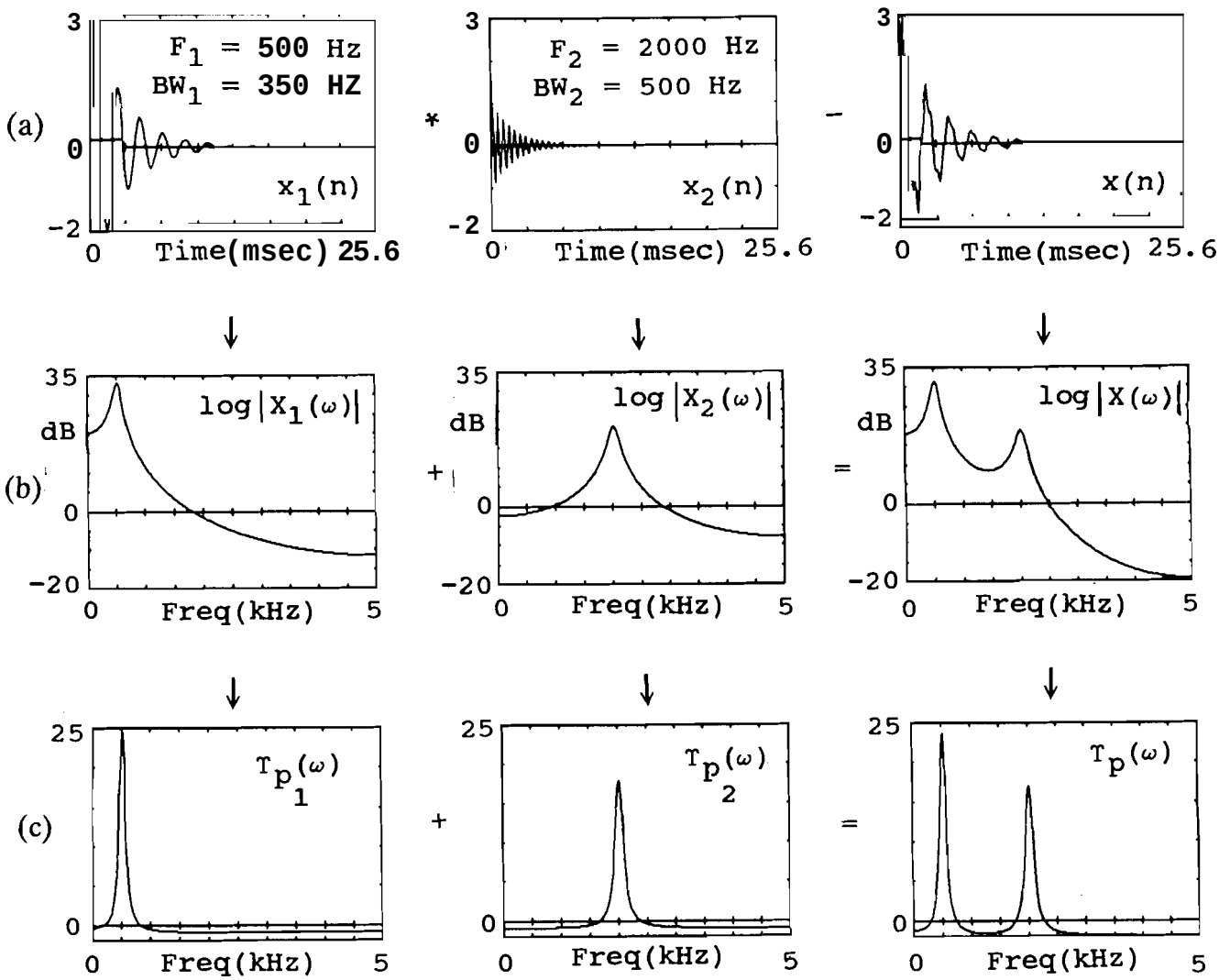
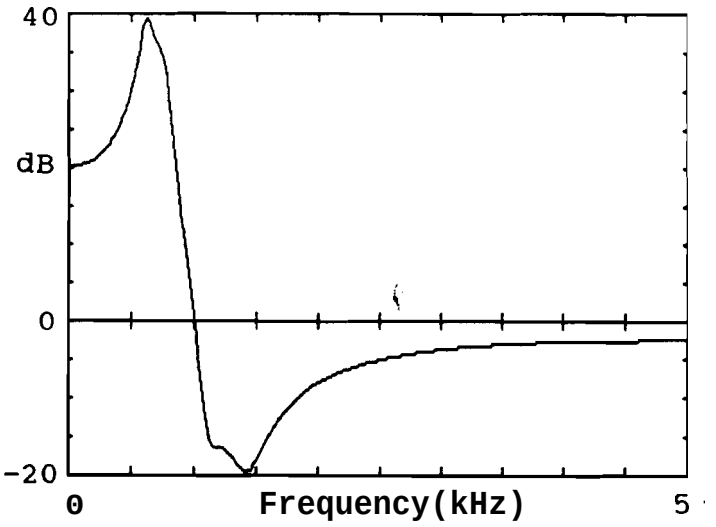
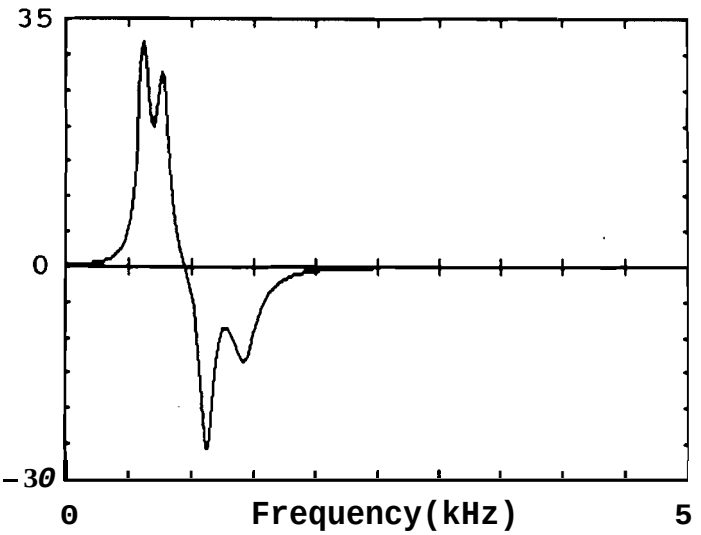


Fig. 2.3. Illustration of the additive property of the group delay functions: (a) Time domain signals; (b) the corresponding FT magnitude spectra, and (c) group delay functions (τ_p).



(a)



(b)

Fig. 2.4. Illustration of the high resolution property of the group delay functions: (a) FT magnitude spectrum showing two poles and two zeros; (b) the corresponding group delay function (τ_p).

peak due to a weak resonance may get completely masked when it is multiplied by small values in the decay region of an adjacent peak due to a strong resonance. Even in the log magnitude function the problem persists even though the contributions due to the resonances are additive here. This is due to the fact that the logarithm of small values result in large negative values and addition of these values again masks the peak due to the weak resonance. In group delay functions the contributions due to resonances are additive in nature. In addition, the signal information due to a resonance is confined to a narrow region around the resonance frequency and the contribution of this resonance in the other regions is uniformly small in the group delay functions. Hence the effect of the contributions due to the resonances on each other is relatively small in the group delay functions. Thus we can isolate the contribution due to a particular resonance and selectively manipulate it by using group delay functions more effectively than by using either FT magnitude or log magnitude functions. In this sense, we say 'thatthe group delay functions possess high resolution property.

The properties of the group delay functions can be exploited for many applications such as design of digital filters and pole-zero modeling [Yegnanarayana **1981A**], [Yegnanarayana **1981B**]. These properties allow manipulation of signal data effectively in many signal processing

situations, like waveform estimation from an ensemble of noisy measurements [Yegnanarayana, Sreekanth, and Anand Rangarajan 1985]. Group delay functions are also used in speech recognition [Itakura and Umezaki 1987] [Yegnanarayana and Raj Reddy 1979] and in formant estimation [Hema A. Murthy, Madhu Murthy and Yegnanarayana 1989b]. It would be interesting to know whether group delay functions and the signal reconstruction algorithms from these functions preserve the complete information of a signal or not. We must know under what conditions loss of information, if any, occurs and how common are these conditions in practice [Yegnanarayana and Madhu Murthy 1987] [Madhu Murthy and Yegnanarayana 1989]. In Chapter 3, we study the effectiveness of representing one dimensional signals through group delay functions.

After gaining insight into signal representation using group delay functions, we attempt to use this knowledge for processing stationary composite signals where time-frequency resolution is the main issue.

2.3. Time-Frequency Resolution Problem

Time-frequency resolution problem is closely connected with the issue of block processing. Windowing in the time and frequency domains affects resolution in the frequency and time domains, respectively. We give the relationship between time and frequency resolutions in this

section. We also consider the time-frequency resolution problem in the context of composite signal decomposition.

2.3.1. The relationship between time and bandwidth

Spectral resolution in Hertz is normally considered as approximately the reciprocal of the signal observation time in seconds. This relationship in the case of deterministic, finite energy signals can be derived as follows [Marple 1987]:

The equivalent time width T_e of a discrete time signal $x(n)$ is defined as

$$T_e = \frac{T \sum_{n=-\infty}^{\infty} x(n)}{x(0)} \qquad (2.3)$$

where T is the time period.

Similarly the equivalent bandwidth B_e of the discrete time Fourier transform $X(f)$ of $x(n)$ is

$$B_e = \frac{\int_{-1/2T}^{1/2T} X(f)\, df}{X(0)} \qquad (2.4)$$

These two measures of temporal concentration and frequency concentration apply to signals such as window functions which are real-valued and symmetric, and integrate to a finite nonzero area. From the above relations, the

time-bandwidth product can be shown to be

$$T_e B_e = 1 \qquad (2.5)$$

Because most resolution measures deal with signals as seen through windows, the time-bandwidth product $T_e B_e$ is most often used to establish the equivalent bandwidth and therefore the resolution.

In the ideal case, that is when we have infinite number of samples in the time domain for band-limited signals, the above equations give the best possible resolution. But in practice, the maximum resolution that can be achieved is limited by the size and the shape of the window, when a priori information is not available about the signal. When additional information about the signal, such as underlying model is available, super resolution techniques promise a much higher resolution [Kay 1988, chapters 13 to 16]. But in general, for practical signals like speech, such information may not be available.

We study the problem of time-frequency resolution in the context of stationary and time varying formulations of speech signals. We use model signals in the form of composite signals in our experiments and study the resolution issues in composite signal decomposition.

2.3.2. Composite signal decomposition

A composite signal is defined as a finite summation of basic wavelets and their echoes. A number of

practical signals we encounter in different fields such as radar, sonar, seismology, biomedical engineering, speech processing etc., are composite in nature. Composite signal decomposition involves identification and separation of the basic wavelets and finding the amplitudes, bandwidths and arrival times of their echoes.

Such a decomposition technique would be useful in application such as finding the different tissue types and their pathological states from ultrasonic characterization in the medical field [Jong-Ho Choi, Si-Whan Kim and Jong-Soo Choi 1986]. Analysis of composite signals is required to understand the nature and path of targets in radar systems [Skolnik 1970]. In seismology, the location of seismic source may be determined by the waves reflected or transmitted by it [Robinson, Durrani, and Peardon 1986]. Here the problem is even more complicated than that in radar systems because the signals normally differ greatly in wave shape. In speech processing, identification of the resonances (also known as formants) [Rabiner and Schafer 1978] and the closed and open phases of glottal source [Larer, Alsaka and Childers 1985] is important.

The main problem in composite signal decomposition is the resolution of wavelets in time and frequency domains [Childers and Pao 1972]. The presence of ambient noise further complicates the decomposition problem. Often little is known about the statistical nature of noise and in a

number of situations it may be correlated to the actual signal itself.

A number of techniques were proposed for composite signal decomposition with varying degrees of success. Using decision theory, the instants of the wavelet arrivals could be estimated in the case of a noisy composite signal, but the estimates are not reliable [Helstrom 1967]. Inverse filtering approach was used in the decomposition of a composite signal consisting of a known signal and its echoes overlapping in time [Childers, Varga and Perry Jr. 1970]. Adaptive techniques for decomposing a composite signal of identical unknown wavelets in noise was described in [Senmoto and Childers 1972]. Cepstrum techniques were used when the composite signal consists of wavelets of a single frequency [Childers and Kemeriat 1972]. But if multiple echoes are present, or if the signal-to-noise ratio (SNR) is less than 20 dB, the cepstrum techniques are not recommended. Linear prediction (LP) techniques can also be used for composite signal decomposition [Ananthapadmanabha 1979], but the technique performs poorly in the presence of noise. Moreover, the basic wavelet has to conform to an all-pole model, and the echoes of the basic wavelets should be well separated. A generalized correlation technique was suggested [de Figueiredo and Gerber 1983] when the composite signal to be processed is made up of wavelets from a finite and known set of elementary wavelets of predefined shapes.

The above techniques cannot be used satisfactorily for decomposing a general case of composite signal, where no a priori information about the signal or the accompanying noise is available. In the general case of composite signal decomposition, the objective is to determine the significant frequencies of the wavelets, and the instants of their occurrence. The main issue is the resolution achievable in the time and frequency domains. This problem arises because of the discrete nature of processing such signals. The problem is compounded by the fact that the signal available for analysis is corrupted by noise.

In this context, we propose to use the high resolution and additive properties of group delay functions for composite signal decomposition [Madhu Murthy and Yegnanarayana 1988]. In Chapter 4, we show that, despite the computational disadvantages, the properties of the group delay functions can be exploited for composite signal decomposition.

2.3.3. Analysis techniques to find changes in bandwidth

Next we consider the problem of decomposing a composite signal consisting of a summation of wavelets and their echoes with the added constraint that the bandwidths of one or more of the wavelets change within the analysis window. In addition to the epochs and resonances, one has to find the instants of bandwidth changes. This is a difficult

problem and the inevitable additive noise makes it more so. This problem, in its entirety, was not considered in literature. Only special cases were considered as in the context of speech [Ananth, Childers and Yegnanarayana **19851.**

Speech is the output of the vocal tract system excited either by a train of periodic glottal pulses or by the turbulence created at a constriction in the vocal tract [Rabiner and Schafer **1978**]**.** Speech analysis aims at finding a set of parameters such as formants, pitch epochs that characterize the speech production system. As the speech production is dynamic in nature, the parameters also change with time. But due to inertia, the characteristics of the system change only slowly with respect to time. Thus over short intervals of time, speech is assumed to be stationary for purposes of analysis [Atal and Hanauer **1971**]**.**

Most speech analysis systems under the quasistationary formulation are based on the linear source-filter model [Rabiner and Schafer **1978**]**.** In general, it is assumed that the source and vocal tract filter are independent of each other. Though this assumption results in computational simplicity, it is not strictly correct [Fant **1979]. ** The self oscillation of the vocal folds depends upon the local aerodynamic forces which in turn are affected by the supraglottal vocal tract. However, it was reported [Krishna Murthy and Childers **1986**] that the vocal fold

vibration is not significantly affected by the vocal tract, although the volume velocity is affected. The interaction causes the volume-velocity signal to be skewed to the right with respect to the glottal area. A ripple in the volume velocity waveform may appear which is thought to be caused by the first formant frequency of the vocal tract. Similarly the vocal tract system also is affected by the source. In the glottal open phase, vocal tract is connected to the subglottal system comprising of lung cavity. Due to the loading effect of the subglottal system, formant frequencies and bandwidths are affected, and especially the damping of the first formant is effectively increased. This mutual interaction of glottal wave and the formant structure is known as source-tract interaction [Childers, Yea and Bocchieri 1983].

It has been conjectured that source-tract interaction is important for high quality speech synthesis [Allen and Strong 1985]. But reliable methods to identify and measure it from the speech signal are not available. Hence there is a need to develop such methods.

To detect source-tract interaction in speech, it is required to find the bandwidth changes within a pitch period. Stationary signal analysis techniques are not suitable in this case. **A.S.Ananth** et **al** have claimed [Ananth, Childers and Yegnanarayana 1985] that it is possible to identify the source-tract interaction from the

actual speech data in a few cases. They have developed a technique utilizing the group delay functions for this purpose. As mentioned by them, the major difficulty in identifying and measuring the source-tract interaction is the short data record length available for analysis.

Detecting the bandwidth changes within a pitch period in the speech signal is required in another context also. It was shown that the LP analysis of speech was affected by the positioning of the analysis window with respect to the pitch pulses in the time domain [Rabiner, Atal, and Sambur 1977][Steiglitz and Dickenson 1977]. Glottal closed phase analysis of speech was found to give more accurate vocal tract parameters [Krishna Murthy and Childers 1986] [Parthasarathy and Tufts 1987]. Identification of glottal open and closed phases can be accomplished by detecting the accompanying bandwidth changes.

We use a time varying formulation to process a model signal to detect the change in the bandwidth representing source-tract interaction in Chapter 5. However, processing real speech signals for measuring the effects of the source-tract interaction is more difficult and we deal with this problem indirectly in Chapter 6 in the context of short data processing.

2.4. Analysis of Speech Signals

We consider the analysis of natural signals such as speech in the context of limitations imposed by time-frequency resolution problem. Speech is generally analyzed under quasistationary formulation as this is simple and mathematically tractable. A highly successful method for analysis of speech under quasistationary assumption is linear prediction (LP) analysis [Makhoul 1975]. This is a model based approach where the model coefficients are obtained by solving a set of linear simultaneous equations characterized by an autocorrelation or a covariance matrix. The LPCs obtained from autocorrelation coefficients are guaranteed to be stable whereas those obtained from covariance coefficients appear to be more accurate [Chandra and Lin 1974]. We discuss the limitations of this approach for accurately capturing the time varying spectral parameters of speech. We review some of the speech analysis methods under time varying formulation that attempt to overcome these limitations [Lee and Silverman 1988] [Grenier 1986] [Casacuberta and Vidal 1987].

Next we consider the problem of noisy speech processing. Though linear prediction analysis gives good frequency resolution for clean speech, it does not work reliably for noisy speech. We review various methods of speech enhancement [Lim 1982] and show that the time-frequency resolution problem plays a vital role in the

performance of these methods.

2.4.1. Speech analysis with time varying formulation: Modified LP method

Quasistationary formulation is not adequate to accurately capture the source and system parameters from a speech signal. Speech signal is essentially time varying in nature. The vocal tract system changes rapidly when speech components such as stops, fricative onsets, codas and vowel-consonant, consonant-vowel transitions are encountered. The formant transitions over short durations give valuable clues for the identification of consonants [Kewley-Port 1983] in the speech signal. We lose this information, and consequently the intelligibility of consonants is affected, if stationarity assumption is imposed using large segments of speech data. We have seen in the last section that even in the steady vowel regions speech cannot be considered as stationary due to the effects of source-tract interaction. The quality of synthetic speech is affected if these changes are not captured by proper analysis and made use of in the synthesis[Childers, Yegnanarayana and Ke Wu 1985]. One way to overcome this difficulty in speech analysis is to perform the identification of the model over short segments, but this requires a compromise between the accuracy that can be achieved with a short data segment and the faithfulness with which a spectrum must be followed. The other alternative is to investigate parametric and nonparametric methods for

nonstationary signal analysis.

The analysis techniques for nonstationary signals may be considered to be broadly based on three assumptions [Grenier 1986]. Assume that $a_i(t)$ and $b_i(t)$ are the time dependent autoregressive and moving average parameters which are to be computed. One assumption is that the process is not too far from stationary, so that the evolution of the parameters being rather smooth can be tracked by adaptive algorithms [Widrow and Stearns 1985]. This is also the assumption implicitly made in sliding DFT [Portnoff 1981]. The tracking ability of these algorithms is prescribed by either the size of a window, or the value of a fading rate and there lies the limitation of these methods. If the process evolves too quickly, the algorithm will not properly track the evolution of the coefficients, unless the window becomes very short, which in turn degrades spectral resolution.

Another assumption might be that the coefficients evolve in a Markovian way. If the parameters of a Markov model are known, the estimation of the parameters $a_i(t)$ and $b_i(t)$ is simply the estimation of the state of the Markov model, a problem that can be solved with the help of the Kalman filter. But Kalman filtering and the state space approach need tremendous amount of computation time [Widrow and Stearns 1985].

A third assumption is that $a_i(t)$ and $b_i(t)$ may be

approximated satisfactorily by a weighted combination of a small number of known functions [Hall, Oppenheim, and Willsky 1983] [Chevalier, Collet, and Grenier 1985]. Here choosing the set of known functions may pose limitations.

Other methods for analysis of time varying speech signals include hierarchical AR modeling [Kakusho and Yanagida 1982], using time varying filter and source [Almeida and Tribolet 1983].

Thus, when we consider short analysis frames, model based approach (AR, ARMA) is of limited value. This is because unless we have a priori knowledge about the signal, so as to choose appropriate type and order of the model, poor spectral estimates will result. Even when we have the above information, due to the short analysis frame size, the estimates may have large biases and variabilities.

Nonparametric methods for nonstationary spectrum estimation, such as narrow band filtering, complex demodulation, short time Fourier transform and various time-frequency representations also suffer from time-frequency resolution problem when short analysis frames are considered.

Within the constraints mentioned above, attempts have been made to incorporate the dynamic spectral variations such as those due to the source-tract interaction in speech synthesis to obtain high quality speech. Fant introduced a ripple on the glottal volume velocity wave to

approximate the local variations[Fant 1982] [Ananthapadmanabha and Fant 1982]. Atal and Remde (1982) developed a multipulse excitation scheme to incorporate the dynamic variations in the excitation signal. In both these cases, vocal tract system is assumed to be stationary within the analysis frame. The source parameters are changed to account for the local spectral variations. But a more reasonable approach would be to incorporate these variations in the vocal tract system also. We propose in Chapter 6 a modified linear prediction method where we incorporate the time varying spectral characteristics in the vocal tract parameters within the analysis frame.

2.4.2. Speech analysis with quasistationary formulation: Weighted LP method

Natural signals such as speech normally occur in association with noise. We consider the problem of enhancement of speech degraded by background noise. We list below some of the methods used for processing noisy speech[Lim and Oppenheim 1979].

In spectral subtraction [Boll 1979] based method, we subtract the estimated noise power spectrum from that of the noisy speech, to obtain an estimate of the power spectrum of the undegraded speech. In Weiner filter based approach [Haykin 19851, we estimate the weights of an optimum filter from the noisy speech. This filter is then

applied to obtain an estimate of the undegraded speech. Adaptive comb filtering [Fraziar, Samsam, Braida, and Oppenheim 1976] exploits the periodicity property of voiced speech for reducing the background noise from the degraded speech. When a reference signal which is uncorrelated with the speech signal but correlated in some way with noise is available, adaptive noise canceling [Sambur 1978] can be used. In all the above methods a priori knowledge about the noise and/or speech signal characteristics is essential. This may not be possible, as often very little is known about the noise statistics in practical situations. Even the characteristics of speech signal can be estimated only approximately as it is time varying.

A number of model based approaches were also proposed for noisy signal processing. In these approaches, speech is modeled as the output of a time varying digital filter that is excited by either a quasistationary train of pulses for voiced sounds or random noise for unvoiced sounds. The digital filter may be characterized by all pole or pole-zero parameters. In the speech enhancement technique based on an underlying speech model, the parameters of the speech model are first estimated and then speech is generated based on the estimated parameters. In all pole modeling or linear prediction modeling of speech, on a short time basis, the speech waveform $s(n)$ is assumed to satisfy a difference equation of the form [Makhoul 1975]

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u(n) \qquad (2.6)$$

where $u(n)$ is a periodic pulse train for voiced speech and random noise for unvoiced speech. The LP coefficients a_1, a_2, ..., a_p characterize the all pole model.

In the noise free case, these coefficients can be obtained by solving a set of simultaneous equations of the form

$$R \cdot a = r \qquad (2.7)$$

where a is the LP coefficient vector given by

$$a = (a_1\ a_2\ a_3\ \ldots\ a_p)^T \qquad (2.8)$$

and R and r are the correlation matrices of orders p x p and p x 1 respectively.

But in the presence of noise, LP modeling, as formulated above, is not effective because the LP coefficients are sensitive to the additive noise [Sambur and Jayanth 1976]. Alternately in the noisy case, equation (2.7) can still be solved after the correlation matrices R and r are estimated taking into consideration the presence of noise. One approach would be to estimate the power spectrum of the undegraded speech from the noisy speech using one of the methods discussed earlier, form R and r from the inverse transform of this estimate and then solve for a in equation (2.7) [Kobatake, Inasi and Kakuta 1978].

A more theoretical approach is to use parameter

estimation rules such as maximum likelihood (ML), maximum a posteriori (MAP) and minimum mean square error (MMSE) for estimating the model parameters. All pole model parameters were estimated from degraded speech based on maximum likelihood approach in [Lim and Oppenheim 1978]. A more general method, where pole-zero modeling of speech was used is discussed in [Musicus and Lim 1979].

In another method, the noisy speech is modeled as an autoregressive and moving average (ARMA) model [Done and Rushforth 1979] where the AR part represents the clean speech signal and the MA part represents the noise. The clean speech is recovered by estimating the poles of the ARMA model. In all these methods the noise is assumed to be white Gaussian process. A technique to process speech degraded by colored noise is proposed recently [Yarman-Vural 1990]. Here the colored noise is first whitened by an estimated all pole noise model. Then the filtered signal in additive white noise is modeled as an ARMA process as mentioned above. Thus the knowledge of the characteristics of the colored noise is essential. computational overhead is also high in the above methods.

We propose a new technique for noisy signal processing, called weighted linear prediction in Chapter 7. Here we do not assume a priori knowledge about the noise statistics. We give more emphasis to the high signal-to-noise ratio (SNR) regions of the noisy signal in the time

domain while computing LP coefficients of the signals. Such modifications were proposed in the LP formulation for estimating reliable model parameters in the context of analyzing speech with short pitch periods [Kakusho, Yanagida and Mizoguchi 1984] [Miyoshi et al 1986]. But in these cases the weights are varied according to the prediction residual of the corresponding data samples.

2.5. Speech Analysis-Synthesis Package

A speech analysis-synthesis package has been developed to test the various speech analysis techniques discussed in this thesis. The synthetic speech obtained from the speech parameters that are extracted taking into consideration the limitations imposed by the time-frequency resolution must be better than that generated using ordinary LP coefficients.

The speech analysis-synthesis package is based on standard LP vocoder [Sambur 1975] and incorporates some ideas expressed in a technical report by B.Yegnanarayana, et al [Yegnanarayana, Childers and Naik 1983]. The LP coefficients are extracted using Durbin's recursive algorithm [Makhoul 1975] on the autocorrelation coefficients of the given data. Pitch is extracted using Average Magnitude Difference Function (AMDF) algorithm [Ross et al. 1974]. We use center clipping [Sondhi 1968] to flatten the spectrum for more reliable pitch extraction before applying

AMDF algorithm. Fant's model is used to generate excitation signal. There is flexibility to choose different frame lengths and frame rates.

2.6. Outline of the Present Work

This review shows that there are limitations in processing speech-like signals with the existing methods to extract formant information as well as excitation. The problem was primarily in the quasistationary assumption and block processing. Block processing restricts the accuracy of formant data that can be extracted using existing methods. Model based methods like LP analysis do overcome the resolution restriction posed by block processing. But the problem of spurious peaks and formant bandwidth estimation still remain. Moreover, rapid changes in the vocal tract characteristics cannot be captured with these methods.

The objective of this thesis is to study systematically some of the issues in digital processing, especially the time-frequency resolution problem which affects analysis of nonstationary signals like speech. It is well known that group delay functions possess excellent properties (additive and high resolution) which are useful in a number of signal processing situations. We first examine in Chapter 3 how digital processing affects the representation of signal information through group delay functions. We show that except for certain types of signals

all other signals can be effectively represented and manipulated through group delay functions.

In Chapter 4, we discuss the time-frequency resolution issues in detail with respect to composite signal decomposition problem. We consider a stationary signal. We show how group delay functions can be used to extract useful information in some difficult cases. Model studies of source-tract interaction are reported in Chapter 5. In particular, changes in the bandwidth of resonances within a short period, as in the case of voiced speech, are modeled. We show how the group delay function can be used to detect bandwidth changes. We investigate methods for processing speech under nonstationary assumption. In this context, we propose a modified autocorrelation based LP method in Chapter 6 to deal with changes over very short (less than 2 msec) segments in speech. The main objective is to capture the dynamic spectral variations such as the effects of the source-tract interaction. In Chapter 7, we attempt to process natural signals such as speech corrupted by background noise. We consider methods for analysis of signals under quasistationary assumption. We show that existing methods and even use of group delay functions do not give satisfactory results, although group delay processing is useful in some select segments. We propose a weighted LP method for analysis of noisy speech data. The performance of these two methods are tested using a

synthesis package that is described in Appendix. We show that the quality of the synthetic speech obtained is significantly better than that obtained from standard LP analysis.

Chapter 3

EFFECTIVENESS OF SIGNAL REPRESENTATION: GROUP DELAY FUNCTIONS

Discretization and truncation operations are essential to implement many digital signal processing algorithms. These operations limit the time and frequency resolutions of the signal. The severity of these limitations depends on the type of representation chosen. We have already studied in the previous chapter the computational problems that result due to discretization and truncation in a common representation, namely the discrete Fourier transform. In this chapter, we study the effectiveness of signal representation through group delay functions. First, we introduce the group delay representation which is useful in certain signal processing situations and study how the above mentioned computational problems manifest in this representation. Next, we describe a series of experiments performed to determine the limitations of the group delay representation. Finally, we summarize the conditions under which the group delay representation may be used for signal representation and processing.

3.1. Algorithms for Signal Representation Using Group Delay Functions

Group delay function is the negative derivative of

the Fourier transform phase of a signal. Precise definitions of the group delay functions and their properties were discussed in the previous chapter.

We describe algorithms for computing the group delay functions as well as for deriving the signal from the group delay functions. These algorithms are based on [Yegnanarayana, Saikia, and Krishnan 1984] and [Oppenheim and Schafer 1989]. An N-point discrete Fourier transform (DFT) is used in the following algorithms. We use the discrete frequency variable k, instead of the continuous variable ω throughout this section. It may be noted that scale factor of the FT magnitude is not captured by $\{\tau_m(k)\}$, and hence this has to be separately stored while computing the group delay. In the case of $\{\tau_p(k)\}$, its average value represents the contribution due to the linear phase term in FT phase spectrum, and this is also separately stored.

3.1.1. Computation of $\{\tau_m(k)\}$

(i) Form $\{x(n)\}$ from the given time domain samples $(s(n))$ such that

$$x(n) = \begin{cases} s(n) & , \quad n = 0, 1, \ldots N/2 - 1 \\ 0 & , \quad n = N/2, N/2 + 1, \ldots N-1 \end{cases}$$

(ii) Obtain the spectral magnitude samples $|X(k)|$, $k = 0, 1, \ldots N-1$ through the DFT of the time

sequence {x(n)}

(iii) Compute the cepstral coefficients {c(n)} through the inverse DFT of {ln|X(k)|}

(iv) Save c(0) (this represents the scale factor) and form the sequence {g(n)}, where

$$g(0) = 0$$

$$g(n) = \begin{cases} nc(n) & , \quad n = 1, 2, \ldots N/2 \\ g(N-n) & , \quad n = N/2 + 1, N/2 + 2, \ldots N-1 \end{cases}$$

(v) Obtain $\{T_m(k)\}$ as the real part of the DFT of {g(n)}

3.1.2. Computation of $\{T_p(k)\}$

(i) Form {x(n)} from the time domain samples {s(n)} such that

$$x(n) = \begin{cases} s(n) & , \quad n = 0, 1, 2, \ldots N/2 - 1 \\ 0 & , \quad n = N/2, N/2 + 1, N/2 + 2, \ldots N-1 \end{cases}$$

(ii) Form the sequence {y(n)}, where

$$y(0) = 0$$

$$y(n) = \begin{cases} n\,x(n) & , \quad n = 1, 2, \ldots N/2 - 1 \\ 0 & , \quad n = N/2, N/2 + 1, N/2 + 2, \ldots N-1 \end{cases}$$

(iii) Let the DFTs of {x(n)} and {y(n)} be {X(k)} and {Y(k)}, respectively.

If

$$X(k) = X_1(k) + j\,X_2(k) \;, \; k = 0, 1, 2, \ldots N-1$$

and

$$Y(k) = Y_1(k) + j\, Y_2(k) \ , \ k = 0, 1, 2, \ldots N-1$$

where $X_1(k)$, $X_2(k)$, $Y_1(k)$ and $Y_2(k)$ are real sequences, then

$$\tau_p(k) = (X_1(k)Y_1(k) + X_2(k)Y_2(k)) \ / \ |X(k)|^2$$

(iv) Compute the average value of $\{\tau_p(k)\}$ and save it. This represents the linear phase component.

We assume in the following reconstruction algorithms that the group delay functions $\{\tau_m(k)\}$ and $\{\tau_p(k)\}$ are available along with the two constants representing the scale factor and linear phase term.

3.1.3. Reconstruction of spectral magnitude from $\{\tau_m(k)\}$

(i) Obtain $\{nc(n)\}$ as inverse DFT of $\{\tau_m(k)\}$.

(ii) Extract the cepstral coefficients $\{c(n)\}$ from $\{nc(n)\}$ and restore the scale factor by assigning it to $c(0)$ as shown below.

$$c(0) = 0$$

$$c(n) = \begin{cases} nc(n)/n \ , & n = 1, 2, \ldots N/2 \\ c(N-n) \ , & n = N/2 + 1, N/2 + 2, \ldots N-1 \end{cases}$$

(iii) Obtain $\{\ln|X(k)|\}$ from DFT of the even sequence $c(n)$.

(iv) Compute $|X(k)|$, $k = 0, 1, 2, \ldots N-1$ as

$$|X(k)| = \exp(\ln|X(k)|)$$

3.1.4. Reconstruction of spectral phase from $\{\tau_p(k)\}$

(i) Obtain $\{nc(n)\}$ from the inverse DFT of $\{\tau_p(k)\}$

(ii) Extract the cepstral coefficients $\{c(n)\}$ from $\{nc(n)\}$ as shown below.

$$c(0) = 0$$

$$c(n) = nc(n)/n \ , \ n = 1, 2, \ldots N/2 - 1$$

$$c(N/2) = 0$$

$$c(n) = - c(N-n) \ , \ n = N/2 + 1, N/2 + 2, \ldots N-1$$

(iii) Find the uncorrected phase $\{\theta_{uc}(k)\}$ from DFT of $\{c(n)\}$.

(iv) Obtain the actual phase $\{\theta(k)\}$ by adding to $\{\theta_{uc}(k)\}$ the contribution due to the linear phase term τ_{p_avg} (average of $\{\tau_p(k)\}$).

$$\theta(0) = 0$$

$$\theta(k) = \theta_{uc}(k) + k \tau_{p\_avg} \ , \ k = 1, 2, \ldots N/2 - 1$$

$$\theta(N/2) = 0$$

$$\theta(k) = - \theta(N-k) \ , \ k = N/2 + 1, N/2 + 2, \ldots N - 1$$

In general, discretization and quantization may bring about partial loss of information in the group delay domain. The severity of the information loss depends on the nature of the signal being processed. For instance, the linear phase term, which is the average of $\tau_{p(w)}$, cannot be computed accurately by averaging the limited number of samples $\{\tau_p(\omega)\}$ when there are large fluctuations in that

function. Similarly, aliasing in the cepstral domain contributes errors in $\tau_m(\omega)$.

In the next section, we investigate the effect of discretization on the representation of signals through the group delay functions and on the reconstructed signals. We identify signal parameters which may affect the signal reconstruction error. Further, we discuss the nature of dependence of this error on these parameters. The consequences of this error in signal processing are discussed in Section 3.3.

3.2. Study of the Nature of Errors in Group Delay Representation

If continuous time and frequency variables are used throughout, then the errors due to aliasing in the cepstral domain are avoided in the computation of the group delay functions. But digital processing of data necessitates discretization, which may result in partial loss of information. The accuracy of representation may depend on the characteristics of the signal itself, such as the number of roots and their locations in the z-plane.

3.2.1. Accuracy of signal representation using group delay functions

We will derive an expression for the magnitude group delay function computed from the algorithm given in [Yegnanarayana, Saikia and Krishnan 1984]. Consider a real,

causal and finite length sequence $x(n)$ whose z-transform is of the form

$$X(z) = \frac{|A| \prod_{k=1}^{m_i} (1-a_k z^{-1}) \prod_{k=1}^{m_o} (1-b_k z)}{\prod_{k=1}^{p_i} (1-c_k z^{-1})} \qquad (3.1)$$

where $|a_k|, |b_k|$ and $|c_k|$ are all less than 1. Here c_ks and a_ks determine the poles and zeros inside the unit circle, respectively, and b_ks determine the zeros outside the unit circle in the z-plane.

The complex cepstrum $\hat{x}(n)$ of $x(n)$ is given by [Oppenheim and Schafer 1989],

$$\hat{x}(n) = \begin{cases} \log |A| & n = 0 \\ \sum_{k=1}^{p_i} \frac{c_k^n}{n} - \sum_{k=1}^{m_i} \frac{a_k^n}{n} & n > 0 \\ \sum_{k=1}^{m_o} \frac{b_k^{-n}}{n} & n < 0 \end{cases} \qquad (3.2)$$

The magnitude cepstrum $c(n)$, which is the inverse Fourier transform of the log magnitude spectrum, is the even part of the complex cepstrum and can be expressed as

$$c(n) = \begin{cases} \log |A| & n = 0 \\ \sum_{k=1}^{p_i} \frac{c_k^{\,n}}{2n} - \sum_{k=1}^{m_i} \frac{a_k^{\,n}}{2n} - \sum_{k=1}^{m_o} \frac{b_k^{\,n}}{2n} & n > 0 \\ -\sum_{k=1}^{p_i} \frac{c_k^{\,-n}}{2n} + \sum_{k=1}^{m_i} \frac{a_k^{\,-n}}{2n} + \sum_{k=1}^{m_o} \frac{b_k^{\,-n}}{2n} & n < 0 \end{cases} \quad (3.3)$$

The magnitude group delay function is given as

$$\tau_m(\omega) = \sum_{n=1}^{\infty} n\, c(n) \cos n\omega$$

$$= \frac{1}{2} \sum_{n=1}^{\infty} \sum_{k=1}^{p_i} c_k^{\,n} \cos n\omega - \frac{1}{2} \sum_{n=1}^{} \sum_{k=1}^{m_i} a_k^{\,n} \cos n\omega$$

$$- \frac{1}{2} \sum_{n=1}^{\infty} \sum_{k=1}^{m_o} b_k^{\,n} \cos n\omega \quad (3.4)$$

Equation (3.4) shows that the magnitude group delay function is the Fourier transform of the summation of the complex exponentials formed by the poles and zeros of the signal. We observe that when any of the roots are close to the unit circle, the corresponding complex exponentials do not decay fast enough with increasing n. Hence, aliasing takes place in the function nc(n), as the DFT size used in the actual implementation is finite. This contributes to errors in the magnitude group delay. We also observe from equation (3.4) that when the number of roots is increased, the total error, which is the summation of error

contributions of each root, also increases. As the DFT size used in the computation is increased, the severity of aliasing in {nc(n)} is reduced and consequently the error in the group delay function would be less. We note here that the magnitude group delay function is more sensitive to aliasing problems than the cepstrum, for the decay of function nc(n) is much slower owing to the multiplication factor n.

The phase group delay function also undergoes errors in the representation of the signal information under a similar set of conditions to those described above. But these errors manifest themselves as undersampling of the group delay function. The algorithm for computing the phase group delay as given in Section 3.1 gives accurate values of the group delay at the sample points. But in the reconstruction algorithm from phase group delay, we have to compute the cepstrum, where aliasing occurs owing to the undersampled phase group delay function. This results in the distortion of the reconstructed signal.

3.2.2. Experimental results

We conducted experiments to study the effects of discretization on group delay representation of signals. Composite signals of the form shown in Fig. 3.1 are used for these experiments. Each signal is a summation of the

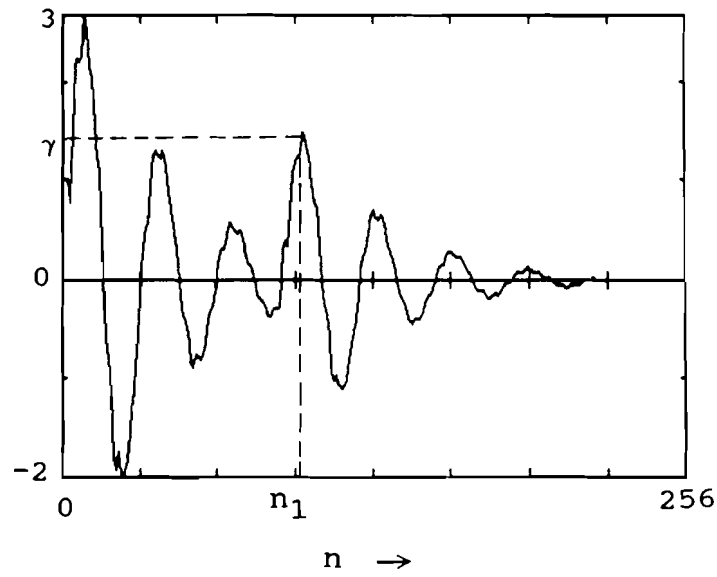


Fig. 3.1. A typical composite signal used in the experiments.

windowed impulse response of a 12th order all-pole system (referred to as the basic signal) and its scaled and shifted version (referred to as its echo). We will not consider the effect of windowing in the following discussion although a single-sided Hamming window is used. This signal is represented in the time domain as

$$y(t) = x(t)u(t) + \gamma x(t-t_1)u(t-t_1) \qquad (3.5)$$

where $u(t)$ is the unit step function and γ is the ratio of the echo amplitude and the basic signal amplitude. The discrete time version of this signal is given by

$$y(n) = x(n)u(n) + \gamma x(n-n_1)u(n-n_1) \qquad (3.6)$$

Taking the z-transform for the equation (3.6), we get

$$Y(z) = (1 + \gamma z^{-n_1}) X(z) \qquad (3.7)$$

where $Y(z)$ and $X(z)$ are the z-transforms of $y(n)$ and $x(n)$ respectively. $X(z)$ contains six pairs of complex conjugate poles located inside the unit circle in z-plane.

From equation (3.7), we observe that in the z-plane the signal contains twelve poles due to the basic

signal and n_1 zeros symmetrically distributed around the origin at a distance of γ from it due to the echo. If $\gamma = 1$, the zeros are on the unit circle and when $\gamma < 1$ or $\gamma > 1$, the zeros are inside or outside the unit circle, respectively.

The following procedure is followed in all the experiments:

A composite time signal y(n) of length less than N/2 samples is taken. An N-point sequence is obtained by padding the above sequence with zeros. The group delay functions τ_m and τ_P are computed. The time domain signal is reconstructed from the group delay functions. Fig. 3.2 shows the signals at various stages in the computation including the reconstruction from the group delay functions. We note that the algorithm for reconstruction of log magnitude from τ_m consists of retracing the steps for the algorithm for computing τ_m from log magnitude. Hence, the reconstructed log magnitude does not show any significant error, even though there may be error in τ_m. But the corresponding algorithms are entirely different for τ_p . Therefore the reconstruction error in the following experiments is predominantly contributed by the error in the phase reconstructed from τ_p.

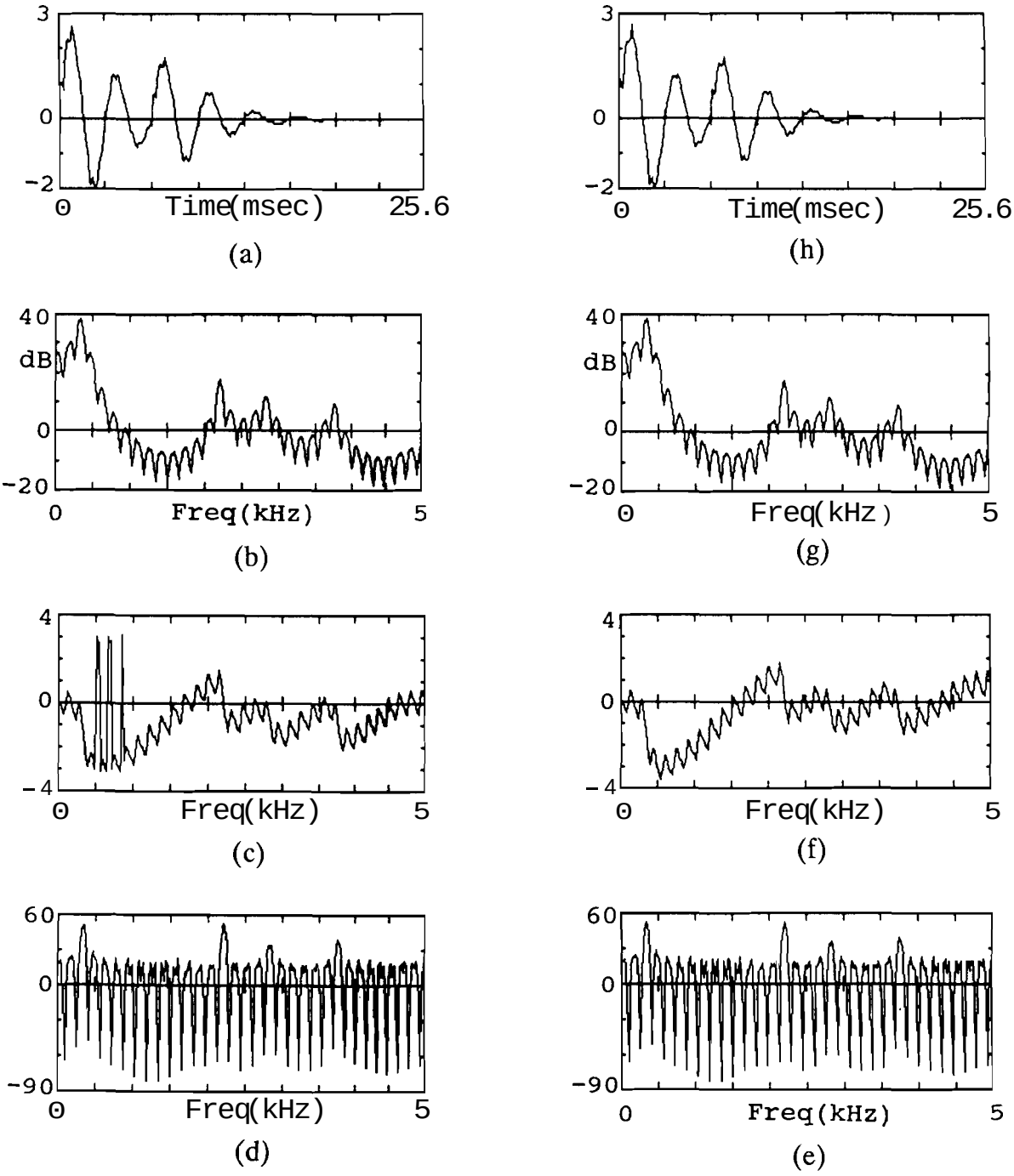


Fig. 3.2. Illustration of the process of group delay computation and reconstruction from the group delay functions: (a) Time domain signal; (b) its FT magnitude; (c) FT phase; (d) magnitude group delay (τ_m); (e) phase group delay (τ_p); (f) reconstructed FT phase from τ_p; (g) reconstructed FT magnitude from τ_m, and (h) the reconstructed time domain signal.

Experiment No.1: Effect of varying the proximity of zeros to the unit circle

From equation (3.7) it is seen that by changing the value of γ, which is the ratio of the amplitudes of the echo and the basic signal, we can move the zeros along the raaial direction in the z-plane. An experiment is conducted in which γ was varied from 0.5 to 2.0. Fig. 3.3 shows the superimposed plots of the original and the reconstructed signals for different values of γ. It is observed from these plots that the reconstruction error is negligible for $\gamma = 0.5$, but increases steadily as the zeros approach the unit circle. As γ is further increased beyond 1.0, the zeros fall outside the unit circle and move away from it. In this case it is observed that the reconstruction error decreases and becomes negligibly small when $\gamma = 2.0$. This shows that proximity of roots (in this case zeros) to the unit circle introduces error in the reconstruction irrespective of whether they are inside or outside the unit circle. This may be attributed to the fact that as the zeros approach the unit circle, abrupt changes occur in the spectral phase at the frequencies corresponding to these zeros. Similarly, in the spectral magnitude the valleys due to the zeros become steeper. These changes in the magnitude and phase result in poor sampling of the corresponding group delay functions and hence loss of information occurs in the group delay

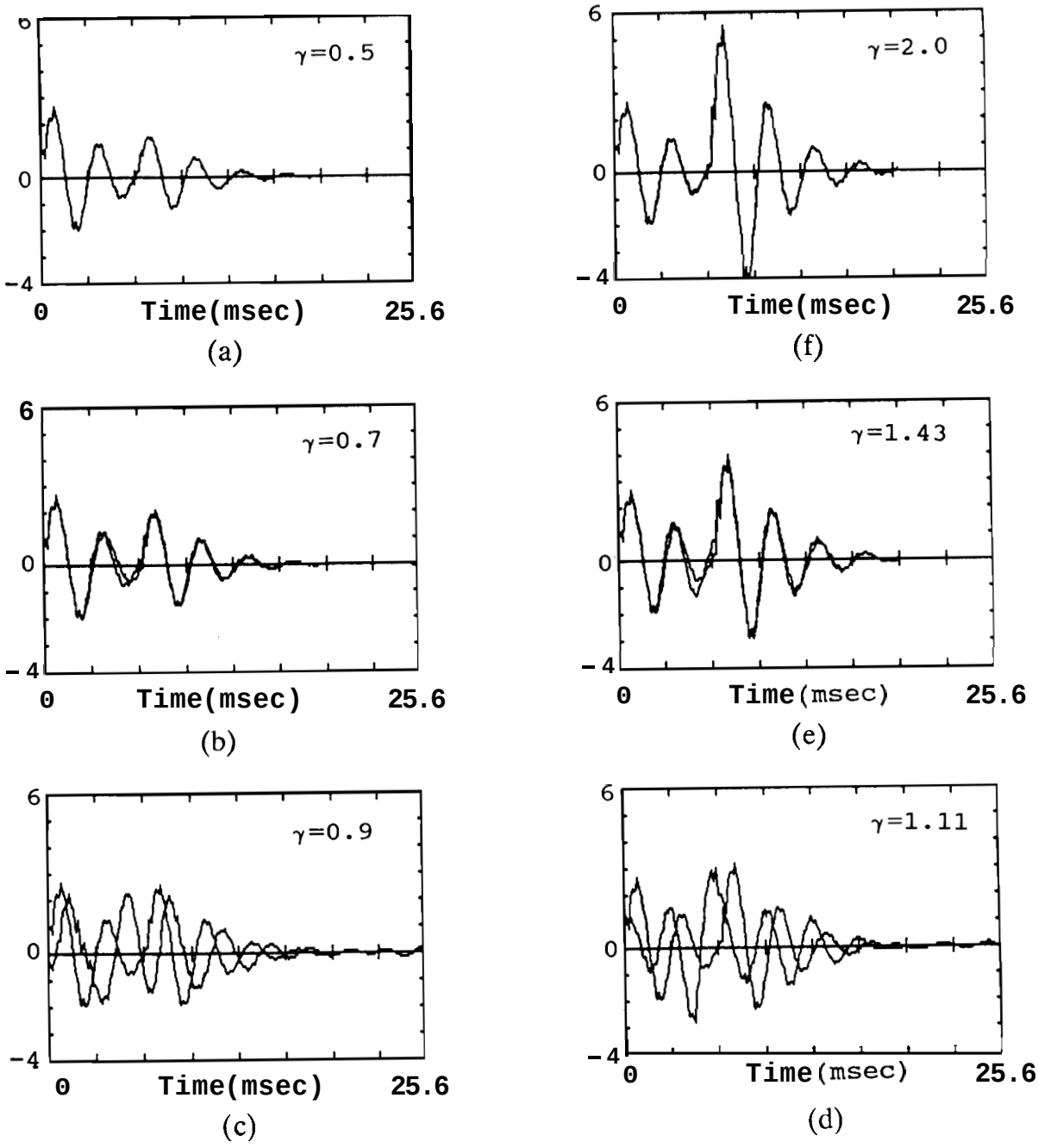


Fig. 3.3. Comparison of the original and reconstructed signals for different values of echo amplitude (γ). Echo shift and DFT size are fixed at 6.2 msec and 512 in the above plots.

transformations.

Experiment No.2: Effect of number of zeros

The effect of varying the number of zeroes equidistant from the unit circle is also studied. The delay n_1 in equation (3.7) is equal to the number of zeros in the z-plane. Hence by varying n_1 in the time domain signal, we can vary the number of zeros. An experiment is conducted in which n_1 is varied from 1 to 20 and the corresponding plots of the original and reconstructed signals are superimposed in Fig. 3.4. It is seen from these plots that as the number of zeros are increased, the reconstruction error also increases.

Experiment No.3: Effect of number of DFT points

The time signal length is fixed at 16 samples and zeros are padded to make up the N points. N is changed from 32 to 1024 in steps, and the superimposed plots of the original and reconstructed signals for different cases are shown in Fig. 3.5. We observe that the reconstruction error decreases as the number of DFT points are increased. For N = 1024, the error is negligibly small. These results also show that poor sampling of the group delay functions causes errors in the reconstruction.

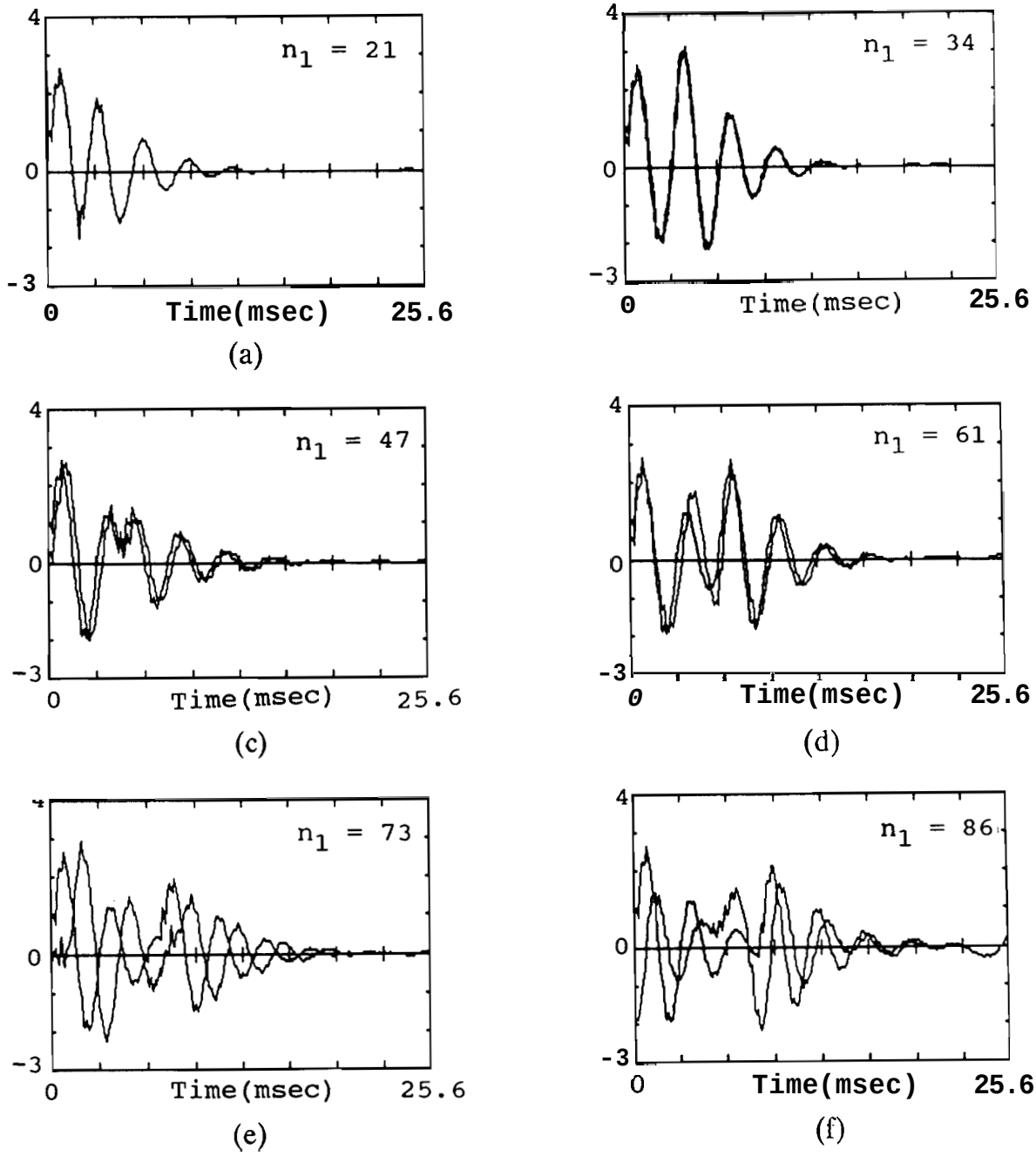


Fig. 3.4. Comparison of the original and reconstructed signals for different values of the number of zeros (n_1) in the z-plane. Echo amplitude and DFT size are fixed at 0.8 and 512 in the above plots.

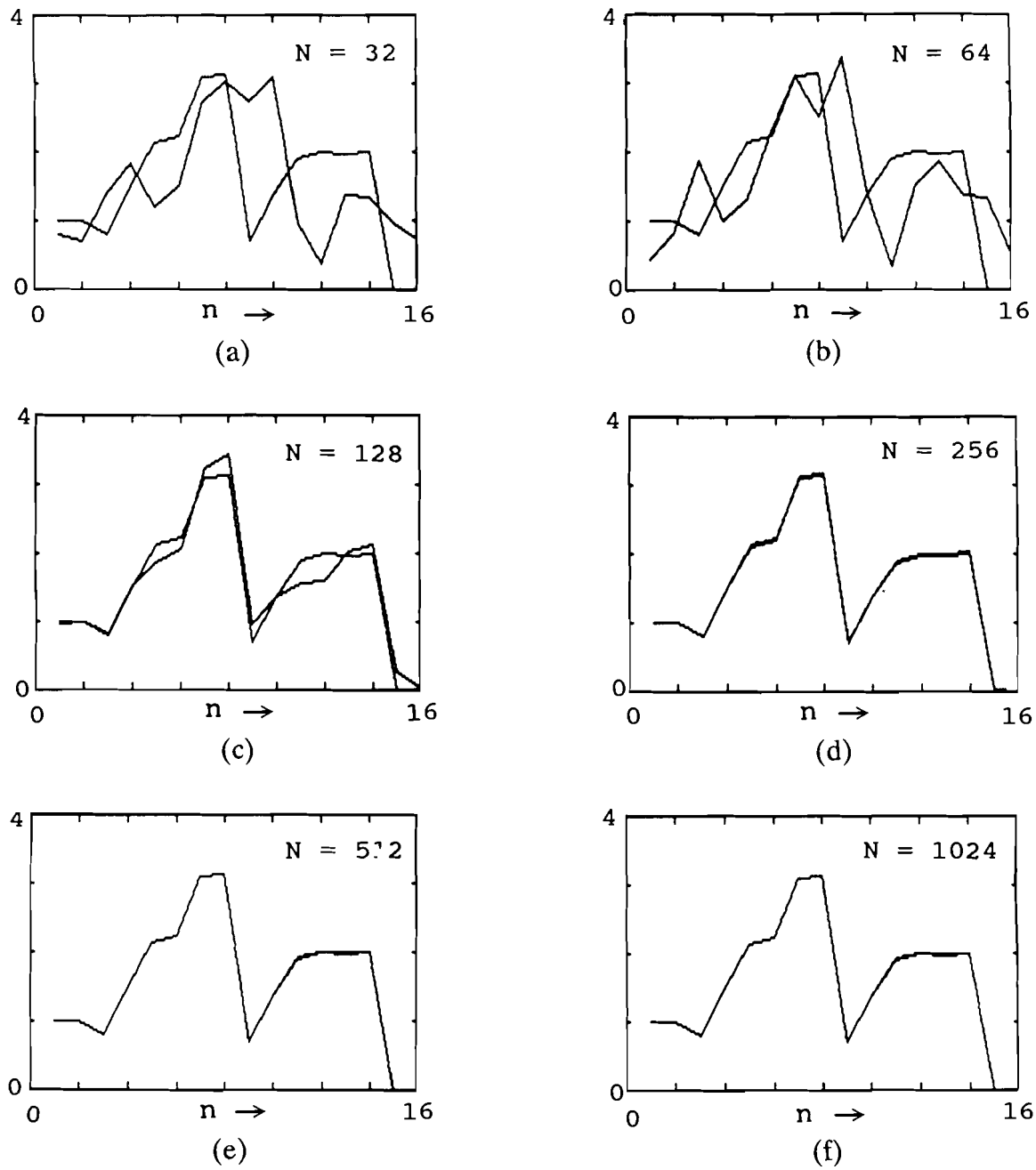


Fig. 3.5. Comparison of the original and reconstructed signals for different values of the DFT size (N). Signal length is 16 samples, echo amplitude is 0.8 and the echo shift is 6 in the above plots.

Experiment No.4: Statistical Studies

We have conducted the above experiments on a large set of data to verify whether the results obtained above are consistent. In each case, mean squared error between the original and the reconstructed time signals is computed. This error is normalized and expressed in percentage. The data is derived from 60 sets of 12th order all-pole systems. The error plots are shown in Fig. 3.6. These plots confirm our earlier observations.

3.3. Discussion

The studies given in this chapter clearly demonstrate the limitation of the group delay representation.

The effects of the characteristics of the signal on the reconstruction error is felt when the number of DFT points chosen is not large. Though phase group delay gives accurate values at the sample points, magnitude group delay is distorted due to aliasing in the cepstral domain. The signal transformation routines given in Section 3.1 for representation of signals in the group delay domain are accurate only when the number of sample points is sufficiently high in the group delay domain. Hence, whenever signal characteristics, such as the number of roots and/or their proximity to unit circle, result in rapid variation in

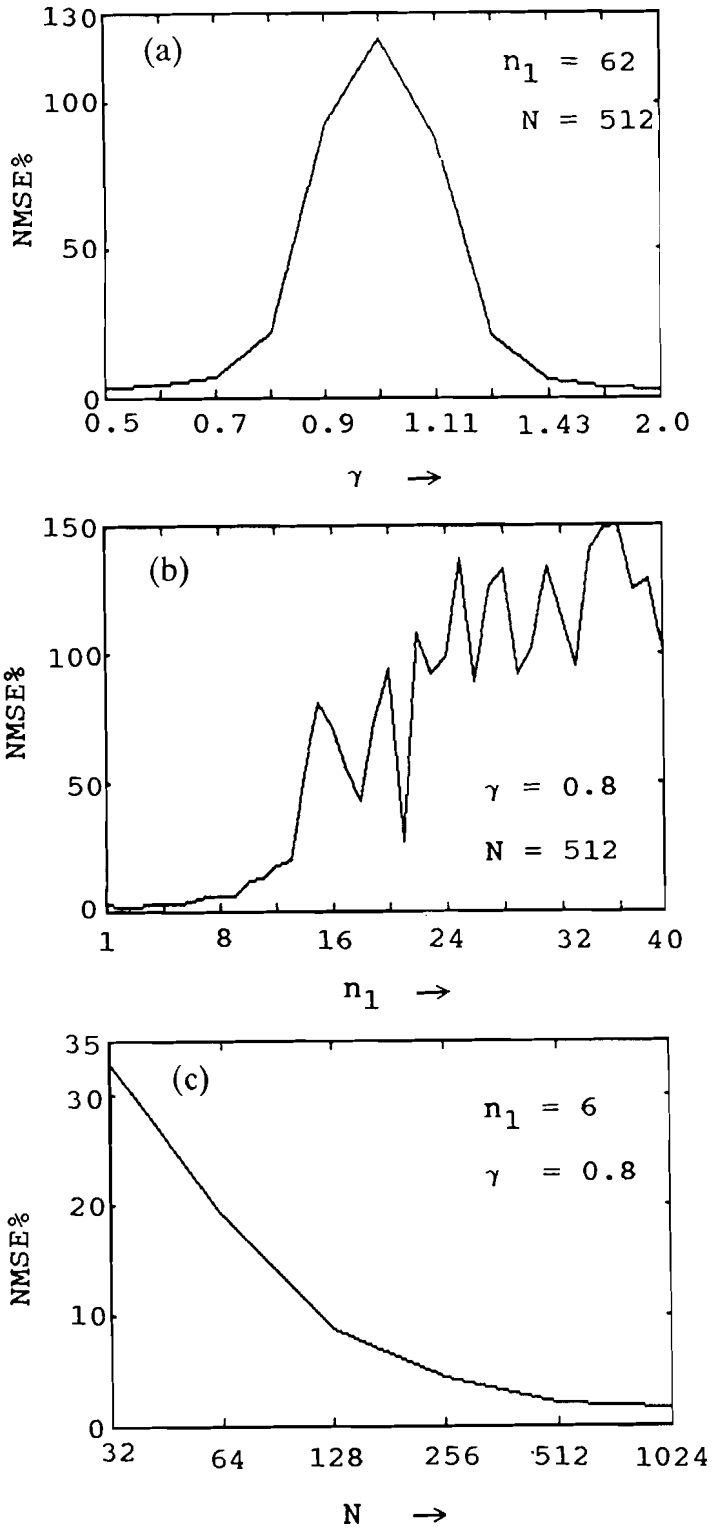


Fig. 3.6. Reconstruction error averaged over 60 frames of data. The variation in the normalized mean squared error (NMSE) with respect to (a) echo amplitude (γ); (b) number of zeros (n_1) in the z-plane, and (c) DFT size (N).

the spectral magnitude or phase, poor sampling occurs in these domains. We note that adequate sampling based on Nyquist criterion in the time domain does not necessarily result in proper sampling in the group delay domain, because of the derivatives involved in the definition of the group delay functions.

The above conclusions are relevant for composite signal decomposition and speech analysis. In the next chapter, we discuss issues involved in the time-frequency resolution problem in the context of composite signal decomposition.

Chapter 4

ANALYSIS OF STATIONARY SIGNALS: COMPOSITE SIGNAL DECOMPOSITION PROBLEM

We have outlined some issues connected with signal representation in the previous chapter. In signal analysis, effects of discretization and truncation manifest as a time-frequency resolution problem. It is relatively easy to obtain a good resolution either in the time or in the frequency domain separately. But if we seek time and frequency resolution together, we encounter limitations imposed by discretization and truncation. Time-frequency resolution stands out as an issue, especially in the problem of composite signal decomposition, where identifying and separating resonances in the spectral domain and epochs in the time domain are the main problems.

A composite signal consists of summation of basic wavelets with different amplitudes occurring at different instants (epochs) of time, together with additive noise. The problem in composite signal decomposition is to determine the shapes of the basic wavelets, their amplitudes and arrival times.

There are several types of composite signals depending on the following situations:

(a) The basic wavelets consist of the impulse response of an all-pole or an all-zero or a pole-zero system.

(b) The basic wavelets are all identical or different.

(c) The noise characteristics are known or unknown.

(d) The characteristics of the system generating the basic wavelets is stationary or nonstationary.

Solutions to the composite signal decomposition problem depend on the type of situation we are considering. Since our objective is to process speech-like signals, we consider a model signal consisting of basic wavelets generated by an all-pole (cascade of resonators) system. Also since our objective in this chapter is to address the issue of time-frequency resolution, we consider only a stationary model of composite signal. Model signals enable us to vary the characteristics of both the basic wavelets and the epochs to simulate signals with desired parameter values. This will help in the development of algorithms to solve the composite signal decomposition problem. Model signals also enable us to study the performance of these algorithms. It is to be noted that it is not possible to have any control on the parameters in the natural signals to enable us to evaluate the algorithms for the composite signal decomposition problem.

This chapter is organized as follows: First we discuss the problem of resolution in both time and frequency

domains in the case of stationary signals and the issues involved in the analysis of such signals. Next we formulate the problem of composite signal decomposition where resolution obtainable in time and frequency domains is studied using various analysis techniques. We explore the possibility of using group delay functions to solve the problem of composite signal decomposition. We develop a method of processing the composite signal using weighting functions derived from a smoothed group delay function to selectively manipulate the resonances. Effects of additive noise on the proposed analysis methods are discussed.

4.1. The Problem of Time-Frequency Resolution for Stationary Signals

Time-frequency resolution is a classic problem in signal analysis. According to the uncertainty principle, the instantaneous frequency and the time of occurrence of a wavelet cannot be specified simultaneously. Hence, we lose time resolution when we try for better frequency resolution and vice versa.

In digital signal processing, a given segment of signal may be represented in both time and frequency domains using the discrete Fourier transform (DFT) pair. Computation of DFT involves windowing the given signal in time and frequency domains and this affects the resolution.

In a number of practical applications, it is

required to resolve signals in both the time and frequency domains. For instance, in speech analysis, it is required to identify and sometimes to separate the resonances of different formants for selective processing. Similarly in the time domain, overlapping responses of the vocal tract due to excitation at the glottal closure and open instances are to be separated. It is easier to resolve signals when they are nonoverlapping either in the time or in the frequency domains. Special analysis techniques are required if the signals overlap both in the time and frequency domains. To study the issues connected with the time-frequency resolution, we formulate a composite signal decomposition problem and attempt to solve it in the following sections.

4.2. Composite Signal Decomposition Problem

We consider a composite signal defined as a finite summation of basic wavelets and their echoes with additive noise. We consider the following model of a composite signal [Yegnanarayana and Madhu Murthy 1987].

$$y(t) = \sum_{i=1}^{p} \sum_{j=1}^{q_i} A_{ij}\, s_i(t-t_{ij}) + N(t) \qquad (4.1)$$

where

s_i is a basic wavelet consisting of concatenation of

resonators,

t_{ij} is arrival time of the wavelet,

A_{ij} is amplitude of the wavelet,

N(t) is additive noise.

First we have to identify the resonances of the composite signal. LP modeling is not always successful due to the possibility of spurious peaks. We use a method based on group delay representation to identify the resonances from a smoothed group delay function derived from the FT magnitude function. Once the resonances are identified, it is required to find the instants of occurrences of these resonances.

4.2.1. Theory of composite signals

One way of looking at the type of composite signal we are considering is that it consists of a spectral envelope part and a fine structure part. Consider the expression for the composite signal as given in equation (4.1). We approximate it in discrete domain as

$$y(n) = \sum_{i=1}^{p} \sum_{j=1}^{q_i} A_{ij}\, s_i(n-n_{ij}) + N(n) \qquad (4.2)$$

Let us assume noise free case. Then the z-transform of y(n) is given by

$$Y(z) = \sum_{i=1}^{p} \sum_{j=1}^{q_i} A_{ij}\, S_i(z)(1-z^{-n_{ij}}) \tag{4.3}$$

Let

$$S_i(z) = \frac{1}{D_i(z)} \tag{4.4}$$

where $D_i(z)$ corresponds to an all-zero system. Then

$$Y(z) = \sum_{i=1}^{p} \sum_{j=1}^{q_i} \frac{A_{ij}\,(1-z^{-n_{ij}})}{D_i(z)}$$

$$= \sum_{i=1}^{p} \frac{\sum_{j=1}^{q_i} A_{ij}\,(1-z^{-n_{ij}})}{D_i(z)}$$

$$= \frac{\sum_{k=1}^{p} \prod_{\substack{i=1 \\ i\neq k}}^{p} D_i(z) \sum_{j=1}^{q_i} A_{ij}\,(1-z^{-n_{ij}})}{\prod_{i=1}^{p} D_i(z)} \tag{4.5}$$

We observe from equation (4.5) that the reciprocal of the denominator part is an all pole system corresponding to the resonances in all the basic wavelets. We call this part of the spectrum as spectral envelope as we do not expect any rapid changes here. The number of peaks in the spectral envelope is equal to the number of distinct resonances in all the basic wavelets. On the other hand the numerator part has many zeros in the z-plane. The number of these zeros depends mostly on n_{ij}. The zeros determine the fine

structure part in the spectrum. It is seen that this part captures the epoch information of the composite signal.

Let us denote the numerator part $Y(z)$ with $E(z)$ and the reciprocal of the denominator part with $S(z)$. That is

$$Y(z) = E(z) \cdot S(z) \tag{4.6}$$

Let

$$e(n) = Z^{-1}[E(z)] \tag{4.7}$$

and

$$s(n) = Z^{-1}[S(z)] \tag{4.8}$$

$e(n)$ is a time sequence which contains the epoch and amplitude information of all the wavelets in the composite signal. $s(n)$ is a time sequence corresponding to the basic wavelets in the composite signal. Thus the composite signal $y(n)$ can be seen as a convolution of $s(n)$ and $e(n)$. That is

$$y(n) = s(n) * e(n) \tag{4.9}$$

4.2.2. Identification of resonances

To decompose the composite signal in the frequency domain, it is necessary to first identify the resonances due to the component wavelets. Here we assume that the energy in each basic wavelet is concentrated around specific narrow regions in the frequency domain. To estimate the resonances in a signal, modeling techniques could be used but they are

not very reliable. In linear prediction (LP) approach [Makhoul 1975] the given signal is characterized by an all pole model. LP analysis requires a choice of the model order, but this is not known as we do not have a priori knowledge of the signal. A low model order may miss some of the low energy resonances, whereas a higher order may generate additional spurious peaks. Also, sometimes closely spaced peaks are not adequately resolved by the model based technique.

We make use of a nonmodel based technique using the group delay functions to overcome the above limitations [Yegnanarayana, Duncan, and Hema A. Murthy 1988]. In this approach, the FT magnitude spectrum of the given composite signal is first computed. The inverse Fourier transform of the magnitude function gives a signal whose causal part is minimum phase. The minimum phase signal is then multiplied by a half Hann window of suitable size. As the magnitude and phase spectra of a minimum phase signal are related, the group delay function derived from this phase retains the resonance information that is originally present in the FT magnitude function of the composite signal.

The group delay function thus obtained contains the prominent resonant peaks without the interference of the associated fine structure. Hence it is easy to identify the peaks from this function. In the case of noisy signals a

number of low level spurious peaks may be present. In such a case suitable thresholding has to be done to select only the prominent peaks. Because of the additive and high resolution properties of the group delay function, the effect of each peak is minimal on the adjacent peaks and hence we get more accurate estimates of the resonant frequencies by this method than by the **LP** modeling. Note that this method may not produce higher resolution than the **LP** method. In fact the resolution in the group delay function is still dictated by the window width used on the original **signal** and on the minimum phase signal.

We apply the above method on the composite signal shown in Fig. **4.1a.** This signal consists of resonances at the following frequencies: 800 Hz, 1300 Hz, and 1800 Hz. In Figs. **4.1b, 4.1c** and **4.1d,** the FT magnitude spectrum, a 12th order **LP** smoothed spectrum and the minimum phase group delay spectrum are shown respectively. We observe that the minimum phase group delay spectrum resolves the resonances much better than the **LP** smoothed spectrum. Note that the poor performance of the **LP** method is also due to the fact that the basic wavelets have different resonance characteristics.

4.2.3. Separation of resonances using band pass filtering

In this method, we make use of the Fourier transform domain to separate the resonances. The time

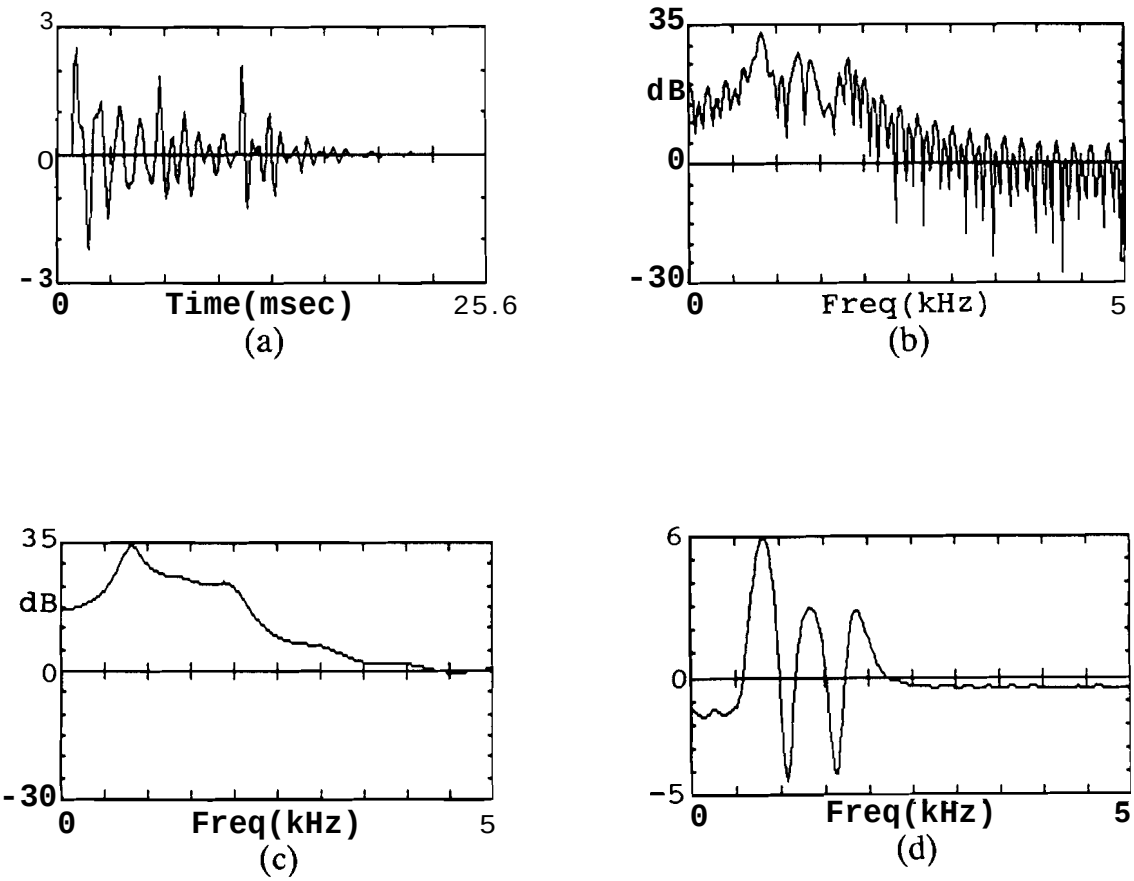


Fig. 4.1. Identification of resonances in a composite signal using minimum phase group delay spectrum: (a) The composite signal; (b) its FT magnitude spectrum; (c) 12th order LP smoothed magnitude; (d) minimum phase group delay spectrum.

signals corresponding to the filtered spectra give epoch information. Here we band pass filter in the FT magnitude spectrum around the selected resonance. The magnitude is set to zero in the unwanted regions and the corresponding time signal is obtained. We observe that the resolution of the epochs is determined by the sharp cutoff window in the spectral domain. The window produces **'ripple'** in the time domain which causes ambiguity in the location of the epochs.

Looking at the spectrum of the composite signal as consisting of an envelope part and a fine structure part, some of the limitations of the above method can be summarized as follows:

(i) The epoch information, which is spread throughout the frequency domain, is partially lost due to the windowing operation. This results in the loss of resolution of epochs in the time domain signal **e(n).**

(ii) Band pass windowing of the spectral envelope introduces ripple in the corresponding time domain function **s(n).** This results in ambiguity about the location of the epoch.

Resolution in both **e(n)** and **s(n)** are thus affected. This causes ambiguity and deformity in the resultant signal **y(n)** which is obtained by convolution of the modified **e(n)** and **s(n).**

4.2.4. Separation of resonances using group delay functions

Simple band pass filtering produces excessive ripple in time domain due to the sharp cut off in the frequency domain. Changing the shape of the frequency window to avoid the sharp cut off causes selection of out-of-band frequencies which interfere with the required frequency. Hence, we need a method where the frequency window must not have sharp cut off, but at the same time there should be no interference from the out-of-band frequencies. These requirements are satisfied if the filtering is done in group delay domain.

We compute the group delay function τ_m from the FT magnitude spectrum of the given composite signal. For group delay filtering, the unwanted regions are to be de-emphasized with respect to the chosen regions. This is done by setting a low value in the unwanted regions. Then the FT magnitude spectrum for the chosen frequency is reconstructed from the filtered group delay function. The FT phase spectrum of the original composite signal is used along with this reconstructed magnitude function to generate the filtered time domain component.

An important issue in group delay filtering is finding the value of the constant which replaces the unwanted regions in the group delay. The value of this constant determines the dynamic range of the filtered signal

in the FT magnitude spectrum. That is the smaller the value of this constant, the larger the dynamic range we get in the magnitude. It is essential to have a large dynamic range to emphasize the selected resonance and to suppress the out of band frequencies. But if the dynamic range is too large, the basic wavelet in the time domain corresponding to the selected resonance may get distorted and affect epoch determination. Hence it is necessary to determine this constant value adaptively based on the information contained in the chosen region. After experimentation with a number of alternatives, we find that a constant value equal to approximately one tenth of the minimum value of the selected portion of the group delay function seem to be adequate.

After group delay filtering is done to separate individual resonances, each of the reconstructed time domain signals contain wavelets of the same frequency. The epochs can be found by peak picking either directly from these time domain signals or from the LP residual derived from them.

Thus the procedure for composite signal decomposition using group delay processing can be described as follows:

(i) Find the FT magnitude and phase spectra of the given composite signal.

(ii) Find the minimum phase group delay function from the FT magnitude spectrum and identify the prominent

resonances that correspond to the basic wavelets.

(iii) Compute the group delay τ_m from the FT magnitude. Select appropriate band limits from step (ii) for filtering wavelets of a particular resonance and set the unwanted regions to a value obtained by using the heuristics described earlier.

(iv) Reconstruct the FT magnitude functions of the component wavelets from the filtered group delay functions.

(v) The time domain signal components for different frequencies are reconstructed from the magnitude spectra from step (iv) and the phase of the original composite signal computed in step (i).

(vi) The epochs are found from the component signals by noting the peak amplitude locations.

The above procedure is applied to the model signal shown in Fig. 4.1a. The group delay function of this signal, given in Fig. 4.2b, does not show the resonant peaks as it is not smoothed, and the fine structure due to epochs is dominant. In Figs. 4.2c and 4.2d the filtered group delay and the reconstructed FT log magnitude for the frequency 1800 Hz are shown. Comparing Fig. 4.2d with the FT log magnitude spectrum of the original composite signal (Fig. 4.2a) is given, we notice that the magnitude function does not contain the out-of-band frequencies in spite of

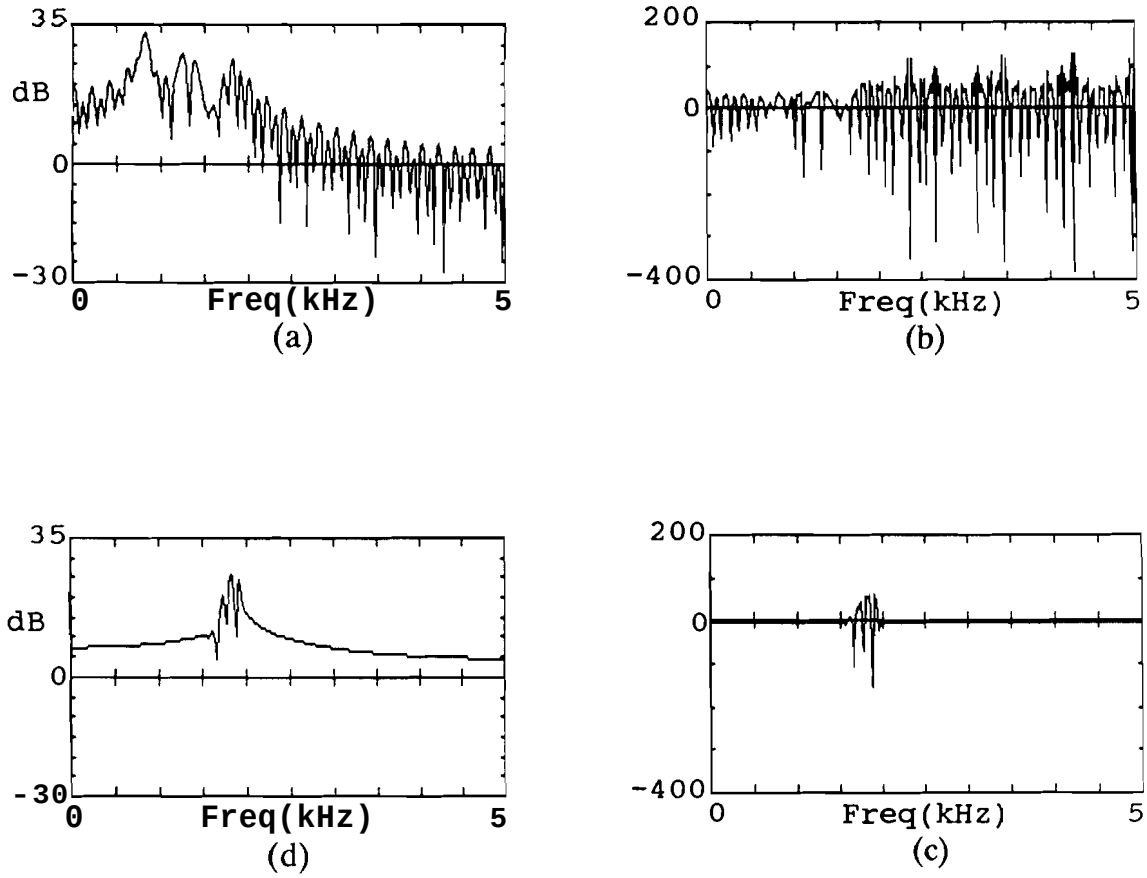


Fig. 4.2. Illustration of the group delay filtering technique: (a) FT magnitude spectrum of the composite signal; (b) its group delay (τ_m) function; (c) group delay filtering to select 1800 Hz component, and (d) reconstructed FT magnitude spectrum.

having a smooth cutoff.

4.2.5. Separation of resonances using weighting functions

Group delay filtering avoids sharp cut off of the spectral envelope component in the frequency domain and hence reduces the corresponding ripple in the time domain. This is clearly an improvement over band pass filtering. But large parts of the epoch information that is spread throughout the spectral domain is removed in the group delay filtering also and smudging of epochs takes place as earlier. We develop a new technique to solve this problem using weighting functions in the spectral domain.

From the above discussion it is evident that the complete spectral fine structure has to be retained to have sharp epochs. However to extract wavelets of a particular frequency, we have to retain the corresponding resonance and suppress all other resonances in the spectral envelope. We propose to use weighting functions in the spectral domain to satisfy this requirement. These weighting functions, when multiplied with the FT, essentially retain the required resonance and suppress the unwanted resonances. The equations for these weighting functions may be represented as follows:

From equations (4.5) and (4.6) we get the z-transform of the spectral envelope component as

$$S(z) = \frac{1}{\prod_{i=1}^{p} D_i(z)} \tag{4.10}$$

A weighting function to retain the k th resonance is formulated as

$$W_k(z) = \prod_{\substack{i=1 \\ i \neq k}}^{p} D_i(z) \tag{4.11}$$

such that when the weighting is applied , we get

$$S(z) \cdot W_k(z) = \frac{1}{D_k(z)} \tag{4.12}$$

Thus the resultant spectral envelope contains only the k th resonance.

$W_k(z)$ is evaluated on the unit circle to obtain $W_k(\omega)$. When $W_k(\omega)$ is used in the spectral domain, the resultant decomposed time signal would contain only the wavelets corresponding to the k th resonance. But, the epochs corresponding to all other wavelets of the composite signal also stand out sharply in the decomposed signal as these are not smudged due to the windowing effects as in the band pass filtering technique. However, these unwanted epochs can be easily removed in the time domain as they are both localized and sharp.

Estimating the parameters of the resonances of the

composite signal is an important issue for designing the weighting functions. LP modeling is not a very reliable technique due to the possibility of either missing closely spaced peaks or generating spurious peaks. Moreover, linear prediction, being an all-pole model, may fail to give good estimates when the composite signal contains poles as well as zeros. These limitations are alleviated considerably by using the minimum phase group delay function. We adopt the following procedure for designing the weighting functions using minimum phase group delay function.

(i) Compute the minimum phase group delay function from the FT magnitude of the given composite signal.

(ii) Threshold the group delay function to retain only the prominent and distinct peaks.

(iii) Invert all the peaks except the peak corresponding to that basic wavelet which is to be filtered from the composite signal.

(iv) Derive the magnitude window function from the modified group delay function of step **(iii)**. Use the algorithm given in section 3.1 assuming that the modified group delay function of step **(iii)** is magnitude group delay.

(v) Derive the phase window function from the modified group delay function of step **(iii)**. Use the algorithm given in section 3.1 assuming that the modified group

delay function of step (iii) is phase group delay.

The weighting function derived by using the above procedure emphasizes the selected peak and flattens the rest of the peaks. We illustrate the results of this procedure in Figs. 4.3 and 4.4. In Fig. 4.3, we illustrate the design of weighting function. Figs. 4.3a and 4.3b show the FT magnitude and minimum phase group delay spectra of the composite signal we have considered earlier. The modified group delay spectrum that emphasizes the 1800 Hz component is given in Fig. 4.3c. The magnitude and phase weighting functions derived from the modified group delay spectrum are given in Figs. 4.3d and 4.3e respectively.

To filter the resonance at 1800 Hz from the composite signal, the weighting functions derived for the purpose are applied on the FT magnitude and phase spectra. This is achieved by multiplying the FT magnitude spectrum with the magnitude weighting function and adding the phase weighting function to the FT phase spectrum. The result of these operations is illustrated in Fig. 4.4. We show the composite signal and its FT magnitude spectrum in Figs. 4.4a and 4.4b respectively. A weighting function is derived from the minimum phase group delay function to pick out only the 3rd resonance as shown in Fig. 4.4c. Fig. 4.4d shows the modified magnitude spectrum after the weighting function is applied. The resulting time domain signal appears in

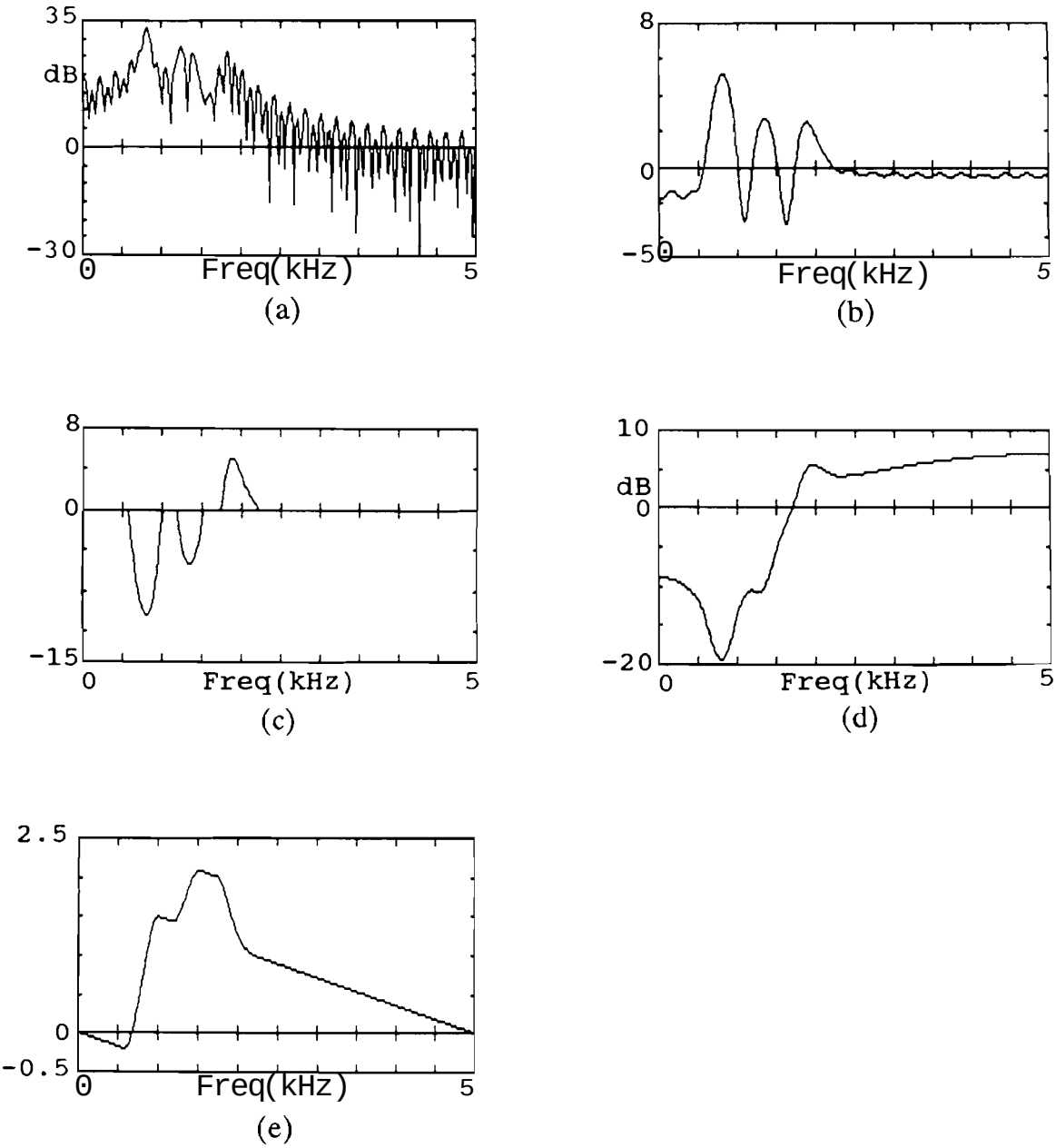


Fig. 4.3. Illustration of the design of the weighting functions from minimum phase group delay: (a) FT magnitude spectrum of a composite signal; (b) the corresponding minimum phase group delay function; (c) modified minimum phase group delay function to filter the 1800 Hz component; (d) magnitude weighting function and (e) phase weighting function derived from the modified minimum phase group delay.

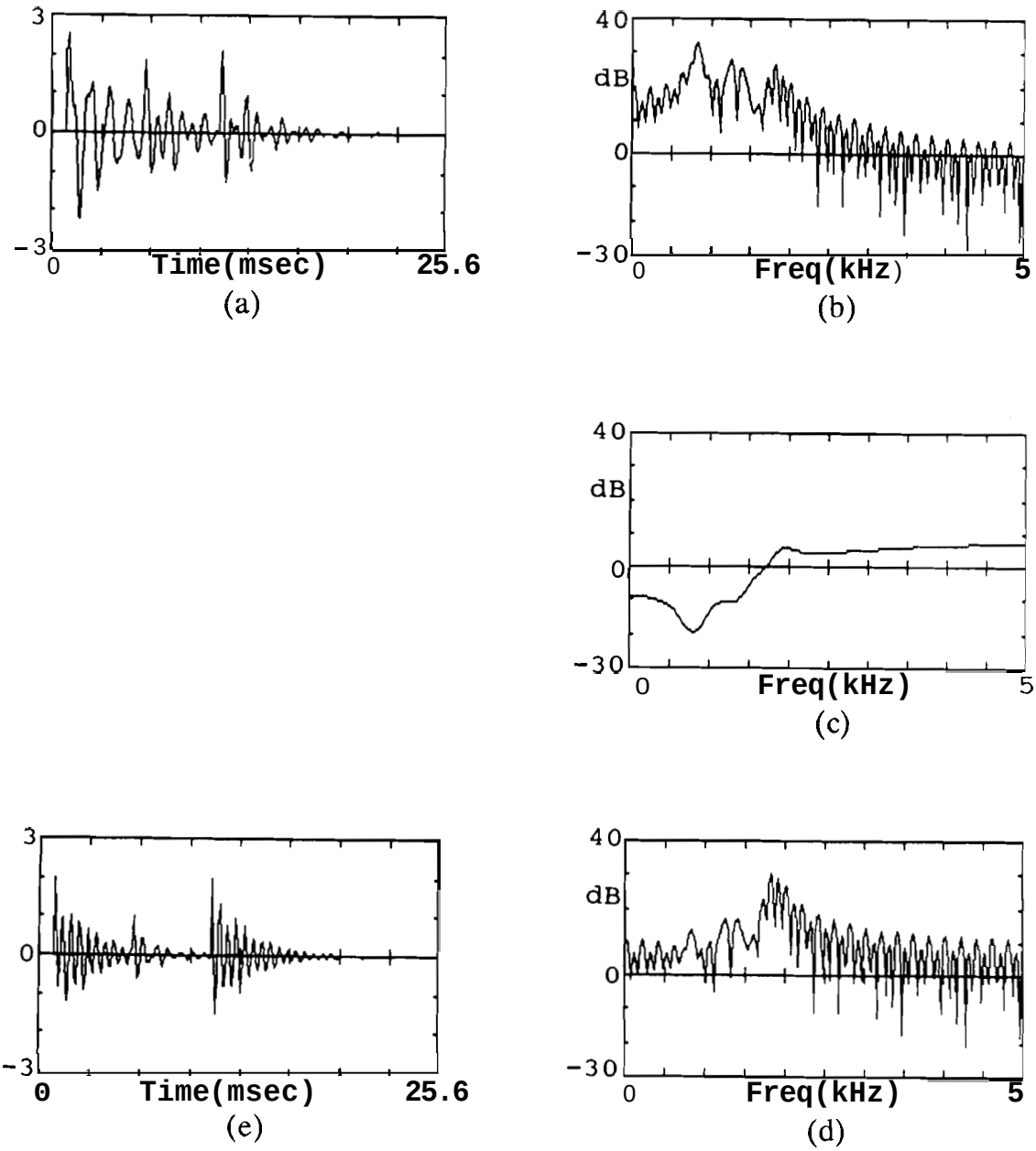


Fig. 4.4. Illustration of application of weighting functions for composite signal decomposition: (a) The composite signal; (b) its FT magnitude; (c) weighting function to filter the 1800 Hz component; (d) modified magnitude after applying the weighting function; (e) the resultant filtered 1800 Hz component.

Fig. 4.4e. We observe, from Fig. 4.4d, that only the selected resonance is emphasized. We also notice that the complete spectral fine structure is retained in contrast to band pass and group delay filtering methods. This prevents smudging of the epochs in the time domain caused by truncation in the frequency domain. Hence we obtain sharp epochs in the time domain as shown in Fig. 4.4e. However, the procedure gives epochs of the unwanted resonances also. But these can be easily identified and removed in the time domain as they are localized.

4.2.6. Experimental results

We used a composite signal of the form given in equation (4.1) where $s_i(t)$ is the i th basic wavelet given by

$$s_i(t) = e^{-\beta t} \sin \omega_i t \qquad (4.13)$$

We have chosen the bandwidths of the damped sinusoids in equation (4.13) to be 10% of their resonant frequencies. The resonances are chosen to be below 5 kHz and the resultant model signal is sampled at 10 kHz. A model signal conforming to the equation (4.1) with the following values for the frequencies and instants (given in sample numbers) of occurrence was chosen :

(i) 800 Hz at 10

(ii) 1300 Hz at 60 and 110

(iii) 1800 Hz at 10 and 110

The amplitudes of all the wavelets are assumed to be unity.

We applied the three decomposition methods described in this chapter on this composite signal. This is illustrated in Fig. 4.5. Figs. 4.5j, 4.5k, and 4.51 show the actual frequency components (i), (ii) and (iii), present in the composite signal. The resonances are identified from 18th order LP spectrum for band pass filtering. The result of band pass filtering to extract the frequency components (i), (ii), and (iii) is shown in Figs. 4.5a, 4.5b and 4.5c, respectively. For the methods using group delay filtering and weighting functions, the resonances are identified using the minimum phase group delay function obtained by choosing Hann window of 30 samples on the minimum phase signal. The results of group delay filtering are given in Figs. 4.5d, 4.5e, and 4.5f. The corresponding plots for the method using weighting functions are shown in Figs. 4.5g, 4.5h, and 4.5i, respectively to extract the frequency components (i), (ii), and (iii).

The reconstructed signals in Fig. 4.5 indicate that the epochs are more sharply defined in the cases of group delay filtering and application of weighting functions

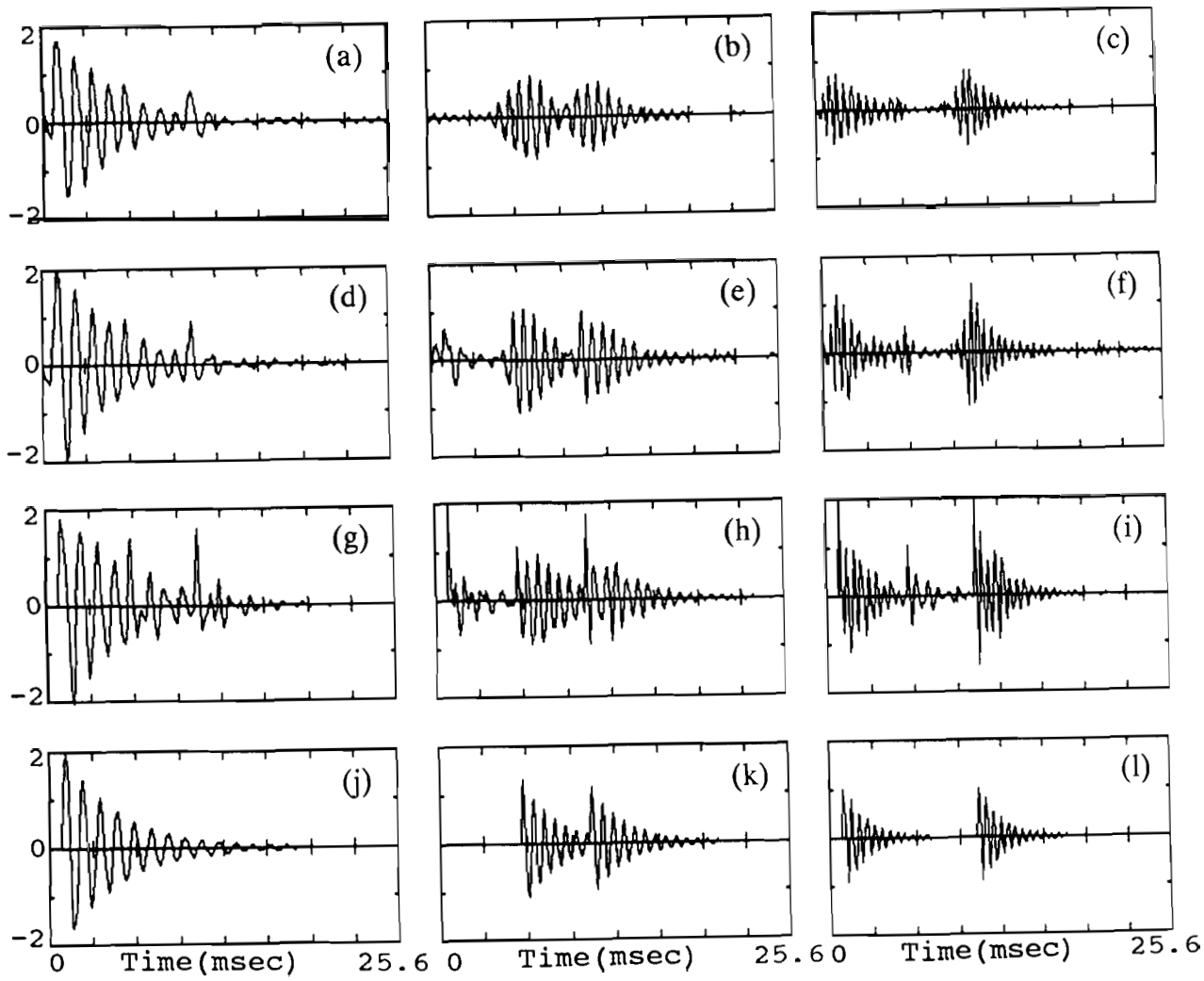


Fig. 4.5. Component wavelets of the composite signal reconstructed by various decomposition techniques: (a)-(c) Component wavelets obtained using band pass filtering; (d)-(f) corresponding plots using group delay filtering; (g)-(i) decomposition using weighting functions; (j)-(l) original component wavelets.

compared to that of band pass filtering. Group delay filtering retains the epoch information of the out of band frequencies also to a large extent as it is contributed by the unfiltered FT phase of the composite signal. However, application of weighting functions also gives all the epochs including those of the out of band frequencies. But here the epochs are much sharper compared to those obtained using group delay filtering. This leads to a more accurate determination of even weak epochs. This may sometimes lead to ambiguity about the presence or absence of the particular frequency wavelets at the indicated epochs. In such cases, simple band pass filtering may be used as an initial step to find an approximate location of the wavelets, and then group delay filtering or weighting function processing is performed to sharply define the epochs.

We have experimented with decomposing noisy composite signals. The main problem here when using conventional approaches is determination of the resonances. Linear prediction performs poorly in the presence of noise and LP smoothed spectrum contains spurious peaks. Some peaks may be missed altogether. **Using** minimum phase group delay processing approach, we get more reliable resonance information. We conducted experiments to decompose the same composite signal used in Fig. 4.1 after adding Gaussian noise at 3 dB SNR. The results are illustrated in Fig. 4.6.

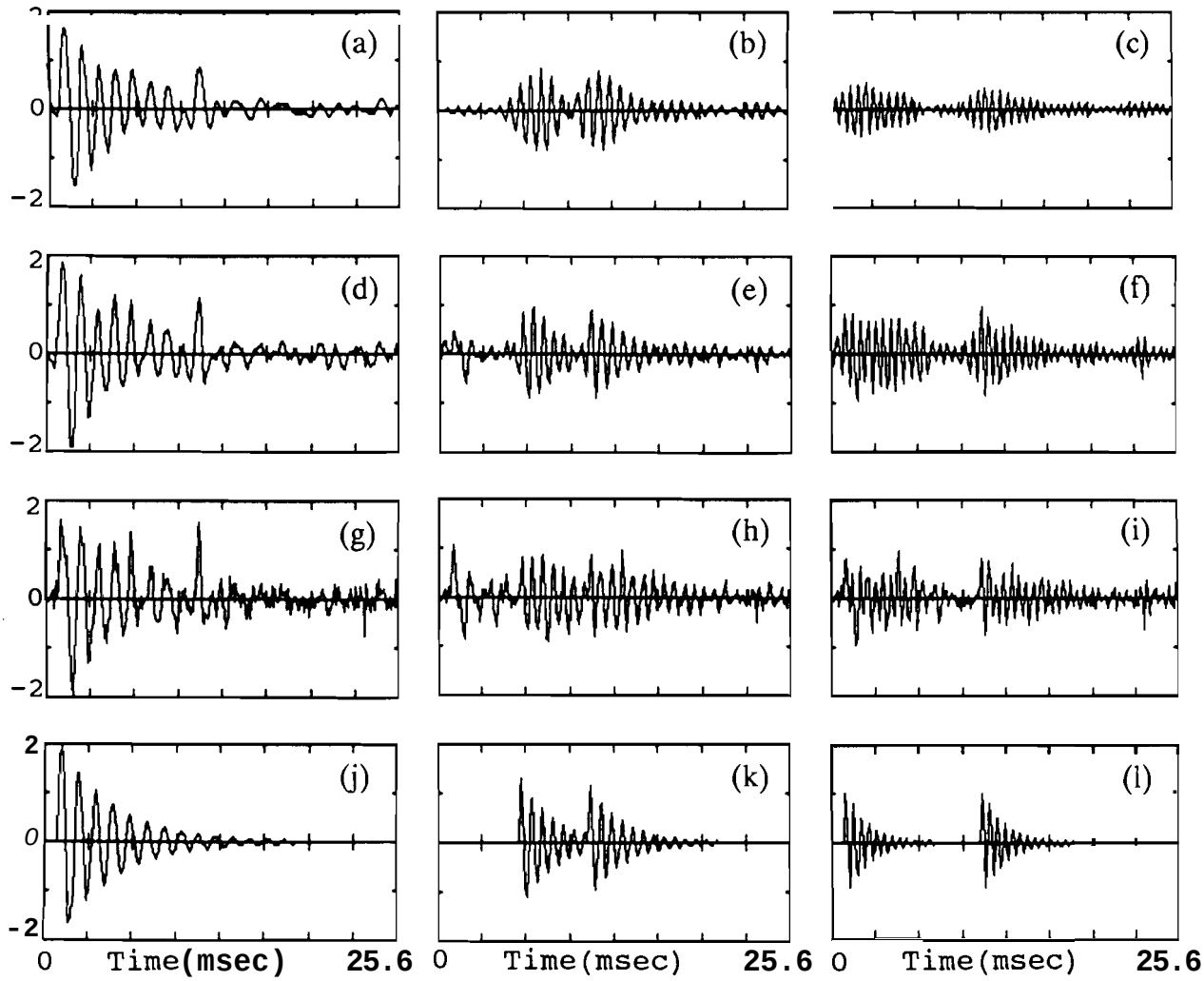


Fig. 4.6. Component wavelets of the noisy (SNR=3dB) composite signal reconstructed by various decomposition techniques: (a)-(c) Component wavelets obtained using band pass filtering; (d)-(f) corresponding plots using group delay filtering; (g)-(i) decomposition using weighting functions; (j)-(l) original component wavelets.

The results of group delay filtering are shown in Figs. **4.6d, 4.6e,** and **4.6f.** We observe that the epochs are much sharper in the group delay filtering than in the band pass filtering (Figs. **4.6a, 4.6b,** and **4.6c)** as in the case of clean signal. Weighting function method also gives sharp epochs as seen from Figs. **4.6g, 4.6h,** and **4.6i** for the three components. But in this case the spurious spikes due to noise are also present and careful interpretation is necessary.

4.3. Discussion

We have studied the time-frequency resolution problem in the case of Fourier transform and group delay representations. We have compared their performance in the context of composite signal decomposition in this Chapter. We find that the band pass filtering in the FT magnitude spectrum suffers from the effects of truncation and consequent time-frequency resolution problem. This problem is less severe in the group delay spectrum due the additive and high resolution properties of the group delay functions. Thus group delay processing facilitates effective identification and separation of component wavelets in a composite signal. They also help in designing suitable weighting functions to decompose the composite signal. We notice that group delay filtering performs well even in the

presence of noise. LP and other model based methods cannot be used for this purpose due to their limitations such as problems with merger of closely spaced peaks and/or spurious peaks. In addition, model based methods require a priori knowledge of the signal and noise statistics to determine the type and order of the model. Group delay processing, being nonmodel based, does not suffer from such limitations. Hence, group delay processing is an effective nonparametric analysis technique when a priori knowledge about the composite signal is not available.

Chapter5

ANALYSIS OF TIME VARYING SIGNALS: TECHNIQUES TO DEAL WITH BANDWIDTH CHANGES

Time-frequency resolution issues under quasistationary assumption were studied in the previous chapter. These were studied in the context of composite signal decomposition. Practical signals are, in general, time varying in nature. Hence there is a need to study and develop techniques to address resolution problem for nonstationary cases also. In this chapter, we address problems encountered in processing time varying signals such as speech. Specifically we study the problem of detecting source-tract interaction in speech.

A speech signal is characterized by a set of resonances called formants in the frequency domain. In the temporal domain, the voiced segments of speech are characterized by pitch periodicity. The instants of glottal closure and opening influence the speech signal. Sharp energy bursts mark the occurrence of consonant sounds. Thus the important parameters for speech analysis are formant frequencies in the frequency domain, and pitch and other epochs in the temporal domain. This suggests that speech can be considered as a special type of composite signal. Thus speech analysis is equivalent to decomposition of a

composite signal to find various resonances and epochs in the frequency and time domains, respectively.

Specifically, we consider the problem of detecting the epoch due to glottal opening within a pitch period in the speech signal as this is related to source-tract interaction. Under stationary assumption, we have to process short segments of speech signal for this purpose. This leads to time-frequency resolution problem. Hence parametric modeling approaches are of limited value. We propose new techniques to tackle this problem in this chapter. We consider time varying formulation where we try to detect system changes within the analysis frame. As in the last chapter, we study these issues using model signals so as to evaluate our techniques quantitatively.

We introduce the source-tract interaction problem in Section 5.1 and explain its importance in speech synthesis. In Section 5.2, we show that group delay function is useful in detecting change in bandwidth in simple model signals. More complex model signals are considered in Section 5.3. We discuss limitations of the proposed methods for processing natural signals in Section 5.4.

5.1. Source-Tract Interaction and Its Importance in Speech Synthesis

Various types of speech processing involve parametric modeling of speech. Most speech analysis systems

are based on the linear source-filter model [Rabiner and Schafer 1978]. In general, it is assumed that the source and vocal tract models are independent of each other. Though this assumption results in computational simplicity, it is not strictly correct. Hence, when these models are used in speech analysis-synthesis systems, they tend to generate synthetic speech with a distinct machine-like quality.

Factors responsible for quality in speech signal are to be identified and incorporated in the synthesis algorithm to obtain natural sounding speech. For example, the formant frequencies and bandwidths of the vocal tract change within a pitch period in the voiced speech segments. This effect is known as source-tract interaction. Though source-tract interaction is conjectured to be important for high quality speech reproduction, reliable methods to identify and measure it from speech signals are not available. The predominant effect in the vocal tract system within a pitch period due to source-tract interaction is the change in bandwidth of the first formant at the instant of glottal opening. In the glottal open phase, the bandwidth of the first formant increases due to loading effect of the subglottal system. Source-tract interaction can be detected if we can identify this change in the bandwidth of the first formant. This information can be used for high quality speech synthesis. We use model signals in these studies.

We consider the problem of identifying bandwidth changes in the following types of composite signals:

(i) A simple wavelet whose bandwidth changes abruptly after a time interval.

(ii) superposition of multiple wavelets, where bandwidth of one of the wavelets changes abruptly after a time interval.

5.2. Identification of Change of Bandwidth in Simple Model Signals

We consider the problem of identifying the change in the bandwidth of a damped sinusoid (basic wavelet). First we show analytically that change in damping produces a ripple in the frequency domain. Next we consider the use of group delay functions to identify this change.

5.2.1. Analysis of a model signal incorporating bandwidth changes

Let us consider a basic wavelet with a frequency ω_o and damping constant β. The equation of this signal can be written as

$$x(n) = e^{-\beta n}\sin\omega_o n \qquad (5.1)$$

This signal can be considered to be the product of two separate components, b(n) and s(n), where

$$b(n) = e^{-\beta n} \qquad (5.2)$$

and

$$s(n) = \sin \omega_o n \tag{5.3}$$

The signals s(n), b(n) and x(n) are illustrated in Figs. 5.1a, 5.1b and 5.1c respectively. We follow the procedure outlined below to analyze the frequency structure of the damped sinusoid given in equation (5.1).

As x(n) is the product of b(n) and s(n) in time domain, $X(\omega)$, the Fourier transform of x(n) can be obtained by the convolution of $B(\omega)$ and $S(\omega)$, where $B(\omega)$ and $S(\omega)$ are the Fourier transforms of b(n) and s(n) respectively. Convolution with $S(\omega)$ merely shifts the roots of $B(\omega)$ by ω_o since $S(\omega)$ is the Fourier transform of a sinusoid. Hence $X(\omega)$ and $B(\omega)$ contain similar root structure except that the roots of $X(\omega)$ are shifted by ω_o with respect to the roots of $B(\omega)$. Thus studying the spectral structure of $B(\omega)$ gives an idea of the structure of $X(\omega)$.

We compute B(z), the z-transform of b(n), as

$$B(z) = \frac{1}{1 - e^{-\beta} z^{-1}} \tag{5.4}$$

$B(\omega)$ can be obtained by evaluating B(z) on the unit circle.

From equation (5.4) we observe that the z-transform of b(n) has a single pole on the real axis and no zeros. Hence we can conclude that spectrum of x(n) has just

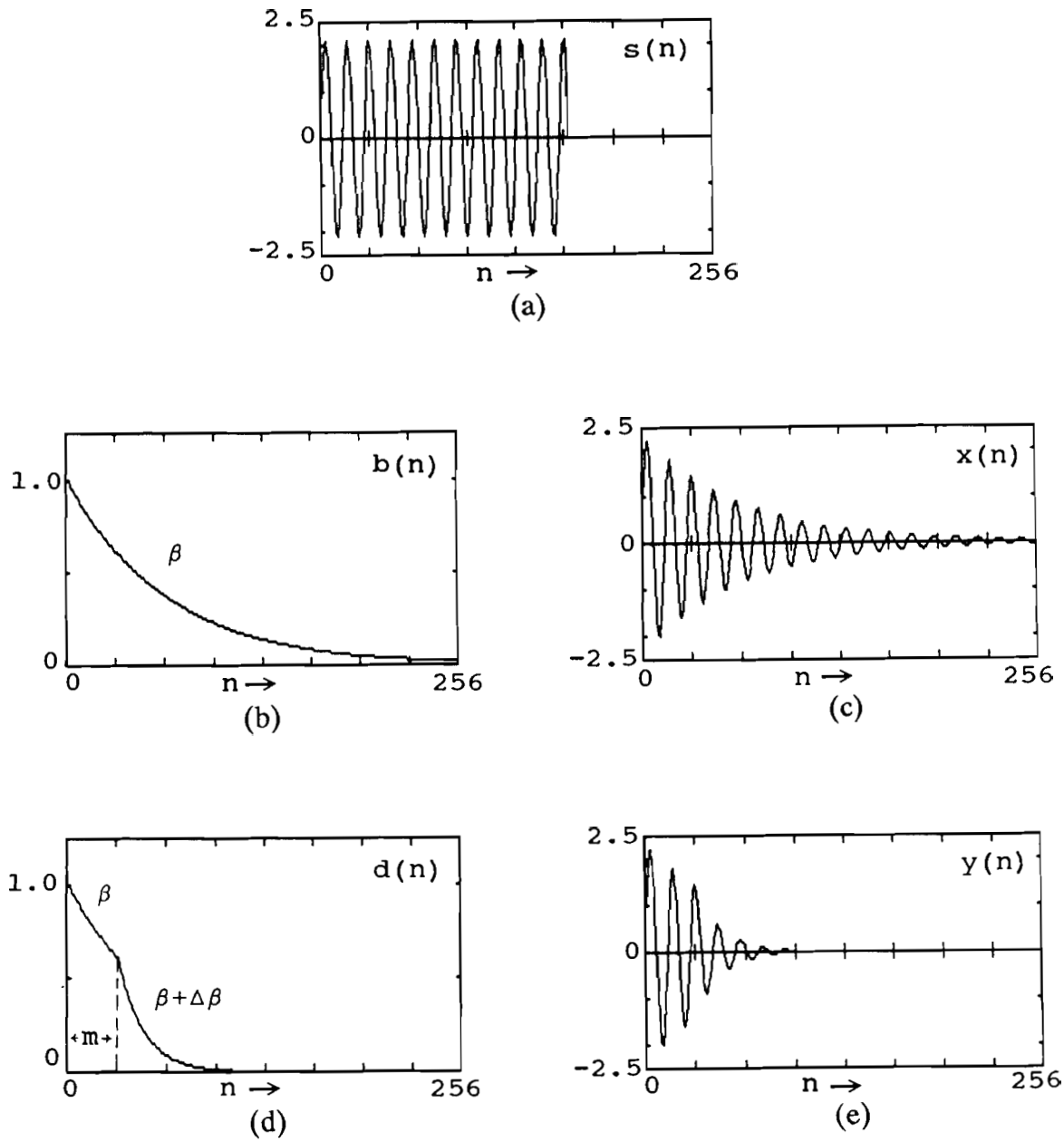


Fig. 5.1. Illustration of the components of a damped sinusoid with and without damping change.

one resonance at ω_o.

Consider another signal **y(n)** that is identical to **x(n)** except that at m **th** sample the damping constant has changed from β to $\beta + \Delta\beta$ as shown in Fig. **5.1e.** The equation for this signal can be written as in the earlier case. This signal also may be considered to be the product of two separate components, namely the damping component **d(n)** (shown in Fig. **5.1d)** and the sinusoidal component **s(n).** The equation for the damping component is given as

$$d(n) = e^{-\beta n - \Delta\beta (n-m)}\, u(n-m) \qquad (5.5)$$

where **u(n)** is a unit step function. **D(z),** the z–transform of **d(n)** can be computed as

$$D(z) = \sum_{n=0}^{a3} e^{-\beta n - \Delta\beta (n-m)}\, u(n-m)\, z^{-n}$$

$$= \sum_{n=0}^{m-1} e^{-\beta n}\, z^{-n} + \sum_{n=m}^{a3} e^{-\beta n - \Delta\beta (n-m)}\, z^{-n}$$

$$= \frac{1 - e^{-\beta m}\, z^{-m}}{1 - e^{-\beta}\, z^{-1}} + \frac{e^{-\beta m}\, z^{-m}}{1 - e^{-(\beta+\Delta\beta)}\, z^{-1}}$$

$$= \frac{e^{-\beta(m+1)}\,(e^{-\Delta\beta} - 1)\, z^{-(m+1)} - e^{-(\beta+\Delta\beta)}\, z^{-1} + 1}{(1 - e^{-\beta}\, z^{-1})\,(1 - e^{-(\beta+\Delta\beta)}\, z^{-1})} \qquad (5.6)$$

We get the Fourier transform **D(ω)** by computing the z-transform **D(z)** on the unit circle. The Fourier transform, Y(ω) of the original signal **y(n)** can be considered to be the result of convolution of **D(ω)** by **S(ω),** where **S(ω)** is the

Fourier transform of **s(n)**. As **s(n)** is a sinusoid, this operation merely results in frequency translation of the spectrum of **D(ω)** by ω_o in **Y(ω)**.

Thus it is seen that **y(n)** contains **(m+1)** zeros distributed in the z-plane, in addition to the poles at $\pm\omega_o$. It is seen from equation (5.6) that the effect of these zeros depends on the extent of damping change. That is, if the change in the damping is more, the zeros move towards the unit circle and hence their effect becomes more predominant in the spectrum. These zeros, which are the result of change in damping, introduce variations in the spectrum which can be seen as a ripple. This ripple is prominent when the zeros are prominent.

5.2.2. Use of the group delay function for identification of change in the bandwidth

We find that ripple is introduced in the spectrum when there is a damping change in the time domain of the basic wavelet. We also observe that the frequency of the ripple, which corresponds to the number of zeros in the z-plane, indicates the instant of damping change; The strength of the ripple gives the extent of damping change. One approach to identify the time instant where the system changes, is to measure the resulting ripple in the FT magnitude spectrum. But here the ripple caused by the

distributed zeros is multiplicative and hence difficult to observe. Moreover, the amplitude of the ripple is dependent upon the extent of bandwidth change that takes place at the m th instant. If this change is small, then the ripple is weak and is difficult to observe.

We show that the group delay function can be effectively used for the detection of the ripple due to the bandwidth change. Let x(n) be a composite signal with a resonant frequency f_o = 1000 Hz, and an initial damping constant β = 0.1 in the first 16 samples. At the 17 th sample the damping constant changes to 0.2; that is $\beta+\Delta\beta$ = 0.2. The signal x(n) is uniformly sampled at 10 kHz. This signal is shown in Fig. 5.2d. The log magnitude and group delay function $\tau_m(\omega)$ derived from the log magnitude are shown in Figs. 5.2e and 5.2f. The corresponding plots for another signal y(n) whose resonance frequency and damping constant are 1000 Hz and 0.1 without any change in damping are shown in Figs. 5.2a, 5.2b and 5.2c for comparison. We observe that the change in damping has introduced a ripple in the log magnitude and group delay functions. We also observe that this ripple is more pronounced in the group delay function due to the special properties of the group delay functions. Frequency of this ripple is equal to the number of zeros in the z-plane, which in turn equals the time instant at which the system change

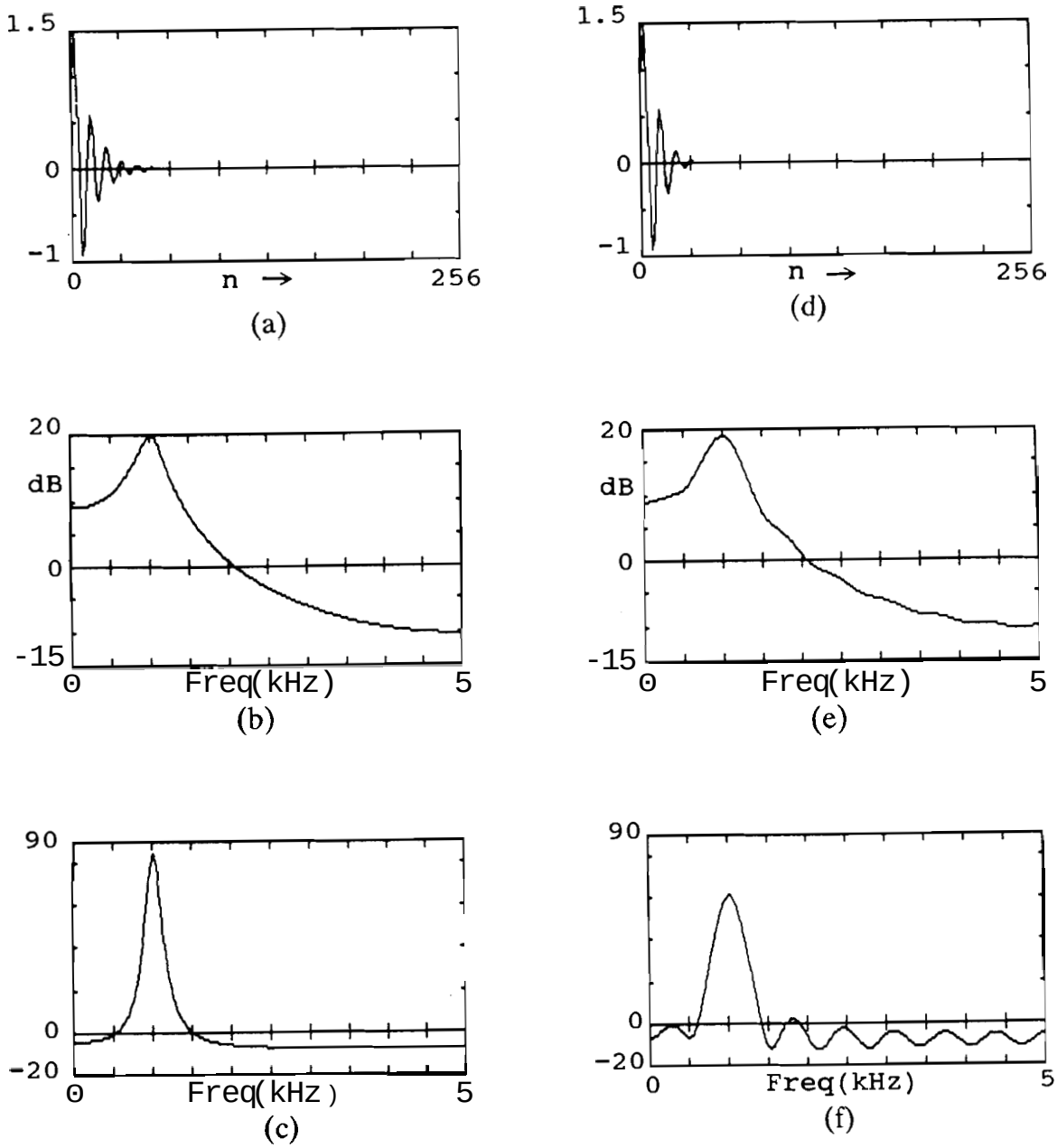


Fig. 5.2. **Illustration** of the appearance of ripple in the group delay function when there is change of damping in the signal: (a)-(c) Wavelet without damping change, its FT magnitude and group delay τ_m respectively; **(d)-(f)** corresponding plots for the same **wavelet** with damping change at 8th sample.

occurs. variation of the ripple with respect to the instant and extent of system change is depicted in Fig. 5.3. We observe that the larger the system change, the larger would be the amplitude of the ripple.

5.3. Detection of Source-Tract Interaction in Model Signals

We have shown in the previous section that it is possible to identify change in bandwidth within the analysis frame in the case of simple model signals such as damped sinusoids. We have observed that the plot of the group delay function shows ripple in such cases. In this section, we consider a model signal consisting of k resonances in which there is a change in the bandwidth in one of the resonances within the analysis frame. This signal approximates a speech signal containing the effects of source-tract interaction. We show this model signal, its log magnitude and group delay function in Figs. 5.4d, 5.4e and 5.4f respectively. Here the bandwidth of the lowest resonance is increased by 50% at sample number 17. The corresponding plots for the same model signal without the bandwidth change are shown in Figs. 5.4a, 5.4b and 5.4c for comparison. We observe from Fig. 5.4f that the ripple in the group delay function, though present, is not apparent for this signal. This is due to the presence of a number of resonances, which mask the ripple. To overcome this difficulty, and to detect the change in bandwidth even

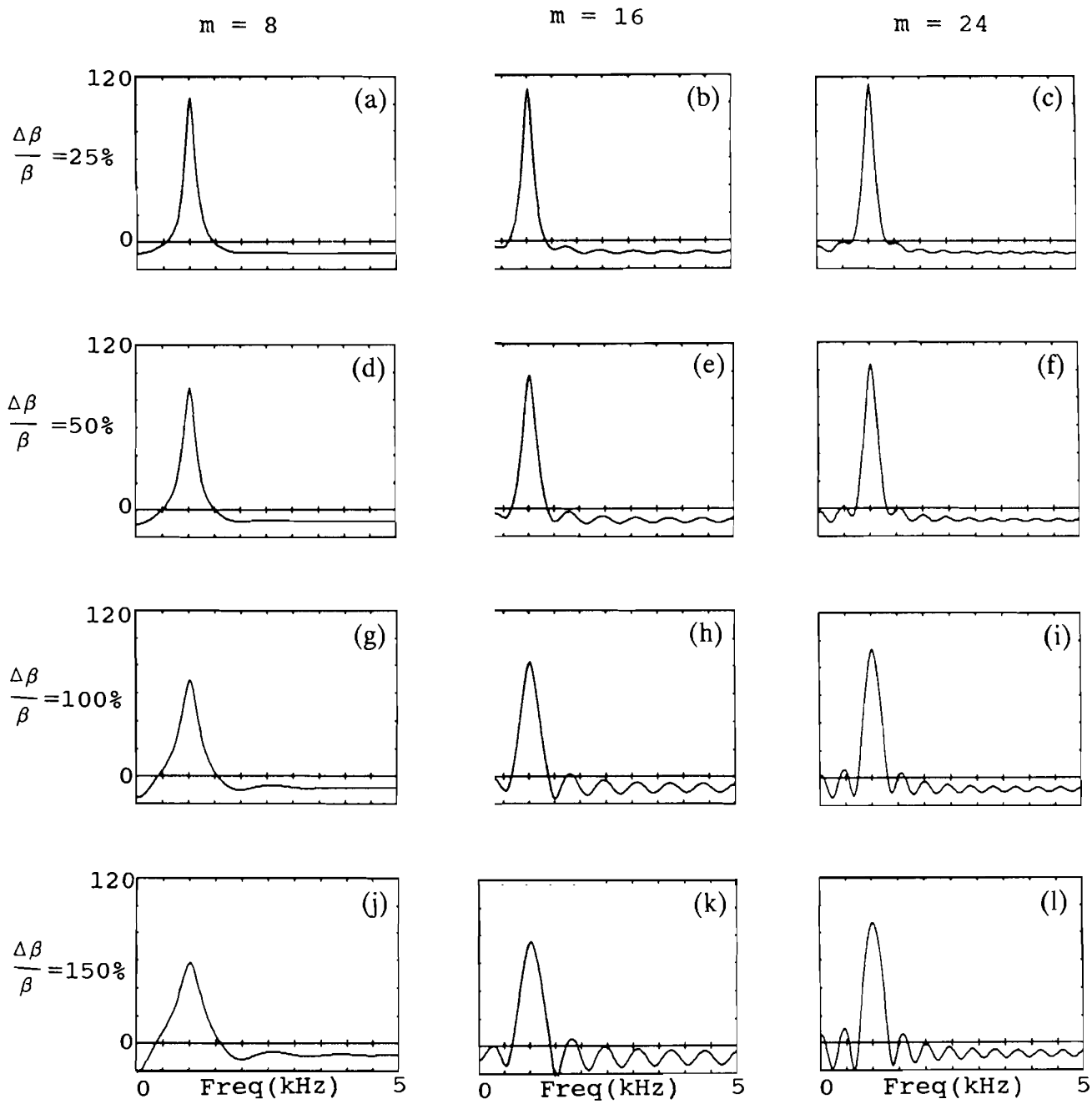


Fig. 5.3. Illustration of the effect of changing the extent ($\Delta\beta/\beta$) and the instant (m) of damping change on the group delay function of the wavelet.

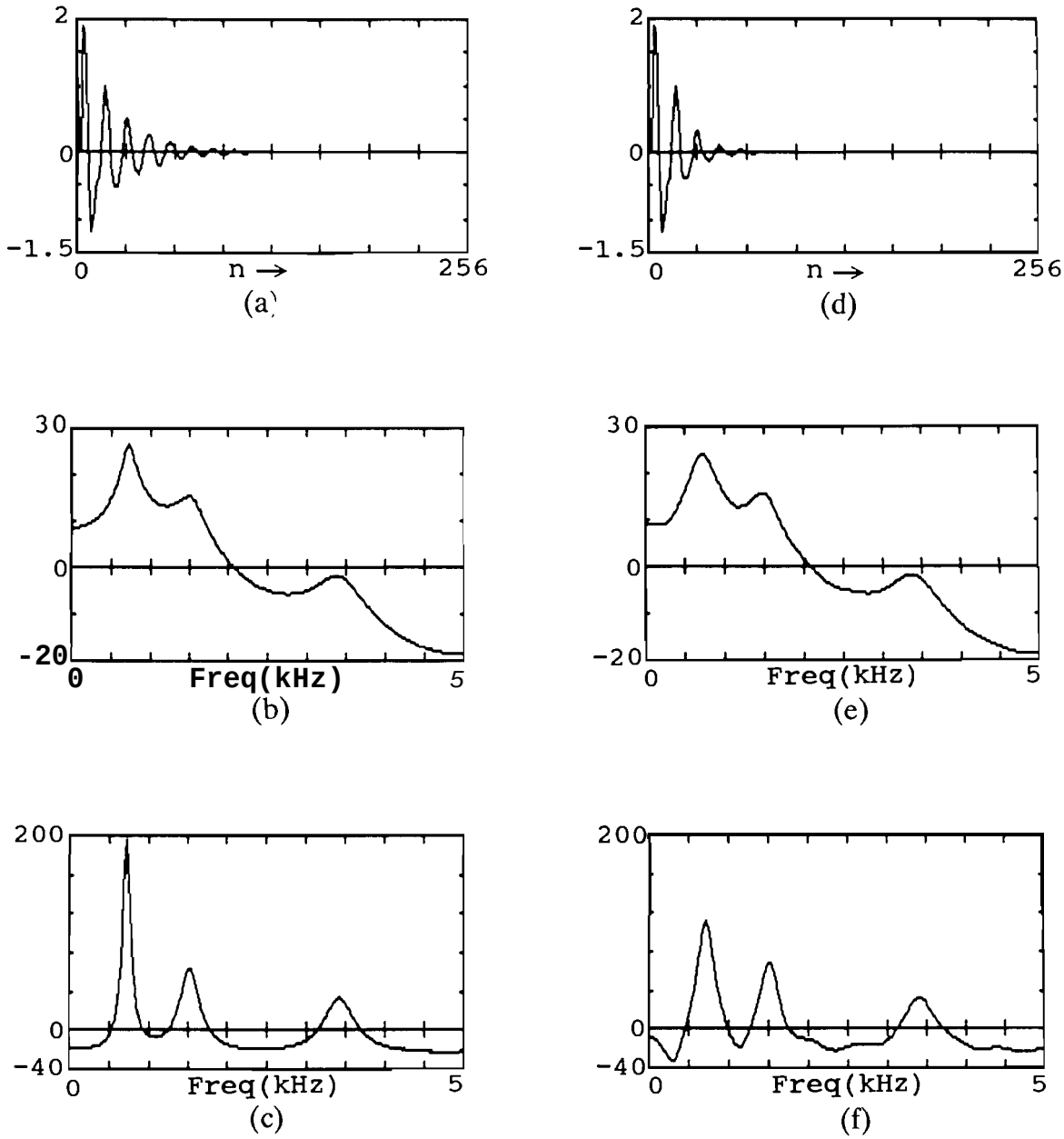


Fig. 5.4. Identification of bandwidth change in a model signal with 3 formants: (a)-(c) Model signal without damping change, its FT magnitude and group delay T_m respectively; (d)-(f) corresponding plots for the same model signal when the bandwidth of the 1st formant changes at 16th sample.

in such cases, we present a technique that utilizes the impulse response of the inverse filter.

5.3.1. Inverse filter based method

We give below the basis for the proposed method. Consider the time signal shown in Fig. 5.4d. This can be looked upon as the impulse response of a system whose characteristics change at the sample number 17. We try to detect this change from the impulse response of another system whose transfer function is inverse of the transfer function of the original system. This is shown analytically as follows:

Let us consider a signal with k resonances. Let the bandwidth of one of the resonances change at the sample number m in the time domain. The z-transform of this signal can be written as

$$V(z) = \frac{1}{\prod_{i=1}^{k-1} D_i(z)} \cdot Y(z) \qquad (5.7)$$

The first term on the right hand side of the above equation is the product of the transfer function of the second order systems due to the (k-1) resonances with constant bandwidths. The second term is due to the resonance where change of bandwidth takes place. Using equation (5.6), Y(z)

can be written as

$$Y(z) = \frac{N(z)}{D(z)} \tag{5.8}$$

where $N(z)$ and $D(z)$ are the contributions due to the zeros and poles respectively in the z-plane.

Hence we can write equation (5.7) as

$$V(z) = \frac{N(z)}{D(z) \prod_{i=1}^{k-1} D_i(z)} \tag{5.9}$$

Here the roots of the numerator part $N(z)$, contain the change of bandwidth information (the instant and extent of bandwidth change).

One approach to extract $N(z)$ is to form an inverse filter of the form

$$A(z) = D(z) \prod_{i=1}^{k-1} D_i(z) \tag{5.10}$$

$N(z)$ can be obtained using this inverse filter.

$$N(z) = V(z) \cdot A(z) \tag{5.11}$$

This approach requires finding $A(z)$ which is difficult. We propose an alternate approach where we need not find the inverse filter explicitly. Consider the components of $V(z)$ in equation (5.9). Due to the m symmetric zeros in $N(z)$, its inverse z-transform $e(n)$ would have significant values only

at n = 0 and at n = m. But due to convolution with the infinite impulse response (IIR) filter 1/A(z), it is not possible to identify the epoch instant m from x(n), as the response of the filter for excitation at n = 0 overlaps with the response for excitation at n = m. We form the reciprocal of V(z) to highlight the epochs as shown below.

$$W(z) = \frac{1}{V(z)} = \frac{D(z)\prod_{i=1}^{k-1} D_i(z)}{N(z)} \qquad (5.12)$$

Consider the nature of the time signal w(n), which is the inverse z-transform of W(z). This can be analyzed by studying the components of W(z). The term 1/N(z) consists of symmetric poles, and its inverse transform $e_1(n)$ has similar epochs as that of e(n). The response of the filter A(z) is confined around the epochs of $e_1(n)$ as it has finite impulse response (FIR). If m is greater than the order of the FIR filter A(z), then there is no overlap of the responses of the epochs of $e_1(n)$ in w(n), and hence it is possible to detect the epoch m from w(n).

5.3.2. Experimental results

We have conducted experiments to detect the bandwidth change in the model signal shown in Fig. 5.4d. We follow the procedure given below:

(i) Find the Fourier transform $V(\omega)$, of the given signal $v(n)$.

(ii) Compute $W(\omega)$ as the reciprocal of $V(\omega)$.

(iii) Find the inverse Fourier transform $w(n)$ of the inverted spectrum $W(\omega)$.

The plots of $v(n)$ and $w(n)$ are shown in Figs. 5.5a and 5.5b respectively. We find that the epoch at $n = 0$ in $w(n)$ is very strong compared to the epoch at $n = m$. Hence we have clipped the impulse response at $n = 0$ and suitably scaled the plot in Fig. 5.5b to observe the impulse response at $n = m$. We see from the plots that it is possible to detect the instant of bandwidth change m, using the method described in this section. The relative amplitude of the impulse response at $n = m$ is directly related to the extent of change in damping. However we note that the epoch at $n = m$ is weak even for a large change in bandwidth compared to the epoch at $n = 0$. Hence this method is not suitable in the case of real signals in which case the epoch due to bandwidth change may get lost in the presence of extraneous peaks.

5.4. Summary

Detection of changes in bandwidths of component wavelets in a composite signal has important applications in speech processing. The effect of source-tract interaction in

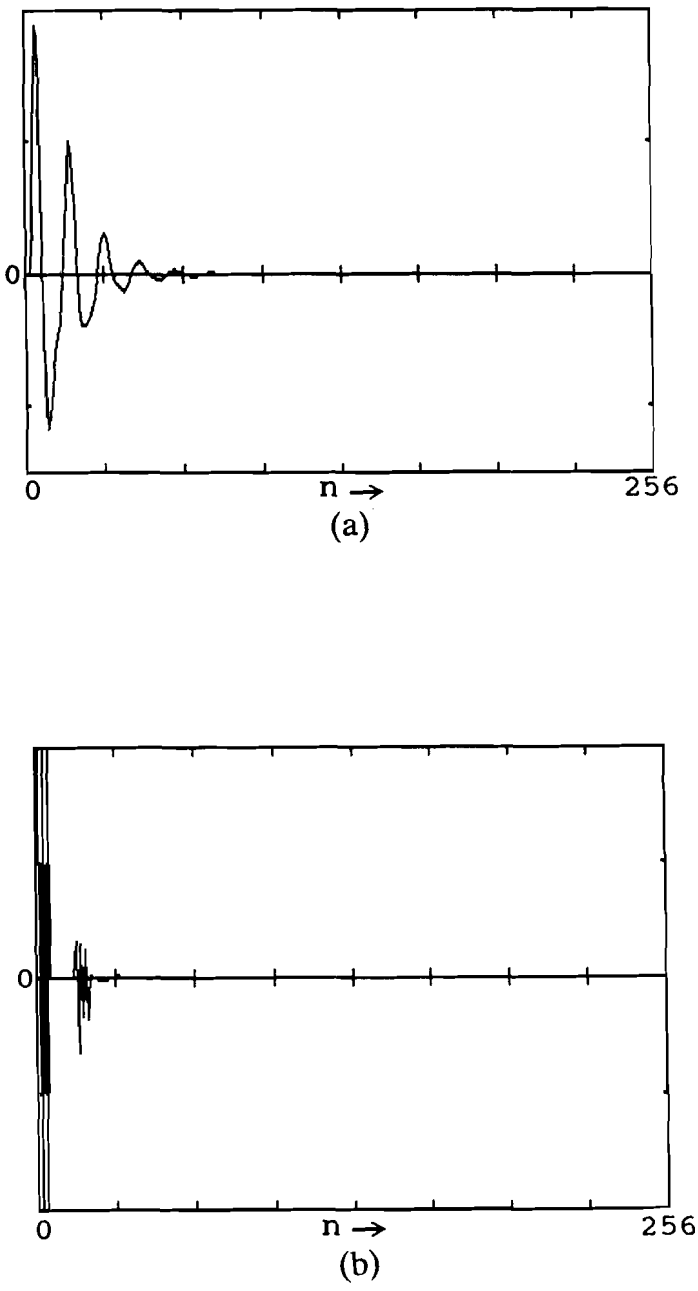


Fig. 5.5. Detection of the instant of damping change using inverse filter based method: (a) A model signal with 3 formants where the bandwidth of the 1st formant changes at the 16th sample; (b) corresponding inverse filter signal.

speech, which is essentially change in the bandwidths of the resonances of the vocal tract system within a pitch period, is conjectured to be related to the natural quality of the speech signal. Thus, if the source-tract interaction is detected in speech analysis, and if its effects are incorporated in speech synthesis, we may be able to generate more natural sounding speech. The main problem in detecting the change in the bandwidths of the formants within a pitch period is the short data length available for processing. In the quasistationary formulation, we may have to analyze speech segments whose length is less than a pitch period. This leads to time-frequency resolution problem and the results of the analysis may not be reliable.

In this chapter, we have outlined techniques for detection of source-tract interaction under a time varying formulation. That is, we do not assume stationarity within the analysis frame. We expect the signal to change within the analysis frame, and try to detect such changes. We have used model signals in our investigations. We have shown that a ripple is introduced in the group delay function of a simple model signal when there is a change in bandwidth. By finding the frequency and amplitude of this ripple, we get the instant and extent of bandwidth change. However in the case of complicated model signals, this ripple is masked by other spectral features and further investigations are

necessary to find a method for detecting this ripple. We have proposed a method based on inverse filtering in such cases.

We have considered only model signals in our studies. Detecting source-tract interaction in speech signals is a more complicated problem. Further investigations are necessary to tackle this problem. We take up this problem in the next chapter where we present a modified LP method which indirectly captures the effects of the source-tract interaction.

Chapter 6

ANALYSIS OF NATURAL SPEECH SIGNALS: MODIFIED LINEAR PREDICTION METHOD

In the previous chapter, we presented methods to detect the effect of source-tract interaction for model signals. In this chapter we describe an analysis technique which captures the effects of source-tract interaction in speech so that the resulting parameters may be used for synthesizing a more natural sounding speech. The standard LP analysis under quasistationary formulation can at best capture only the gross spectral features. This is due to the relatively large analysis window. Short data analysis methods are limited due to the time-frequency resolution problem. Hence there is a need to develop new analysis techniques that capture accurately the spectral variations in speech.

The chapter is organized as follows: First we explain the need for short data analysis and its limitation. We give the basis for a new method called modified linear prediction, and describe the method. We also give analytical formulation for the proposed method and discuss the implementation details. Next we describe experiments performed for speech analysis and synthesis, and present the results. Finally we discuss the effectiveness and

limitations of the modified linear prediction method for speech processing.

6.1. Need for Nonstationary Formulation of Speech

Quasistationary formulation is one of the standard approaches for speech analysis. In this formulation, we assume that the vocal tract system is stationary in each analysis frame. The analysis is performed over frames of 10-20 millisecond intervals. We obtain formant information which gives the average behavior of the system over the entire analysis interval using this method. This assumption simplifies analysis as we get a mathematical formulation involving only linear equations. This is the case with linear prediction (LP) analysis.

6.1.1. Need for short data analysis

Quasistationary formulation for speech analysis has its limitations. It fails to capture the dynamic variations of the continuously changing vocal tract system and excitation. Even within a pitch period, the source and the system change due to the effect of source-tract interaction.

Many attempts have been made to incorporate this source-tract interaction feature in synthesis to improve speech quality. Since it is difficult to apply LP analysis

for very short data records, Fant (1982) simulated the source-tract interaction effect by suitably modifying the glottal pulse shape used for excitation of the LP model in the synthesis. He introduced ripple in the glottal pulse to simulate source-tract interaction. But in these studies the source-tract interaction is not derived from the signal but is assumed. Moreover, the vocal tract system described by the LP model is not varying within the analysis frame due to the quasistationary assumption.

In this chapter we attempt to capture the variations in the system characteristics due to source-tract interaction that occur within a pitch period . As the source and system are interrelated, any effect in the source has a corresponding effect on the system. To capture these variations, ideally, it is necessary to process the speech signals on an instant-to-instant basis. That is we are forced to consider analysis frames as small as 2-3 rnsec. Thus short data record analysis becomes necessary.

6.1.2. Limitations of short data analysis

Short data analysis suffers from time-frequency resolution problem. When the analysis frame is very small (less than 3 msec) the window effects dominate. That is, if we take short segments of speech directly, then the estimates of the correlation parameters become poor.

If we truncate the data, then the estimated correlation coefficients are biased and their variances become large. Hence the LP coefficients computed from these correlation coefficients are not accurate. Adjacent samples are highly correlated in speech signal. But at the edges of the analysis frame the last sample .is followed by zero due to truncation. This leads to errors in the estimation of the correlation parameters. In the case of covariance coefficients stability of the resulting all-pole system is not guaranteed. In addition, the variance of the estimates becomes large even here when we consider short data length.

These problems have arisen mainly due to the quasistationary assumption. We cannot use longer analysis frames to get better estimates as the local variations cannot be captured in that case. Using short segments of speech for analysis leads to the poor estimation problem as mentioned earlier. Hence there is need to consider alternate formulations for speech analysis. We explore merits of nonstationarity assumption in the next section.

6.2. Techniques for Speech Analysis under Nonstationarity Assumption: A Modified Linear Prediction Approach

Analysis methods using nonstationary formulation were discussed in Section 2.4.1 of Chapter 2. These methods are computation intensive and/or the analysis parameters

depend on a set of time functions. There is a need for simpler and general analysis technique.

We propose a method of analysis of short data records which will enable us to process nonstationary segments of speech. Using this method, it is possible to obtain a better estimate of the autocorrelation coefficients for short data records. These coefficients can thus be used to estimate the all-pole model corresponding to the vocal tract system for the region. Thus it is possible to capture the changes in the vocal tract system during production of speech.

6.2.1. Basis for the proposed method

A major problem in estimating the model parameters over short segments of speech signals is the high correlation between the samples. We propose a modified linear prediction method to deal with this problem. Our method is based on the fact that for a given data length the accuracy of estimation depends on the correlation between the samples, especially for large lags. If the correlation between the samples is small, then a short data record is adequate to obtain a good estimate of the autocorrelation coefficients; We have no control on the correlation between speech samples. But we do have some control on the correlation between samples in the LP residual. By using a

higher order model, it is possible to derive an **LP** residual signal where the sample-to-sample correlation is very small. We describe the modified linear prediction method which exploits the above mentioned property of the **LP** residual to obtain better estimates for the model parameters.

6.2.2. Modified linear prediction method

To analyze a short (< 3 msec) speech segment of M samples, the modified linear prediction method uses standard autocorrelation based linear prediction analysis in two phases. In the first phase of the analysis, the first **(p+1)** autocorrelation coefficients $R_s(k)$, of the speech signal are estimated over a large (10-30 msec) frame of N samples centered around the short data record of interest. These coefficients pick up the gross features of the spectral envelope of the signal. We compute the **LP** coefficients and obtain the **LP** residual for this frame. In the second phase, we estimate the autocorrelation coefficients $R_e(k)$, of the **LP** residual corresponding to a short data record within the large frame chosen above. These coefficients capture the local variations of the vocal tract system. This is possible because in addition to source information, the residual contains system changes that are not captured by the **LP** filter. We modify the first set of autocorrelation coefficients $R_s(k)$ using $R_e(k)$ to obtain modified

autocorrelation coefficients R(k). R(k) can be computed from $R_s(k)$ and $R_e(k)$ by convolving them in the time domain. Alternatively the spectrum of R(k) can be obtained by multiplying the spectra of $R_s(k)$ and $R_e(k)$ in the frequency domain from which R(k) can be found. We show later in this chapter that convolution is a better choice for implementation. New LP coefficients can be obtained from R(k). These coefficients give a better representation of the dynamic vocal tract system over the short analysis frame than the standard LPCs.

6.2.3. Procedure for implementing modified linear prediction method

Let s(n) be the short segment of speech of M samples centered around the sample point L. It is required to compute the modified linear prediction coefficients for s(n). To obtain the gross spectral characteristics we consider a large frame of N (>>M) samples y(n), centered around the point L as shown in Fig. 6.1.

Steps leading to the computation of modified linear prediction coefficients are as follows:

(i) Find the autocorrelation $R_s(k)$ of y(n) as

$$R_s(k) = \frac{1}{N} \sum_{n=0}^{N-1} y(n)\, y(n+k), \quad k = 0, \pm 1, \ldots, \pm p \qquad (6.1)$$

where p is the order of linear prediction. Compute

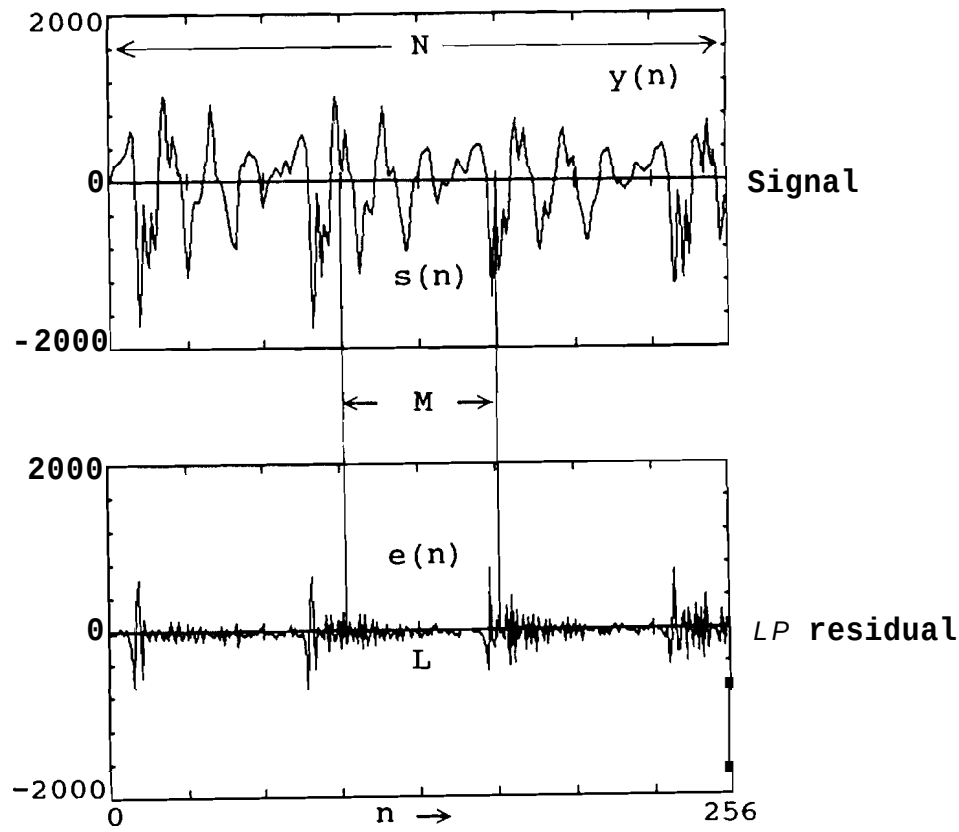


Fig. 6.1. Illustration of the short and large data frames used in the modified linear prediction method.

the standard LP coefficients a(i), i = 1, 2, ..., p, from $R_s(k)$ using Durbin's recursion.

(ii) Use the corresponding LP residual e(n) in the short data frame to compute $R_e(k)$ as

$$R_e(k) = \frac{1}{M} \sum_{n=0}^{M-1} e(n)\, e(n+k), \quad k = 0, \pm 1, \ldots, \pm(M-1) \qquad (6.2)$$

(iii) Extrapolate $R_s(k)$ to $\pm q$ values where $q > p+M-1$, using the LP coefficients a(i)'s

$$R_s(k) = \sum_{i=1}^{p} a(i)\, R_s(k-i), \quad k = \pm(p+1), \pm(p+2), \ldots, \pm q \qquad (6.3)$$

This is done to reduce truncation effects in the subsequent convolution.

(iv) Convolve $R_s(k)$ and $R_e(k)$ to get the modified autocorrelation R(k).

$$R(k) = \sum_{i=-(M-1)}^{M-1} R_e(i)\, R_s(k-i), \quad k = 0, \pm 1, \ldots, \pm p \qquad (6.4)$$

(v) Compute modified LP coefficients from the modified autocorrelation coefficients using Durbin's recursion.

6.2.4. Implementation issues

We have outlined a procedure to compute the modified autocorrelation coefficients R(k) in the last

subsection. An alternate procedure for computing R(k) which may be computationally less costly is by multiplying the power spectra of the LP model of the signal over the large analysis frame and the corresponding LP residual over the short analysis frame. R(k) can be obtained from the product spectrum. But this procedure leads to aliasing in the time domain. Zero-padding in the frequency domain is not a practical solution for this problem. Hence we chose convolution in the time domain to get R(k).

We consider the accuracy and computational overhead involved in obtaining R(k). The computational overhead depends on the lengths of the sequences $R_s(k)$ and $R_e(k)$ and the number of lags for which R(k) has to be computed. To obtain accurate R(k), the sequences $R_s(k)$ and $R_e(k)$ must not be truncated. For a short data record of M samples we get $R_e(k)$ of size (2M-1) samples. When we take these autocorrelations and convolve, we get an R(k) which is closer to the true autocorrelation sequence. This is because it is equivalent to the inverse discrete Fourier transform of a spectrum derived from multiplying two power spectra, since there is no aliasing in the direct computation by convolution . As R(k) is also an autocorrelation, stability of the resulting model is guaranteed.

Accuracy of the values of R(k) depends on the

accuracies of the values of $R_s(k)$ and $R_e(k)$. For large lags near k = M, $R_e(k)$ is a poor estimate of the true autocorrelation due to fewer samples (1 or 2) used in the computation of $R_e(k)$. Due to the uncorrelated nature of the LP residual, correlation values at higher lags quickly taper off to zero for most of the frames. But in those frames where the values for higher lags are significant, this could cause instability problems in the LP filter computed from R(k). If $R_e(k)$ is truncated to avoid the values for large k, then the resulting $R_e(k)$ is not a complete autocorrelation sequence and hence may result again in instability problems in the computation of LP coefficients from R(k). If $R_e(k)$ is extended by using an LP filter of order much smaller than M, the size of $R_e(k)$ may become large and indefinite, since even a small correlation in $R_e(k)$ can produce large number of significant extended samples. This also increases the computational overhead. Another alternative is to use covariance of the short segment of the residual. This can also cause instability problems. We have chosen a truncated version of $R_e(k)$ for our studies. To reduce the possibility of instability due to truncation effect, we multiply this $R_e(k)$ with a Bartlett window before computing R(k) through convolution. Though this does not guarantee stability, we have not encountered any problems due to this in our experimental studies that are described in the next section.

6.3. Experimental Results

We have experimented with natural speech data to capture the dynamic variations of the spectrum using the modified LP method. We can see from Fig. 6.2 that significant values of the autocorrelation function of the LP residual are generally concentrated around the zero lag. Figs. 6.2a and 6.2b show a segment (256 samples) of speech signal and its LP residual, respectively. The autocorrelation of the speech signal over a frame of 256 samples is shown in Fig. 6.2c. The autocorrelation of the LP residual over a short frame of 32 samples corresponding to region (A) in Fig. 6.2b is given in Fig. 6.2d. We observe that due to low correlation between the samples in the LP residual, the higher lags taper off to zero quickly. Hence the window effects which essentially affect the higher lags are small in the autocorrelation estimate of the LP residual.

The following procedure is used to obtain the modified autocorrelation coefficients at the instant of interest:

(i) Select a long frame of 256 samples around the instant of interest in the given speech signal.

(ii) Compute the autocorrelation function $R_s(k)$ for p lags after applying a double sided Hamming window to the

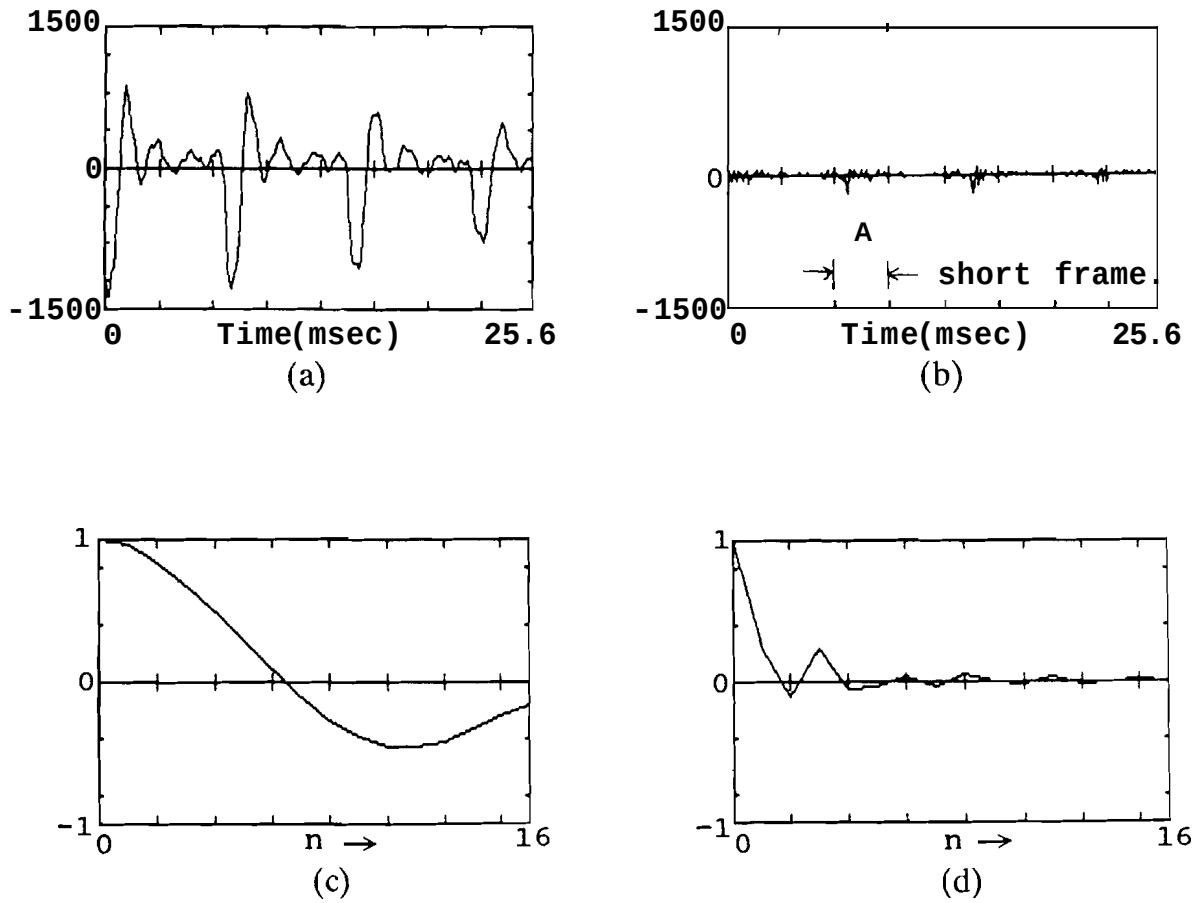


Fig. 6.2. Plots showing (a) speech signal; (b) its LP residual; (c) autocorrelation of the speech signal and (d) auto-correlation of a short frame (3.2 msec) of LP residual.

signal.

(iii) Compute the LP coefficients and the LP residual.

(iv) Choose a short (1-3 msec) frame of M samples of LP residual around the instant of interest.

(v) Compute the autocorrelation $R_e(k)$ of the short data frame for all lags (M-1).

(vi) Extrapolate the autocorrelation function $R_s(k)$ computed in step(ii) from p to p+M-1 lags using the LP filter. (This is done to avoid window effects in the convolution).

(vii) Convolve $R_e(k)$ from step (v) and $R_s(k)$ from step (vi) to obtain the modified autocorrelation function R(k) for p lags.

(viii) Compute the modified LP coefficients from R(k) using Durbin's recursion.

6.3.1. Experiments in speech analysis

We have conducted experiments to analyze speech to capture time-varying spectral features. We have used the procedure outlined above to obtain the modified LP coefficients in these experiments. We show in Fig. 6.3 the continuous variation of the spectral information even over short segments.

The speech signal, its autocorrelation $R_s(k)$ and its LP spectrum for order 12 are shown in Figs. 6.3a, 6.3b

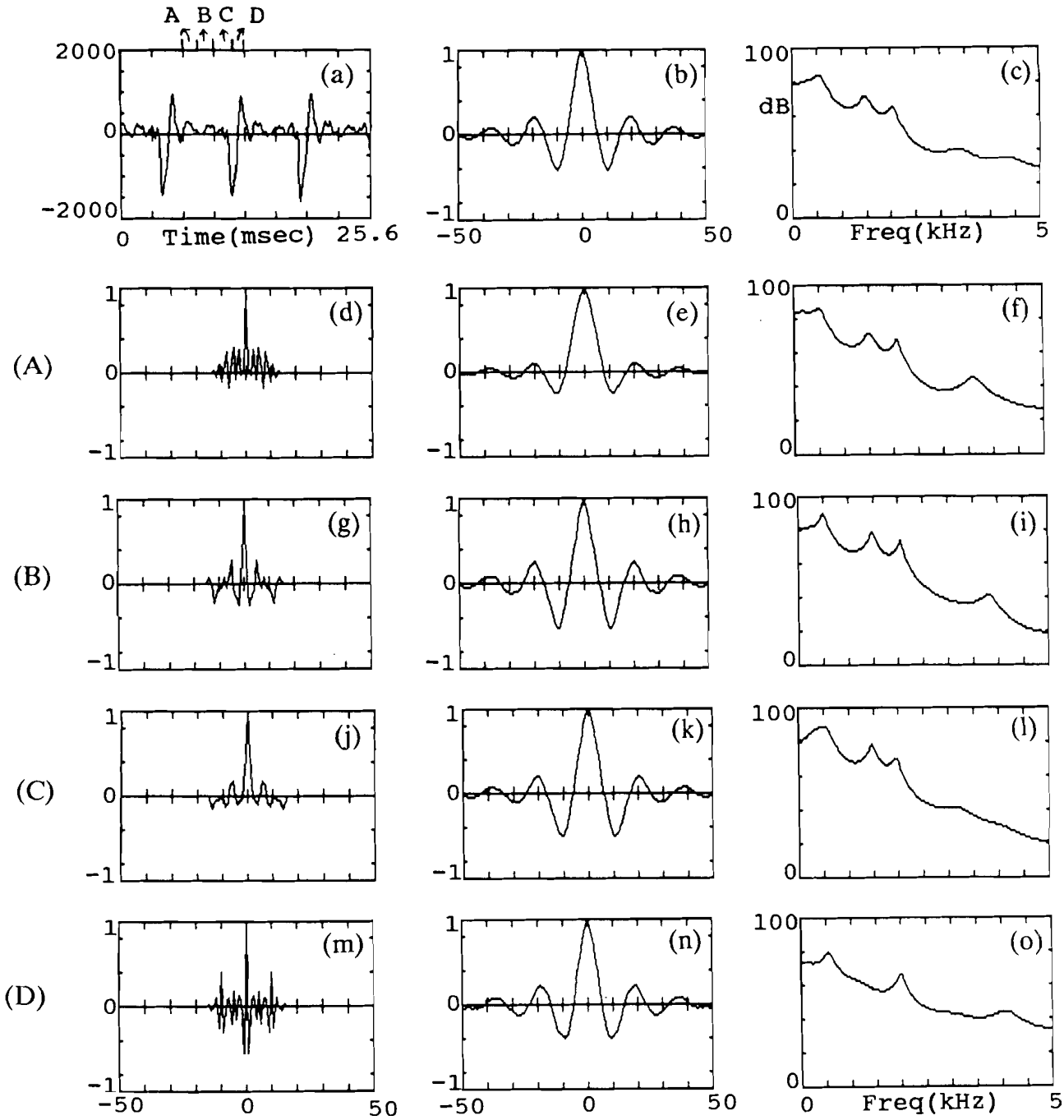


Fig. 6.3. Illustration of the effectiveness of modified linear prediction in capturing local variations in short segments: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.3a), the corresponding modified autocorrelation and LP spectrum; (g)-(i) (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 1.6 msec (16 samples).

and 6.3c respectively. A few consecutive short data frames of 16 samples each labeled as (A), (B), (C) and (D) are chosen from the speech signal in Fig. 6.3a for which we find the modified LP spectra. The autocorrelation $R_e(k)$ of LP residual from frame (A), the .corresponding modified autocorrelation R(k) and the modified LP spectrum are shown in Figs. 6.3d, 6.3e and 6.3f respectively. The corresponding plots for frames (B), (C) and (D) are shown in Figs. 6.3g - 6.3o. We observe that the modified autocorrelation and LP spectrum in all the cases are only slightly different from the corresponding plots for standard autocorrelation and LP spectrum shown in Figs. 6.3b and 6.3c. The slight differences we notice are due to the local effects caused by the dynamic variation of the vocal tract in the speech signal. We notice that these differences cause essentially slight shifts in the bandwidths and frequencies of the formants. In Figs. 6.4 and 6.5 we illustrate the same plots for short segments of size 24 and 32 samples respectively. We observe the above mentioned dynamic variation in all these plots. These variations may be due to source-tract interaction among other things.

The speech signal chosen in all these experiments is from a voiced region. We show that similar results are obtained in the case of signals from unvoiced and transition regions. We investigated whether a transition from an

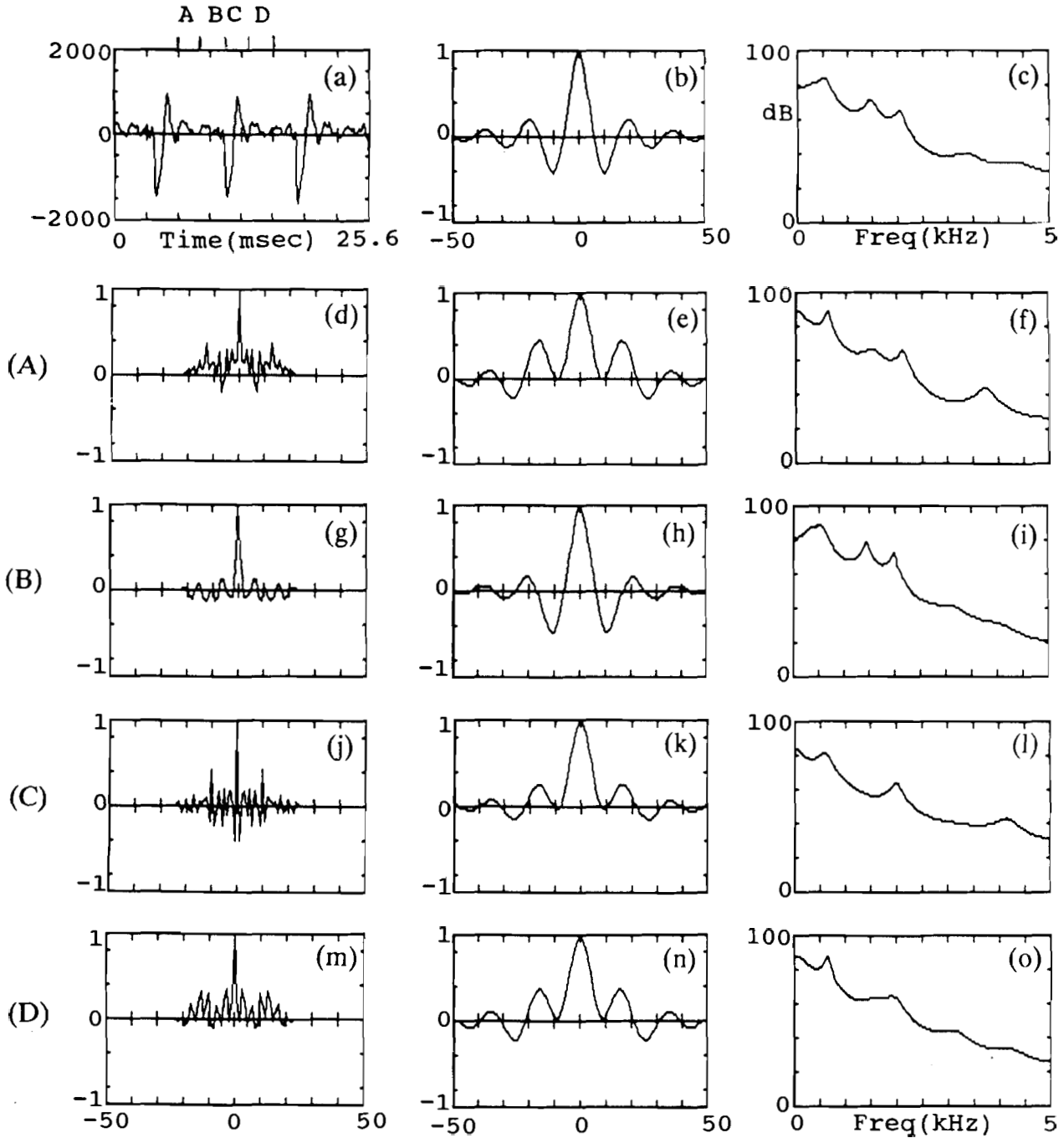


Fig. 6.4. Illustration of the effectiveness of modified linear prediction in capturing local variations in short segments: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.4a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 2.4 msec (24 samples).

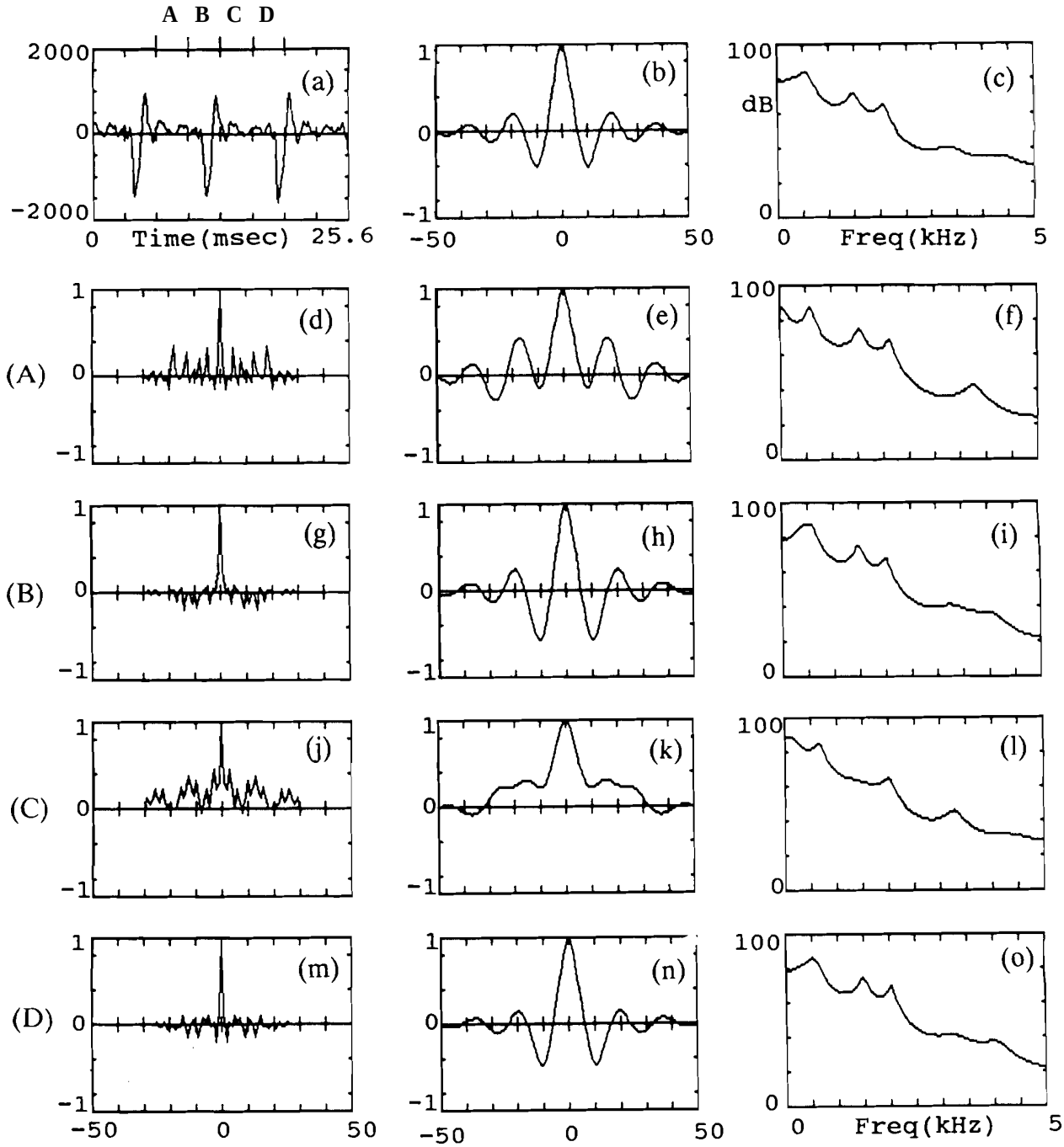


Fig. 6.5. Illustration of the effectiveness of modified linear prediction in capturing local variations in short segments: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.5a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 3.2 msec (32 samples).

unvoiced to voiced region in a speech signal is captured by the modified linear prediction method. This is illustrated in Fig. 6.6. The transition frame taken from the sound 'Sha' is shown in Fig. 6.6a. Its autocorrelation and LP spectrum are shown in Figs. 6.6b and 6.6c. We have plotted the autocorrelation of the LP residual, the modified autocorrelation and the modified LP spectrum corresponding to short segments (A), (B), (C), and (D) in Figs. 6.6d - 6.6f, 6.6g - 6.6i, 6.6j - 6.61, and 6.6m - 6.60, respectively. The short segments are taken from different portions of the transition frame as shown in Fig. 6.6a in such a way that segment (A) is in the unvoiced portion, segments (B) and (C) are in the transition portion and segment (D) is in the voiced portion. We see from the plots of modified LP spectra that the dynamic range of the modified LP spectrum corresponding to short segment (D) is distinctly larger than that of the modified spectrum corresponding to segment (A). This indicates that modified LP method has captured the local variations at unvoiced to voiced transitions.

We have conducted experiments to see whether rapid formant transitions in the voiced portions of the speech signal can be captured using the modified linear prediction method. We have taken a transition frame corresponding to the sound 'We'. This sound starts with the vowel sound /u/

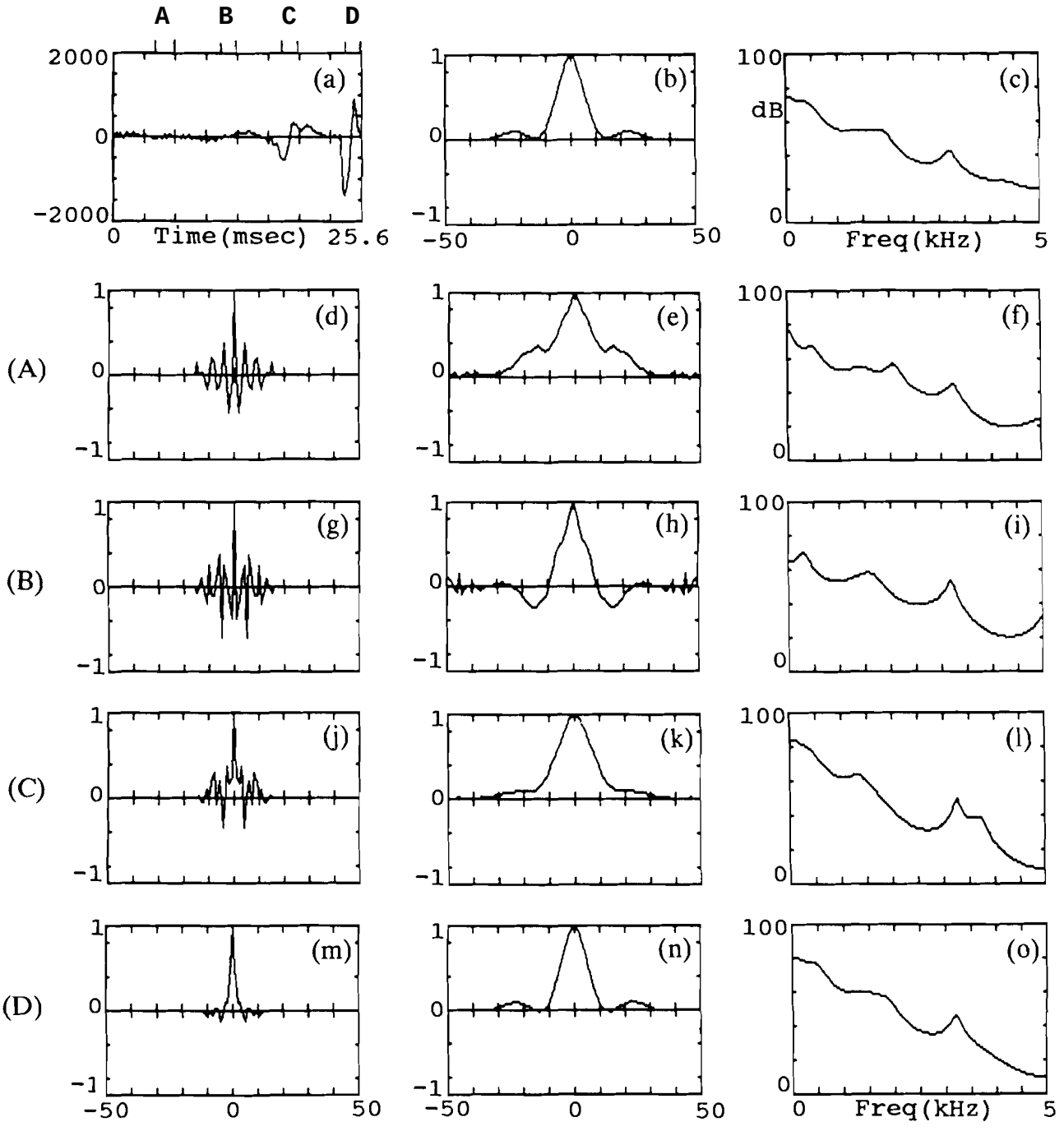


Fig. 6.6. Illustration of the effectiveness of modified linear prediction in capturing local variations in a frame containing an unvoiced to voiced transition: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment **A** (indicated in Fig. 6.6a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(1) and (m)-(o) show similar plots for the short segments B, **C** and **D** respectively. The short segment size is 1.6 msec (16 samples).

and ends with the vowel sound /i/. We show the plots of the transition frame, its autocorrelation, and LP spectrum in Figs. 6.7a, 6.7b, and 6.7c, respectively. We have chosen the short segments (A), (B), (C) and (D) from the transition frame shown in Fig. 6.7a such that (A) is in the initial portion of the frame where the sound /u/ is expected to be dominant, (B) and (C) are in the transition portion and (D) is in the final portion where the sound /i/ is expected to be dominant. We plot the autocorrelation of the LP residual, modified autocorrelation and modified LP spectrum corresponding to short segments (A), (B), (C) and (D) in Figs. 6.7d - 6.7f, 6.7g - 6.7i, 6.7j - 6.71, and 6.7m - 6.7o, respectively. We observe from the modified LP spectrum of segment (A) that the first and second formant peaks are nearer indicating the presence of the vowel sound /u/. Modified spectrum of segment (D) shows that the second formant peak has moved away from the first formant peak and is closer to the third formant peak. This indicates the presence of the vowel sound /i/. These results show that modified linear prediction method captures the formant transitions effectively over short segments.

These experiments show that it is possible to capture the dynamic variations of the vocal tract system from the speech signal using modified linear prediction method. We investigate in the next section the effect of the

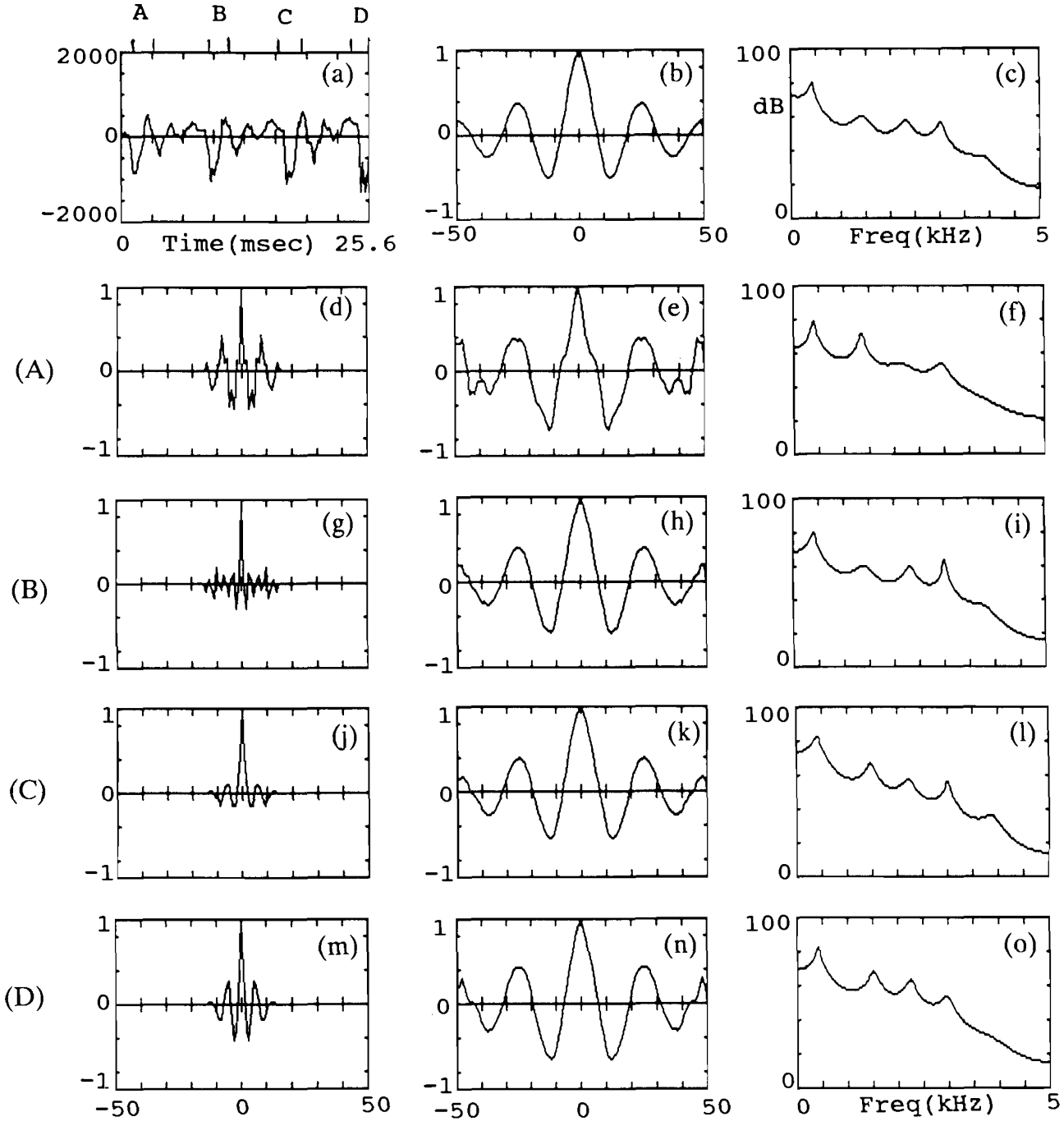


Fig. 6.7. Illustration of the effectiveness of modified linear prediction in capturing local variations in a frame containing formant transitions: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.7a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 1.6 msec (16 samples).

modified LP method on the quality of synthetic speech.

6.3.2. Experiments in speech synthesis

We have conducted experiments to study whether the dynamic variations of the vocal tract system that are captured by the modified LP method impart naturalness to the synthetic speech signal. We compare the performances of standard LP method and modified LP method with regard to the quality of the synthesized speech.

For standard LP analysis, a frame width of 256 samples and LP order of 12 was used. For modified LP analysis, a frame width of 256 was chosen to obtain the gross autocorrelation, and a short frame width of 16 was chosen to extract the dynamic variations and incorporate them in the gross autocorrelation to obtain the modified autocorrelation. The modified LP coefficients are obtained from the modified autocorrelation. We compute the standard and modified LP coefficients once in 16 samples.

We have chosen the sentence **'We** were away a year ago° uttered by a male speaker and digitized it at 10 kHz sampling rate. The formant tracks for different cases are shown in Fig. 6.8. Fig. **6.8a** shows the formant tracks obtained using the standard LP of order 12 with frame size of 256 samples and frame shift of 16 samples. Figs. **6.8b** and **6.8c** correspond to the modified LP with short frame width of

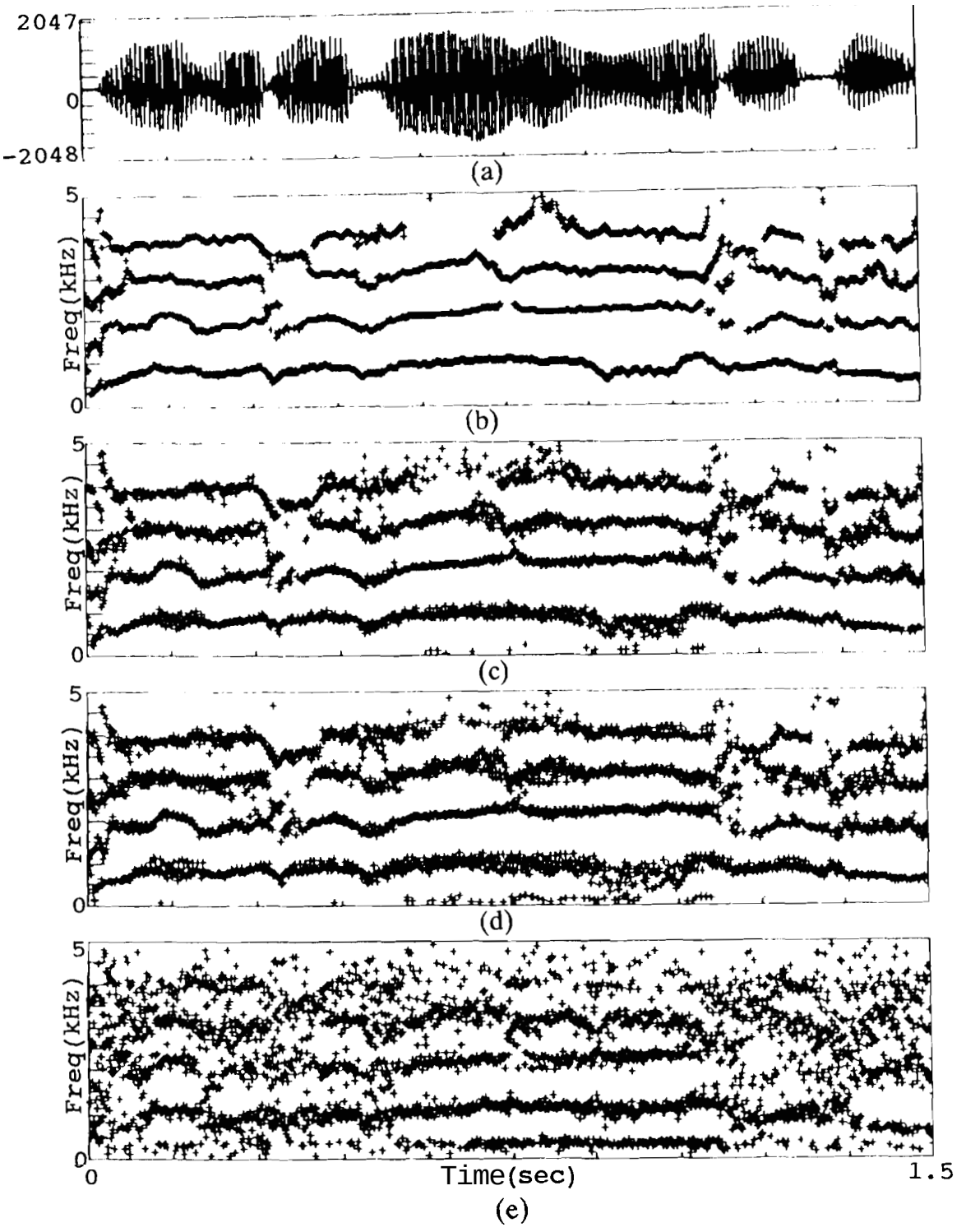


Fig. 6.8. Illustration of the ripple introduced in the formant tracks by the modified linear prediction method: (a) The waveform of the sentence **"We** were away a year **ago"** spoken by a male speaker; its formant tracks obtained by (b) standard LP, frame **size=25.6** msec; (c) modified LP, frame **size=25.6** msec, short frame **size=3.2** msec; (d) modified LP, frame **size=25.6** msec, short frame **size=1.6** msec; (e) standard LP, frame **size=3.2** msec. In all the above cases the parameters are computed once in every 1.6 msec.

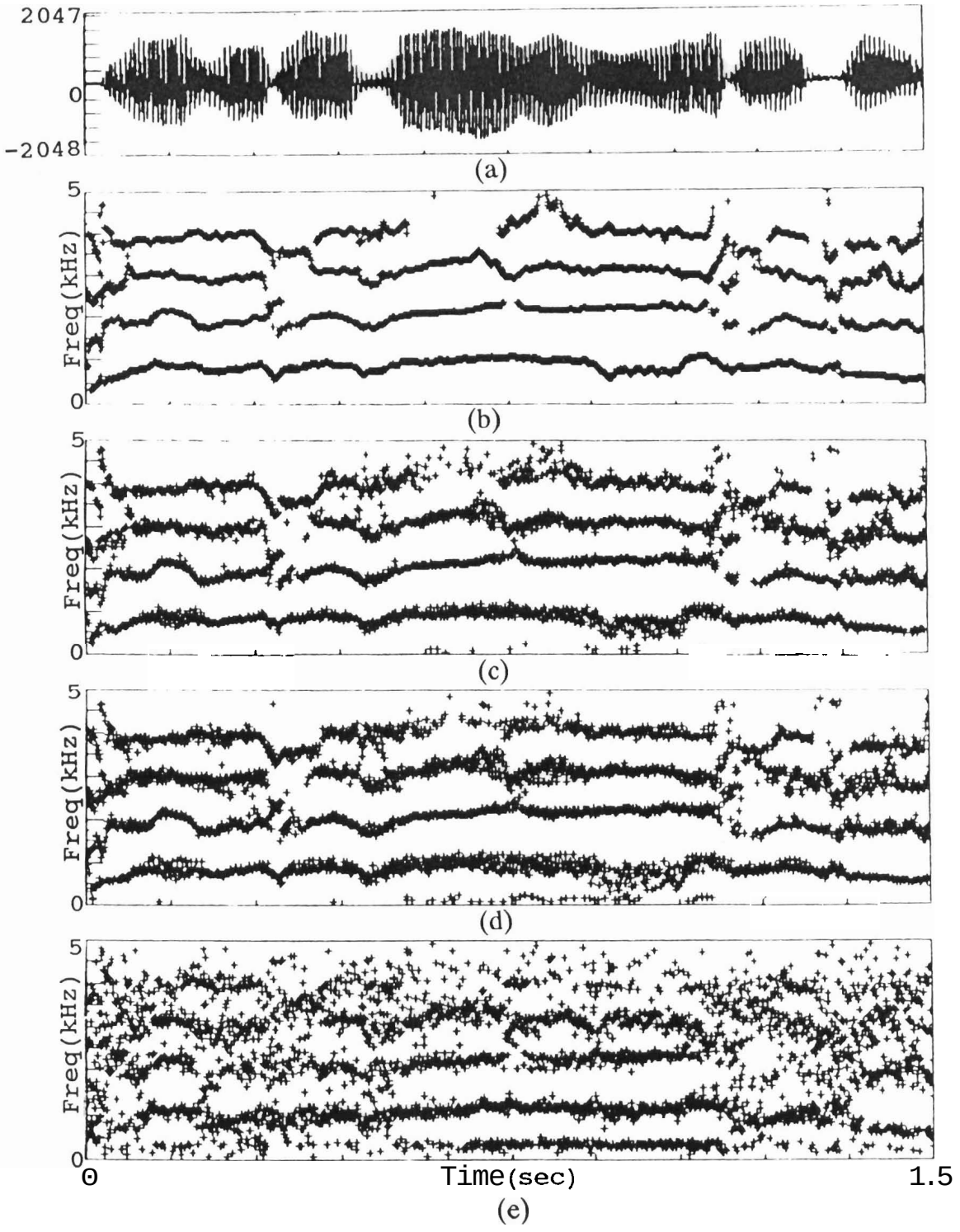


Fig. 6.8. Illustration of the ripple introduced in the formant tracks by the modified linear prediction method: (a) The waveform of the sentence **"We** were away a year **ago"** spoken by a male speaker; its formant tracks obtained by (b) standard LP, frame **size=25.6** msec; (c) modified LP, frame **size=25.6** msec, short frame **size=3.2** msec; (d) modified LP, frame **size=25.6** msec, short frame **size=1.6** msec; (e) standard LP, frame **size=3.2** msec. In all the above cases the parameters are computed once in every 1.6 msec.

32 samples and 16 samples, respectively. Fig. **6.8d** shows formant tracks for the standard LP order 10 with frame size of 32 samples and frame shift of 16 samples. We observe that in all cases, the modified LP method gives similar formant tracks as those of the standard LP. But in the formant tracks obtained by the modified LP analysis, the formants fluctuate around a mean value. From Figs. **6.8b** and **6.8c** we notice that as the short frame size decreases, the fluctuations in the formant tracks increase. This could be due to averaging of the time varying spectral information that takes place within the frame for larger frame sizes. The occurrence of fluctuations in the formant tracks in the case of modified LP can be explained as follows: Speech signal is the output of human speech production mechanism comprising of the vocal tract system and the excitation system. These systems, being made up of living tissue, undergo subtle variations continuously. We may expect such changes to be reflected as continuous variation of the frequencies and bandwidths of the formants. In addition, the source-tract interaction affects the formants within a pitch period. When a large analysis frame width is used, as in the case of standard LP, only gross features are captured as is evident from the formant tracks shown in Fig **6.8a.** Short analysis frames are not suitable for standard LP analysis. In this case the estimated LP spectrum is inaccurate due to

time-frequency resolution problem introduced by the window effects. We have given in Fig 6.8d formant tracks obtained from standard LP with 32 sample analysis frame. We notice that several spurious peaks are present.

We have plotted the formant tracks for female voice in Fig. 6.9. These plots also confirm our earlier observation. But we notice from Figs. 6.8 and 6.9 that the formant fluctuations from modified LP analysis are less in the case of female voice compared to male voice. This could be due to differences in the source-tract interaction. The effect is felt more in male speech than in female speech. This may be because glottal opening and closed phases within a pitch period are comparatively well defined in the male speech. In the case of female speech the glottal closure phase may be insignificant in most of the frames.

The formant tracks given in Figs. 6.8 and 6.9 do not provide full information about the analysis parameters from which they are derived. Bandwidth information, which is perceptually significant, is not captured by the formant tracks. This is to be noted especially as the effect of source-tract interaction is mostly in terms of bandwidth changes. We wanted to investigate whether these changes can be captured by the modified LP coefficients. Thus formant tracks derived from the modified LP coefficients indicate that the formant frequencies follow those that are obtained

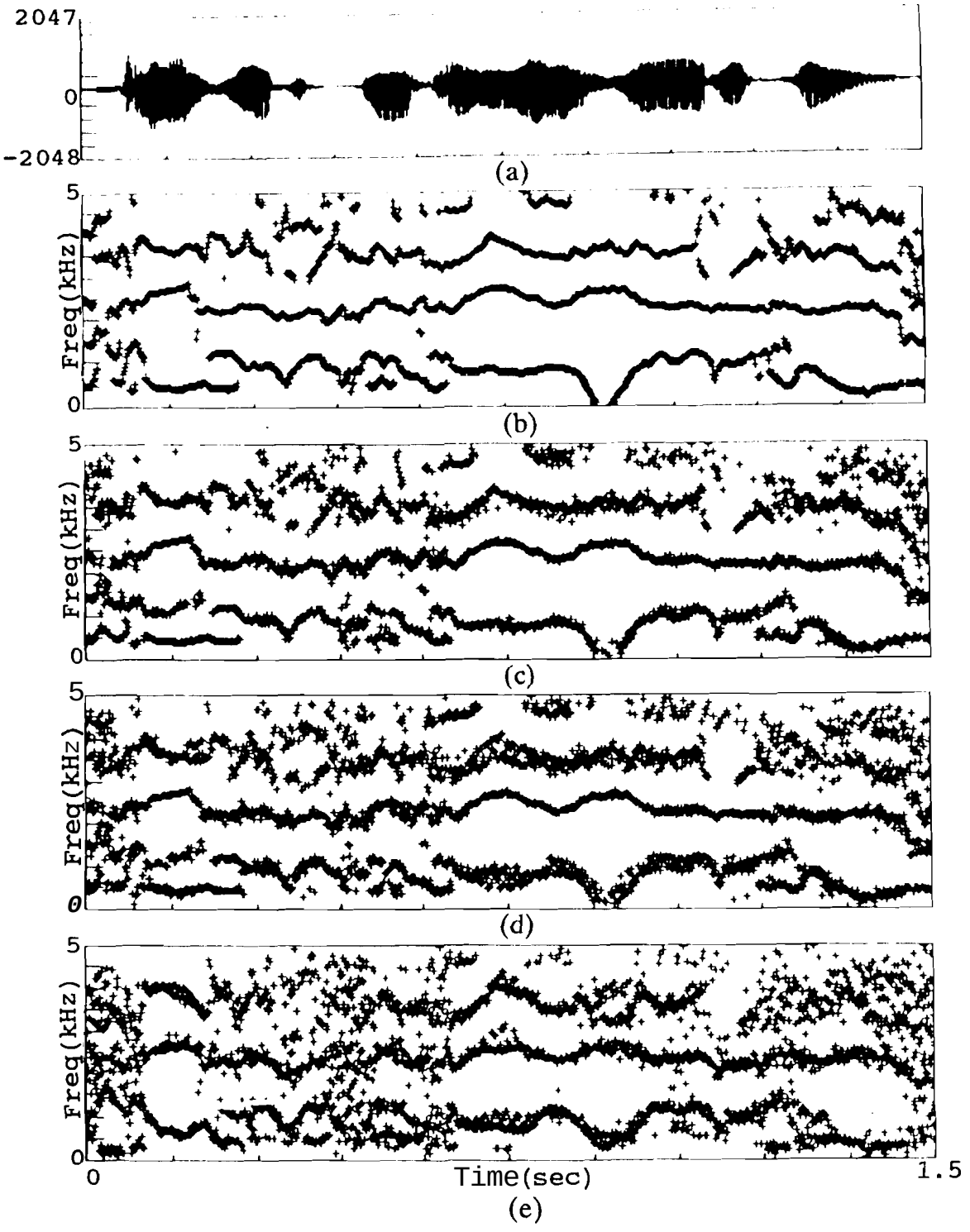


Fig. 6.9. Illustration of the ripple introduced in the formant tracks by the modified linear prediction method: (a) The waveform of the sentence **"We** were away a year **ago"** spoken by a female speaker; its formant tracks obtained by (b) standard **LP,** frame **size=25.6** msec; (c) modified LP, frame **size=25.6** msec, short frame **size=3.2** msec; (d) modified **LP,** frame **size=25.6** msec, short frame **size=1.6** msec; (e) standard LP, frame **size=3.2** msec. In all the above cases the parameters are computed once in every 1.6 msec.

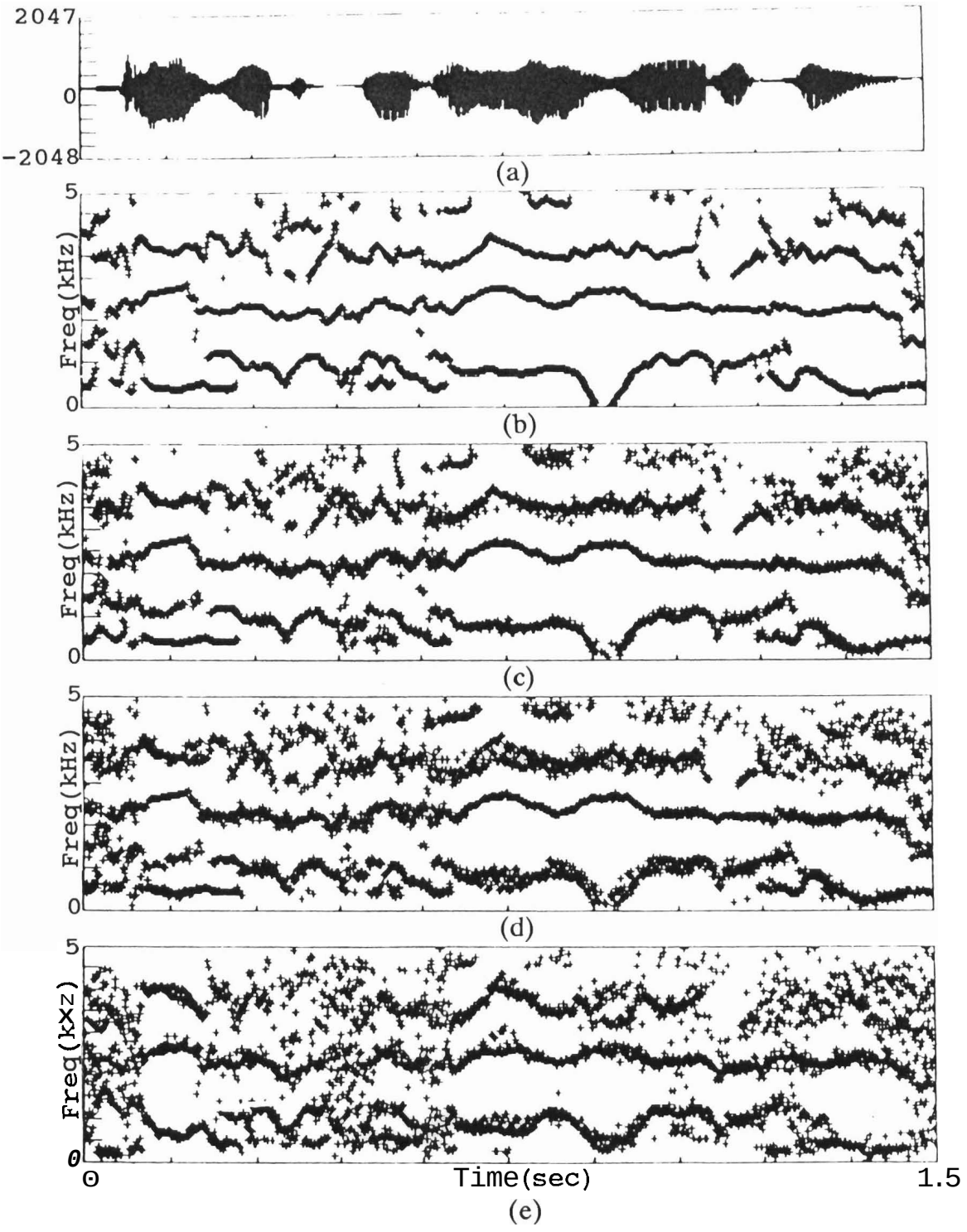


Fig. 6.9. Illustration of the ripple introduced in the formant tracks by the modified linear prediction method: (a) The waveform of the sentence **"We** were away a year **ago"** spoken by a female speaker; its formant tracks obtained by (b) standard LP, frame **size=25.6** msec; (c) modified LP, frame **size=25.6** msec, short frame **size=3.2** msec; (d) modified LP, frame **size=25.6** msec, short frame **size=1.6** msec; (e) standard LP, frame **size=3.2** msec. In all the above cases the parameters are computed once in every 1.6 msec.

by the standard LP method. We have to test the perceptual quality of the synthesized speech from both the methods to determine whether modified LP method has captured any additional information that enhances the natural quality.

In the speech synthesis experiments, we have used an excitation signal that is derived from the smoothed pitch and gain contours of the speech signal that is being analyzed. **Fant's** model is used to approximate the glottal pulse. Speech is synthesized using both standard and modified LP coefficients. Informal listening tests show that a distinctive shiver is added to the synthetic speech generated from modified LP coefficients compared to that from standard LP. This effect is found to improve the naturalness of the synthetic speech.

6.4. Discussion

We have proposed a new technique called modified linear prediction in this chapter. This is used for processing speech over short data records, under time varying formulation. The studies given here demonstrate the usefulness of this technique, for capturing the dynamic variations of the vocal tract system from the speech signal.

Earlier attempts for capturing the temporal variations of the speech spectrum concentrated on the improving the excitation signal as in the case of multipulse

excitation based speech analysis and synthesis. In the present study we have attempted to incorporate such changes in the model of vocal tract system.

In the modified **LP** analysis we have circumvented the limitations caused by the time-frequency resolution problem. This is achieved by noting that the error in estimating autocorrelation parameters due to the short window effects is more when the correlation between the signal samples is more. Hence we have developed a two pass **LP** based algorithm. **In** the first pass, we capture the gross spectrum by using autocorrelation parameters estimated from a relatively long data frame. As the window length is large we can get good estimates of the autocorrelation. In the second pass, we use a much shorter window length; but we use the **LP** residual of the signal to reduce the correlation between the samples. Thus the autocorrelation parameters estimated from the short data are still good as the correlation between the data samples is less in this case. The autocorrelation in the second pass captures the local variations. We combine the autocorrelations obtained in the first and second passes through convolution operation to obtain an improved estimation of autocorrelation coefficients. The modified **LP** coefficients are derived from these autocorrelation coefficients.

We have shown experimentally that the modified **LP**

method captures the local variations in the spectrum. We notice from the plots of the spectra corresponding to consecutive short frames that the formant frequencies and bandwidths undergo continuous variation. We have illustrated that the variations associated with unvoiced to voice transition, formant transitions are captured. The formant tracks obtained from the analysis of natural speech using modified LP method indicate that the method is stable and compares well with the standard LP method. The formant tracks show the presence of a ripple which is due to the local fluctuations of the formant frequencies. Speech synthesized using the modified LP coefficients sounds more natural than that using the standard LP coefficients. This may be due to the fact that modified LP method captures the effect of source-tract interaction, namely perceptually significant bandwidth changes within a pitch period.

Here the issue of source-tract interaction requires further explanation. In each pitch period, the bandwidths of the formants are relatively low in the initial part (closed glottis phase) and high in the later part (open glottis phase). This due to source-tract interaction. Thus when we take several pitch periods, the bandwidths change from low to high and high to low, alternately. We may consider the two cases separately. When the change is from high bandwidth to low bandwidth, the interaction between

these two regions is minimum and a short data processing technique like modified LP captures the change adequately. But in the case of change from low bandwidth to high bandwidth, the high bandwidth region is affected by the low bandwidth region and hence it is impossible to capture the system change accurately.

So even when the low and high bandwidth regions are known within a pitch period of a speech signal, either because we have taken a synthetic signal or we have made use of additional information like electro-glotto-graph (EGG) signal along with the natural speech, we are unlikely to accurately capture the changes due to bandwidth. This is because of the influence of the high amplitude (low bandwidth) signal on low amplitude (high bandwidth) signal. Thus even if we find a reliable method to identify the low to high bandwidth transition (instant of glottal opening) as we have attempted in the last chapter, it may be of limited significance. It is difficult to isolate the high bandwidth region due to the influence of the slow decay from the adjacent segment. Still the modified LP method captures changes in an automatic way, though the changes may not be accurate.

We also note that only small changes like those due to bandwidth variations can be captured effectively in the second pass of the modified LP algorithm. This is

because the LP residual from the first pass is normally less correlated and its autocorrelation over a short frame is nearly an impulse especially in high amplitude regions (closed glottis regions). Thus its spectrum is almost flat. This shows that large changes like in transition segments cannot be effectively captured, since a short segment can only make small changes in the spectrum. But there are other methods to deal with such large changes as in transitions.

There is scope for further investigations on the modified LP approach. It is still not clear what are the optimum sizes of the short and the long frames used in the method. A careful study may be required to identify what must be the most appropriate model orders to be chosen in the first and the second passes. Another important area that needs further investigation is concerned with the choice of the excitation signal that is to be used along with the modified LP coefficients for synthesis. Updating the excitation signal information once in several pitch periods and choosing Fant's model for the excitation pulse as we have done in the studies presented here may not be the best approach. Updating the excitation information in intervals shorter than a pitch period so as to synchronize it with the updation of system information and choosing a multipulse model may prove to be more effective in generating natural sounding speech. This may result in combining the benefits

of the multipulse LP and modified LP methods.

We have not considered the effect of noise in the speech analysis-synthesis, especially in the context of time-frequency resolution problem in our studies until now. We discuss this issue in the next chapter.

Chapter 7

ANALYSIS OF NOISY SPEECH SIGNALS: WEIGHTED LINEAR PREDICTION METHOD

Natural signals are, in general, noisy. For example, speech may be corrupted either by background noise at the stage of collecting the data or by channel noise when it is transmitted through a communication channel. We have considered only clean speech in the modified linear prediction technique discussed in the previous chapter. In this chapter we propose a new technique, called weighted linear prediction, for processing noisy speech .

Processing noisy signals is essential in a number of situations. The background noise in environments such as the shop floors, streets and cockpits degrades speech and reduces its intelligibility and quality. Other contexts where speech enhancement is required include reduction of noise in speech to improve the performance of speech coders, voice and/or speaker recognition systems. In speech analysis-synthesis systems, the quality and intelligibility of the synthetic speech is severely affected in the presence of noise.

If the speech signal and the noise do not overlap either in the time or in the frequency domains, then simple band pass filtering techniques can be used for speech

enhancement. But that is not the case in most of the practical situations and hence more sophisticated signal processing techniques are required. Some of the important methods for speech enhancement were reviewed in Chapter 2. Almost all these methods require some a priori knowledge about the noise and signal characteristics. Some methods require the noise statistics such as the estimate of noise power spectrum. Some others depend on the availability of a reference signal which is in someway correlated with noise and uncorrelated with the signal. Most of the methods assume that the corrupting noise is white Gaussian. In a number of situations, this kind of a priori knowledge may not be available. Our aim in this chapter is to use modeling approach to enhance speech corrupted by background noise of unknown statistics. Ideally, we must be able to estimate model parameters representing the undegraded speech from the noisy speech, so that we can use these parameters for speech synthesis.

We describe a technique for noisy signal processing where a priori knowledge about the noise and the speech signal is not essential. This technique is based on a modification to the standard linear prediction technique to give more weightage to high signal-to-noise ratio (SNR) portions of the noisy speech signal. We can get better estimates of the model parameters using this new technique.

We show this with the help of experiments on model signals and also by noting the improvement in quality of the speech synthesized from parameters obtained using the new technique.

This chapter is organized as follows: In Section 7.1 we study limitations of the standard LP technique for noisy signal processing. We introduce the basis for the proposed weighted linear prediction technique in Section 7.2. We present its analytical formulation and discuss implementation issues. In Section 7.3 we present experimental results obtained using model signals as well as natural signals for analysis and synthesis. We discuss the implications of these results. In Section 7.4 we summarize the weighted linear prediction based approach and compare this formulation with other model based methods given in literature.

7.1. Problems in Noisy Signal Processing

Noisy signal processing implies either to obtain an enhanced signal directly or to extract features of the signal accurately in the presence of noise. We consider LP modeling of noisy speech in this section. We discuss limitations of the LP method in the presence of noise. We also discuss problems in choosing only high SNR regions of the noisy speech signal for processing.

7.1.1. Limitations of LP method

Standard linear prediction modeling is one of the widely used methods for speech processing. It is successfully used for applications such as speech coding, speaker and voice recognition, speech analysis and synthesis systems, voice conversion and time scale modification. But it is found that the performance of these **LP** based systems quickly degrades in the presence of background noise. This is mainly because the linear prediction formulation is based on an all-pole model. In the presence of noise this model is no longer valid. In the following subsections we consider the effect of noise on the **LP** parameter estimation and methods of reducing this effect.

Background noise in a speech signal introduces stray zeros in the z-plane representation of the signal. Hence the actual behavior of the noisy speech is pole-zero in nature. But when we still apply all-pole model as in the **LP** method, the estimated all-pole parameters may no longer be closer to that of the corresponding clean speech. The peaks in the **LP** spectrum correspond to formant frequencies and hence determination of their locations accurately is important. Due to the effect of the zeros introduced by the noise, the bias and variance of the estimated formant frequencies will be large. Thus the estimates of the formant frequencies derived from the **LP** spectrum of the noisy speech

may not be accurate. As the noise spectrum is wide band, it may mask some of the low intensity higher formants of the speech signal. Hence the higher formants may be completely lost in the LP spectrum. If we increase the order of the LP to capture the higher formants, we may get spurious peaks due to the effect of the noise. The presence of noise also affects the dynamic range of the magnitude spectrum. This is reflected in the LP smoothed spectrum as well. Effect of spurious peaks in the synthetic speech is the appearance of ringing sound in the processed signal. Reduction in the dynamic range has the effect of giving undue emphasis to higher frequencies.

Thus the main limitations of the standard LP formulation for noisy signal processing are the following:

(i) Errors in the estimation of formant frequencies.

(ii) Reduction in the dynamic range.

7.1.2. Processing of noisy speech in high SNR regions: Time-frequency resolution issues

We notice in general that in the time domain the background noise energy is more or less evenly distributed, whereas the energy of the speech signal is not. At a global level the signal energy is high in the voiced regions and low in the unvoiced regions and zero in the silence regions. We find such variations in the signal energy at a local

level also. For example within a pitch period in the voiced speech, the signal energy is generally high just after the major excitation epochs (glottal open-to-close transition) and then tapers off to a minimum until the subsequent excitation occurs. If we compute the SNR of the noisy signal, we may find some regions where SNR is high and other regions where it is low. This suggests an approach to obtain better estimates of the model parameters from the noisy speech in which we consider only high SNR regions for parameter estimation. This must give relatively accurate model parameters as the signal is dominant in the high SNR regions compared to noise.

Thus the analysis intervals for obtaining model parameters from the noisy speech may be chosen from the high SNR regions. The length of the chosen analysis interval assumes significance here. At a global level, the SNR ratio is high in the voiced portions of a noisy signal. One approach is to choose an analysis window that encompasses the entire voiced portion. But this could be several pitch periods long. In that case the underlying quasistationary assumption within the analysis interval for parameter estimation is no longer valid due to the nonstationary nature of the speech. Hence the estimated parameters may not capture the time varying spectral information. Besides, within the voiced portion itself, the SNR varies locally.

Hence the effect of the noise from the relatively low SNR regions within the analysis frame may still dominate. We notice that within a pitch period the SNR is maximum after the glottal closure and then gradually tapers off to a small value. This suggests an alternative where the analysis interval is chosen to encompass only the high SNR regions within a pitch period using an appropriate threshold. But here the analysis interval may become too short especially in the case of female speech, and the time-frequency resolution problem may dominate. This affects the frequency resolution of the resulting model signal.

7.2. Weighted LP Formulation

We have suggested in the last section that it is reasonable to process noisy speech only in the high SNR regions to obtain better estimates of the model parameters. But there is no convenient signal processing technique to accomplish this task without suffering from the effects of short data record processing. In this section, we propose a new technique, called weighted linear prediction, in which we give more emphasis to the high SNR regions while estimating the model parameters. First we give the basis and description of the proposed technique. Then we present an analytical formulation of the technique and. Finally we discuss implementation details.

7.2.1. Basis and description of the proposed method

In the standard linear prediction formulation we assume that sample values in each analysis interval can be approximated by a linear weighted sum of a certain number of previous samples. The estimation error for each sample is computed as the square of the difference between the actual sample value and the estimated value. The sum of estimation errors for the entire frame is minimized to obtain the linear prediction coefficients corresponding to that frame. But in the case of noisy signal, the noise corrupted sample cannot be estimated as a weighted sum of a certain number of previous samples, because the noise associated with each sample is uncorrelated with that of the other samples. If such a formulation is forced, the resulting model parameters will be inaccurate. We have already seen in the previous section that we cannot consider each high SNR region of the noisy signal separately for parameter estimation as that will lead to time-frequency resolution problem. Hence in our proposed method, we consider duration of the analysis interval in the range of 20-30 msec to avoid the short window effects and at the same time put more emphasis in the high SNR regions of the analysis interval. This reflects in the error minimization criterion we considered previously in the following way. If the signal is noisy, then samples in those regions where SNR is low are not so reliable and error

in those regions need not be minimized to the extent it is done in other regions. We use a weight function which is related to the SNR at different instants of the speech signal. This weight function is applied to the error function before the error minimization criterion is applied to obtain the model parameters.

An important issue here is how to know the SNR at different instants of the noisy speech signal, as we do not have a priori knowledge about the noise and the signal. We use a simple estimation rule. Wherever the instantaneous energy of the noisy signal is high, we label it as high SNR region. This is a reasonable assumption as the signal energy fluctuates in time whereas the energy of the background noise is approximately constant. Thus an estimate of the SNR profile of the noisy signal can be obtained as a function of instantaneous energy. The weight function can be chosen to be either equal to or in some way related to the instantaneous energy function of the noisy signal.

7.2.2. Analytical formulation

Let us assume that $s(n)$ is a clean speech signal of M samples which is corrupted by additive background noise $N(n)$ to result in the noisy signal $y(n)$. That is

$$y(n) = s(n) + N(n), \quad n = 0, 1, \ldots M-1 \qquad (7.1)$$

Let $\hat{s}(n)$, a weighted sum of p previous noisy samples, be the

estimate of the clean signal . That is

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k y(n-k) \tag{7.2}$$

where a_k, k = 1, 2, ... p are the weighted LP coefficients to be computed.

Let us formulate an error function e(n) as

$$e(n) = y(n) - \hat{s}(n) \tag{7.3}$$

We denote the total error energy E as

$$E = \sum_{n=0}^{M-1} w(n)\ (e(n))^2 \tag{7.4}$$

where w(n) is a weight function. From equations (7.4) , (7.3) and (7.2), we get

$$\begin{aligned} E &= \sum_{n=0}^{M-1} w(n)\ (y(n) + \sum_{k=1}^{p} a_k y(n-k))^2 \\ &= \sum_{n=0}^{M-1} w(n)\ ((y(n))^2 + 2y(n) \sum_{k=1}^{p} a_k y(n-k) \\ &\qquad + (\sum_{k=1}^{p} a_k y(n-k))^2) \end{aligned} \tag{7.5}$$

We minimize E by setting

$$\frac{\partial E}{\partial a_k} = 0, \quad 1 \le k \le p \tag{7.6}$$

From equations (7.5) and (7.6), we obtain a set of simultaneous equations given as follows:

$$\sum_{k=1}^{p} a_k \sum_{n=0}^{M-1} w(n)y(n-k)y(n-i) = -\sum_{n=0}^{M-1} w(n)y(n)y(n-i),$$

$$i = 1, 2, ... p \qquad (7.7)$$

This can be written in a simple form as

$$\sum_{k=1}^{p} a_k \phi_w(k,i) = -\phi_w(0,i), \quad i = 1, 2, ... p \qquad (7.8)$$

where $\phi_w(k,i)$ is a weighted correlation function defined as

$$\phi_w(k,i) = \sum_{n=0}^{M-1} w(n)y(n-k)y(n-i) \qquad (7.9)$$

We notice that the values of $\phi_w(k,i)$ for k = 1, 2, ... p and i = 1, 2, ... p in equation (7.8) form a weighted correlation matrix. Thus we can obtain the weighted LP coefficients a_k, k = 1, 2, ... p by first computing $\phi_w(k,i)$ and $\phi_w(0,i)$ from y(n) and then solving the set of simultaneous equations given in equation (7.8).

The weight function w(n) plays an important role in the above formulation. If w(n) = 1, then the estimated signal $\hat{s}(n)$ is closer to the noisy signal y(n). The problem is how to choose w(n) such that the estimated signal $\hat{s}(n)$ is closer to the clean signal s(n) in the mean square sense. We follow an approach outlined below to solve this problem.

The error function e(n), defined in equation (7.3), can be considered to have two components as follows:

$$\begin{aligned} e(n) &= N(n) + s(n) - \hat{s}(n) \\ &= e_1(n) + e_2(n) \end{aligned} \qquad (7.10)$$

where

$$e_1(n) = N(n) \quad (7.11)$$

and $$e_2(n) = s(n) - \hat{s}(n) \quad (7.12)$$

We do not know the values of the error components $e_1(n)$ and $e_2(n)$ separately, but we can assume that relatively $e_1(n)$ dominates in the low SNR regions of the noisy speech signal $y(n)$ and $e_2(n)$ in the high SNR regions. Hence one way of weighting $e(n)$ would be to give more emphasis to those regions of $e(n)$ corresponding to high SNR regions of $y(n)$. This has the effect of minimizing, to a larger extent, $e(n)$ dominated by $e_2(n)$ contribution in these regions. In the low SNR regions, less weightage is given to $e(n)$ so that $e(n)$ which is dominated by $e_1(n)$ contribution (noise contribution) in these regions is minimized to a lesser extent. Thus we have to choose a weight function $w(n)$ which gives more emphasis to the high SNR regions of $y(n)$.

7.2.3 Implementation issues

We have described the weighted LP technique and have given an analytical formulation in the previous subsection. We have to solve the set of simultaneous linear equations given in (7.8) to obtain the weighted LP coefficients, Using general methods like Gauss elimination for this purpose requires a computational overhead of $(p^3/3 + O(p^2))$ operations. We notice, from equation (7.9)

that the weighted correlation matrix is symmetric. Therefore we have used Choleskey decomposition method [Whilkinson and Reinsch 1971] to solve equation (7.8) more effectively. This method requires about half the computation of $(p^3/3 + O(p^2))$.The exact form of the weight function that is to be used is also important. We have tried several alternatives. We have taken the absolute magnitude function, $|y(n)|$ of the noisy signal $y(n)$ as the basis for all these alternatives, since it reflects the estimation rule for the SNR profile of $y(n)$. We list below some of the weighting functions that are considered in our experiments.

$$w(n) = |y(n)| \tag{7.13}$$

$$w(n) = \frac{1}{2K+1} \sum_{i=-K}^{K} |y(n+i)| \tag{7.14}$$

$$w(n) = |y(n)|^{\alpha}, \quad 0 \leq \alpha \leq 1 \tag{7.15}$$

$$w(n) = LPF_{\omega_c}(|y(n)|) \tag{7.16}$$

where $LPF_{\omega_c}(\cdot)$ is a low pass filter operator with cut off frequency ω_c. The value of ω_c is chosen such that only that portion of the frequency spectrum corresponding to the first formant frequency region of $y(n)$ is included.

$$w(n) = 1 \quad \text{for} \quad |y(n)| \geq \psi$$
$$= 0 \quad \text{otherwise} \tag{7.17}$$

The absolute magnitude of the noisy signal itself can be taken as the weighting function as given in equation

(7.13). Various smoothing techniques on the absolute magnitude result in the weighting functions given in equations (7.14), (7.15), and (7.16). Equation (7.17) gives a weighting function which selects only those portions of the noisy signal where its absolute magnitude is greater than or equal to a given threshold ψ.

We illustrate the different types of weight functions described above in Fig. 7.1. A noisy speech signal at 3 dB SNR is shown in Fig. 7.1a. Fig. 7.1b shows the weight function derived as the absolute magnitude of the noisy signal (equation (7.13)). A 11 point averaged version of Fig. 7.1b is given in Fig. 7.1c (equation (7.14) with K = 5). Figs. 7.1d and 7.1e show weight functions corresponding to equation (7.15) with a = 0.4 and equation (7.17) with ψ = 50.0 respectively. We found through our experiments that the averaged absolute value weight function shown in Fig. 7.1c gives the best results. Using this weight function results in stable weighted LP filters compared to the other weight functions. Moreover, the low pass filter based weight function (equation (7.17)) also gives acceptable results. But it is computationally expensive. Hence we have used only the averaged absolute value based weight function throughout our experiments.

In the next section we present the results of applying the weighted LP method on model signals as well as

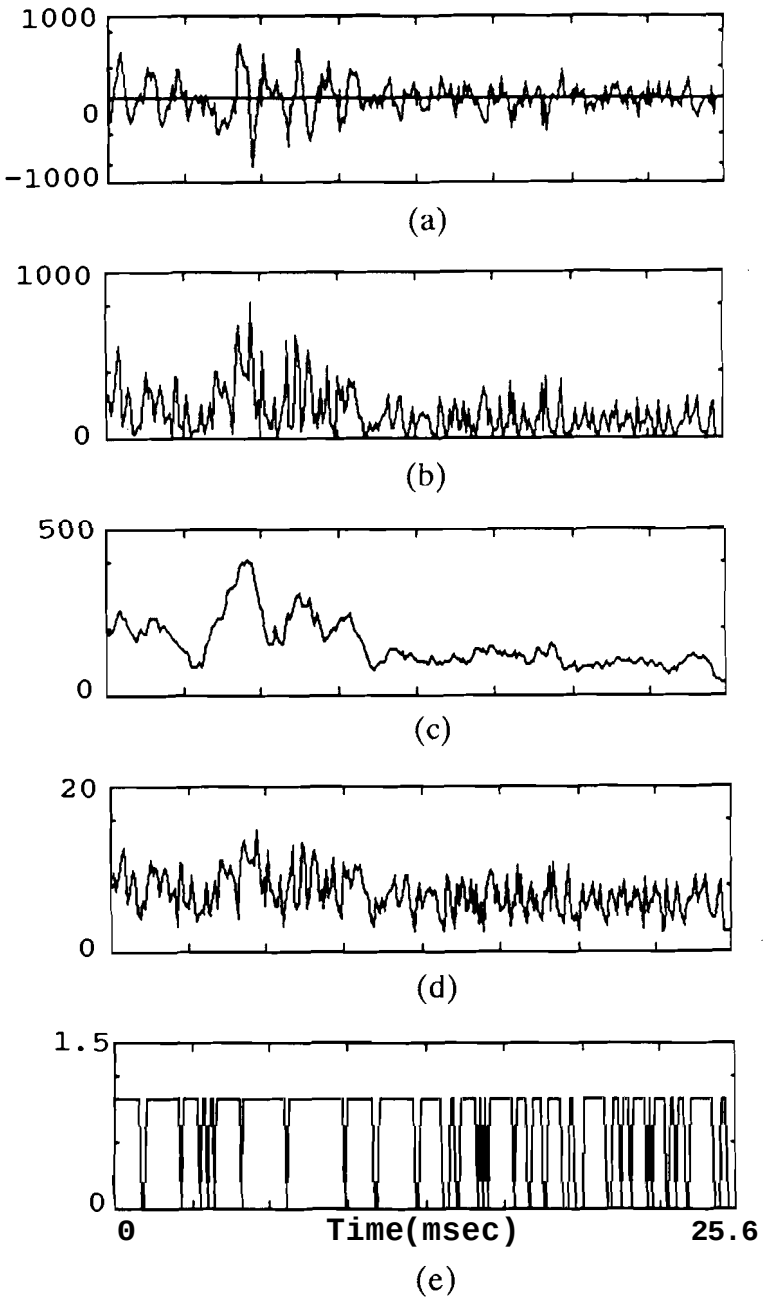


Fig. 7.1. Formulation of different weighting functions based on the SNR criterion: (a) A typical noisy speech signal (SNR=3dB) and (b), (c), (d), and (e) corresponding weighting functions based on the absolute magnitude of the noisy signal. (See text for the equations of the weighting functions)

natural signals.

7.3. Experimental Results

We have conducted a series of experiments to show the effectiveness of the weighted LP method in the estimation of analysis parameters. We first describe experiments involving model signals. We show using bias and variance computations that weighted LPCs are more consistent estimations of the pole locations of an all pole system than the standard LPCs. Next, we describe experiments on speech enhancement using LP synthesis. We show from the plots of formant tracks and informal listening tests on the synthesized speech that the effect of noise is reduced by using the weighted LP method.

7.3.1. Experiments using model signals

In these studies we have used synthetic vowel sounds with additive noise at different signal to noise ratios. In general, the effect of noise on the spectrum of a speech signal manifests as masking of the formant information and reducing the dynamic range. Due to these effects the smoothed spectrum obtained from the noisy signal will have small dynamic range. There is also the possibility of missing formant peaks and occurrence of spurious peaks. This is illustrated in Fig. **7.2b** where the LP spectrum of

the synthetic sound /a/ corrupted by additive noise at 3 dB SNR is shown. Its FT magnitude spectrum is shown in Fig. 7.2a. In Fig. 7.2c we show the weighted LP spectrum of the same signal. We notice that the weighted LP spectrum has an increased dynamic range and a relatively well defined formant structure. To determine whether the above result is consistent and whether the estimates of the formant frequencies are accurate, we have conducted the following experiment.

We have taken 50 realizations of the synthetic sound /a/ corrupted by additive noise at 10 dB SNR. The synthetic sound is made up of five formants whose frequencies are 813, 1313, 2688, 3438, and 4438 Hz respectively. The bandwidths are chosen based on the following formula [Dunn 1961]:

$$B_i = 50\ (1 + (F_i/1000)^2/6)\ \text{Hz} \qquad (7.18)$$

where B_i = bandwidth of the i th formant

and F_i = frequency of the i th formant.

Fant's model is used as excitation with approximately 80 msec pitch period. The additive noise is white Gaussian. Each of the 50 realizations differs from the others in the additive noise component, although all have the same SNR. We have superimposed the weighted LP (order = 8) spectra computed from each one of the 50 realizations in Fig. 7.3c. The average of these spectra is shown in Fig. 7.3d. The

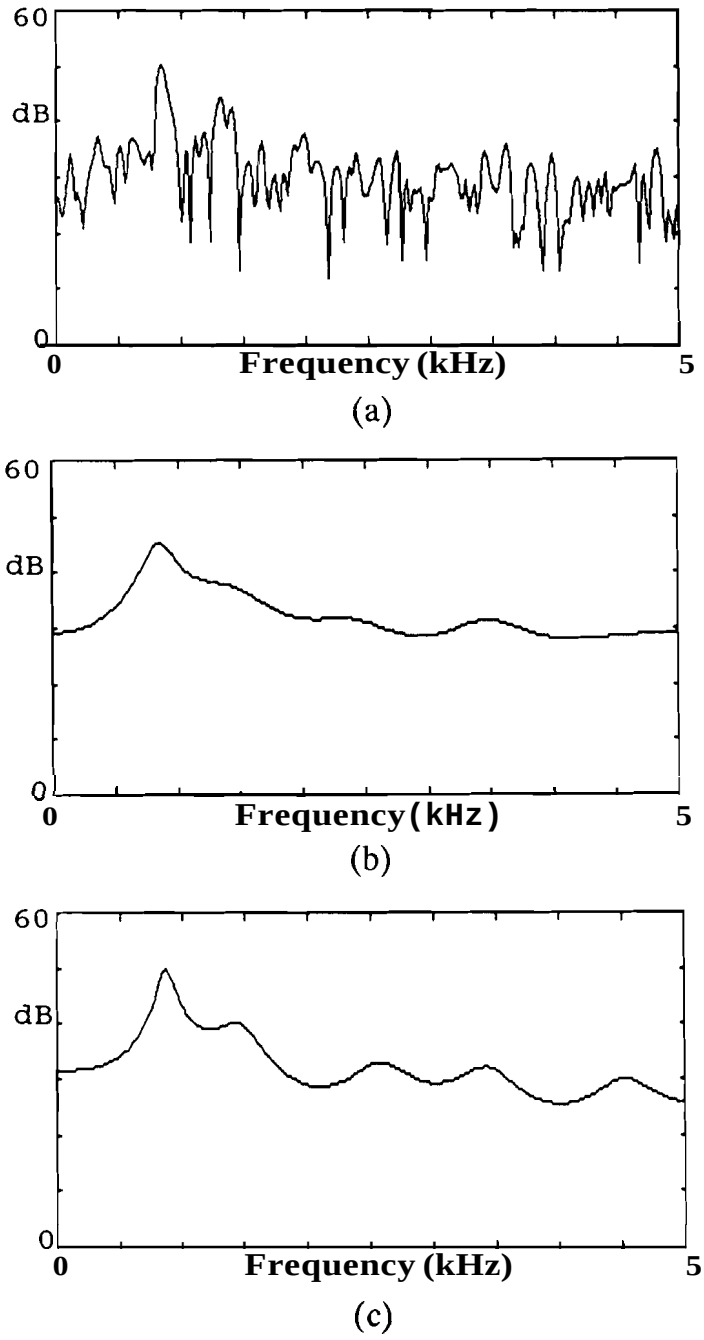


Fig. 7.2. Illustration of the effectiveness of the weighted LP method in enhancing the smoothed spectrum of the synthetic sound /a/ corrupted by additive noise at 3dB SNR: (a) FT magnitude spectrum; (b) standard LP spectrum; and (c) weighted LP spectrum.

corresponding superimposed and average spectra using standard LP (order = 8) are shown in Fig. 7.3a and Fig. 7.3b respectively. Comparing Figs. 7.3a-7.3b with Figs. 7.3c-7.3d, we observe that weighted LP method gives better performance in terms of consistency and improvement in dynamic range. The bias and variance of the formant frequency estimates using standard LP and weighted LP methods are given in Table 7.1. We notice that bias and variance for the weighted LP method are generally small compared to those for the standard LP method.

We have repeated the above experiment with SNR of 3dB to investigate whether the superiority of the weighted LP method is maintained at lower SNR levels also. We illustrate the superimposed and average spectra for the standard LP (Figs. 7.4a and 7.4b) and for the weighted LP (Figs. 7.4c and 7.4d). We notice that the spectra obtained by weighted LP method are distinctly better. The second formant which is difficult to notice in the standard LP spectra is clearly seen in the weighted LP spectra. The bias and variance of the formant estimates for both the cases are given in Table 7.2. We notice that except for the first formant, the weighted LP estimator is better than the standard LP estimator.

In Figs. 7.5 and 7.6 we show the results of repeating the above experiments for the synthetic vowel /e/

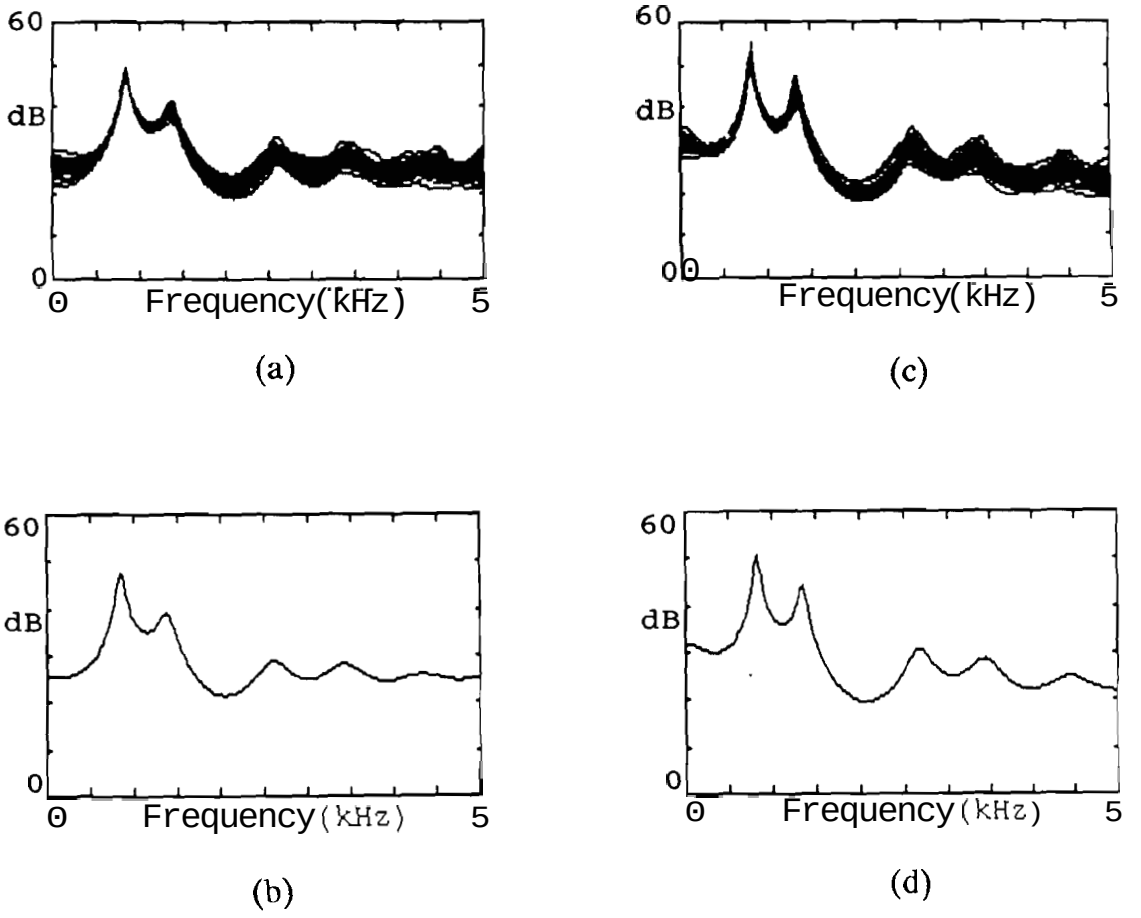


Fig. 7.3. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound */a/* corrupted by additive noise at **10dB** SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

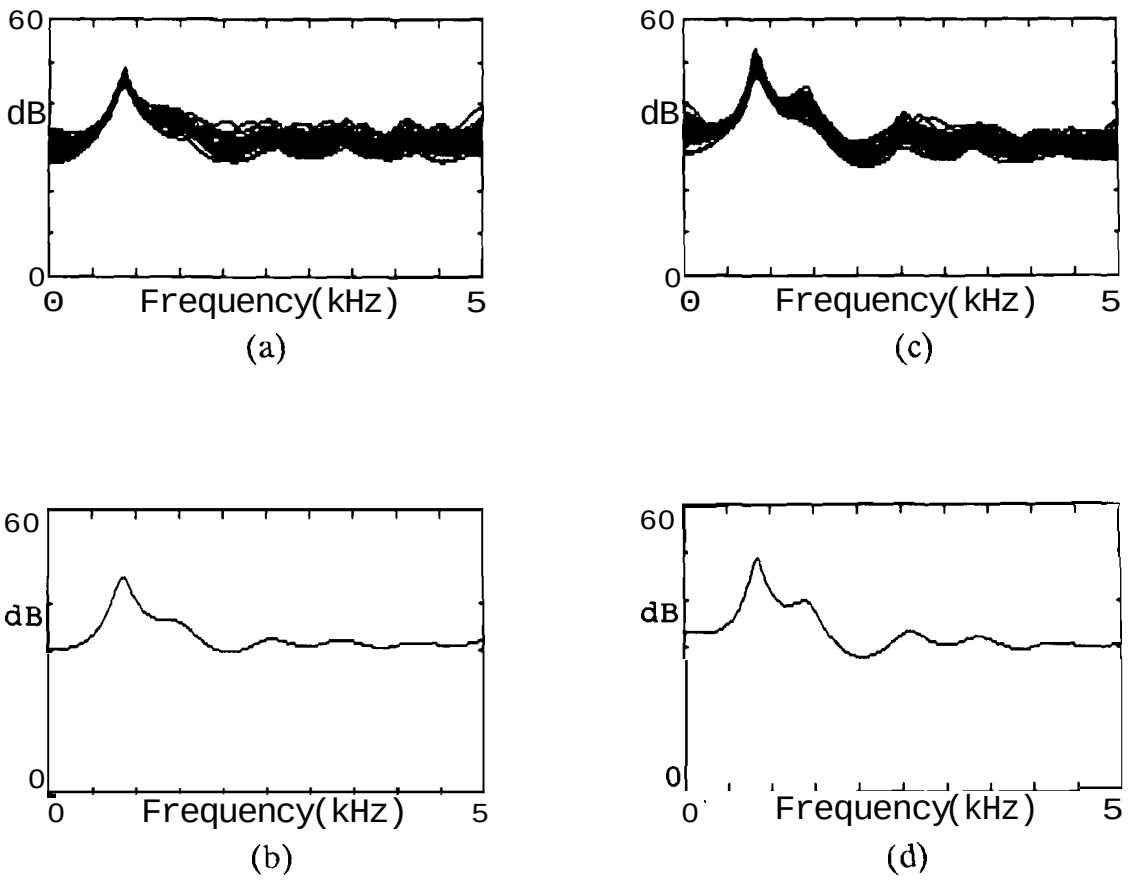


Fig. 7.4. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /a/ corrupted by additive noise at 3dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

TABLE 7.1

COMPARISON OF THE RESULTS FOR ESTIMATING THE FORMANT FREQUENCIES OF THE VOWEL /a/ WITH ADDITIVE NOISE AT SNR=10dB USING STANDARD LP AND WEIGHTED LP METHODS

Formant	LP		Weighted LP	
Freq.	Bias	Variance	Bias	Variance
1	33.09	1178.2	19.42	467.08
2	55.36	3238.76	34.27	1349.46
3	-80.97	10755.77	-22.77	2993.4
4	-4.16	3619.93	-17.69	5571.64
5	-126.48	40256.5	-34.98	19776.95

TABLE 7.2

COMPARISON OF THE RESULTS FOR ESTIMATING THE FORMANT FREQUENCIES OF THE VOWEL /a/ WITH ADDITIVE NOISE AT SNR=3dB USING STANDARD LP AND WEIGHTED LP METHODS

Formant	LP		Weishted LP	
Freq.	Bias	Variance	Bias	Variance
1	46.77	2408.12	38.95	1698.77
2	154.9	37060.56	72.29	6814.76
3	-116.52	36421.8	-98.32	16328.75
4	-51.72	32417.05	-43.47	17785.8
5	-145.03	57665.33	-148.39	51551.12

with five formants. The frequencies of the formants are 438, 1813, 2688, 3438, and 4438 Hz respectively. The bandwidths are computed using equation (7.18). The excitation and other conditions are same as in the previous case. The bias and variance values for 10 dB and 3 dB SNRs are given in Tables 7.3 and 7.4. We notice that weighted LP method performs better than standard LP method in this case also.

From the above experiments we may conclude that the weighted LP gives more consistent formant peaks and a better dynamic range for a noisy speech signal compared to that of the standard LP. Next we investigate the effect of the weighted LPCs in speech synthesis.

7.3.2. Experiments using natural speech signals

We have used the utterance "we were away a year ago" spoken by a male speaker in our experiments. The signal is low pass filtered to 5 kHz and sampled at 10 kHz. We have added noise samples at an average SNR level of 10 dB to this signal. The noisy speech signal and the local SNR plot are shown in Figs. 7.7a and 7.7b, respectively. (Local SNR plot gives local SNR in contrast to the average SNR which is 10 dB). We have used a window size of 256 samples for analysis. The standard LP and the weighted LP coefficients are computed once in every 6.4 msec. In both the cases an LP order of 10 was chosen. For computing standard LPCs, we use

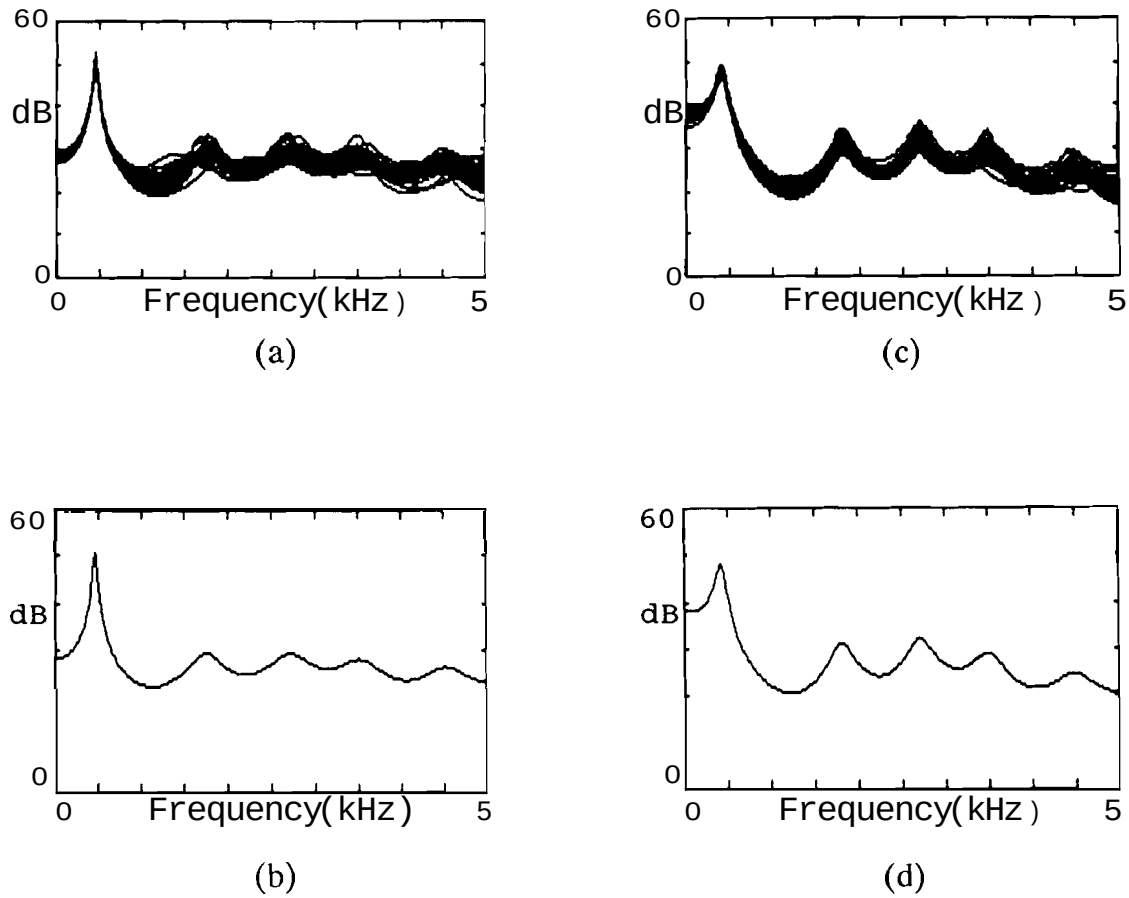


Fig. 7.5. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /e/ corrupted by additive noise at 10dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

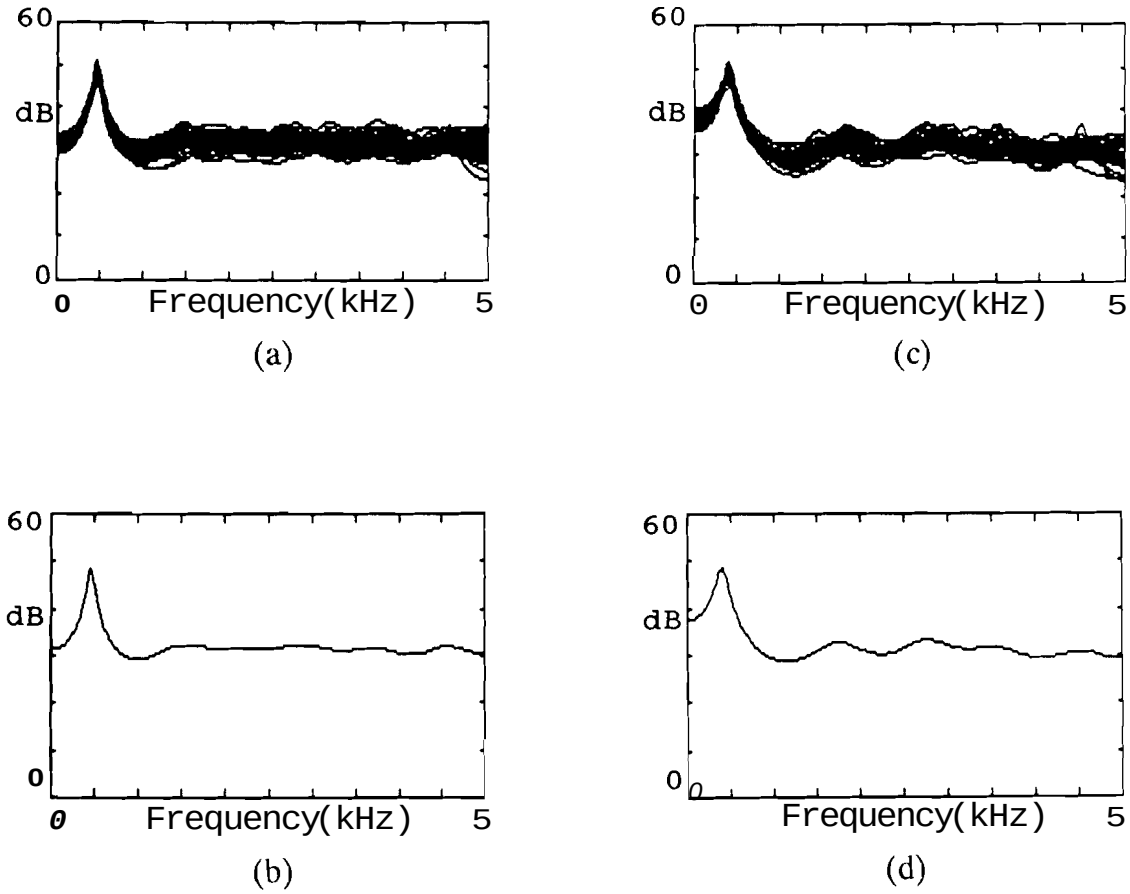


Fig. 7.6. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /e/ corrupted by additive noise at 3dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

TABLE 7.3

COMPARISON OF THE RESULTS FOR ESTIMATING THE FORMANT FREQUENCIES OF THE VOWEL /e/ WITH ADDITIVE NOISE AT SNR=10dB USING STANDARD LP AND WEIGHTED LP METHODS

Formant	LP		Weighted LP	
Freq.	Bias	Variance	Bias	Variance
1	17.08	371.77	-20.42	537.45
2	-79.02	13709.44	-6.36	1173.41
3	18.25	6713.68	14.34	1621.15
4	83.44	14590.72	41.69	6549.86
5	63.95	21542.24	56.0	9627.04

TABLE 7.4

COMPARISON OF THE RESULTS FOR ESTIMATING THE FORMANT FREQUENCIES OF THE VOWEL /e/ WITH ADDITIVE NOISE AT SNR=3dB USING STANDARD LP AND WEIGHTED LP METHODS

Formant	LP		Weighted LP	
Freq.	Bias	Variance	Bias	Variance
1	19.81	517.05	-40.0	1882.22
2	-207.85	104714.42	-74.72	22812.61
3	-73.92	98368.36	44.21	25676.78
4	101.5	91069.52	97.16	39956.92
5	98.62	51039.05	36.09	36125.51

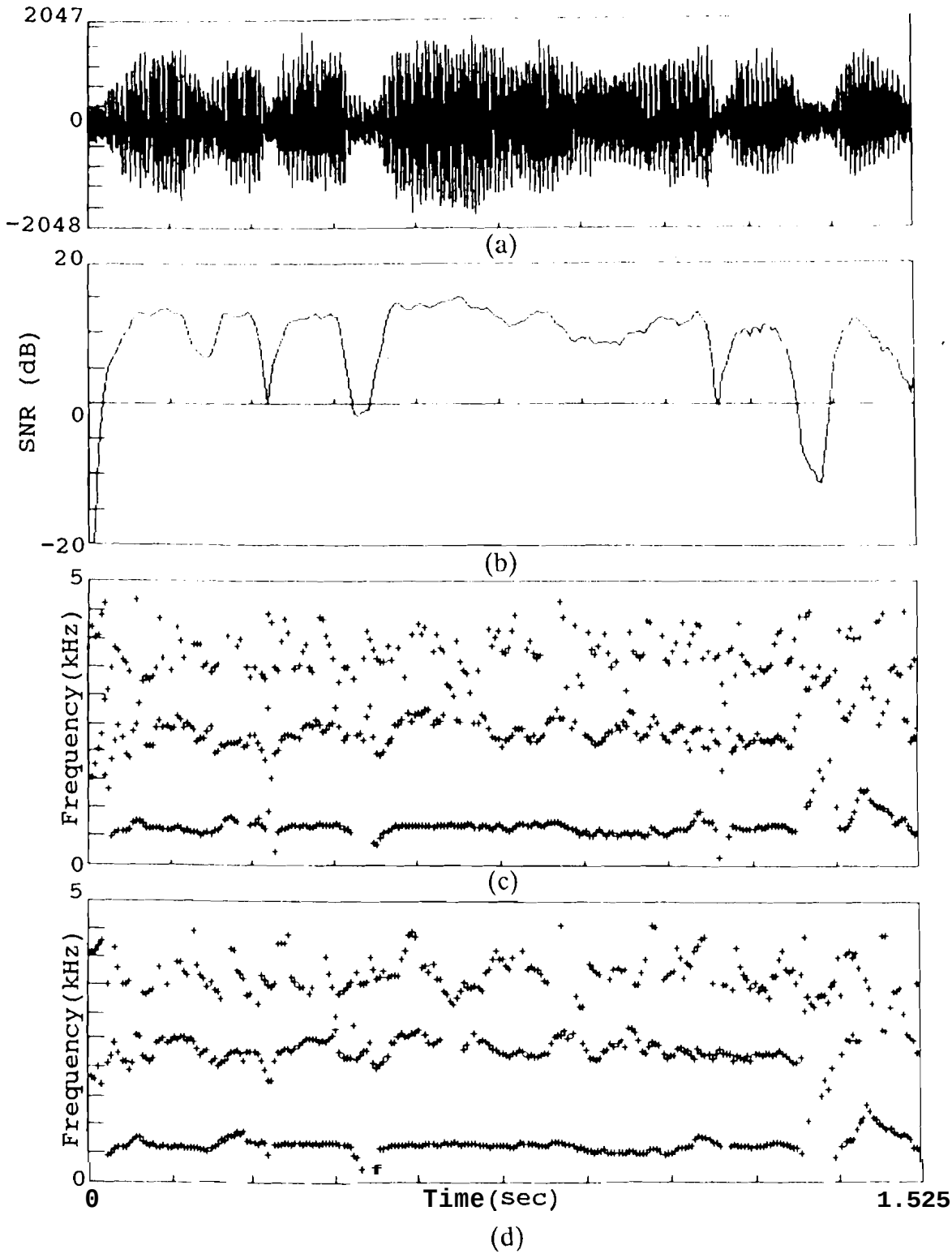


Fig. 7.7. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=10dB); (b) SNR plot; *(c)* formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

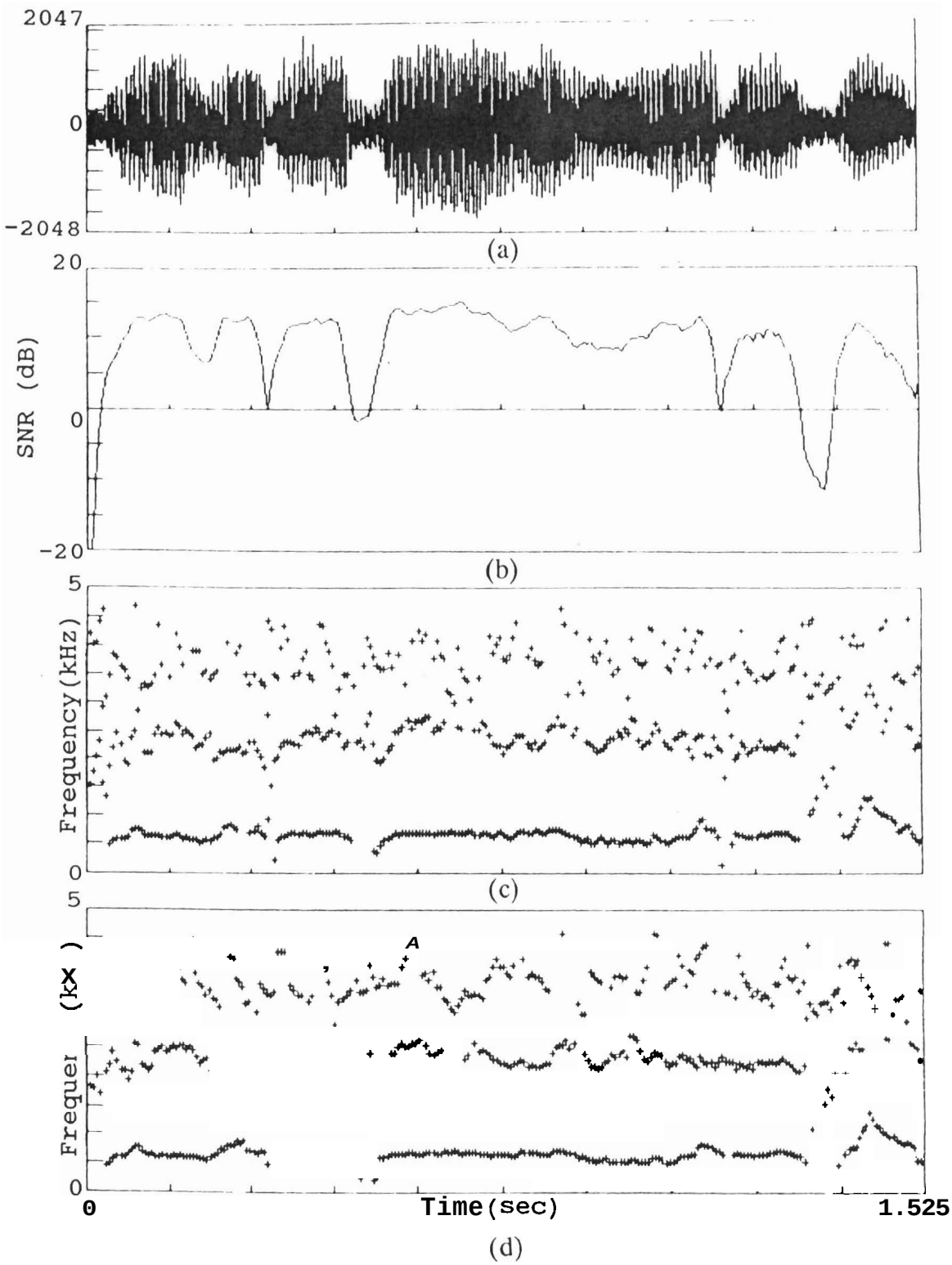


Fig. 7.7. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=10dB); (b) SNR plot; (c) formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

autocorrelation coefficients computed from hamming windowed input signal. For computing weighted LPCs, we use a 11-point smoothed absolute magnitude of the input signal as the weight function. The formant tracks obtained from the standard LP and weighted LP methods are shown in Figs. **7.7c** and **7.7d** respectively. We observe that the formant tracks from the weighted LP are better since (i) the discontinuities in the tracks are less and (ii) the spurious formant peaks are also reduced.

The above experiment is repeated with the background noise increased to make overall SNR equal to 3 dB. The corresponding plots are shown in Fig. 7.8. We observe that even in this case the weighted LP method gives superior formant tracks compared to standard LP method.

Though all these experiments show the advantages of the weighted LP method, ultimately the effectiveness of the method depends on the perceptual quality and intelligibility of the speech synthesized using this method. Hence we have conducted experiments in which the synthetic speech obtained from the standard LP method and the weighted LP method are compared. We have used an excitation function derived from the clean speech and **Fant's** model for glottal pulse in these experiments. The standard and weighted LPCs obtained in the previous experiments for SNR of 10 dB and 3 dB are used. The informal listening tests show that

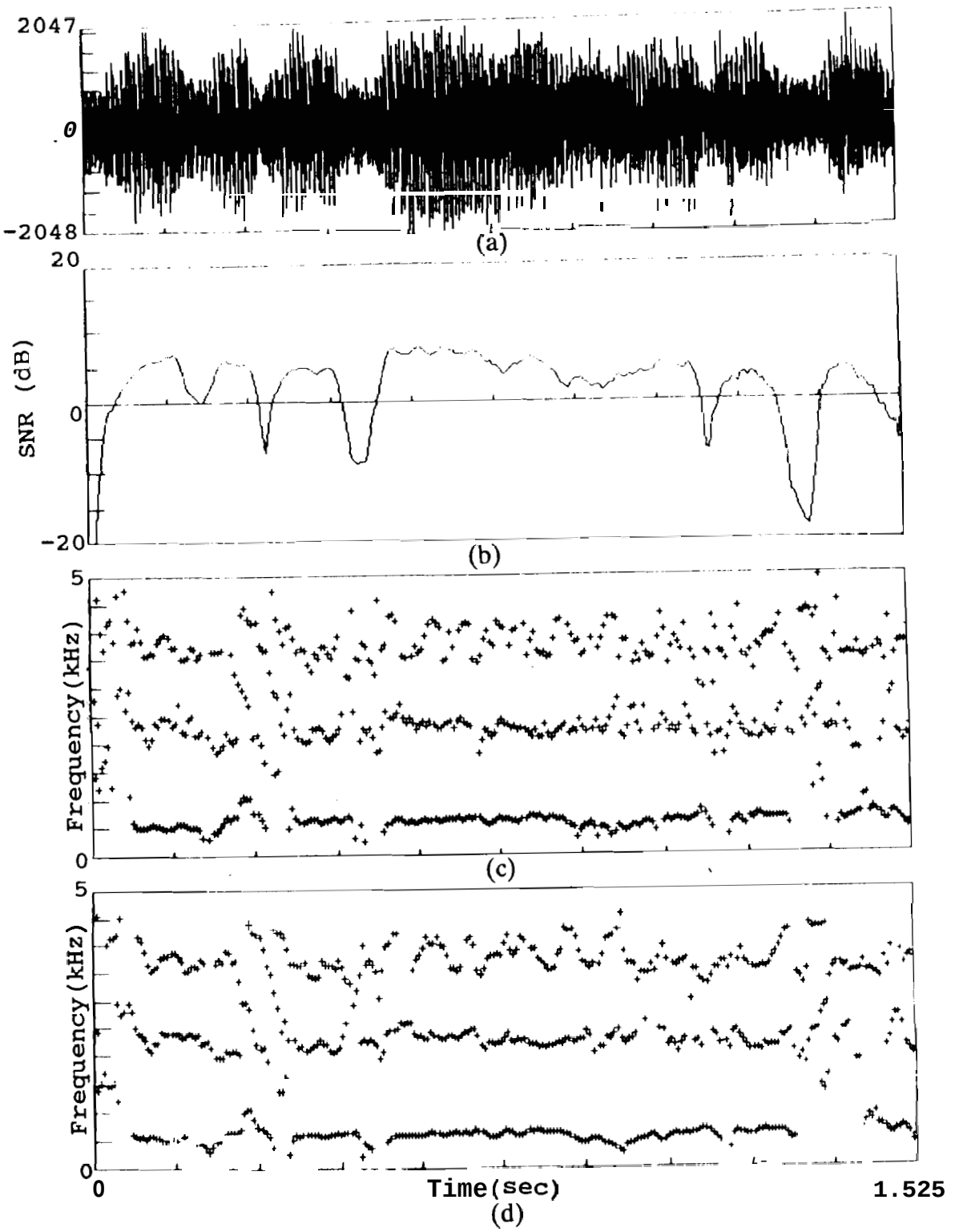


Fig. 7.8. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=3dB); (b) SNR plot; (c) formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

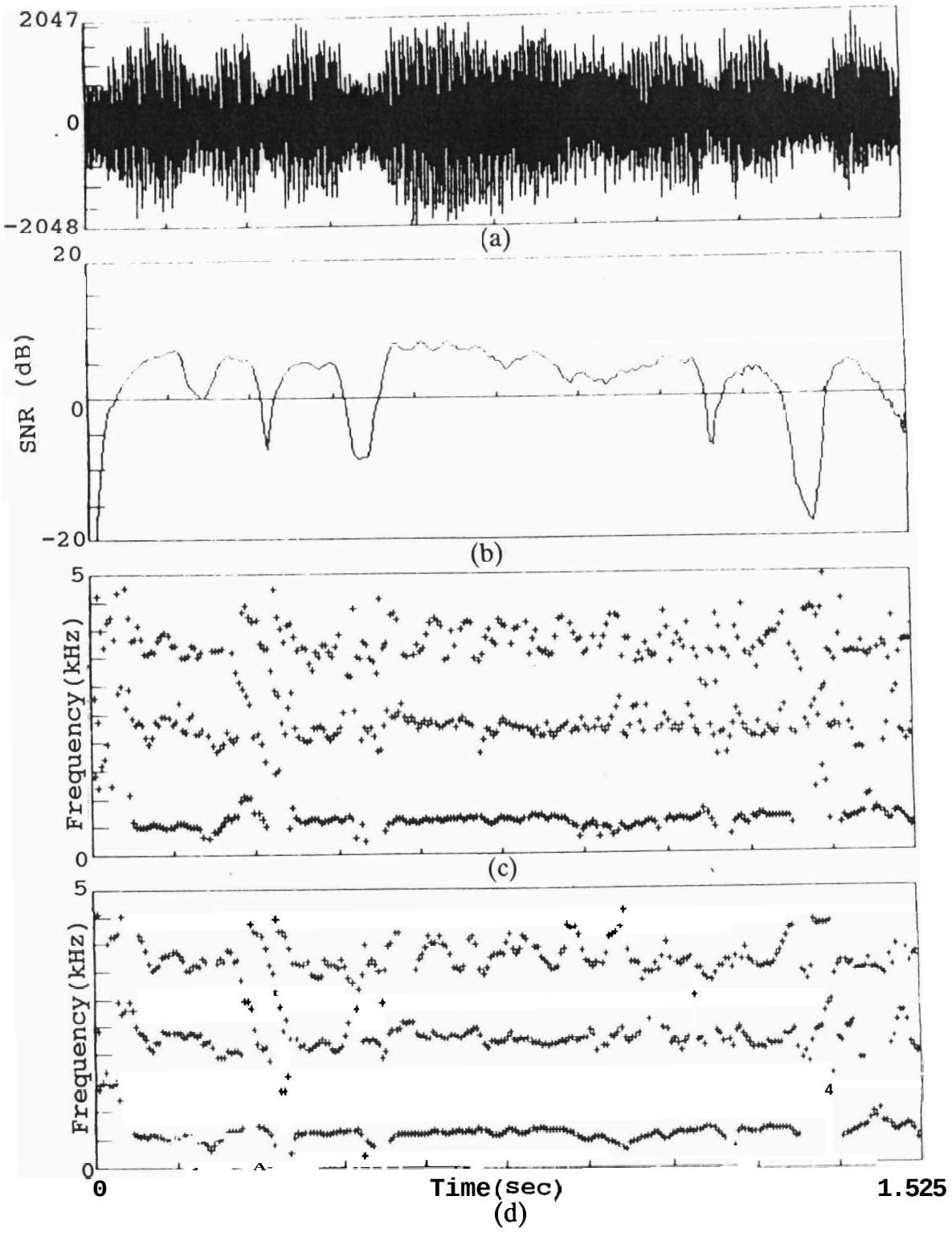


Fig. 7.8. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=3dB); (b) SNR plot; (c) formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

synthetic speech generated from weighted LPCs is distinctly better than that from standard LPCs for both the noise levels.

7.4. Discussion

In this chapter we have attempted to tackle the problems of noisy signal processing in the context of limitations imposed by the effects of time-frequency resolution. We have presented a new method based on weighted linear prediction to obtain an improved set of formant estimates from noisy speech signals.

Presently available methods for noisy signal processing require a priori information about the signal and/or noise. But such information may not be available in general. Hence the weighted LP method presented here is relevant as it does not require any a priori information.

In the weighted LP method we have used the principle that high SNR regions of a given noisy signal must be given importance in any signal processing algorithm. Choosing only high SNR regions of the noisy signal may require processing short data records. This results in time-frequency resolution problem. Hence we have used the above principle in the context of linear prediction analysis. In the linear prediction the analysis parameters are obtained by minimizing the sum of prediction errors. But in low SNR

regions the prediction error consists of mainly noise contribution and this fact must be taken into consideration while minimizing the prediction error. This is achieved in the weighted LP algorithm by using a weight function in the minimization step to emphasize prediction error in the high SNR regions. This results in a set of robust prediction coefficients as they are obtained by minimizing the prediction error predominantly in high SNR regions. We have given an analytical formulation for weighted LP method.

We have applied the weighted LP method on model signals as well as natural signals to evaluate its performance. We have found that the smoothed spectra obtained using this method have higher dynamic range compared to that obtained from the standard LP method. We have also noticed that the weighted LP method emphasizes formant peaks in the smoothed spectra. The formant estimates, in general, are more consistent. Bias and variance measurements also confirm this fact. The only exception to the above rule is the first formant in some cases.

We have experimented with noisy speech signals using the weighted LP method to find its effectiveness for speech enhancement. Comparison of the formant tracks show the superiority of the weighted LP method over the standard LP method. Informal listening tests of the synthetic speech

generated from the standard LP and the weighted LP coefficients confirm that the weighted LP method produces distinctly better quality speech.

One practical consideration that should be mentioned with regard to the weighted LP method is the stability of the weighted LP filter. We recall that the correlation matrix obtained from the weighted LP algorithm is symmetric but not Toeplitz. Hence the weighted LP filter is not guaranteed to be stable. In practice we found that the weighted LP filter tends to be unstable when (i) the weight function has large and rapid fluctuations and (ii) when the bandwidth of the first formant is low.

Chapter 8

SUMMARY AND CONCLUSIONS

8.1. Summary of the Thesis

We have studied in this thesis issues related to time-frequency resolution tradeoff for processing short segments of speech signals. Speech signals are processed over short segments to capture the time varying vocal tract system characteristics. Short segment processing results in time-frequency resolution problem. We have investigated the limitations of digital processing techniques which have given rise to this problem.

Discretization and truncation operations, which are necessary for digital processing lead to time-frequency resolution problem. The severity of this problem varies with the representation and processing algorithms chosen to process the signals. We have studied these limitations in the case of group delay function representation which is useful in a variety of signal processing applications due to its additive and high resolution properties. We have found that the signal information can be adequately represented and processed in the group delay domain except in those situations when (1) the roots of the time domain signal are too close to the unit circle in the z- plane; (2) there are too many roots; and (3) DFT size used in the processing is

small. In these cases the discretization and truncation affects dominate. Some signal information is lost in the group delay representation.

We have studied the time-frequency resolution problem in the case of decomposing model composite signals under stationary formulation. Here the problem is to separate the basic wavelets of different resonant frequencies and their epochs from a composite signal which consists of a summation of basic wavelets and their scaled and shifted replicas. We have compared the performance of simple band pass filtering and group delay filtering for composite signal decomposition. We have found that the group delay filtering performs better, indicating that choosing an appropriate representation for signal analysis helps in reducing the effects of time-frequency resolution problem.

Next we considered the time-frequency resolution problem for a time varying signal formulation. Most of the natural signals that we wish to process are time varying. In this connection we have studied the problem of source-tract interaction in speech where it is required to capture bandwidth changes of the vocal-tract formants within a pitch period. This can be considered as a special case of composite signal decomposition problem where it is required to find the instant and extent of bandwidth change of the basic wavelets within the analysis frame. We have

experimented with model composite signals in a time varying formulation. We have found that a bandwidth change in a basic wavelet in the time domain introduces a ripple in the group delay domain. The frequency and amplitude of this ripple are directly related to the instant and extent of bandwidth change. We proposed an inverse filter based method to detect the instant of bandwidth change in simple model signals that simulate the effects of source-tract interaction.

Next we have proposed a new method called modified LP method to process natural speech signals in a nonstationary formulation. We have studied the issues involved in capturing local variations such as source-tract interaction, of a time varying speech signal. Though we are interested in dynamic variations over short time intervals, we cannot consider short analysis segments as this would adversely affect the accuracy of the autocorrelation or covariance estimates. In the modified LP method, autocorrelation estimate over a long time interval is modified by using the autocorrelation of the LP residual over a short analysis frame to obtain a better estimate of the autocorrelation in that frame. We have found that this method gives consistent parameter estimates over short time intervals and captures the dynamic variations of the vocal tract system from the speech signal. Synthetic speech

generated using this method sounds more natural.

After getting insight into the issues relating to time-frequency resolution problem, we have investigated techniques for processing natural signals such as speech corrupted by background noise where time-frequency resolution stands out as a main problem. We have proposed a weighted LP analysis technique for processing noisy speech signal under quasistationary assumption. In this method certain regions (in the present case high SNR regions) within the analysis frame were emphasized for parameter estimation. The weighted LP analysis method has been found to give consistent formants in general and to improve the dynamic range of the estimated spectrum from the noisy speech. Synthetic speech generated using the weighted LP method sounds better than that using the standard LP method.

A speech analysis-synthesis package was developed to test the performance of the above techniques.

We have studied the issues connected with time-frequency resolution while processing speech signals in this thesis. We have observed that by choosing appropriate domains for signal representation and innovative processing techniques, it is possible to lessen the effects of time-frequency resolution problem which dominates the conventional representations and processing techniques.

8.2. Scope for Future Work

Our investigations in this thesis suggest that further work needs to be done in several directions. A rigorous theoretical formulation for the effectiveness of signal representation using group delay functions is yet to be evolved. Detection of source-tract interaction using group delay functions holds much promise. We have considered only synthetic signals in our investigations. If source-tract interaction can be detected in natural signals and incorporated in speech analysis-synthesis systems, we may be able to generate improved quality synthetic speech. In the weighted linear prediction method, we have favourably weighted high SNR regions only in the time domain. Similar type of weighting can be done in the frequency domain also or in both the time and frequency domains simultaneously.

Further investigations are required to decide the optimal shape of the weight function. We have shown that the dynamic variations of the vocal tract system are captured by the modified LP method. It may be interesting to study the effect of relative positioning of the short data segment with respect to glottal open and closure instances on the synthesized data. it is also worthwhile to investigate whether speech quality can be improved by combining the modified linear prediction with a better model of excitation like multipulse excitation.

Appendix

DESCRIPTION OF SPEECH ANALYSIS-SYNTHESIS PACKAGE

We describe in this Appendix, a simple analysis and synthesis system that we have used in our experiments. This is based on the principle of linear prediction (LP). The salient feature of this package is the flexibility incorporated in it to independently manipulate the parameters of the excitation and the vocal tract system.

In the linear prediction based approach, we impose on the speech signal a linear source-filter model [Makhoul 1975], where the source and the filter represent the excitation function and the vocal tract system respectively. But in practice, the filter represents not only the vocal tract system but also some contribution of the excitation and the radiation. The source is modeled in a simple way as a series of quasi-periodic impulses in the voiced frames and wide band random noise in the unvoiced frames. Thus the difficult problem of separation of the contributions of source function from that of the vocal tract system is avoided. In the synthesizer, the control parameters are reset to new values at the beginning of every analysis frame. Since one of our objectives is to determine the factors responsible for producing natural sounding synthetic speech, we have chosen to implement a scheme which controls

independently, the vocal tract system characteristics such as formant frequencies and bandwidths, and the excitation characteristics such as pitch, pitch contour and glottal wave shape, gain and gain contour. Hence we have followed an approach similar to that of a standard LPC-type vocoder.

We have developed a package of subroutines which is useful for speech analysis and synthesis. Using the analysis routines we can extract LP coefficients/formant frequencies, pitch and gain contours and other related parameters from a speech signal. The synthesis routines are useful in generating synthetic speech. This package has been used to perform various experiments connected with the investigations reported in this thesis.

In section A.l, we describe the speech analysis algorithms which include algorithms for linear prediction parameter extraction, pitch and gain extraction. Speech synthesis algorithms for generating the excitation function and synthesizing speech from the excitation function and LP filter are described in section A.2.

A. 1. Package for Speech Analysis

We analyze speech signal to extract the linear prediction coefficients corresponding to the vocal tract system, voiced/unvoiced decision and pitch & gain parameters corresponding to the excitation function. We discuss the

procedures used for extracting analysis parameters.

For speech analysis, a frame size of 25.6 msec (256 samples at 10 kHz sampling rate) and a frame rate of 100 frames per second are used. There is flexibility for changing both frame size and frame rate.

A.1.1. Extracting vocal tract parameters

For each frame of data, p autocorrelation lags are found. These lags form the coefficients of a set of linear simultaneous equations in p linear prediction parameters. These equations are solved by Durbin's recursion method to compute LPCs. Speech signal is first pre-emphasized and then windowed to prevent spurious peaks and to maintain the stability of the filter. We use a double sided Hamming window on the signal. Pre-emphasis is done using the following equation:

$$s(n) = s(n) - 0.9 * s(n-1).$$

A.1.2. Extraction of excitation parameters

Pitch extraction: There are a number of pitch extraction algorithms in the literature. We use a fairly simple one that computes the absolute magnitude difference function (AMDF) in this package. For voiced segments, the AMDF shows a prominent dip corresponding to a lag equal to the pitch period. For unvoiced segments, the AMDF does not show a

show a single prominent dip, but there may be several smaller dips at random lags. In the pitch extraction algorithm using AMDF, we check whether there is a dip smaller than a threshold value. If such a dip is found, then the frame is labeled as a voiced frame and the position of the dip gives the pitch period. The frames where all the dips are above the threshold are labeled as unvoiced frames. Though the AMDF based algorithm for pitch extraction is adequate in many cases, it may fail to accurately label the frame as voiced/unvoiced in certain situations such as when the voicing is weak or when the pitch is changing rapidly. Hence a certain amount of manual editing is required to correct such errors before further processing of pitch can take place.

Energy computation: The residual energy of an analysis frame after removing the predictable part is obtained as a by-product in the Durbin's recursion algorithm for computing LPCs. The energy can also be computed directly from the LP residual. Gain per sample is computed from this energy value.

A.1.3. Generation of pitch and gain contours

The pitch and gain contours are used for computing the excitation signal. The pitch and gain per sample values are interpolated to generate these contours. Abrupt

variations in the pitch and gain contours may appear at some places due to the limitations of the analysis procedures. Some form of smoothing has to be used as outlined below to smooth out these abrupt changes as they are undesirable.

The pitch values collected from each frame are interpolated to generate pitch values at intervals of pitch period to form a pitch contour. In addition, an m point median filter is applied to smooth the pitch contour in the intervals corresponding to the voiced regions. The value of m is normally taken as either 3 or 5. A value of 0 is set in the intervals corresponding to the unvoiced regions in the pitch contour.

The gain per sample values for each frame are interpolated to generate gain per sample for each pitch period to form a gain contour. An m point (m = 3 or 5) averaging algorithm is used to smooth the gain contour.

A.2. Package for Speech Synthesis

We synthesize speech using the excitation signal generated from the pitch and gain contours and LPCs. We describe the procedures used for generating the excitation signal and for synthesizing speech.

A.2.1. Generation of excitation signal

Three sets of parameters are required to generate

the excitation signal. They are voiced/unvoiced decision, gain per sample, pitch period. In addition, the shape of the excitation pulse is to be chosen. The unvoiced regions are identified by looking for zero values in the pitch contour. Energy per frame is computed using gain contour in those regions, and random noise at that energy level is generated as excitation signal. In the voiced regions, pitch is determined from the pitch contour, and the frame width is taken as equal to the pitch period. The energy per frame is computed from the gain contour and the number of samples in that frame. We have used two alternate waveforms for the pitch pulse: an impulse at the beginning of each pitch period or Fant's model. In both the cases the dc component is removed from the signal.

A.2.2. Generation of synthetic speech

The excitation signal, which is generated as described in the last section is convolved with the LP filter to generate the synthetic speech signal. The LP filter which is formed using the LP coefficients, is updated at the same rate at which the LPCs are obtained from the speech signal at the time of analysis. The synthetic signal thus obtained is de-emphasized to compensate for the pre-emphasis given at the time of analysis.

REFEREN(

1. D.R.Allen and W.J.Strong, "A model for the synthesis of natural sounding vowels," J. Acoust. Soc. Am., Vol.78, pp.58-69, 1985.

2. L.B.Almeida and J.M.Tribolet, "Nonstationary spectral modeling of voiced speech," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-31, pp.664-678, Jun. 1983.

3. A.S.Ananth, D.G.Childers and B.Yegnanarayana, "Measuring source-tract interaction from speech," Proc. IEEE ICASSP-85, pp.1093-1096, 1985.

4. T.V.Ananthapadmanabha, "Composite signal decomposition by digital inverse filtering," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-27, pp.95-97, Feb. 1979.

5. T.V.Ananthapadmanabha and G.Fant, "Calculation of the true glottal flow and its components," Speech Communication, Vol.1, pp.167, 1982.

6. B.S.Atal and S.L.Hanauer, "Speech analysis and

synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., Vol.50, No.2, pp.637-655, 1971.

7. B.S.Atal and J.R.Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," Proc. IEEE ICASSP-82, pp.614-617, 1982.

8. S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-27, pp.113-120, Apr. 1979.

9. E.O.Brigham, The Fast Fourier Transform, Prentice-Hall, Englewood Cliffs, NJ, 1974.

10. F.Casacuberta and E.Vidal, "A nonstationary model for the analysis of transient speech signals," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-35, pp.226-228, Feb. 1987.

11. S.Chandra and W.C.Lin, "Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-22, pp.403-415, Dec. 1974.

12. M.C.Chevalier, C.Collet, and Y.Grenier, "Speech analysis and restitution using time dependent auto regressive models," Proc. IEEE ICASSP-85, pp.501-504, 1985.

13. D.G.Childers, R.S.Varga and N.W.Perry Jr., "Composite signal decomposition," IEEE Trans. Audio Electro Acoust., Vol.AU-18, pp.471-477, Dec. 1970.

14. D.G.Childers and R.C.Kemeriat, "Signal detection and extraction by cepstrum techniques," IEEE Trans. Inform. Theory, Vol.IT-18, pp.745-759, Nov. 1972.

15. D.G.Childers and M.T.Pao, "Complex demodulation for transient wavelet detection and extraction," IEEE Trans. Audio Electro Acoust., Vol.AU-20, pp.295-308, Oct. 1972.

16. D.G.Childers, D.P.Skinner and R.C.Kemeriat, "The cepstrum: A guide to processing," Proc. IEEE, Vol.65, pp.1428-1443, Oct. 1977.

17. D.G.Childers, J.J.Yea and E.L.Bocchieri, "Source/vocal-tract interaction in speech and singing synthesis," Proc. Stockholm Music Acoustics Conference, Vol.1,

pp.125-141, 1983.

18. D.G.Childers, B.Yegnanarayana and Ke Wu, "Voice conversion: factors responsible for quality," Proc. IEEE ICASSP-85, pp.748-751, 1985.

19. J.W.Cooley and J.W.Tukey, "An algorithm for the machine computation of complex Fourier series," Math. Computation, Vol.19, pp.297-381, Apr. 1965.

20. R.J.P. de Figueiredo and A.Gerber, "Separation of superimposed signals by a cross correlation method," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-31, pp.1084-1089, Oct. 1983.

21. W.J.Done and C.K.Rushforth, "Estimating the parameters of a noisy all pole process using pole zero modeling," Proc. IEEE ICASSP-79, pp.228-231, 1979.

22. H.K.Dunn, "Methods of measuring vowel formant bandwidths," J. Acoust. Soc. Am., Vol.33, No.12, pp.1737-1746, 1961.

23. G.Fant, "Glottal source and excitation analysis," STL - QPSR KTH, No.1, Stockholm, Sweden, pp.85-107, 1979.

24. G.Fant, "The voice source-acoustic modeling," STL - QPSR 4/1982, pp.28-48, 1982.

25. S.T.Fathima, "Studies on a class of information recovery problems," M.S. Thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India, 1986.

26. R.H.Fraziar, S.Samsam, L.D.Braida, and A.V.Oppenheim, "Enhancement of speech by adaptive filtering," Proc. IEEE ICASSP-76, pp.251-253, 1976.

27. Y.Grenier, "Nonstationary ARMA models via simultaneous AR and MA estimates," Proc. IEEE ICASSP-86, pp.2339-2342, 1986.

28. M.G.Hall, A.V.Oppenheim, and A.S.Willsky, "Time-varying parametric modeling of speech," Signal Processing, Vol.5, pp.267-285, 1983.

29. M.H.Hayes, J.S.Lim and A.V.Oppenheim, "Signal reconstruction from phase or magnitude," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, pp.672-680, Dec. 1980.

30. S.Haykin, Adaptive Filter Theory, Prentice-Hall, Englewood Cliffs, NJ, 1985.

31. C.W.Helstrom, Statistical Theory of Signal Detection, New York:Pergamon, 1967.

32. Hema A. Murthy, A.A.Babu, B.Yegnanarayana and K.V.Madhu Murthy, "Speech coding using Fourier transform phase," Proc. EUSIPCO-88, Grenoble, France, pp.879-882, Jul. 1988.

33. Hema A. Murthy, K.V.Madhu Murthy and B.Yegnanarayana(1989a), "Formant extraction from Fourier transform phase," PROC. IEEE ICASSP-89, 1989, pp.484-487.

34. Hema A. Murthy, K.V.Madhu Murthy and B.Yegnanarayana(1989b), "Formant extraction from phase using weighted group delay function," Electronic Letters, Vol.25, No.23, 9th Nov. 1989, pp.1609-1611.

35. F.Itakura and T.Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum," PROC. IEEE ICASSP-87, pp.1257-1260, 1987.

36. Jong-Ho Choi, Si-Whan Kim, and Jong-Soo Choi, "Measurement methods of ultrasonic velocity by dip point analysis from echo spectrum and ultrasonic attenuation considering a dispersion," Proc. IEEE ICASSP-86, pp.1777-1780, 1986.

37. O.Kakusho and M.Yanagida, "Hierarchical AR model for time varying speech signals," Proc. IEEE ICASSP-82, pp.1295-1298, 1982.

38. O.Kakusho, M.Yanagida and R.Mizoguchi, "A sample-selective linear prediction of speech," PROC. IEEE ICASSP-84, 1984.

39. S.M.Kay and S.L.Marple, "Spectrum analysis- A modern perspective," Proc. IEEE, Vol.69, No.11, pp.1380-1418, Nov. 1981.

40. S.M.Kay, Modern Spectral Estimation, Theory & Application, Prentice-Hall, Englewood Cliffs, NJ, 1988.

41. D.Kewley-Port, "Time varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am., Vol.73, pp.322-335, 1983.

42. H.Kobatake, J.Inasi, and S.Kakuta, " Linear prediction coding of speech signals in a high ambient noise environment," Proc. IEEE ICASSP-78, pp.472-475, Apr. 1978.

43. A.K.Krishna Murthy and D.G.Childers, "Two channel speech analysis," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-34, pp.730-743, Aug. 1986.

44. J.N.Larar, Y.A.Alsaka and D.G.Childers, "Variability in closed phase analysis of speech," PROC. IEEE ICASSP-85, pp.1089-1092, 1985.

45. Y.T.Lee and H.F.Silverman, "On a general time varying model for speech signals," Proc. IEEE ICASSP-88, pp.95-98, 1988.

46. J.S.Lim and A.V.Oppenheim, "All pole modeling of degraded speech," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-26, pp.197-210, Jun. 1978.

47. J.S.Lim and A.V.Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, Vol.67, pp.1586-1604, Dec. 1979.

48. J.S.Lim, (Ed.), Speech Enhancement, Prentice Hall, Inc., 1982.

49. K.V.Madhu Murthy and B.Yegnanarayana, "Composite signal decomposition using group delay functions," Proc. Int. Conf. on Communication Systems (ICCS'88), Singapore, pp.1036-1040, 1988.

50. K.V.Madhu Murthy and B.Yegnanarayana, "Effectiveness of representation of signals through group delay functions," Signal Processing, Vol.17, No.2, pp.141-150, Jun. 1989.

51. J.Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, vol.63, No.4, pp.561-580, Apr. 1975.

52. S.L.Marple, Digital Spectral Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1987.

53. J.V.McCanny and J.C.White (Eds.), VLSI Technology and Design, Academic Press, Orlando, Florida, 1987.

54. Y.Miyoshi, K.Yamato, M.Yanagida and O.Kakusho, "Analysis of speech signals of short pitch period by the sample selective linear prediction," PROC. IEEE

ICASSP-86, pp.1245-1248, 1986.

55. B.R.Musicus and J.S.Lim, "Maximum likelihood parameter estimation of noisy data," Proc. IEEE ICASSP-79, pp.224-227, 1979.

56. A.V.Oppenheim and R.W.Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

57. S.Parthasarathy and D.W.Tufts, "Excitation synchronous modeling of voiced speech," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-35, pp.1241-1249, Sep. 1987.

58. M.R.Portnoff, "Short-time Fourier analysis of sampled speech," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-29, pp.364-373, Jun. 1981.

59. L.R.Rabiner, B.S.Atal, and M.R.Sambur, "LPC prediction error analysis of its variation with the position of the analysis frame," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-25, pp.434-442, Oct. 1977.

60. L.R.Rabiner and R.W.Schafer, Disital Processinq of

Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

61. E.A.Robinson, T.S.Durrani, and L.G.Peardon, Geophysical Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1986.

62. M.Ross, H.Schafer, A.Cohen, R.Freuberg, and H.Manley, "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-22, pp.353-362, 1974.

63. M.R.Sambur, "An efficient linear prediction vocoder," Bell Syst. Tech. J., Vol.54, pp.1693-1723, Dec. 1975.

64. M.R.Sambur and N.S.Jayanth, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-24, pp.488-494, Dec. 1976.

65. M.R.Sambur, "Adaptive noise canceling for speech signals," IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-26, pp.419-423, Oct. 1978.

66. S.Senmoto and D.G.Childers, "Adaptive decomposition of

a composite signal of identical unknown wavelets in noise," IEEE Trans. Syst. Man, Cybern., Vol.SMC-2, pp.59-66, Jan. 1972.

67. M.I.Skolnik (Ed.), Radar Handbook, McGraw-Hill, New York, 1970.

68. M.M. Sondhi, @@Newmethods of pitch extraction," IEEE Trans. Audio Electro Acoust., Vol.AU-16, pp.262-266, 1968.

69. K.Steiglitz and B.Dickenson, "The use of time-domain selection for improved linear prediction,@@IEEE Trans. Acoust. Speech and Signal Processing, Vol.ASSP-25, pp.34-39, Feb. 1977.

70. B.Widrow and S.D.Stearns, Adaptive Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1985.

71. J.H.Wilkinson and C.Reinsch, Linear Alsebra, Vol.II, New York: Springer-verlag, 1971.

72. F.T.Yarman-Vural, @@Enhancementof speech in additive, locally stationary and colored noise, using linear prediction,@@Signal Processing, Vol.20, No.3, pp.211-

217, Jul. 1990.

73. B.Yegnanarayana, "Formant extraction from linear-prediction phase spectra," J. Acoust. Soc. Am., Vol.63, pp.1638-1640, 1978.

74. B.Yegnanarayana and D.Raj Reddy, "A distance measure based on the derivative of linear prediction phase spectrum," Proc. IEEE ICASSP-79, pp.744-747, 1979.

75. B.Yegnanarayana(1981a), "Speech analysis by pole-zero decomposition of short-time spectra," Signal Processing, Vol.3, No.1, pp.5-17, Jan. 1981.

76. B.Yegnanarayana(1981b), "Design of ARMA digital filters by pole- zero decomposition," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-29, pp.433-439, Jun. 1981.

77. B.Yegnanarayana, D.G.Childers and J.M.Naik, "A flexible analysis-synthesis system for studies in speech processing," Technical Report, University of Florida, Gainesville, 1983.

78. B.Yegnanarayana, "Significance of group delay functions

in signal processing," Proc. IEEE Symposium on Circuits and Systems, Montreal, Canada, 1984.

79. B.Yegnanarayana, D.K.Saikia and T.R.Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-32, pp.610-622, Jun. 1984.

80. B.Yegnanarayana, J.Sreekanth and A.Rangarajan, "Waveform estimation using group delay processing," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-33, pp.832-836, Aug. 1985.

81. B.Yegnanarayana and K.V.Madhu Murthy, "Effectiveness of representation of signals through group delay functions," ISSPA-87, Brisbane, Australia, pp.280-285, Aug. 1987.

82. B.Yegnanarayana, S.T.Fathima and Hema A. Murthy, "Reconstruction from Fourier transform phase with applications to speech analysis," PROC. IEEE ICASSP-87, pp.301-304, 1987.

83. B. Yegnanarayana, K.V. Madhu Murthy and Hema A. Murthy,

"Processing of noisy speech using partial phase," Proc. European Conference on Speech Technology, Edinburgh, pp. 203-206, Sep. 1987.

84. B.Yegnanarayana, K.V.Madhu Murthy and Hema A. Murthy, "Applications of group delay functions in speech processing," JIETE Special issue on Speech Processing, Vol.34, No.1, pp.20-29, Jan-Feb 1988.

85. B.Yegnanarayana, G.Duncan, and Hema A. Murthy, "Improving formant extraction from speech using minimum phase group delay spectra," IEE Colloquim on Speech Processing, 1988.

LIST OF FIGURES

Fig. 3.3. Comparison of the original and reconstructed signals for different values of echo amplitude (γ). Echo shift and DFT size are fixed at 6.2 msec and 512 in the above plots.

Fig. 3.4. Comparison of the original and reconstructed signals for different values of the number of zeros (n_1) in the z-plane. Echo amplitude and DFT size are fixed at **0.8** and 512 in the above plots.

Fig. 3.5. Comparison of the original and reconstructed signals for different values of the DFT size (N). Signal length is 16 samples, echo amplitude is **0.8** and the echo shift is 6 in the above plots.

Fig. 3.6. Reconstruction error averaged over 60 frames of data. The variation in the normalized mean squared error (NMSE) with respect to (a) echo amplitude (γ); (b) number of zeros (n_1) in the z-plane, and (c) DFT size (N).

Fig. 4.1. Identification of resonances in a composite signal using minimum phase group delay spectrum: (a) The composite signal; (b) its FT magnitude spectrum; (c) 12th order LP smoothed magnitude; (d) minimum phase group delay spectrum.

Fig. 4.2. Illustration of the group delay filtering

technique: (a) FT magnitude spectrum of the composite signal; (b) its group delay (τ_m) function; (c) group delay filtering to select **1800** Hz component, and (d) reconstructed FT magnitude spectrum.

Fig. **4.3**. Illustration of the design of the weighting functions from minimum phase group delay: (a) FT magnitude spectrum of a composite signal; (b) the corresponding minimum phase group delay function; (c) modified minimum phase group delay function to filter the **1800** Hz component; (d) magnitude weighting function and (e) phase weighting function derived from the modified minimum phase group delay.

Fig. **4.4**. Illustration of application of weighting functions for composite signal decomposition: (a) The composite signal; (b) its FT magnitude; (c) weighting function to filter the **1800** Hz component; (d) modified magnitude after applying the weighting function; (e) the resultant filtered **1800** Hz component.

Fig. **4.5**. Component wavelets of the composite signal reconstructed by various decomposition techniques: (a)-(c) Component wavelets obtained using band pass filtering; **(d)-(f)** corresponding

for the short segments B, C and D respectively. The short segment size is 1.6 msec (16 samples).

Fig. 6.4. Illustration of the effectiveness of modified linear prediction in capturing local variations in short segments: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.4a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 2.4 msec (24 samples).

Fig. 6.5. Illustration of the effectiveness of modified linear prediction in capturing local variations in short segments: (a)-(c) The speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.5a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)- 1 and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 3.2 msec (32 samples).

Fig. 6.6. Illustration of the effectiveness of modified

linear prediction in capturing local variations in a frame containing an unvoiced to voiced transition: (a)-(c) the speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.6a), the corresponding modified auto-correlation and LP spectrum; ( )-( ) (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 1.6 msec (16 samples).

Fig. 6.7. Illustration of the effectiveness of modified linear prediction in capturing local variations in a frame containing formant transitions: (a)-(c) the speech signal, its autocorrelation and the LP smoothed spectrum; (d)-(f) the autocorrelation of the LP residual corresponding to the short segment A (indicated in Fig. 6.7a), the corresponding modified autocorrelation and LP spectrum; (g)-(i), (j)-(l) and (m)-(o) show similar plots for the short segments B, C and D respectively. The short segment size is 1.6 msec (16 samples).

Fig. 6.8. Illustration of the ripple introduced in the formant tracks by the modified linear prediction

method: (a) The waveform of the sentence "We were away a year ago" spoken by a male speaker; its formant tracks obtained by (b) standard LP, frame size=25.6 msec; (c) modified LP, frame size=25.6 msec, short frame size=3.2 msec; (d) modified LP, frame size=25.6 msec, short frame size=1.6 msec; (e) standard LP, frame size=3.2 msec. In all the above cases the parameters are computed once in every 1.6 msec.

Fig. 6.9. Illustration of the ripple introduced in the formant tracks by the modified linear prediction method: (a) The waveform of the sentence "We were away a year ago" spoken by a female speaker; its formant tracks obtained by (b) standard LP, frame size=25.6 msec; (c) modified LP, frame size=25.6 msec, short frame size=3.2 msec; (d) modified LP, frame size=25.6 msec, short frame size=1.6 msec; (e) standard LP, frame size=3.2 msec. In all the above cases the parameters are computed once in every 1.6 msec.

Fig. 7.1. Formulation of different weighting functions based on the SNR criterion: (a) A typical noisy speech signal (SNR=3dB) and (b), (c), (d), and (e) corresponding weighting functions based on the absolute magnitude of the noisy signal. (See

text for the equations of the weighting functions).

Fig. 7.2. Illustration of the effectiveness of the weighted LP method in enhancing the smoothed spectrum of the synthetic sound /a/ corrupted by additive noise at 3dB SNR: (a) FT magnitude spectrum; (b) standard LP spectrum; and (c) weighted LP spectrum.

Fig. 7.3. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /a/ corrupted by additive noise at 10dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

Fig. 7.4. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /a/ corrupted by additive noise at 3dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

Fig. 7.5. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /e/ corrupted by additive

noise at 10dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations; (c) fifty overlaid realizations for weighted LP; (d) average of realizations.

Fig. 7.6. Illustration of the effectiveness of weighted LP method in estimating the smoothed spectrum of the synthetic sound /e/ corrupted by additive noise at 3dB SNR: (a) Fifty overlaid realizations for standard LP; (b) average of realizations;.(c) fifty overlaid realizations for weighted LP; (d) average of realizations.

Fig. 7.7. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=10dB); (b) SNR plot; (c) formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

Fig. 7.8. Illustration of the effectiveness of weighted LP method in tracking formants of noisy speech: (a) Noisy speech signal (average SNR=3dB); (b) SNR plot; (c) formant tracks obtained using standard LP; (d) formant tracks obtained using weighted LP.

LIST OF TABLES

LIST OF SPECIAL SYMBOLS

$\Delta_o(t)$	Sampling function with a period T.
ω	Angular frequency.
$\theta(\omega)$	Fourier transform phase function.
$\theta_u(\omega)$	Fourier transform phase function in unwrapped form.
$\tau_m(\omega)$	Group delay function derived from FT magnitude.
$\tau_p(\omega)$	Group delay function derived from FT magnitude.
γ	The ratio of the echo amplitude and the basic signal amplitude.
β	Damping constant.
$\Delta\beta$	The amount of change in the damping constant.
$\phi_w(k,i)$	Weighted correlation function.

LIST OF PUBLICATIONS

1. B. Yegnanarayana and K. V. Madhu Murthy, "Effectiveness of representation of signals through group delay functions," ISSPA-87, Brisbane, Australia, pp. 280-285, 1987.
2. B. Yegnanarayana, K.V. Madhu Murthy, and Hema A. Murthy, "Processing of noisy speech using partial phase," European Conference on Speech Technology, Edinburgh, pp. 203-206, 1987.
3. B. Yegnanarayana, K.V. Madhu Murthy and Hema A. Murthy, "Applications of group delay functions in speech processing," JIETE Special issue on Speech processing, vol. 34, No. 1, pp. 20-29, Jan-Feb, 1988.
4. Hema A. Kurthy, A.A. Babu, B. Yegnanarayana, and K.V. Madhu Murthy, "Speech coding using Fourier transform phase," in Proc. EUSIPCO-88, pp. 879-882, 1988.
5. K.V. Madhu Murthy and B. Yegnanarayana, "Composite signal decomposition using group delay functions," in Proceedings of International Conference on Communication Systems (ICCS'88), Singapore, pp. 1036-1040, 1988.
6. K.V. Madhu Murthy and B. Yegnanarayana, "Effectiveness of

representation of signals through group delay functions," Signal Processing, vol.17, No.2, pp.141-150, June 1989.

7. Hema A. Murthy, K.V.Madhu Murthy, and B.Yegnanarayana, "Formant extraction from Fourier transform phase," in Proc. IEEE ICASSP-89, pp. 484-487, 1989.

8. Hema A. Murthy, K.V.Madhu Murthy, and B.Yegnanarayana, "Formant extraction from phase using weighted group delay function," Electronics Letters, vol.25, No.23, pp. 1609-1611, 9th Nov. 1989.