SHORT SEGMENT ANALYSIS OF SPEECH FOR ENHANCEMENT

A THESIS

submitted by

P. SATYANARAYANA

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

FEBRUARY 1999

To my

Parents, Teachers

and well-wishers

THESIS CERTIFICATE

This is to certify that the thesis entitled **SHORT SEGMENT ANALYSIS OF SPEECH FOR ENHANCEMENT** submitted by **P. Satyanarayana** to the Indian Institute of Technology, Madras for the award of the degree of Doctor of Philosophy is a bonafide record of research work carried out by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. K. M. M. Prabhu
Dept. of Electrical Engg.

Prof. B. Yegnanarayana
Dept. of Computer Science and Engg.

Madras 600 036

Date:

ACKNOWLEDGEMENTS

I thank Prof. Yegnanarayana for all his personal interest in me and for the excellent research environment he has created for all of us to learn. His critical feedback during the *Monday seminars* actually taught me how to make a presentation before an audience. It was a pleasure to learn from him, both in the classroom as well as in the technical discussions we used to have. I thank him for all the valuable time he spent in explaining difficult concepts to my average mind. I thank him for patiently correcting several drafts of my thesis.

Prof. Prabhu has been more like a friend to me and has been a constant source of moral support and encouragement. I am thankful to him for his help and cooperation throughout this work and for his valuable advice on several occasions.

I am grateful to Dr. Carlos Avendaño, now in UC Davis, for all the Matlab programming he taught me and for allowing me to use some of his programs.

I learnt a lot in my informal discussions with Dr. Hema, Dr. Chandrasekhar, Dr. Sundar and Dr. Sarnudravijaya (TIFR). I am grateful to them for all their personal interest in me and for their encouragement.

My special thanks to Mr. Rajendran who has been like an elder brother to me, and who has helped me immensely on numerous occasions. I will miss his reassuring bass voice. I thank Tamil for all his help while he was here.

Thank you, Manish, for all your help and concern, and for all that you taught me.

I am extremely thankful to all my bright and young friends in the lab: Anitha, Anjani, Hemant, Ikku, Jyotsna, Kishore, Mathew, Pragathi and Sivaramakrishna, who competed with my own brothers in being affectionate to me. I thank you all for your help, cooperation and encouragement besides all that I learnt from you. Thank you, Anjani, for patiently going through several chapters of the thesis and giving your feedback. I thank my friends Kavitha, Sarathy and Srikanth for their moral support and encouragement, during their stay here. Thank you, Karna, for your timely help on many occasions.

I will fondly remember for a long time to come the wonderful time I had with my friends

in the hostel: Sarma, Ranjit, Panku (Pankaj), Avssp (Prasad), Sivaramakrishna, Roopie (Rupesh), Ranu (Rana), Bucket (Amit), Dinnu and the Marathi Junta'97. Thank you, Wilson, for all the hours I spent listening to you playing the piano and for all that you taught me. It was sheer pleasure.

I can never forget the immense help extended to me by Mr. Rajendran in the Research section (Ad block). His calm voice cooled my nerves on so many occasions. I am also thankful to Mr. Anbarasan (Research section), Mr. Krishnan (EE Dept.) and Mr. Osman (Scholarship section) for their help. I also thank the mess staff of Krishna hostel who used to go out of their way to be kind and helpful to me.

I thank my uncle and aunt: Dr. and Mrs. Kalidas, my aunt Mrs. Venku Bai and all my cousins for their love and affection, during my stay here.

Needless to mention the love and moral support of my father, mother and my two brothers, Raman and Ravi. This work would not have been possible but for their support.

Finally, I thank my fellow citizens of India for contributing either directly or indirectly to make education at IITs possible. It is a luxury, considering the fact that there are millions of children who never get an opportunity to even attend school in their life.

P. Satyanurayana

ABSTRACT

KEYWORDS: speech analysis; speech enhancement; linear prediction residual signal; reverberation; instants of significant excitation.

Speech is the most natural means of communication among human beings. Speech signals are nonstationary in nature. Processing nonstationary signals to capture their dynamic behaviour requires analysis of short data records. Additionally, in practical environments speech signals are often degraded due to ambient noise, telephone channel distortion, reverberation in rooms and speech of competing speakers. Processing degraded speech is required for enhancement and for systems such as speech recognition and speaker verification systems. Analysis of speech is normally performed at the suprasegmental **(100–300** ms) level and at the segmental **(10–30** ms) level. These analysis methods do not exploit the characteristics of the short **(1–3** ms) segments of speech, such as the presence of high Signal–to–Noise Ratio (SNR) regions. The present work focusses on analysing short segments of speech, which we refer to as *subsegmental analysis*. Methods based on Fourier transformation cannot be used for the subsegmental analysis due to problems of frequency resolution. The low correlation between the samples of the linear prediction residual signal is exploited for the analysis of short **(1–3** ms) segments of speech. Methods based on subsegmental analysis are developed for enhancement of degraded speech. The methods proposed for speech enhancement are based on the fact that the correlation between the samples of the LP residual signal is low and can therefore be manipulated to sorne extent without introducing distortion into the processed speech signal.

The major contributions of the thesis are:

- A method for enhancement of noisy speech is proposed which performs emphasis

of high SNR segments of the speech signal relative to the low SNR segments.

- A method for enhancement of reverberant speech is proposed for emphasis of high Signal–to–Reverberant component Ratio (SRR) regions relative to the low SRR regions.
- For speech data collected in practical environmental conditions, a method based on the averaged normalized prediction error is proposed to identify the type and level of degradation.
- Robustness of the group–delay–based method for extraction of the instants of significant excitation is illustrated for degradations in practical conditions such as ambient noise, telephone channel distortion and reverberation in small–rooms.
- A method for enhancement of speech degraded by speech of a competing speaker is proposed. The method uses the instants of significant excitation and subsegmental analysis for enhancement.
- The concept of source–system windowing for speech signals is proposed for analysis of short (1–3 ms) segments of the signal.

TABLE OF CONTENTS

List of Tables

List of Figures

ABBREVIATIONS

CP	–	closed phase of a glottal cycle in voiced speech
FIR	–	Finite Impulse Response
LP	–	Linear Prediction
LPCs	–	Linear Prediction Coefficients
OP	–	open phase of a glottal cycle
SDR	–	Signal–to–Degradation component Ratio
SNR	–	Signal–to–Noise Ratio
SRR	–	Signal–to–Reverberant coniponent Ratio
SVD	–	Singular Value Decomposition

NOTATION

Lower case boldface letters are used to denote vectors and upper case boldface letters to denote matrices. In addition, the following convention is used throughout the thesis:

English Symbols

a_k – kth linear prediction coefficient

$\mathbf{a}$ – vector of linear prediction coefficients: $[a_1\ a_2 \cdots a_p]^T$

$\mathbf{a}_a$ – augmented vector of linear prediction coefficients: $[1\ a_1 \ldots a_p]^T$

f – frequency variable in Hertz for a discrete time signal

F_0 – fundamental frequency

F_1, F_2, F_3 – first, second and third formants, respectively

H_k – Entropy for the kth frame

H_0 – threshold of entropy for mapping function

$\mathbf{I}_p$ – Identity matrix of order p x p

j – square root of -1

n – discrete time index

p – linear prediction order

$\mathbf{R}_{ss}$ – Symmetric Toeplitz autocorrelation matrix for signal s_n

$s(n)$, s_n – clean speech signal as a function of time index n

$\mathbf{S}$ – Toeplitz signal prediction matrix

T_{60} – reverberation time

$v(n)$, $w(n)$ – random noise variables as functions of n

$x(n)$ – signal variable

$y(n)$, y_n – noisy speech signal

z – coniplex argument of a z–transform

Greek Symbols

α_f – slope parameter for fine weight mapping function

α_g – slope parameter for gross weight mapping function

β – attenuation factor due to echo

γ – weight value

Γ – diagonal matrix of weight values

$\delta(n)$ – unit pulse at n = 0

ζ – reciprocal flatness measure

ζ_0 – threshold of reciprocal flatness measure for mapping function

η – normalized prediction error

$\overline{\eta}$ – average normalized prediction error

λ_k – kth eigenvalue

$\mathbf{\Lambda}$ – diagonal matrix of eigenvalues

ξ_n – damping factor of LP all–pole filter at instant n

π – the angle in radians corresponding to 180^o

$\rho(\mathbf{a})$ – Rayleigh quotient of a matrix for the vector a

σ_k – kth singular value

σ_v^2, σ_w^2 – noise variances

$\mathbf{\Sigma}$ – diagonal matrix of singular values

$\tau(\omega)$ – group–delay as a function of frequency w

$\overline{\tau}$ – average group–delay

ϕ, θ – angle measured in radians

ω – frequency variable in radians for a discrete time signal

Ω – frequency variable in radians/second for a continuous time signal

Miscellaneous Symbols

$(.)^*$ – complex conjugate

$(.)^T$ – transpose

$(.)^H$ – conjugate/Hermitian transpose

$tr(.)$ – trace of a matrix

$\mathbf{diag(x)}$ – diagonal matrix whose diagonal elements are the elements of vector x

$\|.\|_2$ – Euclidean norm of a vector

$\|.\|_F$ – Frobenius norm of a matrix

$|.|$ – absolute value

$Re(.)$ – real part of a complex number

$Im(.)$ – imaginary part of a complex number

$\mathcal{E}(.)$ – statistical expectation operator

$\mathcal{L}_2$ – Euclidean norm

$*$ – convolution operator

$\log(.)$ – natural logarithm

$\log_{10}(.)$ – logarithm to base 10

"Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals, the power of speech is intended to set forth the expedient and the inexpedient, and therefore likewise the just and the unjust. And it is a characteristic of man that he alone has any sense of good and evil, of just and unjust, and the like, and the association of living beings who have this sense makes a family and a state." - Aristotle, ***Politics***

"Human subtlety ... will never devise an invention more beautiful, more simple, or more direct than does nature, because in her inventions nothing is lacking and nothing is superfluous" - Leonardo da Vinci

Chapter 1

ISSUES IN SPEECH PROCESSING

1.1 OBJECTIVES OF THE THESIS

Speech is the most natural means of communication among human beings. Speech signal is produced as a result of excitation of the time–varying vocal tract system. Speech signal is processed for estimation of the time–varying characteristics of the speech production mechanism for several applications like automatic speech recognition, enhancement of degraded speech and speaker recognition/verification. In this thesis we propose methods to process short (1–3 ms) segments of the speech signal, which we refer to *as subsegmental analysis.* We show that the subsegmental analysis helps to enhance degraded speech. In this research subsegrnental analysis is used for the following studies:

1. To study the changes in the characteristics of the vocal tract within a glottal cycle in voiced speech
2. For enhancement of speech degraded by additive random noise
3. For enhancement of speech degraded by room reverberation
4. For enhancement of speech degraded by speech of a competing speaker
5. To study practical issues in the implementation of the methods based on subsegmental analysis for speech enhancement.

Direct processing of short (1–3 ms) segments of the speech signal results in severe windowing effects. Therefore, in this work we propose methods based on the characteristics of the linear prediction residual signal to perform subsegmental analysis.

1.2 SPEECH PRODUCTION

Speech can be considered as a sequence of sound units like phonemes, syllables, etc., both at the production and perceptual levels. The speech signal is produced as a result of time–varying excitation of the time–varying vocal tract system. The different blocks in a model of the speech production mechanism are shown in Fig. 1.1. The speech

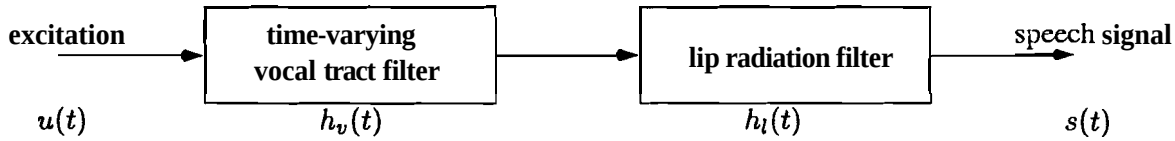


Fig. 1.1: Different blocks in a model of the human speech production mechanism.

signal $s(t)$ in the continuous time domain can be expressed as

$$s(t) = u(t) * h_v(t) * h_l(t) \tag{1.1}$$

where $u(t)$ is the excitation signal, $h_v(t)$ is the time–varying impulse response of the vocal tract and $h_l(t)$ is the response due to lip radiation. One of the objectives in speech analysis is to derive the time–varying characteristics of the speech production mechanism from the speech signal, in order to identify the sound units in speech. In this section we briefly discuss the speech production mechanism.

Different segments of a speech signal can be broadly classified into voiced speech, unvoiced speech and silence. Voiced speech is produced by the periodic excitation of the vocal tract due to vibrations of the vocal cords at the glottis (see Figs. 1.1 and 1.2). Typical glottal volume velocity pulses, which constitute the excitation for the production of voiced speech, are shown in Fig. 1.3(a) (also see Fig. A.1 in Appendix–A). The duration of a glottal cycle (T_0) is indicated in Fig. 1.3(a) which is 10 ms in

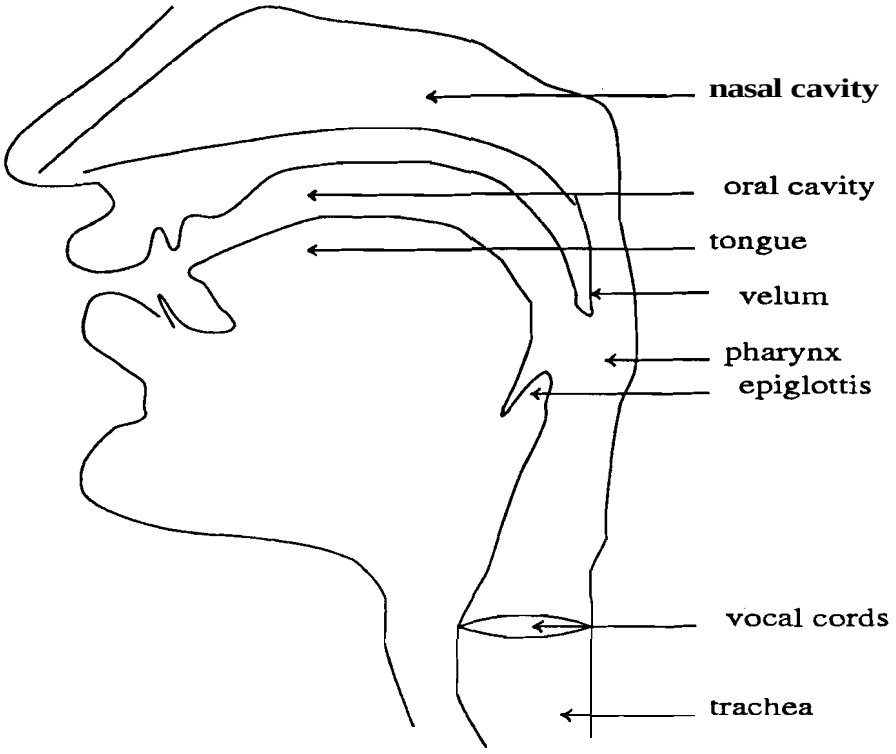


Fig. 1.2: The human speech production mechanism.

this case. The reciprocal of T_0 is called the *fundamental frequency* or *pitch* (F_0) which is 100 Hz. The spectrum of the periodic pulse sequence is shown in Fig. 1.4(a) (see also Fig. A.1(b) in Appendix–A). The spectrum exhibits line–spectral characteristics with the power residing only at the harmonics of the pitch frequency (F_0). We also observe that the spectrum has a roll–off of approximately 12 dB/octave. The roll-off is inversely proportional to the slope of the trailing edge of the glottal pulse in the open phase (i.e., the slope of the pulse in the transition region between the open and the closed phase). The rest of the vocal tract (the oral and nasal passages) acts as a filter. Typical frequency response of the vocal tract for vowel $/a/$ is shown in Fig. 1.4(b), where F_1, F_2, F_3, F_4 and F_5 are the first five natural resonances or *formants* of the vocal tract. The modulated air flow so produced radiates from the lips as a pressure

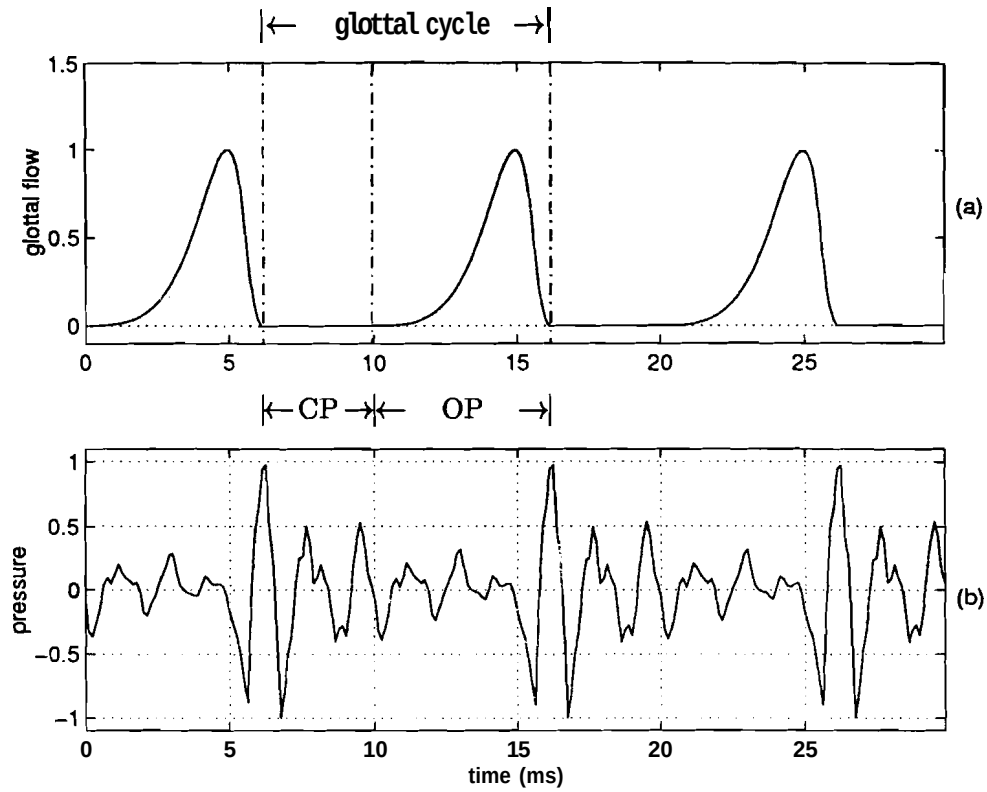


Fig. 1.3: (a) glottal volume velocity, (b) radiated speech pressure wave. CP – closed phase, OP – open phase.

wave. The radiation characteristic of the lips has a differentiating action on the volume velocity at the lips and produces a 6 dB/octave lift in the short–time spectrum of the radiated speech pressure. The lip radiation effect and the voice source are generally considered together as an effective source in modeling the voiced speech. If $u_g(t)$ is the glottal volume velocity then (1.1) can be written as

$$\begin{aligned} s(t) &= [h_l(t) * u_g(t)] * h_v(t) \\ &\approx \frac{du_g(t)}{dt} * h_v(t) \\ &= u'_g(t) * h_v(\mathrm{t}) \end{aligned} \tag{1.2}$$

where $u'_g(t)$ is the first derivative of the glottal volume velocity (see Fig. A.1(c) in Appendix–A). A typical voiced speech waveform (pressure) is shown in Fig. 1.3(b),

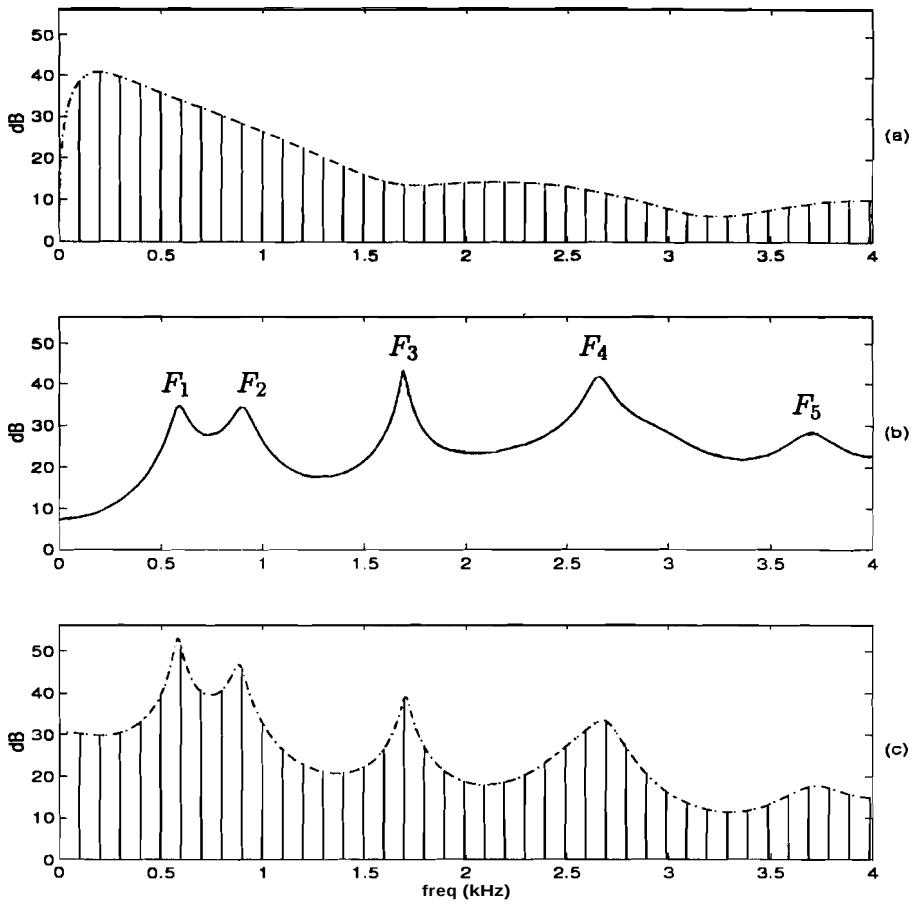


Fig. 1.4: Typical spectra of (a) quasiperiodic glottal excitation, (b) vocal tract and (c) voiced speech waveform for vowel /a/.

synchronized with the glottal flow. We observe from the figure that significant excitation of the vocal tract system takes place at the instant of glottal closure [1], since the maximum change in the glottal volume velocity occurs at this instant. We refer to the instant of glottal closure as the *instant of significant excitation.* We also observe that the speech signal has a large amplitude in the closed glottis interval (indicated as CP in Fig. 1.3), immediately after the instant of significant excitation, and a relatively low amplitude in the open glottis interval (indicated as OP in Fig. 1.3). The low amplitude in the region OP is both due to the natural decay of the waveform as well as due to increased damping in the open glottis interval. In the open glottis interval the trachea

is coupled to the vocal tract leading to increase in the formant frequencies and bandwidths relative to their values in the closed glottis interval [2]. Thus there are changes in both formant frequencies and bandwidths from closed to open glottis intervals in voiced speech. The spectrum of the signal in Fig. 1.3(b) is shown in Fig. 1.4(c), which is the product of the spectrum of the periodic glottal source in Fig. 1.4(a), the spectrum of the vocal tract in Fig. 1.4(b) and a 6 dB/octave highpass filter response of the lip radiation effect (recall Fig. 1.1). The spectrum in Fig. 1.4(c) exhibits line-spectral nature due to the periodicity of the signal. The energy of the signal lies mainly at the harmonics of the pitch frequency (F_0). In the case of ***unvoiced speech*** the excitation to the vocal tract is due to turbulent air flow at a constriction created somewhere along the vocal tract. Unvoiced speech has a low amplitude relative to the amplitude of voiced speech, and is a noise-like signal.

The time-varying nature of the speech production mechanism necessitates windowing the speech signal, so that the relatively steady segments of the signal can be analysed. For analysis purposes, a linear source-system model is assumed for speech production. The source and system characteristics are assumed quasistationary in the analysis interval [3]. This simple model does not give an accurate representation of the speech signal in each frame. For example, in consonant to vowel (CV) transitions (e.g., /*dz a*/ as in jar) and vowel to consonant (VC) transitions (e.g., /*ak*/ as in *tack*), rapid changes occur in the vocal tract shape and excitation characteristics, within a duration of about 50-100 ms [4-8]. Hence, an analysis window of duration 10-30 ms does not provide adequate temporal resolution. The changes in the vocal tract characteristics from one glottal cycle to another in the voiced regions of such CV or VC transitions are smeared. Fig. 1.5(a) shows the time domain waveform of an utterance of the CV /*dz a*/ (as in jar) sampled at 11.025 kHz. The point of consonant release is indicated

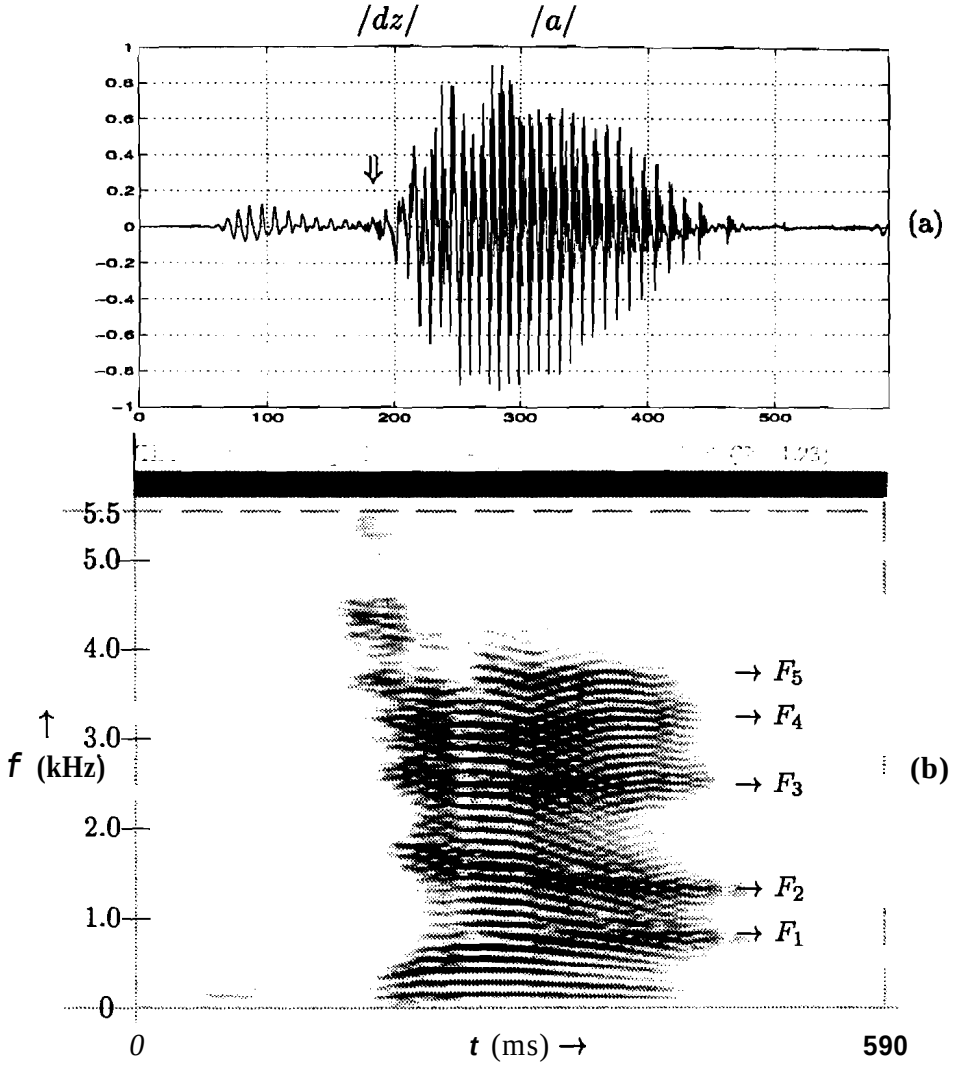


Fig. 1.5: Spectral transition for CV */dz a/* as in *jar*. (a) waveform in the time domain, (b) spectrogram. F_1 – F_5 are the formants for vowel */a/* in the steady region.

by an arrow pointing downward in Fig. 1.5(a). In Fig. 1.5(b) the spectral changes that occur as a function of time in the CV are shown as a *spectrogram*. The spectrogram is a plot of a sequence of short–time Fourier transform magnitude spectra obtained using overlapping windows of duration 20 ms. The spectral amplitude at any instant of time t and frequency f is represented on a gray scale, with the dark bands representing high energy regions. The dark horizontal striations in Fig. 1.5(b) are due to the concentration of energy at the pitch harmonics. The changes in the formant frequencies (F_1–F_5)

in the CV transition region can also be seen in the figure. The formant changes in the region of consonant release are smeared due to the large (20 ms) duration of the analysis window. The quasistationary assumption over a 10–30 ms duration is not valid even for steady voiced regions in the speech signal. This is because the excitation characteristics change within each glottal cycle due to glottal vibrations, and the vocal tract system changes due to coupling and decoupling of the trachea during the open and closed phases of the glottal excitation, respectively [2, 9–12].

The positioning of the analysis window relative to the signal has significant influence on the results obtained [13–19]. Pitch–synchronous analysis is reported to give better results than pitch–asynchronous analysis [20–22]. Due to changes in the vocal tract system within a glottal cycle, even pitch synchronous placement of a 10–30 ms analysis window does not guarantee that the signal within the window corresponds to a steady system [23]. For example, in a CV or a VC situation the variations in the characteristics of the vocal tract from one glottal cycle to another are smeared. Pitch synchronous analysis of the steady voiced regions also smears the variations in the characteristics of the vocal tract within a glottal cycle. The derived characteristics will be the same for each glottal cycle.

1.3 SPEECH PROCESSING

Methods proposed in the literature for processing speech can be broadly classified into three categories:

1. *Suprasegmental* level (100–300 ms of signal for analysis),
2. *Segmental* level (10–30 ms of signal for analysis) and
3. *Subsegmental* level (1–3 ms of signal for analysis).

Methods for processing speech at the suprasegmental level are guided mainly by perception [24–26]. Features such as intonation [27–32], stress [33], duration [33, 34] and speaking rate [35] are studied at the suprasegmental level. Methods for processing speech at the segmental level (10–30 ms) such as the short–time spectrum and Linear Prediction (LP) analysis [36] are dictated more by signal processing considerations such as window effects [37–39] and time–frequency resolution rather than by the characteristics of the signal. Methods for processing speech at the subsegmental level are primarily guided by the characteristics of the speech signal and the speech production mechanism [1, 23].

1.4 SUBSEGMENTAL ANALYSIS

If short (1–3 ms) segments of the speech signal are analysed pitch synchronously, then one may capture the consistent variations in similar segments in successive glottal cycles. However, both the size and location of the short analysis window are crucial for accurate analysis of changes of the vocal tract system. The advantages of analysing short (1–3 ms) segments of the speech signal instead of the usual 10–30 ms segments are manifold.

- Firstly, the changes within a glottal cycle and from one glottal cycle to another can be tracked by analysing short (1–3 ms) segments of the data.
- Secondly, the influence of the fundamental frequency on the LP spectrum derived from the speech signal is avoided [40, 41].
- Signal samples due to heavily damped formants exist only for a few (1–3) milliseconds immediately after the instant of significant excitation of the vocal tract. They may not be captured well when larger analysis frames are used [19].

- In practical environmental conditions, where the speech signal is corrupted by noise and/or reverberation, the high signal–to–noise ratio (SNR)/signal–to–reverberant component ratio (SRR) segments within a glottal cycle of the voiced speech can be exploited for analysis when short (1–3 ms) segments of the speech signal are used.

For short (1–3 ms) analysis windows, the short–time Fourier transform gives poor spectral resolution [42]. Linear Prediction (LP) analysis by the autocorrelation method [40] also performs poorly due to severely biased estimates of the autocorrelation coefficients as a result of the short duration of the window. The effects of the short window are due to the high correlation between samples in the speech signal [43–45]

The issues that arise in the subsegmental analysis of speech are:

(a) The duration and positioning of the short analysis frame should be guided by the characteristics of the speech production mechanism. For example, analysis of the speech signal in the closed glottis region in voiced speech provides an accurate estimate of the frequency response of the vocal tract system [1, 46,471. This is because the speech signal in the closed glottis interval represents the force–free response of an all–pole system. However, for such an analysis to be possible, knowledge of the closed glottis region as well as its duration is necessary.

(b) The instants of significant excitation of the vocal tract, which correspond to the closing instants of the glottal cycles [1] in the case of voiced speech, need to be identified. These instants enable us to identify similar regions in the voiced parts of the speech signal (e.g., the closed glottis region or the open glottis region) for analysis.

(c) The method for identification of the instants of significant excitation should be robust to degradations.

(d) Short window effects severely bias the results of analysis. These effects can be reduced to some extent by using a model based analysis (e.g., LP analysis using the covariance method).

In practical conditions, speech signals are often degraded by noise and characteristics of the transmission medium (e.g., reverberation in rooms, telephone channel). These degradations compound the problem of speech analysis. When speech is corrupted by additive random noise, the signal–to–noise ratio (SNR) varies with time and also varies as a function of frequency in the spectral domain. This is because the speech signal has a large (30–60 dB) dynamic range in the temporal and spectral domains. In the spectral domain the SNR is high in the formant regions, and it is low in the valley regions (see Fig. 1.4(c)). In the time domain, due to the damped sinusoidal nature of the speech signal within a glottal cycle of voiced sounds (see Fig. 1.3(b)), the signal energy is usually higher in the vicinity of the instant of significant excitation of the vocal tract system. Hence, the SNR varies within a glottal cycle of voiced sounds. Similarly, when speech is corrupted by reverberation in a small room, the signal–to–reverberant component ratio (SRR) varies over short (1–3 ms) segments in the time domain. Fig. 1.6 shows the plots of SRR and SNR as a function of time. Fig. 1.6(a) shows a clean speech signal. The energy of the clean speech and the SRR for the reverberant speech are computed for every 2 ms frame shifted by one sample (8 kHz sampling rate) and the plots are shown in Figs. 1.6(b) and 1.6(c), respectively. The reverberant speech signal was generated by convolving the clean speech signal in Fig. 1.6(a) with the impulse response of a room collected at a distance of 1.5m from the source in a normal officeroom. Likewise, the SNR is computed for the noisy speech obtained by adding white noise to the clean speech signal in Fig. 1.6(a) (overall SNR = 10 dB) and is plotted in Fig. 1.6(d). It is obvious that SRR and SNR vary with time

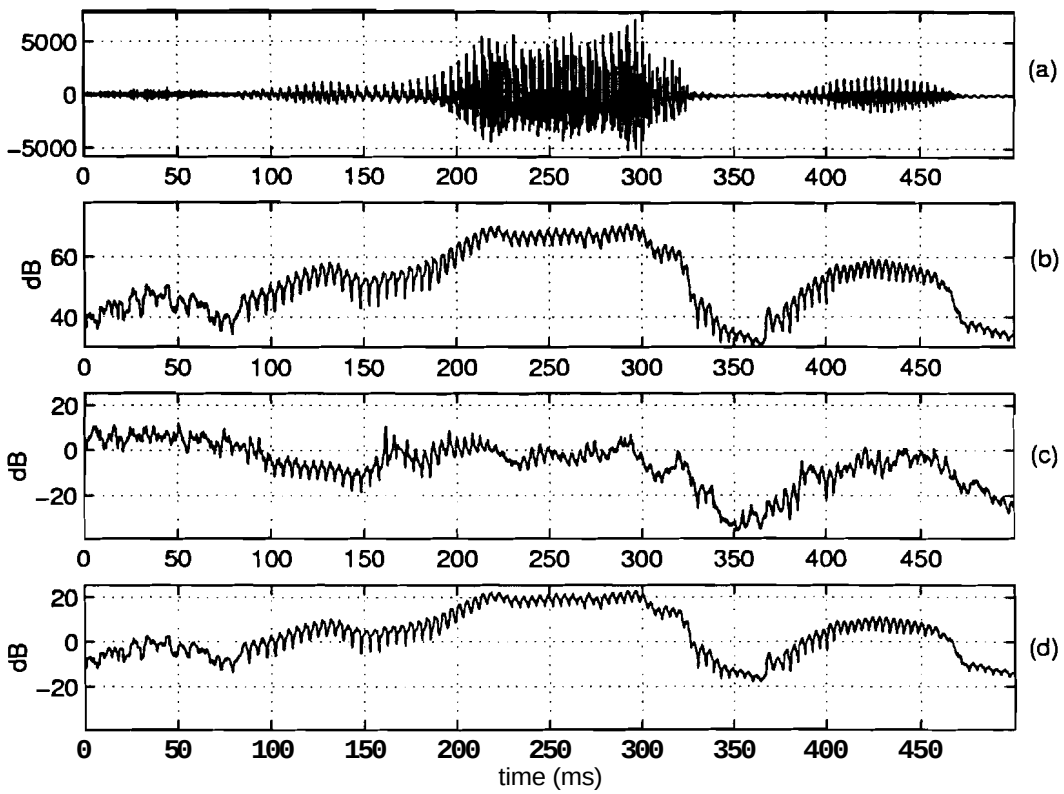


Fig. 1.6: Variation of short–time SRR and SNR with time for degraded speech. (a) Clean speech signal. (b) Short–time energy of clean speech computed using 2 ms frames. (c) Short–time signal–to–reverberant component ratio (SRR). (d) Short–time signal–to–noise ratio for an average SNR of 10 dB.

since the signal energy is also a function of time. In the case of reverberant speech both the signal energy and the energy of the degrading component are time–varying. The finer variations (ripple) in the SRR and SNR plots in Figs. 1.6(c) and 1.6(d) are due to the variation of the signal energy within a glottal cycle (see Fig. 1.3(b)). Analysis of the signal for short (1–3 ms) segments corresponding to high SNR/SRR regions may be useful to process degraded speech for enhancement.

Issues for Enhancement

When speech is corrupted by additive noise and/or room reverberation, larger (10–

30 ms) analysis frames do not exploit the high SNR/SRR segments within the frame for enhancement. Instead, most of the methods for speech enhancement estimate the spectral characteristics of noise (either in the silence regions of the same channel or on a separate channel) and subtract it from the spectrum of speech estimated using a 20–30 ms analysis frame [48–50]. Moreover, the spectral characteristics of the noise cannot be assumed to be stationary in many practical situations. In the case of speech degraded by room reverberation, the impulse response of the transmission medium between the source and the receiver is estimated, and inverse filtering of the degraded speech is performed to achieve dereverberation [51, 52]. Thus the focus of most of the methods proposed for speech enhancement is on suppression/cancellation of the degradation, rather than on enhancement of the characteristics of the speech signal.

The methods proposed for speech enhancement address the following issues:

(a) Definition of SNR

(b) Identification of high SNR regions

(c) Processing of the segments

(d) Objective and subjective criteria for enhancement

1.5 SCOPE OF THE PRESENT WORK

One of the problems addressed in this thesis is to study the changes in the vocal tract characteristics within a glottal cycle of voiced speech. During the closed phase of the glottal cycle the vocal tract is closed at one end. In the open phase, the relation between the acoustic pressure over the vocal folds and the airflow is in general nonlinear [53–55]. The characteristics of the system during the open phase are not constant, but signal dependent. In our research the nonlinear effects are assumed to

be insignificant. Secondly, the characteristics of the vocal tract system in the open phase are represented by an all–pole model, which will always be some kind of time average of the varying vocal tract system. Thirdly, the closed phase of a glottal cycle can be very short and sometimes may not even exist (e.g., high pitched female voices). However, we assume that a short (1–3 ms) segment of the signal immediately after the glottal closure in the voiced speech corresponds to the closed phase.

In the case of speech degraded by reverberation, the dimensions of the room significantly influence the perceived degradation. For example, in a large concert hall–like room, the reverberation time (T_{60}) [56] may be of the order of several seconds. In small rooms the reverberation time may be of the order of a few hundred milliseconds. In this thesis we address the problem of enhancement of speech corrupted by additive random noise and speech corrupted by reverberation in small rooms. These degradations are typical in a speakerphone situation [57].

1.6 ORGANIZATION OF THE THESIS

The focus of the work presented in this thesis is on subsegmental analysis of speech and its application to enhancement of degraded speech. The evolution of ideas presented in this thesis is given briefly in Table–1.1. The contents of the thesis are organized as follows:

A review of several methods for processing speech is presented in Chapter 2. The chapter is primarily divided into three sections. Sections 2.2, 2.3 and 2.4 review methods for processing speech at the suprasegmental, segmental and subsegmental levels, respectively.

In Chapter 3, we present a new windowing method for subsegmental analysis of speech. The method overcomes, to some extent, the limitations of short (10–30 ms)

Table 1.1: Evolution of ideas presented in the thesis.

Short Segment Analysis of Speech for Enhancement

- Nonstationarity of speech signals
- Degradation of speech signals in practical conditions due to ambient noise, telephone channel distortion, reverberation in rooms and speech of competing speakers
- Methods for processing speech
  - ***Suprasegmental*** level: using 100–300 ms of signal for analysis; motivated mainly by human aural perception
  - ***Segmental*** level: using 10–30 ms of signal for analysis; guided mainly by signal processing considerations
  - ***Subsegmental*** level: using 1–3 ms of signal for analysis; motivated primarily by the characteristics of the speech signal and the speech production mechanism
- Traditional methods for processing speech at the suprasegmental and segmental levels
  - may not provide adequate temporal resolution to capture the rapid changes in the speech signal such as in a CV or VC situation
  - do not exploit the short (1–3 ms) high SNR/SRR segments in the signal

segine analysis of speech

- Using pitch synchronous analysis of short (1–3 ms) segments of the speech signal, variations in these segments in successive glottal cycles can be captured
- Effects of truncation are severe when processing short (1–3 ms) segments using the traditional waveform windowing, hence the need for a new windowing procedure for processing short segments
- LP residual signal is used for subsegmental analysis due to the low correlation between samples of the residual signal

Subsegmental analysis for enhancement of degraded speech

- Time–varying SNR/SRR of degraded speech over short (1–3 ms) durations due to the time–varying amplitude of speech in general, and the damped sinusoidal nature of voiced speech in particular
- Human beings perceive speech by capturing some features from the high SNR regions in the spectral and temporal domains, and then extrapolating tlie features in the low SNR regions
- Therefore, emphasis of high SNR/SRR regions of the speech signal relative to the other regions is achieved by modifying the LP residual signal.
- Practical issues in the implementation of the proposed methods for speech enhancement:
  - Identification of the type of degradation (noise/reverberation) so that suitable enhancement method can be employed for processing the speech signal
  - Identification of the level of degradation so that proper settings of the parameters of tlie enhancement method can be chosen automatically
- To identify the high SNR segments in degraded speech, which usually correspond to the 1–3 ms segments immediately after the instants of significant excitation of the vocal tract, a robust method for determining these instants is required.

window.analysis.

As applications of the subsegmental analysis, we present methods for enhancement of speech degraded by additive noise and reverberation in Chapters 4 and 5, respectively. The key idea is to modify the LP residual signal for enhancement.

In Chapter 6, we present a method for determining whether a given speech signal is degraded or not; if degraded, whether the source of degradation is additive noise or reverberation, and the level of degradation. This aspect is important in practical conditions. Depending upon the type of degradation, suitable method can be employed for enhancement.

In Chapter 7, we investigate the robustness of a group-delay-based method for determining the instants of significant excitation in speech signals. Knowledge of the instants of significant excitation is useful in identifying accurately the short (1–3 ms) high SNR/SRR segments in degraded speech.

In Chapter 8, we propose methods which use the knowledge of instants of significant excitation and the subsegrnental analysis for speech enhancement. In particular, we address the problem of enhancement of speech degraded by speech of a competing speaker.

Chapter 9 summarizes the work presented in this thesis. We highlight the contributions of this research and give some directions for future work.

Chapter 2

METHODS FOR PROCESSING SPEECH – A REVIEW

In this chapter a review of the methods for processing speech is presented. The chapter is organized as follows. Section **2.1** gives an overview of the information present in the speech signal at the three different levels, namely suprasegmental, segmental, and subsegmental levels. Section **2.2** presents a review of the methods for processing speech at the suprasegmental level. This includes methods for' processing noisy speech and reverberant speech for enhancement. Section **2.3** presents a review of the methods for processing speech at the segmental level, their advantages and limitations. Section **2.4** presents a review of the methods for processing speech at the subsegmental level and the motivation for the work presented in this thesis. Section **2.5** presents an overview of the thesis.

2.1 INTRODUCTION TO SPEECH PROCESSING

The speech signal carries speaker-specific information besides the linguistic message [58]. The information about the speaker and speech resides at the suprasegmental level, segmental level and at the subsegmental level. At the suprasegmental level the information bearing features are the variation of pitch with time (intonation), the duration of syllables and words, the stress on certain syllables, the number of syllables spoken

per second (syllable rate) and the influence of one sound unit on another in continuous speech (warticulation). At the segmental and subsegmental levels the information is in the features such as the characteristics of the vocal tract, the voice source, the pitch etc. It is also well known that the glottal wave is highly characteristic of the speaker whereas the vocal tract parameters are mainly characteristic of the speech [59]. The features at the suprasegmental level significantly influence the features at the segmental and subsegmental levels. For example, intonation and stress are known to influence the characteristics of the voice source in continuous speech [60–62]. Thus the features at different levels are not completely independent of each other. In the sections to follow, speech processing methods at the three different levels are discussed.

2.2 PROCESSING OF SPEECH AT THE SUPRASEGMENTAL LEVEL

Processing speech at the suprasegmental level using more than 200 ms of the signal for analysis is motivated by the human auditory perception. Some of the important reasons cited for emulating the humari auditory perception are that irrelevant acoustic variabilities will be reduced and that the phonetic contrasts will be enhanced [63]. Another important reason is the inherent robustness of the human auditory perception to environmental degradations [64]. Secondly, "it appears that the short–term memory of the auditory periphery in mammals (exhibited, e.g., by forward masking [65]), the firing rate adaptation constant and the buildup of loudness is of the order of 200 ms. This means that the peripheral human auditory system can effectively utilize rather large (about syllable sized) time–spans of the audio signal". Based on previous experimental studies, it is learnt that sensitivity of humari hearing to both the amplitude and frequency modulation is highest for frequency of modulation at about 4–6 Hz. As a likely consequence of this sensitivity, the modulation spectrum of speech is character-

ized by a dominance of components around 4 Hz [66]. This approximately corresponds to the syllable rate in continuous speech.

Based on the above results there is extensive literature on methods to process speech at the suprasegmental level. In [67] a computational auditory model based on the temporal characteristics of the information in the auditory nerve fiber firing patterns was presented. The model, called *Ensemble Interval Histogram* **(EIH)** model, has essentially two stages: the cochlear filtering stage and the multi–level crossing detector stage. **Each** cochlear filter is followed by a multi–level crossing detector. A histogram of intervals between successive level crossings is computed for each level. The cochlear filters are equally spaced on a log–frequency scale. The output of the model is a frequency domain representation of the input signal in terms of the ensemble histogram of firing patterns. The output of the auditory model was converted to linear prediction coefficients which were input to a speech recognizer. Recognition experiments [24, 68] showed that this representation was robust with respect to noise contamination. In order to investigate the role of cochlear filters, these filters were replaced by bandpass filters whose frequency responses are the Fourier transform of Hamming windows with suitable duration. It was pointed out that the robustness was mainly due to timing–synchrony analysis and not due to the shape of the cochlear filters [24, 69]. The closed loop EIH model was proposed in [70]. It is constructed by adding a feedback system to the former open loop EIH system. While the open loop EIH system is a computational model based on the ascending path of the auditory periphery, the feedback system is motivated by the descending path. A comparison of the performance of Mel cepstra and EIH as input features to a phone classification task was presented in [71]. The EIH demonstrated improved performance under adverse conditions (telephone channel degradation).

The *RelAtive SpecTrA*, abbreviated RASTA [72–74], is based on processing the temporal contours of each frequency component in the short–time power spectrum/critical band power spectrum [75,76]. The power spectrum is computed using a 25 ms analysis window overlapped by 12.5 ms. The different steps in the RASTA technique are as follows: the short–time power spectrum/critical band power spectrum is first computed. The spectral amplitude is compressed using a static nonlinear transformation. The time trajectory of each transformed spectral component is filtered. The filtered spectral components are expanded using a static nonlinear transformation which is usually, but not necessarily, the inverse of the compressive transformation function. In the conventional RASTA technique, a fixed IIR bandpass filter is used for all the frequency components. The filter has a lower cut off frequency of 0.26 Hz and an upper cut off frequency of about 13 Hz with sharp zeros at 28.9 Hz and 50 Hz. The RASTA filter mentioned above has a long (about 500 ms) time constant. The filter eliminates frequency components which do not vary with time. This is typically the case with the channel distortions, for example the telephone channel characteristics. The bandpass filter emphasises the frequency components which are due to the spectral transitions in speech, for example due to the frequency components corresponding to the syllable rate (4–6 Hz, depending upon the speaker). The RASTA technique has been extensively used for automatic speech recognition. Perceptual Linear Prediction (PLP) analysis [77] is performed on the filtered and uncompressed spectral components. The parameters obtained from PLP have been used for matching, and as a result reduced recognition error rates were reported [26]. Several variations of the RASTA approach have also been proposed. The J–RASTA technique introduces a parameter J for improved robustness against channel distortions and additive noise [78].

In the J–RASTA technique the nonlinear function used for compression is

$$St(w) = \log[1 + JS(\omega)] \quad (2.1)$$

where $S(\omega)$ is the input spectral value, $S_t(\omega)$ is the transformed spectral value and J is a signal dependent positive constant. This transformation is approximately linear for small spectral values and nearly logarithmic for large spectral values. It is reported that J–RASTA performs better than both RASTA and PLP techniques, in terms of recognition error rate [26]. The RASTA technique has been used for channel normalization in automatic speech recognition. In [79], the filters for trajectories of critical band energy for channel normalization are derived from the data using a constrained optimization procedure. In [80], an extensive discussion on the conditions under which the RASTA filtering is useful as a channel normalization technique is given.

2.2.1 **Enhancement** of Noisy **Speech**

There exist several practical situations in which enhancement of speech degraded by additive interfering signals is required or is desirable. For example, in telephony it may be necessary to bring down the level of the additive noise in the channel to improve the comfort level for the listener. Listening to noisy speech for extended periods of time can cause listener fatigue. It is also well known that under noisy environments, the speaker's vocal apparatus is strained resulting in an effect called the Lombard *effect* named after Etienne Lombard who first investigated this phenomenon in 1911. Both the voice level and the pitch increase when noise is delivered to a speaker's ear [81]. However, this aspect is not discussed further as this is beyond the scope of the work presented here. In this section a review of the methods at the suprasegmental level to process noisy speech for enhancement is presented.

When a speech signal is degraded (by room reverberation or additive noise) its

modulation depth [82], i.e., the relative variation of the envelope, decreases. There is a relation between the Modulation Transfer Function (MTF) for certain modulation frequencies (e.g., in the range 0.4 to 20 Hz) and speech intelligibility [83, 84]. Thus, according to an idea suggested by M. R. Schroeder, an artificial increase of the modulation depth of the degraded speech in a certain range of modulation frequencies may perhaps achieve a better preservation of intelligibility. The MTF of the envelope filter required for this purpose should, however, not be defined for the total signal, but for critical bands corresponding to the analysis in the human ear. Based on this, Langhans and Strube [85] have proposed the decomposition of the speech signal into many frequency bands and an envelope filtering performed on each one of them. The results are summed up. However, for more flexibility for experimentation, the *overlap add* (OLA) method [86] was used in [85]. Overlapping segments of the signal are Hamming windowed, and after appending zeros symmetrically, are transformed by the FFT. The squared envelope of the short–time spectrum is obtained. Power summation over critical bands is performed. After some multiplicative modification of the component signals, an inverse FFT is performed for each frame and the resulting overlapping segments are added up to an output signal. The filtering of the tracks was performed by the inverse of the MTF of noise–corrupted speech. It was implemented by a 63rd order FIR filter. Processing was performed both prior to degradation and after degradation to test for improvement in intelligibility. It is reported that although the objective SNR was increased by 3 dB, there was no noticeable improvement perceptually. The filtering was then performed on the logarithm of the power spectrum. Irnproved intelligibility was reported, if the processing was done prior to degradation by noise.

A similar approach was suggested by Hermansky et al. [87] based on the RASTA

technique. The cubic–root compressed short–term power spectrum is processed using the RASTA technique. The processed signal is obtained via OLA using the phase [1] of the original noisy speech. However, the enhanced speech was found to contain musical noise. Note that the traditional RASTA method uses **fixed,** data–independent filters for processing the temporal tracks. Later, a modification of the above method was proposed in [89] which uses noncausal FIR Wiener–like filters to process the cubic root of the estimated power spectrum. Each filter is designed to optimally map a time window of the noisy speech spectrum of a specific frequency to a single estimate of the short–term magnitude spectrum of clean speech. The design is carried out on parallel recordings of clean and noisy data. Thus, the designed filter bank is noise–specific and the algorithm is most efficient on disturbances similar to those present in the training data. To circumvent the noise–dependency of the method, in [90] an adaptive speech enhancement technique based on selecting a set of pre-computed FIR filters to process the compressed short–time power spectral trajectories of noisy speech was proposed. The shape of the frequency response of the pre–computed filters depends only on the signal–to–noise ratio and does not depend on the center frequency of the channel. This allows for a compact design in which the filter selection criterion is the estimate of the signal–to–noise ratio at the particular frequency channel. The heart of the system is the filter table, which has the filter coefficients along with their corresponding frequency–specific signal–to–noise ratios. To derive the set of signal–to–noise ratio specific filters the magnitude frequency responses of filters derived at a given signal–to–noise ratio are averaged. A non–causal linear phase FIR filter is designed to match the averaged

[1] Wang and Lim [88] concluded that an effort to accurately estimate the phase from the noisy speech is unwarranted in the context of speech enhancement if the estimate is used to reconstruct a signal by combining it with an independently estimated magnitude or to reconstruct the signal using the phase–only signal reconstruction algorithm.

response. During the operation of the speech enhancement system on data corrupted by unknown noise, the signal–to–noise ratio is estimated for each frequency band using the method in [91]. An appropriate filter bank is then constructed by selecting those filters from the table whose frequency specific signal–to–noise ratio labels are closest to the estimated values. It is reported that a noticeable suppression of the perceived noise is achieved by this method, although there is a residual noise which has a different character than the input noise. The residual noise exhibits a periodic fluctuation due to the emphasis of certain modulation frequencies.

2.2.2 Enhancement of Reverberant Speech

When the pressure fluctuations due to speech travel in a closed space, the finer details of its time–intensity distribution are blurred before reaching the listener. This blurring results from the superposition of the reflected sound waves with different delays and intensities to the original (direct path) waveform [66]. The early echoes introduce zeros in the speech spectrum and the speech is perceived as *hollow.* Late echoes are perceived as distinct repetitions of the previous sounds [92]. The problem of enhancement of speech degraded by reverberation appears in applications such as hands–free telephony when the microphone is placed away from the speaker (especially in phones in automobiles) and audio-conferencing in small rooms. It is also well known that noise and reverberation degrade the performance of automatic speech recognition and speaker verification systems.

Mitchell and Berkley [93, 94] were among the first to address the problem of reduction of the effects of room reverberation on speech signals. They suggested a center–clipping process for removing the reverberant tails of speech produced in a room with long reverberation time. The different steps in their method are as follows:

The input speech is divided into several channels by a set of contiguous band filters each less than one octave wide, and the output of each filter is passed through independent center-clippers. The instantaneous output of each center-clipper is made zero unless the absolute value of the input exceeds a threshold value and otherwise varies linearly with the input. Harmonic distortions introduced by the center clippers are then removed by an output filter bank identical to the input set of filters. A six-channel system using two to three octave filters (250–3500 Hz) was used to process input speech recorded in an auditorium. Clipping levels used were such that the output of each center-clipper was zero approximately 50% of the time. A reduction in the perceived reverberation is reported.

As mentioned earlier, both additive noise and reverberation reduce the modulation depth of a speech signal. Assuming a room with an idealized exponential reverberation (without discrete echoes), corresponding to an impulse response

$$h_r(t) = w(t) \exp(\frac{-6.9t}{T_{60}}), \quad t \geq 0 \tag{2.2}$$

where T_{60} is the 60 dB decay time and $w(t)$ is a stationary white noise, then the MTF is given by

$$M(\Omega) = \left[1 + \left(\frac{\Omega\, T_{60}}{13.8}\right)\right]^{-\frac{1}{2}} \tag{2.3}$$

i.e., a first–order low–pass characteristic. This implies that the room reverberation smears the envelope of the speech signal resulting in hollow quality and the effects due to reverberation tails. Therefore, Langhans and Strube [85] tried to use the inverse characteristic for the envelope filter, but limited to 9.5 dB above 10 Hz and decaying to zero above 40 Hz in order to avoid strong enhancement of rapid fluctuations. The different steps in their method are as given in Section 2.2.1 above. The principle behind the method is the recovery of the average envelope modulation spectrum of

the anechoic speech from the reverberated speech. It is reported in [85] that the above mentioned high–pass filtering of the critical band frequency tracks degraded the quality of processed speech. This is because the intrinsic stochastic modulation of the noise $w(t)$ was increased and became audible as an annoying irregular fluctuation.

Previous work using the above principle has also been reported by Hirsch in [95]. Hirsch reported an improvement on the automatic recognition of reverberant speech by high pass filtering the temporal tracks of the short–time Fourier transform power spectrum. Improvement of the quality of the reconstructed speech after filtering was also reported.

Filtering the temporal tracks of the short–time spectrum was also suggested in [96] for enhancement of reverberant speech. The different steps in the method are as explained in Section **2.2.** As in the case of the method proposed for processing noisy speech, the filters for processing the temporal tracks are non–causal FIR filters derived by solving the Wiener–Hopf equations obtained by minimizing the $\mathcal{L}_2$ norm of the difference between the filtered time trajectories of the power spectrum of the corrupted speech and the corresponding desired trajectories of clean speech. A filter is designed for each frequency channel. The lengths of the filters were chosen so that they are greater than the reverberation time T_{60}. It is reported that a reduction of reverberation was audible in the processed speech. The modulation frequencies attenuated by reverberation were also restored, to some extent, after processing. However, it is concluded that the restoration of the modulations alone does not guarantee a good quality speech. This is because the OLA resynthesis procedure, which uses the phase of the corrupted signal, contributes to the perceived artifacts in the processed speech.

Mourjopoulos and Hammond [97] suggested enhancement of reverberant speech by recovering the envelope of the anechoic speech from each sub–band of the reverberated

speech and recombine them to obtain enhanced speech. The different steps in the algorithm are as follows: the speech signal is filtered into a number of contiguous frequency bands (N=5 bands were used in this method). In each band n, the signal is expressed as a product of two terms: the slowly varying, positive, envelope function $A_{sn}(t)$ and the cosine modulated instantaneous phase $p_{sn}(t)$, that describes the fine structure of the signal in that sub–band. Hence the speech signal $s(t)$ is expressed as:

$$s(t) = \sum_{n=1}^{N} A_{sn}(t) \cos[p_{sn}(t)] \tag{2.4}$$

The impulse response of the room $h_r(t)$ is separately measured for the given experimental set up, using the swept sine technique. It has its corresponding envelope and phase components $A_{hn}(t)$ and $p_{hn}(t)$, respectively. If $A_{rn}(t)$ and $p_{rn}(t)$ are the envelope and phase components of the reverberant speech, it is shown in [97] that

$$A_{rn}(t) \approx \frac{1}{2} A_{hn}(t) * A_{sn}(t) \tag{2.5}$$

In each band, an inverse operator $A_{hn}^{-1}(t)$ is designed. It is convolved with the envelope of the reverberant speech in that band to recover the anechoic speech envelope:

$$A_{rn}(t) * A_{hn}^{-1}(t) \approx \frac{1}{2} A_{sn}(t) \tag{2.6}$$

The recovered envelope and the phase extracted from the reverberant speech are combined according to (2.4) to obtain enhanced speech. It is reported that for T_{60} values of up to 5 s, the envelope deconvolution scheme achieves considerable enhancement. For higher values of T_{60}, no significant improvement in intelligibility was found. Note that this method requires the knowledge of the room impulse response for the given relative positions of the source and the receiver.

2.3 PROCESSING OF SPEECH AT THE SEGMENTAL LEVEL

Most of the methods proposed in literature for processing speech at the segmental (10–30 ms) level are driven by signal processing considerations and have their origin in spectrum estimation methods proposed earlier. A review of the developments in the field of spectrum estimation is presented in [98,99]. Some of the popular methods for processing speech are reviewed in this section.

One of the most popular speech processing tools is the sound spectrogram. It is a time–frequency representation of the speech signal [100,101] obtained by computing the discrete Fourier transform (DFT) for overlapping 20–30 ms frames using the FFT algorithm [102]. The Fourier spectrum computed from each such frame is generally referred to as the short–time spectrum. Depending upon the desired time and frequency resolution, the broadband or the narrowband spectrogram could be used.

The short–time Fourier transform (STFT)–based spectrogram has a fixed time and frequency resolution, which is decided a priori. But in practice, the signals like speech are time–varying. Czerwinski and Jones [103] proposed a method of adaptively adjusting the window length used in short–time Fourier analysis. They have chosen a Gaussian curve as the window function, whose variance parameter is varied depending upon the signal characteristics. A short window allows the STFT to show the quickly changing signal structure at the expense of poorer frequency resolution, while a longer window provides better frequency resolution at the expense of blurring out signal transitions. The wavelet transform [104], which facilitates multi–resolution analysis of signals, has also been used for analysis of speech signals.

Another important speech processing tool, the group–delay spectrum, which is based on the STFT but different from the STFT magnitude has been proposed by Yegnanarayana [105]. Although the Fourier transform magnitude and phase spectra

are independent functions of frequency domain features of a signal, most of the techniques for feature extraction from a signal are based upon manipulating the Fourier transform magnitude only. The phase spectrum of the signal corresponds to phase delay corresponding to each of the sinusoidal components of the signal. However it is difficult in practice to process the Fourier transform phase of signals for the extraction of features due to the inevitable wrapping of the phase spectrum. An alternative to processing the phase spectrum is processing the group–delay function. The group–delay function is the negative derivative of the (unwrapped) Fourier transform phase. The group-delay function can be computed directly from the time domain signal. The group–delay function possesses additive and high resolution properties [106, 107]. The high resolution property results because around each resonance frequency the group–delay function behaves like a squared magnitude response [105]. In the Fourier transform magnitude spectrum the tails of the stronger formants attenuate the weaker formants due to the multiplication of the individual magnitude spectra. This effect is avoided in case of group-delay spectrum due to the additive property of the peaks of the individual resonances in the group–delay spectra. However the group-delay function in general is not well behaved for all classes of signals. For example, the zeroes of the z–transform of the excitation signal in voiced speech which are close to the unit circle produce large amplitude spikes in the group–delay function. The large spikes mask the details of the peaks due to the vocal tract resonances. Hence, Hema and Yegnanarayana [108] proposed a modified group–delay function–based method for formant extraction to handle the practical difficulties encountered in using group-delay functions. The modified group–delay function is obtained by multiplying the group–delay function computed from the speech signal with an estimate of the rapidly fluctuating component $(\hat{Z}(\omega))$ of the group–delay function of the excitation signal. $\hat{Z}(\omega)$ is ob-

tained by dividing the short-time squared magnitude spectrum of the speech signal by its cepstrally smoothed spectrum. In [108], a method for obtaining the log magnitude spectrum from the modified group-delay function has also been proposed.

McAulay and Quatieri [109-112] proposed a sum of sinusoids model to represent speech signals. The method represents the glottal excitation in terms of a sum of sine waves of arbitrary amplitudes, frequencies and phases. This model is written as

$$u(t) = \mathrm{Re} \sum_{l=1}^{L(t)} \mathcal{A}_l(t) \exp\left(j \left[\int_0^t \omega_l(\theta)\, d\theta + \phi_l \right] \right) \tag{2.7}$$

where, for the lth sinusoidal component, $\mathcal{A}_l(t)$, $\omega_l(t)$ and ϕ_l represent the amplitude, frequency and a fixed phase, respectively. The sum of sine waves, when applied to a time-varying filter, leads to the desired sinusoidal representation for speech waveforms. If the time-varying impulse response of the vocal tract filter is $h_v(\tau; t)$, then the speech signal $s(t)$ is given by

$$s(t) = \int_0^t h_v(t-\tau; t)\, u(\tau)\, d\tau \tag{2.8}$$

The amplitudes, frequencies and phases are estimated from the short-time Fourier transform (STFT). For a given frequency track a cubic function is used to unwrap and interpolate the phase such that the phase track is maximally smooth. This phase function is applied to sine wave generator, which is amplitude modulated and added to the other sine waves to give the final speech output. It is reported that the resulting synthetic waveform preserves the general shape of the original waveform and is perceptually indistinguishable from the original speech. Since the representation is general, high-quality reproduction was obtained for superposed speech waveforms, music waveforms, speech in musical backgrounds and certain marine biological sounds.

One of the most popular methods for speech analysis and coding is the Linear Prediction (LP) analysis of speech. The concept of linear prediction was originally in-

troduced by G. Udny Yule in *1927* to obtain a finite parameter model for a stationary random process. His objective was to investigate the periodicities in time series with special reference to Wolfer's sunspot numbers. Given an empirical time series $s(n)$, Yule used the method of regression analysis to find the coefficients of the model. Since the regression of $s(n)$ is on its own past instead of on other variables, it is called *self*–regression or autoregression. The regression analysis results in the normal equations involving empirical autocorrelation coefficients of the time series, now popularly called the Yule–Walker equations [99]. The linear prediction formulation is briefly presented below, since this forms the basis for most of the work presented in this thesis.

Linear Prediction analysis of speech

In LP analysis of speech, the speech signal $s(n)$ is assumed to be the output of an all–pole system. Let $\hat{s}(n)$ be the output of such an all–pole model

$$\begin{aligned} \hat{s}(n) &= -\sum_{k=1}^{p} a_k \, s(n-k) \\ &= -\mathbf{a}^T \, \mathbf{s}_{n-1} \end{aligned} \tag{2.9}$$

where $\mathbf{a} = [a_1 \; a_2 \cdot \cdot \cdot a_p]^T$ is the vector of Linear Prediction Coefficients (LPCs) and $\mathbf{s}_{n-1} = [s(n-1) \; s(n-2) \ldots s(n-p)]^T$ is the vector of past signal samples used to predict the nth sample. The instantaneous error $e(n)$ is given by

$$\begin{aligned} \mathrm{e(n)} &= s(\mathrm{n}) - \hat{s}(n) \\ &= \sum_{k=0}^{p} a_k \, s(n-\mathrm{k}), \qquad a_0 = 1 \\ &= s(n) + \mathbf{a}^T \, \mathbf{s}_{n-1} \end{aligned} \tag{2.10}$$

The expected value of error energy is given by

$$\begin{aligned} E &= \mathcal{E}\left\{e^2(n)\right\} \\ &= \mathcal{E}\left\{e^T(n) \; e(n)\right\} \end{aligned} \tag{2.11}$$

Using *(2.10)* in *(2.11)*, we have

$$\begin{aligned} E &= \mathcal{E}\left\{s^2(n)\right\} + 2\,\mathbf{a}^{\mathsf{T}}\,\mathcal{E}\left\{\mathbf{s}_{n-1}\,s(n)\right\} + \mathbf{a}^{T}\,\mathcal{E}\left\{\mathbf{s}_{n-1}\,\mathbf{s}_{n-1}^{T}\right\}\,\mathbf{a} \\ &= \mathcal{E}\left\{s^2(n)\right\} + 2\,\mathbf{a}^{\mathsf{T}}\,\mathbf{r}_{ss} + \mathbf{a}^{\mathsf{T}}\,\mathbf{R}_{ss}\,\mathbf{a} \end{aligned} \tag{2.12}$$

where

$$\mathbf{R}_{ss} = \begin{bmatrix} r_{ss}(0) & r_{ss}(1) & \cdots & r_{ss}(p-1) \\ r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(p-2) \\ \vdots & & \ddots & \vdots \\ r_{ss}(p-1) & & \cdots & r_{ss}(0) \end{bmatrix} \tag{2.13}$$

is a Toeplitz symmetric autocorrelation matrix,

$$\mathbf{r}_{ss} = [r_{ss}(1)\; r_{ss}(2) \cdots r_{ss}(p)]^{T} \tag{2.14}$$

and

$$r_{ss}(k-i) = \mathcal{E}\left\{s(n-k)\; s(n-i)\right\}, \qquad i,k = 1,2,\cdots p \tag{2.15}$$

are the autocorrelation coefficients. Minimising E in *(2.12)* w.r.t. the *LPC* vector **a**

$$\frac{\partial E}{\partial \mathbf{a}} = 0 \tag{2.16}$$

yields the following Yule–Walker equations

$$\mathbf{R}_{ss}\,\mathbf{a} = -\mathbf{r}_{ss} \tag{2.17}$$

or alternatively,

$$\sum_{k=1}^{p} a_k\, r_{ss}(k-i) = -r_{ss}(i), \qquad i = 1,2,\cdots p \tag{2.18}$$

The error signal $e(n)$ obtained using the optimal coefficients a is the ***Linear prediction residual*** signal. Using *(2.17)* in *(2.12)* the least squared error for the frame of samples is obtained as

$$\begin{aligned} E_{min} &= \mathcal{E}\left\{s^2(n)\right\} + \mathbf{a}^{T}\,\mathbf{r}_{ss} \\ &= r_{ss}(0) + \sum_{k=1}^{p} a_k\, r_{ss}(k) \end{aligned} \tag{2.19}$$

The normalized linear prediction error η is defined as the ratio of the energy of prediction error to the energy of the signal in the frame [40]

$$\eta = 1 + \sum_{k=1}^{p} a_k \frac{r_{ss}(k)}{r_{ss}(0)} \tag{2.20}$$

Since η is a ratio of energy values, it is always positive. Since the energy of prediction error cannot exceed the energy of the signal, η is upper bounded by one. It is shown in [40] that the same Yule–Walker equations can be obtained by a frequency domain approach, in which the short–time power spectrum of the speech signal $P_s(\omega)$ is approximated by the spectrum of the all–pole model $\hat{P}_s(\omega)$ by minimising the following cost function

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}_s(\omega)} \, d\omega \tag{2.21}$$

where

$$\hat{P}_s(\omega) = \frac{G^2}{| 1 + \sum_{k=1}^{p} a_k \exp(-j\omega k) |^2} \tag{2.22}$$

and G = $(E_{min})^{\frac{1}{2}}$ is the gain of the all–pole model. E_{min} is the minimum value of E. The form of the above cost function in **(2.21)** results in the positive deviations of the signal spectrum $P_s(\omega)$ from the model spectrum $\hat{P}_s(\omega)$ contributing to the integrand more than the negative ones. The $\hat{P}_s(\omega)$ therefore approximates the $P_s(\omega)$ best at the peaks of the signal spectrum. The normalized prediction error η can be written in the frequency domain as [40]

$$\eta = \frac{\exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \hat{P}_s(\omega) \, d\omega\right)}{\frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{P}_s(\omega) \, d\omega} \tag{2.23}$$

From **(2.23)** above we see that when $\hat{P}_s(\omega)$ is flat, $\eta = 1$. On the other hand, if the spread of $\hat{P}_s(\omega)$ is large, then η becomes close to zero. Thus η is a spectral flatness

measure [113].

In 1951, Wadsworth and Robinson [114] used linear prediction for processing seismic signals. In seismic signal analysis, since the interest is in determining the direct arrival times and strengths of the deep reflections of the dynamite explosion, the impulse response of the earth's crust has to be deconvolved from the recorded seismic traces. This deconvolution was achieved by performing LP analysis on the seismic traces. Later, Atal and Schroeder [115] used LP analysis for low bit rate speech coding. An efficient method to solve the Yule–Walker equations is facilitated by the Levinson–Durbin recursion method [116] which made real time implementation of LP analysis for speech coding possible. The Levinson–Durbin method exploits the Toeplitz and symmetric properties [117, 118] of the autocorrelation matrix to reduce the $\mathcal{O}(p^3)$ complexity of matrix inversion to $\mathcal{O}(p^2)$. The concept of linear prediction has also been extensively used for adaptive equalization to compensate for the time dispersion introduced by bandwidth constrained channels for digital data transmission [119]. An error analysis of the effects of data perturbations on the estimates of LPCs and signal parameters obtained from LP analysis is given in [120]. Some of the pitfalls of LP analysis are given in [40,121]. Its sensitivity to additive noise in the signal is discussed in [122–124]. Several methods have also been suggested to make the linear prediction analysis technique more robust to noise and outliers in the data [125–128].

Burg proposed the *maximum entropy spectral analysis* method [129, 130] in 1967 for processing geophysical signals. Burg posed the question that if the finite autocorrelation sequence $r_{ss}(0)$, $r_{ss}(1)$, $r_{ss}(2) \ldots r_{ss}(\mathrm{p})$ is assumed known or can be reliably estimated, how should the remaining unknown lags $r_{ss}(\mathrm{p}+1)$, $r_{ss}(p+2) \cdots$ be specified in order to guarantee that the entire autocorrelation sequence is positive semidefinite.

There are an infinite number of possible extrapolations that will yield valid autocorrelation sequences. Burg argued that the extrapolation should be made in such a way as to maximize the entropy of the time series characterized by the extrapolated autocorrelation sequence. This time series would then be the most random, in an entropy sense, of all series that have known autocorrelation sequence for lags 0 to p. A spectral estimate produced from this extrapolated autocorrelation sequence would then be for a process of maximum entropy. For the case of a Gaussian random process, the entropy rate was shown to be proportional to the integral of the natural logarithm of the power spectrum of the process. Therefore, for a Gaussian process, the maximum entropy spectrum is found by maximizing the entropy rate subject to the constraints that it satisfy the Wiener–Khintchine relationship for the $p+1$ known autocorrelation values. The solution, found by the Lagrange multiplier technique, is identical to the solution found by the least squares method. Thus, the maximum entropy method and the least squares all–pole modeling technique are identical for Gaussian random processes [131].

Hermansky et al. [132–135] proposed the Perceptually based Linear Predictive (PLP) analysis of speech. The PLP method of speech analysis models the speech auditory spectrum by the spectrum of a low order all–pole function. The auditory spectrum is obtained by critical–band spectral analysis which integrates the speech energy spectral density over 18 bands in the 0–5 kHz frequency range. The spectral analysis is followed by equal–loudness pre–emphasis which ernphasizes the middle and the upper part of the speech spectrum. To reduce the dynamics of the speech spectrum, an intensity–to–loudness cubic compression is performed. Eighteen samples of the auditory spectrum, obtained in this way, are transformed through the inverse DFT into the autocorrelation domain. Five coefficients of a 5th order all–pole model

are computed using the Yule–Walker equations. It was shown, through analysis of both synthetic and natural speech, that by applying the PLP analysis, some inherent problems of the standard LP analysis of speech are mitigated. The PLP method is less sensitive to the value of the fundamental frequency (F_0) of voiced speech and approximates spectral envelope zeros better than the standard LP analysis method.

2.3.1 Enhancement of Noisy Speech

In this section we present a review of methods at the segmental level for processing noisy speech for enhancement. We review only those enhancement methods which process speech after degradation. In the literature methods have also been suggested to process speech prior to degradation to combat the effects of noise [136–138], in situations where we have access to speech prior to degradation.

Weiss et al. [139–141] proposed a generalized ***correlation subtraction*** method for enhancement of speech degraded by additive noise. The short–time spectral magnitude (STFT) of speech is estimated and raised to the power ν, where ν is a positive constant. This is inverse Fourier transformed to obtain a generalized correlation function of the noisy speech signal. An estimate of the generalized correlation function of the background noise is obtained from the silence intervals. Assuming that the noise characteristics are stationary, the generalized correlation function of the background noise is subtracted from the generalized correlation function of noisy speech. This is transformed back into Fourier domain and raised to the power $1/\nu$. This is combined with the original noisy phase to obtain enhanced speech spectrum. The enhanced time–domain waveform is obtained using the overlap–add technique.

In 1979, Boll [142] extended the ideas proposed by Weiss et al. by performing the subtraction in the spectral domain directly, setting the negative spectral values to

zero after subtraction and reconstructing the enhanced speech. Hence this modified method is called the ***spectral subtraction*** method. One of the serious drawbacks of the above method is that it produces a ***musical noise*** in the enhanced speech. This noise arises because of peaks randomly spaced in the time–frequency plane due to the deviation of the estimated (smoothed) spectrum of noise from the instantaneous noise spectrum. The results of intelligibility tests on noisy speech enhanced using the spectral subtraction method are presented in [143]. The speech was corrupted by additive Gaussian noise. Results indicate that the method does not increase speech intelligibility. However, the subjective impression indicates that the processed speech distinctly sounds less noisy. A review of the spectral subtraction method and its variations and a unifying framework for several methods for speech enhancement is given in [144]. Later, McAulay and Malpass [145] proposed a variation of the spectral subtraction method which attenuates a particular spectral line depending on how much the measured speech plus noise power exceeds an estimate of the background noise. The spectral subtraction method has also been used to enhance speech for robust word recognition in a car environment [146].

Previous versions of the spectral subtraction method used uniformly spaced frequency transformations (e.g., DFT). Gulzow et al. [147] proposed application of filterbanks with bark–scaled frequency bands to perform spectral subtraction for speech enhancement. They investigated a discrete wavelet transformation and nonuniform polyphase filterbank. A filterbank with 70 channels for the case of wavelet transformation and a filterbank with 256 channels for the case of polyphase filterbank were used. It is reported that the amount of residual noise perceived is lower compared to the spectral subtraction method due to Boll.

As mentioned previously, one of the serious drawbacks with the spectral subtrac-

tion method is the overriding musical noise in the enhanced speech. This is due to the appearance of randomly spaced peaks in the time–frequency plane. There have been attempts to alleviate this problem by post–processing the speech in the time–frequency plane after spectral subtraction. Whipple [148] proposed removing spectral peaks in the time–frequency plane if there is no precedent or no spectral peak succeeding the current spectral peak in a fixed area around the current instant of time. Samudravijaya and Rao [149] proposed filtering the frequency tracks, after spectral subtraction, using linear phase FIR bandpass filters. The FIR filters have a passband in the range 1–16 Hz based on the knowledge that spectral transitions in continuous speech seldom lie beyond this range. Thus randorn peaks due to spectral subtraction are attenuated to a large extent. Clearly, the price paid, as in the case of RASTA like approaches, is that there will be smearing of rapid spectral transitions.

Speech enhancement has also been performed using the signal and noise subspace concepts. This is based on the Singular Value Decomposition (SVD) [150] of the noisy signal prediction matrix. The signal prediction matrix has a Hankel structure. A new noise reduced data matrix is obtained by truncating the singular values corresponding to the noise subspace. Since the enhanced data matrix will not retain the Hankel structure, antidiagonal averaging of this matrix is performed to restore the Hankel structure. Jensen et al. [151] proposed the use of Quotient Singular Value Decomposition (QSVD) for signal subspace–based speech enhancement. This formulation has prewhitening operation (for non–white noise situations) as an integral part of the algorithm. The interpretation of the SVD truncation as a zero–phase filtering operation is given in [152].

Bouquin–Jeannes et al. [153] proposed a two channel speech enhancement method for hands–free communication. The method computes the coherence function between

the signals recorded on two microphones for each frame [154,155]. At any frequency, a high coherence value indicates it is a component predominantly due to speech. Otherwise it is due to noise. The assumption made here is that the cross–correlation between the noises recorded on the spatially separated microphones is zero. The short–time spectrum of noisy speech is multiplied by the magnitude of the coherence function to obtain an enhanced spectrum. This is used to reconstruct enhanced speech. A minor variation of this method has been presented in [156]. Yet another two channel noise reduction method has been proposed in [157].

A Minimum Mean Square Error (MMSE) estimator of the short–term spectrum of the speech signal based on Hidden Markov Modeling (HMM) [158,159] of the clean speech signal as well as the noise process was proposed by Ephraim [160]. Xie and Compernolle [161] later investigated a MMSE estimator of the speech spectrum in the logarithmic domain. The ***a priori*** probability distribution functions (PDF) for speech and noise are assumed to be log–normal. The method estimates the short–time spectral magnitude in a frame calculated using an N point DFT. The enhanced speech is obtained by using the estimated spectrum with noisy phase and reconstructing the signal using the overlap–add method.

The periodicity of voiced speech has also been exploited for enhancement of speech corrupted by additive interference. Lim et al. [162] have studied the effect of ***adaptive*** comb ***filtering*** of noise–corrupted speech on intelligibility. The adaptive comb filtering operation passes only the harmonics of speech. Since interfering signals will, in general, have energy in the frequency regions between the speech harmonics, the comb filtering operation can reduce noise while preserving the periodicity of speech. Comb filtering operation in the frequency domain translates to symmetric FIR filtering in the time domain. The taps of the FIR filter will be a glottal cycle apart. The comb filter is

implemented in the time domain. It is reported that there is a marginal improvement in the intelligibility at very low SNR (< 0 dB) using a 3–tap filter. The intelligibility decreases with increasing number of taps. However, the processed speech "sounds" less noisy due to the dominance of pitch harmonics over the interference signal.

Yet another technique based on the periodicity of voiced speech has been proposed by Sarnbur [163]. This technique is based upon the principles of *Least Mean Square* (LMS) adaptive filtering [164]. While the classical LMS adaptive filtering requires the reference noise recorded on a separate channel, the method due to Sambur does not require reference noise. Instead it uses the knowledge of the duration of the glottal cycle, which is separately estimated using the Average Magnitude Difference Function (AMDF) algorithm. The speech signal samples which are a glottal cycle length prior to the present sample are used as the reference input of the original speech signal. Thus the reference input is generated using a linear weighted sum of samples taken from the previous glottal cycle. The difference between the current sample and the linear weighted sum of samples taken from the previous glottal cycle is the error signal. This error signal is used to update the weights on a sample by sample basis according to the LMS algorithm. It is reported that this technique improves the quality of noisy speech after processing. The technique also improves the performance of LP analysis/synthesis of noisy speech. A generalized comb filtering technique has been proposed in [165].

Wavelet transform–based methods for noise reduction have also been proposed. Donoho and Johnstone [166–168] proposed thresholding in the wavelet domain to reduce the effects of noise. Let $y(n)$ represent the samples of a noisy signal:

$$y(n) = s(n) + w(n), \quad n = 0, 1, ..., N-1 \tag{2.24}$$

where $s(n)$ are samples of the clean signal and $w(n)$ are samples of a zero–mean white

Gaussian noise process $\mathcal{N}(0, \sigma_w^2)$. The method proposed in [166] obtains the wavelet coefficients by performing the wavelet transformation on the noisy observations $y(n)$. The wavelet coefficients are thresholded using σ_w as the threshold. The enhanced signal is obtained by performing inverse wavelet transformation on the thresholded wavelet coefficients. Both ***hard thresholding*** and ***soft thresholding*** strategies have been proposed. In hard thresholding, all the wavelet coefficients less than σ_w are set to zero. In soft thresholding, all the wavelet coefficients below σ_w are set to zero while the other coefficients are shrunk by a value a. Of the two thresholding schemes, hard thresholding yields the smaller mean square error. However, hard thresholding exhibits spurious oscillations. Similar to classical noise reduction methods, the wavelet–based method too has a tradeoff between noise reduction and oversmoothing of signal details. The method due to Donoho uses an orthogonal wavelet basis. Lang et al. [169] have proposed a similar method but using an undecimated, nonorthogonal wavelet basis.

Neural network based approaches have also been proposed for speech enhancement. Knecht et al. [170] used a neural network architecture to implement a nonlinear Volterra filter for adaptive noise canceling. The motivation for the nonlinear filtering comes from the fact that when the observed data and the data to be estimated are jointly Gaussian, then the linear filters perform optimally. Since acoustic signals cannot generally be modeled as Gaussian processes, nonlinear filters have been employed for noise canceling. The method employs two microphones to record noisy speech data. The speech source is assumed to be equidistant from the microphones while the interference signal is assumed to be off–axis. The scaled difference between the two microphone signals contains no signal components and forms the reference input to the canceler. The scaled sum of the two microphone signals is the primary input to the canceler. The nonlinear filter is reported to have shown improved performance in

terms of intelligibility–weighted gain compared to an adaptive linear filter.

Yet another neural network–based approach has been proposed by Tamura and Waibel [171]. In [171], noise reduction is viewed as a mapping from a noisy signal space to a noise–free signal space. The problem is finding such a complex nonlinear mapping. The proposed noise reduction method uses a four–layered feedforward neural network to capture the nonlinear mapping. Using the back propogation learning algorithm [172, 173], the network is trained with noisy speech signals as input and the corresponding noise–free speech signals as target output. The trained network is reported to have produced noise–suppressed signals even for signals that differed from the training data in both the original speech input as well as the type of environmental noise. An analysis of the method and improvements to the method mentioned above have been presented in [174] and [175], respectively.

Anitha and Yegnanarayana [176] proposed a similar neural network–based approach for capturing the correlated features in the speech signal in both time and frequency domains. The network is a three layered feedforward auto–association network [177, 178] which is trained on clean speech using 10 ms segments of the speech signal as input. When the trained network is presented with 10 ms segments of noisy speech as input, the network gives only the correlated part of the signal as output and uncorrelated part is suppressed thus achieving speech enhancement. The network exhibited good generalization capability since enhancement was achieved for noisy speech data which was not part of the training data. The processed speech was found to be less noisy. The higher formants, which are generally weaker, were however found to be attenuated.

Shen et al. [179] have proposed an H_∞ norm–based filtering method [180] for speech enhancement. The least squares approximation suffers from the disadvantage that an

estimate of only the total error in an approximation problem is available; there is no estimate for the accuracy of the approximation at each of the data points. On the other hand, the H_∞ norm based approximation provides such a bound on the error. The minimisation of the H_∞ norm minimises the maximum of the approximation error rather than the total error, thus bounding the maximum error possible at any data point [181,182]. The other advantages are that no a *priori* knowledge of the noise statistics is required. It is robust to modeling errors. The method proposed by Shen et al. uses the following state-space formulation of the problem. The coefficients of a 10th order linear prediction model estimated from 16 ms segments of noise corrupted speech form the state–transition matrix. A linear transform of the state vector corrupted by additive noise gives the observed noisy speech sample.

2.3.2 Enhancement of Reverberant Speech

There are several techniques proposed in literature, which process the signal at the segmental level, for enhancement of reverberant speech. Both single microphone and multi–microphone approaches have been proposed. A brief review of these methods is presented below.

The speech signal corrupted by reverberation recorded on a microphone can be considered to be the convolution of the original sound with the impulse response of the room between the source and the receiver. Schafer [183] assumed the above model and proposed a homomorphic deconvolution method to retrieve clean speech. The underlying motivation is the fact that deconvolution in the time domain corresponds to subtraction in the cepstrum domain. Since the complex cepstrum of a speech signal is usually concentrated around the cepstral origin, while that of the echoes is composed of pulses extending far away from the origin, it follows that low–time filtering

in the cepstral domain can be used to remove the echo's cepstrum. For the case of a simple echo, such a deconvolution can be achieved. However, in real environments the situation is more complex due to multiple echoes and diffused sounds due to reverberation tails. Neely and Allen [184] have also proposed a cepstrum based method which estimates the room impulse response and inverse filters the reverberant speech signal to cancel the effects of reverberation. They assumed that the room impulse response is minimum phase. In real situations, this is again not true.

Flanagan and Lummis [185] proposed a multi–microphone approach for processing reverberant speech. The speech signal from each microphone is separated into several subbands. Among all the microphone outputs in each subband the maximum is chosen. The subbands so chosen are recombined to obtain enhanced speech. Allen et al. [186] have proposed a two–microphone approach for processing reverberant speech which essentially combines the methods in [185] and [93]. Their method uses two spatially separated inputs (one at each ear) to enable measurement of interaural correlation/coherence in each band of an analysis filter bank. In bands with high levels of interaural coherence, which implies the presence of a strong, direct component, the signal is passed relatively unaltered to a synthesis operation. Bands with low levels of coherence (containing mainly reverberation) are attenuated. Bloom and Cain [187] suggested modifications to the method proposed by Allen et al. The modifications suggested are frequency domain smoothing of coherence measures on a critical–band basis and suppression rules based on threshold coherence estimates. Improved quality of processed speech was reported, compared to that obtained by using the method due to Allen et. al.

Farrell et al. [188] proposed a microphone array-based beamforming technique [189] for speech enhancement. The beamforming technique exploits the directional property

of a microphone array, when the different microphone signals are combined, to reject signals whose direction of arrival does not coincide with the ***look direction*** of the array. The ***look direction*** of the array is normal to the array and thus only the signal arriving along the normal is treated as a desired signal. By steering the array in the direction of the source, the interference signals (both noise and reverberation) arriving from other directions are suppressed.

As mentioned above, the impulse responses of typical rooms are non–minimum phase and have therefore unstable inverses. Therefore inverse filtering based methods have a limited scope in practice. Liu et al. [190] proposed a microphone array processing technique for blind dereverberation of speech signals. This method factors the signal received at each microphone into the minimum phase and all–pass components. To recover the minimum phase component of the original speech, spatial averaging followed by low–time filtering in the cepstral domain is applied to the minimum phase components of the individual microphone signals. The phase information of the microphone signals is preserved in their all–pass components. The final dereverberated speech is obtained from the synthesis of the recovered minimum phase and all–pass components.

Subramaniam et al. [52] have proposed a two–microphone cepstrum based processing method. The method reconstructs the room impulse response associated with each microphone using cepstral operations. The estimated impulse responses are used to perform deconvolution. The dereverberation achieved has been demonstrated by a comparison of segments of clean, reverberant and dereverberated speech waveforms in a voiced region. Several methods for enhancement of speech corrupted by noise and reverberation, especially for telephony applications, can be found in [191].

It is clear from the above discussion in Section 2.3 that several of the methods

presented above for processing speech at the segmental level are driven by signal processing considerations. In the following section, we see how subsegmental speech processing addresses the problems specific to the speech signal, i.e., takes into account the characteristics of the speech production mechanism.

2.4 PROCESSING OF SPEECH AT THE SUBSEGMENTAL LEVEL

Using an analysis window of duration **10–30** ms does not provide adequate temporal resolution. The rapid changes that occur in the characteristics of the vocal tract in the case of CVs and also the changes within a glottal cycle of voiced speech are smeared. Subsegmental analysis enables us to track these changes. However, the positioning of the analysis window becomes critical when such short speech segments are processed. In this section, a review of methods proposed for pitch synchronous analysis of short **(1–3** ms) segments of the speech signal is presented. Very few methods have been proposed in the literature for processing speech at the subsegmental level.

Ananthapadmanabha and Yegnanarayana [1] have proposed the identification of epochs (instants of significant excitation of the vocal tract system) by applying the epoch filter theory [192] to linear prediction residual. The accuracy of epoch identification was tested by performing LP analysis of the speech signal in the closed glottis interval. The covariance formulation of LP analysis [193] was used for estimating the linear prediction coefficients from short **(1–3** ms) segments [194]. In the case of synthetic vowels, the estimated frequency response was found to compare well with the actual frequency response, when the short analysis interval was chosen in the closed glottis region. Bandwidths of formants were also estimated correctly. For other positions of the analysis interval the estimated frequency response was strongly influenced by the position.

The performance of conventional covariance method of linear prediction analysis deteriorates rapidly in the presence of noise. In addition, the number of data samples in the open or closed glottis intervals is usually small which makes accurate estimation of the parameters of the signal difficult. The problem of estimating the exponential parameters from short, noisy observations has been studied extensively [195–199]. Parthasarathy and Tufts [23] proposed a pitch synchronous modeling technique for voiced speech which analyses the speech signal in the closed and open glottis intervals. The glottal closure was estimated by computing the energy in the frequency band containing the first formant (300 Hz–1000 Hz approx.) using 3–4 ms segments of the speech signal and identifying the peaks of the energy contour. Two all–pole models were used in each glottal cycle, and the model parameters were changed at estimated times of transitions from open–to–closed and closed–to–open glottis. For estimating the parameters of the all–pole models, the linear prediction analysis formulation was used. The predictor length used was more than the minimum required and the singular value decomposition (SVD) was used to obtain a low–rank approximation of the data matrix [195] in the linear prediction analysis. This method was found to yield accurate estimates of the poles with reduced variance.

Nathan and Silverman [7] have extended the linear prediction model based on time–dependent poles proposed in [200, 201]. The speech signal is analysed pitch synchronously in the closed glottis interval. The 4 ms segment immediately after glottal closure in each glottal cycle is assumed to be the closed glottis interval. The locations of glottal closures are estimated using the algorithm proposed in [23] discussed above. An autoregressive model with time–varying coefficients is used to represent the vocal tract system. The time–varying coefficients of the autoregressive model are assumed to vary linearly with time. Each coefficient is specified by two parameters, an initial value

and a rate of variation. These parameters are estimated using an iterative algorithm to obtain the maximum–likelihood estimate. The results of analysis of several cases of diphthongs and vowel–to–stop consonant transitions are presented. It was found that the method detected consistent changes in steady–state formant values, preceding final stops, in a region where the vocal tract changes shape very rapidly. It is also reported that the estimates of time–varying coefficients did not yield stable trajectories for the formants for the glottal stop /k/. This is indicative of the fact that the assumed linear model for the parameters is not appropriate in all cases.

Yegnanarayana and Veldhuis [19] have also suggested pitch synchronous analysis of voiced speech, performed in the closed and open glottis intervals. The closed and open glottis intervals were identified based on the knowledge of the instants of significant excitation of the vocal tract system [202, 203]. In the case of analysis of voices with higher fundamental frequency, the analysis frame constrained to lie within the closed glottis interval may become too short for reliable extraction of parameters. To alleviate this problem, they proposed an averaging technique called the multi–cycle covariance method, which averages covariance estimates over a number of consecutive glottal cycles. The covariance estimates are obtained from pitch synchronous speech segments in the closed/open glottis interval. The formant frequencies F_k and their bandwidths B_k were derived from the roots of the prediction polynomial using the formulas [15]:

$$F_k = \frac{f_s}{2\pi}\theta_k \tag{2.25}$$

$$B_k = -\frac{f_s}{\pi}\log(\rho_k) \tag{2.26}$$

where f_s is the sampling frequency, k is the index of the particular formant, θ_k, $-\pi < \theta \leq \pi$, is the normalized formant frequency and ρ_k, $0 \leq \rho_k < 1$, is the pole radius. It was found that consistent estimates of formant frequencies and to a lesser extent the formant bandwidths, could be derived by analysing the speech signal in the closed

glottis region. The formant frequency tracks obtained by analysing the signal in the open glottis region were found to be less consistent.

2.4.1 Enhancement of Speech using Subsegmental Processing

To the best of our knowledge there are no methods which process 1–3 ms of the speech signal for enhancement. In this work we propose methods for processing speech degraded by noise and reverberation based on the subsegmental processing of speech. Unlike the traditional approaches whose objective is noise subtraction/dereverberation, the focus of the proposed methods is on emphasis of high SNR/SRR segments of speech. This approach is motivated by the fact that human beings perceive speech by capturing some features from high SNR regions in the spectral and temporal domains, and then interpolate the features at various levels in the low SNR regions [204]. Direct manipulation of the samples of the speech signal will result in distortion. Hence we propose weighting the samples of the LP residual signal to give more emphasis to high SNR/SRR segments of speech relative to the other segments. The weighting is done at the global (40–50 ms) level as well as at a short (1–2 ms) level. The weighted residual signal is used to excite the time–varying LP all–pole filter to produce enhanced speech.

2.5 OUTLINE OF THE WORK PRESENTED IN THIS THESIS

Previous work in speech processing mainly focussed on processing the speech signal at the suprasegmental level using 100–300 ms or more of the signal for processing and at the segmental level using 10–30 ms of the signal. In the work presented in this thesis, the focus is on subsegmental processing of speech, using 1–3 ms of the signal for analysis.

Subsegmental processing is advantageous for various reasons. If short (1–3 ms)

segments of the speech signal are taken pitch synchronously for analysis, then one may capture the consistent variations in similar segments in successive glottal cycles of voiced speech. In practical conditions, where the speech signal may be corrupted by noise and/or reverberation, the high SNR/SRR segments within a glottal cycle of voiced speech can be exploited for reliable analysis when short segments of the speech signal are used.

Chapter 3 discusses the issues in the analysis of short (1–3 ms) segments of speech. Chapters 4 and 5 discuss subsegmental analysis methods for speech enhancement. The practical issues in the implementation of these methods are considered in Chapter 6. Chapters 7 and 8 discuss methods for identifying high SNR segments in degraded speech and using them for enhancement.

2.6 SUMMARY

In this chapter, we have presented a review of three important speech processing paradigms, namely processing at the suprasegmental level, the short–time or segmental level and the subsegmental level. We have also discussed the issues that arise in these methods. We have discussed an outline of the work presented in this thesis based on subsegmental processing of speech.

Chapter 3

SOURCE–SYSTEM WINDOWING FOR SPEECH ANALYSIS AND SYNTHESIS

In the previous chapters we have discussed the issues in processing speech at the suprasegmental, segmental and subsegmental levels. We have also seen the merits of subsegmental processing of speech, which is primarily motivated by the characteristics of speech production mechanism. In this chapter, a method for subsegmental analysis of speech is proposed to bring out variations in the vocal tract system characteristics in short (1–3 ms) segments. To reduce the effects of truncation of conventional waveform windowing, the concepts of windowing the source and system components of the speech signal are introduced. The individual windowed components are combined to generate a signal whose characteristics correspond mostly to the short region of interest. Through this analysis, the vocal tract system characteristics within a glottal cycle can be more accurately represented by modeling the system in the closed and open glottis regions separately, than by the conventional short–time (10–30 ms) analysis.

Before we present the proposed method, a brief discussion is given in Section 3.2 on various windowing options available. The proposed method is presented in Section 3.3. Results of analysis of different vowels and other types of speech segments are presented in Section 3.4 to demonstrate the effectiveness of the method for obtaining an estimate of the characteristics of the vocal tract from short (1–3 ms) segments of speech. Finally

the significance of this analysis for speech synthesis is discussed briefly in Section 3.5.

3.1 INTRODUCTION

One of the objectives in speech analysis is to derive the time–varying characteristics of the speech production mechanism from the speech signal. For analysis purposes, a linear source–system model is assumed for speech production, and the source and system characteristics are assumed quasistationary in the analysis interval [3]. Obviously, this simple model does not give an accurate representation of speech in each frame. Even in the steady voiced sounds, the excitation characteristics change within each glottal cycle due to glottal vibrations, and the vocal tract system changes due to coupling and decoupling of the trachea during the open and closed phases of the glottal excitation, respectively [10]. The linear prediction analysis [40] of short (10–30 ms) segments of speech capture only the averaged behaviour over the analysis frame. The detail lost in the LPC modeling cannot easily be compensated for by using a glottal pulse model [205] for excitation.

The difficulty in determining the characteristics of the vocal tract system from short (1–3 ms) segments of the speech signal is that in the short segment analysis the samples outside the chosen interval are either assumed to be zero (short–time spectrum, LP analysis by the autocorrelation method [36]) or assumed to belong to the same stationary region (covariance method [42, 206]). For example, using 1–3 ms (8–24 samples at 8 kHz sampling rate) of speech, the short–time spectrum (STFT) gives a poor resolution of the resonance frequencies, and the estimated autocorrelation coefficients for LP analysis are biased. The covariance method is sensitive to the window positioning. In the covariance method, a necessary but not sufficient condition for the correlation rnatrix to be non–singular is that the model order must be no greater than half the data length. The resulting all–pole model is not guaranteed to

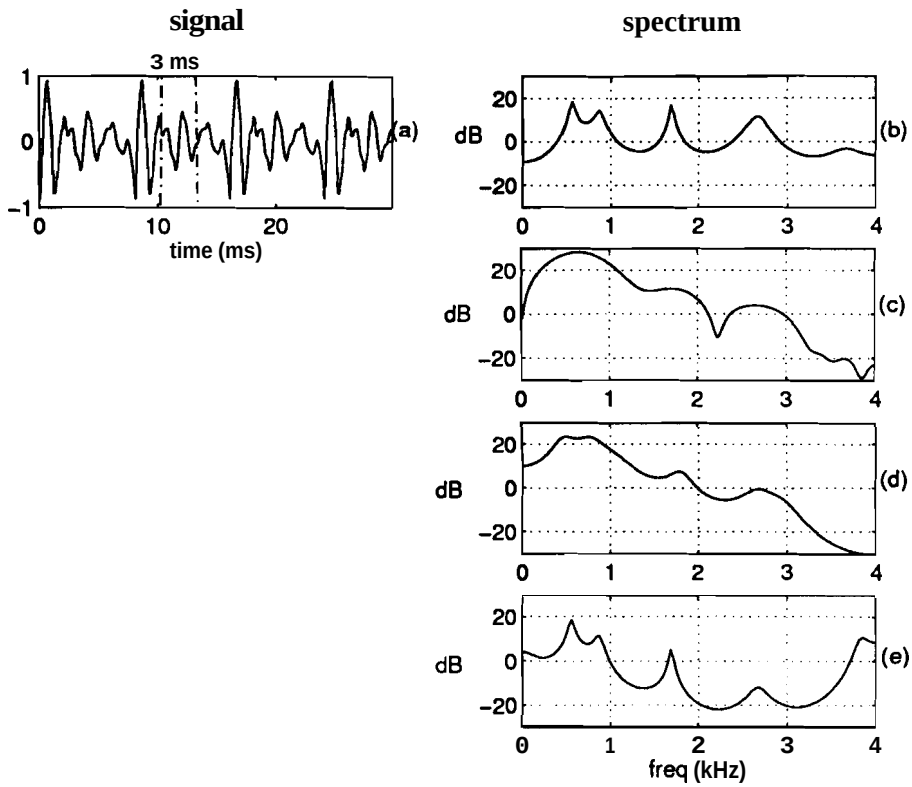


Fig. 3.1: Estimated spectra from short (3 ms) data of a synthetic signal by different methods. (a) Signal generated by exciting a 10th order all–pole model using periodic differentiated glottal pulses. (b) Spectrum of the all–pole model. (c) Short–time spectrum using a 3 ms Hamming window. (d) LP spectrum by the autocorrelation method. (e) LP spectrum by the covariance method.

be stable. Fig. 3.1 illustrates some of the problems with the methods mentioned above. **A** synthetic signal is used for analysis. The synthetic signal is obtained by exciting an all–pole model by a periodic sequence of glottal pulses. The short–time spectrum reflects the poor spectral resolution due to the small (3 ms) size of the window. The other methods yield poor results due to biased estimation of the autocorrelation values from the short data. The objective of this work is to explore methods to reduce the effects of the short window in the analysis. It is to be noted that windowing the signal causes discontinuity at the edges of the window. At the same time, windowing is essential to capture the dynamic characteristics of the source and system in the

speech production mechanism. While several methods have been proposed earlier for windowing the signal directly, we propose a new approach in this chapter, which we call ***Source-System windowing*** [43,44]. The central idea is to explore methods where the source and system components of a speech signal are independently modified to confine their effects to the selected analysis window region. The windowed components are then used to regenerate a speech signal corresponding to the source and system in that region, although the generated signal itself may extend beyond the selected window length. The generated signal is analysed to extract the system characteristics more accurately. In this chapter we show that even an approximate decomposition of the original speech signal into source and system components will be adequate to implement the proposed source–system windowing.

3.2 WINDOWING OPTIONS FOR SPEECH ANALYSIS

As mentioned earlier, windowing is essential for analysing the speech signal to extract the characteristics of the vocal tract system in a quasistationary state, as both the excitation source and the vocal tract shape change with time during speech production. Fig. 3.2 gives a summary of various options available for windowing the speech signal. Broadly, one can view windowing either of the signal waveform or in the source–system components. Waveform windowing is straightforward, where a region in the signal is selected and multiplied by a suitable window function. The choice of the region and the window function could be done in a signal–independent manner, as for example in Hamming or Hanning window of fixed duration at any arbitrary position in the signal. The window position and shape could also be signal–dependent. In the latter case, the window could depend on the signal characteristics, like low amplitude regions at the ends, or synchronizing the end points with zero crossings, or synchronizing with a glottal cycle. Choice of the signal–dependent window could also be influenced by

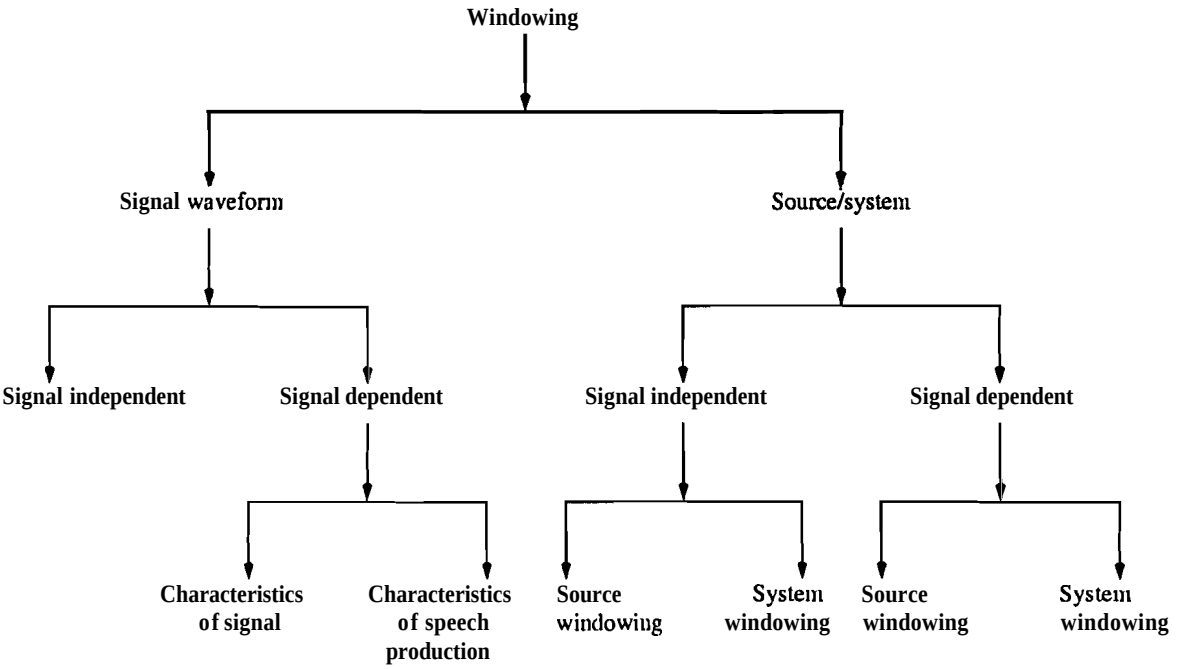


Fig. 3.2: Various windowing options for speech analysis.

the speech production, as for example, depending on whether the analysis segment is voiced or unvoiced; if voiced, whether the region is an open or closed glottis region.

For windowing in the source–system components, there are again choices such as signal–independent and signal-dependent windowing. Typically, windowing the source component refers to windowing some kind of a residual signal, such as linear prediction residual. Likewise windowing the system component refers to modification of the modeled vocal tract system characteristics outside the region of interest. In the signal–independent windowing, the position and size of the source and system window functions are selected arbitrarily, and implemented uniformly throughout. On the other hand, in the signal-dependent windowing, the position and size are dictated by the signal and the speech production characteristics. For example, the source window is placed on the residual signal in a pitch synchronous manner, and the size is chosen to avoid abrupt changes at the ends of the window. In the next section, we discuss

effects of these choices for windows on the analysis of speech signals.

3.3 SOURCE–SYSTEM WINDOWING

In the waveform windowing, the shape of the signal waveform is altered. Consequently, the source and system characteristics derived from the signal may not represent the speech production system well. The effects of waveform windowing will be severe when the window size is small (1–3 ms). In these cases the system model tries to fit the zero value samples outside the window, assuming them as natural extension of the samples within the window region. This leads to bias in the estimated autocorrelation or spectral values, and consequently results in errors in the parameters of the model derived from these values. The main reason for the bias is the correlation between samples in natural speech. The high correlation between signal samples can be seen from the autocorrelation function of a segment of speech (Fig. 3.3(a)) as shown in Fig. 3.3(b) for a window size of 30 ms. Truncation of a signal with significant correlation between samples would result in bias in the estimated values. The errors in correlation estimates get worse as the window size is reduced. The bias can be seen by a comparison of the autocorrelation values in Figs. 3.3(b) and 3.3(d). The autocorrelation values in Fig. 3.3(d) are computed from the 3 ms Hamming windowed segment shown in Fig. 3.3(c).

The correlation between samples is reduced significantly in the residual signal derived from LP analysis. The LP residual signal for the signal segment in Fig. 3.3(a) is shown in Fig. 3.3(e). The autocorrelation function of the LP residual signal for the same two window sizes is shown in Figs. 3.3(f) and 3.3(h). The values of the normalized autocorrelation function of the residual signal are small and they remain small even when the window size is reduced. That is, the short window effects are much less severe for the residual signal than for the original signal. The system characteristics

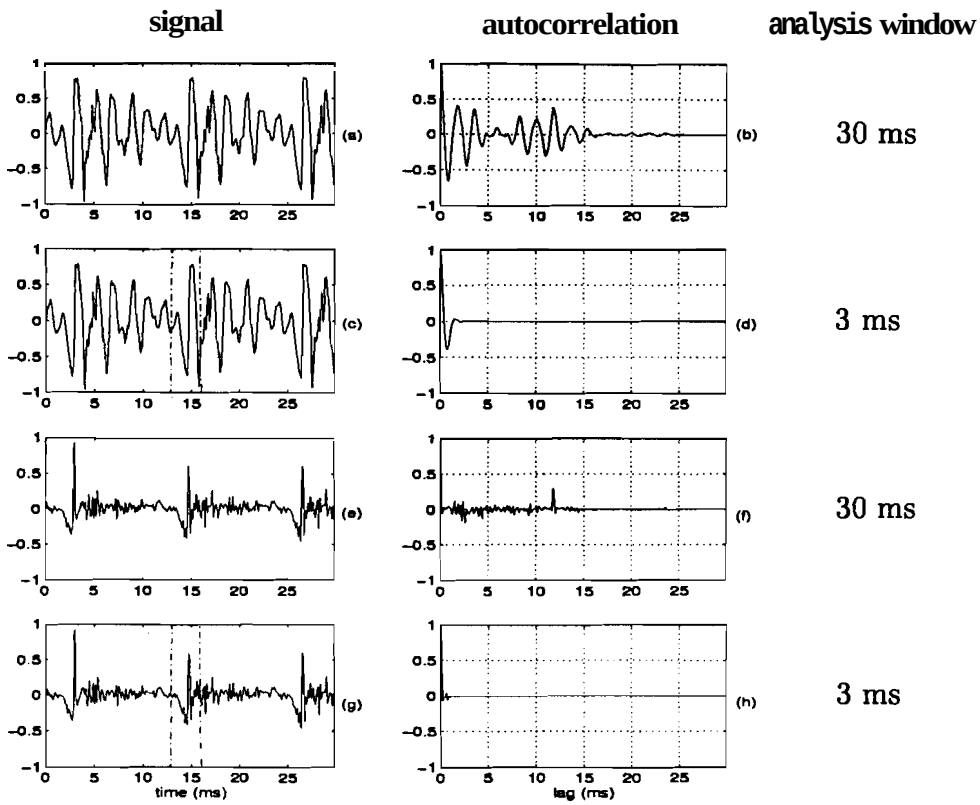


Fig. 3.3: The effect of window duration on autocorrelation estimates. (a) Segment of a natural vowel. (b) Autocorrelation of the signal in (a). (c) A **3** ms segment of the signal in (a) is enclosed between two dotted lines. (d) Autocorrelation of the **3** ms segment in (c). (e) LP residual signal corresponding to the signal in (a). (f) Autocorrelation of the residual signal in (e). (g) A **3** ms segment of the residual'signal is enclosed between two dotted lines. (h) Autocorrelation of the **3** ms segment in (g).

in the signal not captured in the LPCs appear in the LP residual signal. Moreover, these characteristics are typically reflected in small durations of the residual signal due to the finite impulse response nature of the inverse filter. Therefore, selecting a short window in the residual signal and reexciting the all–pole system would generate a signal whose characteristics in the selected window will be similar to those in the corresponding window in the original speech signal. Due to the all–pole filtering, the signal so generated extends beyond the chosen window with its natural decay, even though there is no excitation. This signal may be called *source windowed signal.* Using

a nonrectangular tapered window on the residual signal will reduce the edge effects, without significantly affecting the source characteristics. Note that the generated signal beyond the residual signal window region is due to the all–pole filtering. Since the signal generated beyond the residual signal window region is not influenced by the excitation, we are not likely to get significantly new information other than what is present in the all–pole filter.

In order to reduce the dominance of the system (all–pole filter) on the residual signal excited waveform (source windowed signal), we propose a modification of the system in the regions outside the chosen window. We call this bandwidth (BW) win-*dowing*, by which we mean a modification of bandwidths of the resonances to produce a tapering window effect. A bandwidth function is used to increase the bandwidth of the poles of the original all–pole system significantly, beyond the selected window region. The resulting waveform, which we will refer to as source–system windowed signal, will have nearly the same vocal tract system characteristics as that of the original speech signal within the window. The samples beyond the selected region taper faster than the signal in the short region of interest, thus enhancing the characteristics of the signal within the window. The steps in the algorithm for the source–system windowing are given in Table–3.1.

The BW windowing may reduce the frequency resolution slightly, but would still bring out the characteristics of the system in the analysis window. The results of analysis of a synthetic signal, generated by exciting a 10th order all–pole model by a periodic sequence of Liljencrants–Fant (LF) model glottal pulses [205], are shown in Fig. 3.4. The signal is preemphasized before analysis. To generate the synthetic signal in the closed glottis region, the all–pole rnodel spectrum in Fig. 3.4(a) is used, and for the open glottis region the same rnodel with its poles damped, as shown in Fig. 3.4(b),

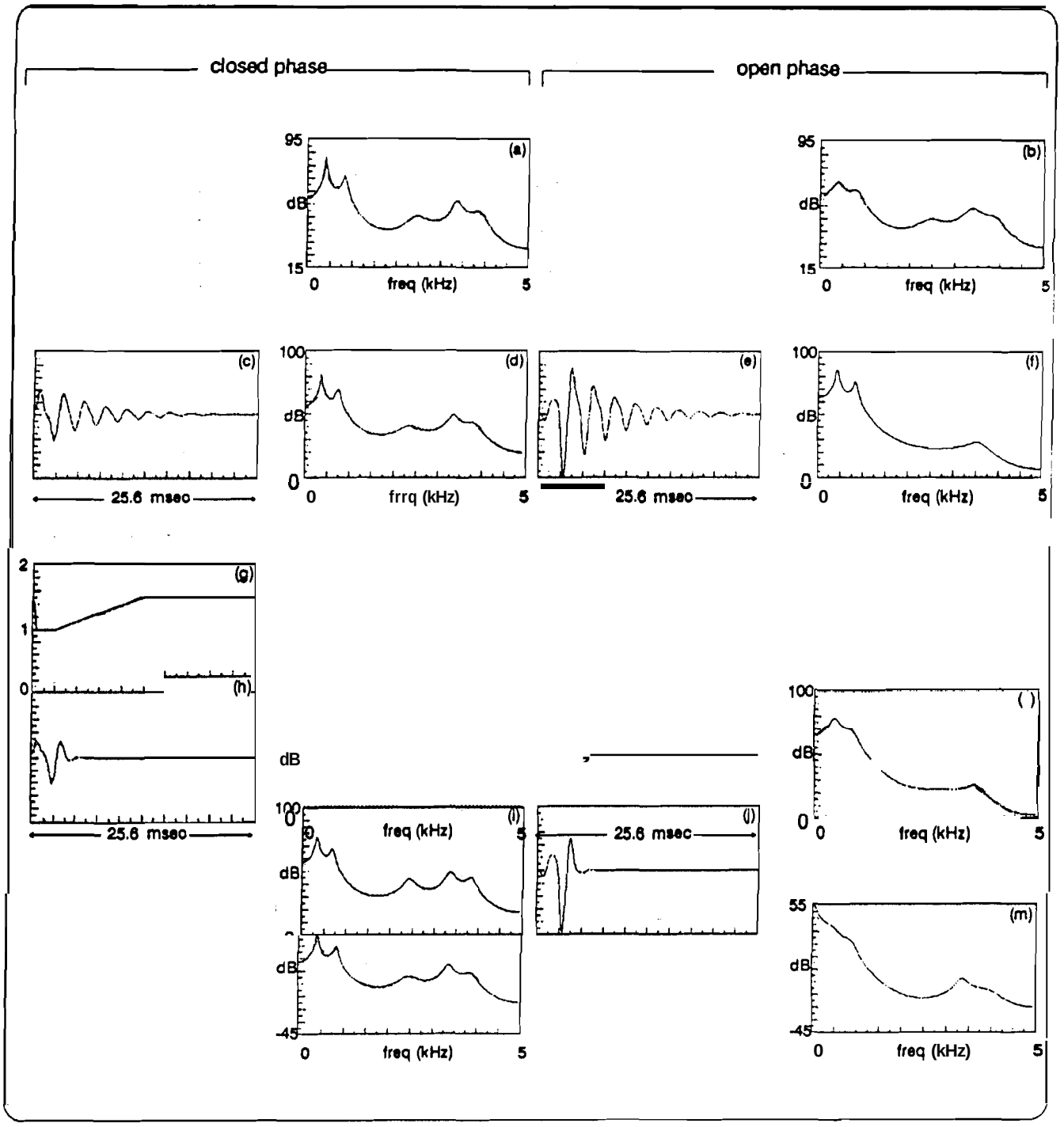


Fig. 3.4: Analysis of a synthetic signal using source–system windowing. (a), (b) 10th order dl–pole model spectra used for synthesis in the closed and open glottis intervals, respectively. (c) Signal generated without BW windowing in the closed glottis interval. (d) 10th order LP spectrum of the'signal in (c). (e) Signal generated in the open glottis interval without BW windowing. (f) The corresponding 10th order LP spectrum. (g) BW window function. (h), (i), (j) and (k) are the figures corresponding to (c), (d), (e) and (f) for the case with BW windowing. (1) 10th order covariance analysis LP spectrum in the closed glottis region of the signal. (m) 10th order covariance analysis LP spectrum in the open glottis region of the signal.

Table 3.1: Algorithm for source–system windowing.

- A 10th order autocorrelation LP analysis is performed using 25 ms Hamming windowed segments of the speech signal. The frames are overlapped by 15 ms.
- The linear prediction residual signal is computed for the entire speech signal using the LPCs obtained above.
- A short (3 ms) window with tapered edges is applied to the residual signal in the region of interest.
- The windowed residual signal is used to excite the time–varying all–pole filter obtained above to generate the ***source–system windowed signal.*** The original all–pole filter is used in the short region of interest where the excitation is nonzero. The filter is damped outside the region of the source window using $a_k^{mod} = a_k(\xi_n)^{-k}$, k = $0, 1 \cdots p$ and n = $0, 1, \cdots, N-1$. ξ_n is the bandwidth function of the shape shown in Fig. 3.4(g), a_k is the kth LPC and N is the number of samples in the frame.
- The signal so generated is analysed using a 10th order LP analysis to derive the LP spectrum corresponding to the short region of interest.

is used. Since all the poles are damped, some higher formants may be lost in the LP analysis of the damped signal for the open glottis region. The need for bandwidth windowing is evident from Figs. 3.4(e) and 3.4(f) which show the signal generated using source windowing in the open glottis interval and the corresponding 10th order LP spectrum, respectively. Clearly, the oscillations due to the first formant outside the selected window dominate the analysis. This influence is reduced significantly when the BW windowing is used, as shown in Fig. 3.4(k). The bandwidth function used for damping the poles is shown in Fig. 3.4(g). Comparison with covariance analysis is illustrated in Figs. 3.4(l) and 3.4(m) for the closed and open glottis regions, respectively. The covariance analysis gives the correct LP spectrum, as shown in Fig. 3.4(l), in the closed phase, as it does not have any effects of glottal source. Fig. 3.4(m) shows that the covariance analysis fails in the open glottis region.

We have also observed that the significant features, such as formant peaks and

their bandwidths, are preserved to a large extent even for very sharp changes in the bandwidth function. There will be slight increase in the bandwidths of the formants as the bandwidth window is sharpened. The effectiveness of BW windowing is illustrated through the symmetric Itakura distances given in Table–3.2. The symmetric Itakura distance is given by [207]

$$d_{ij} = \frac{1}{2}\left[\frac{\mathbf{a}_{a\_j}^T \mathbf{R}_i \mathbf{a}_{a\_j}}{\mathbf{a}_{a\_i}^T \mathbf{R}_i \mathbf{a}_{a\_i}} + \frac{\mathbf{a}_{a\_i}^T \mathbf{R}_j \mathbf{a}_{a\_i}}{\mathbf{a}_{a\_j}^T \mathbf{R}_j \mathbf{a}_{a\_j}} \right] \tag{3.1}$$

where

$$\mathbf{a}_{a\_i} = [1 \; a_{i1} \; \cdots \; a_{ip}]^T \tag{3.2}$$

is an augmented vector of one set of LPCs, $\mathbf{a}_{a\_j}$ is an augmented vector of another set of LPCs, $\mathbf{R}_i$ is the signal autocorrelation matrix:

$$\mathbf{R}_i = \begin{bmatrix} r_{ss}(0) & r_{ss}(1) & \cdots & r_{ss}(p) \\ r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(p-1) \\ \vdots & & \ddots & \vdots \\ r_{ss}(p) & & \cdots & r_{ss}(0) \end{bmatrix} \tag{3.3}$$

corresponding to the set $\mathbf{a}_{a\_i}$ and $\mathbf{R}_j$ is the signal autocorrelation matrix corresponding to the set $\mathbf{a}_{a\_j}$. The distances are computed between the spectra in the closed and open glottis regions in a glottal cycle of a natural voiced speech signal, for cases with and without bandwidth windowing. The distances are also computed between spectra for three different bandwidth windows. The three bandwidth windows (denoted as A,B,C in Table–3.2) differ in their sharpness of the taper outside the duration of the residual signal window, with sharpness increasing from A to C. The spectrum obtained in the closed glottis region without BW windowing is consistent with the spectra obtained using BW windowing, for the three BW windows, as seen by the closeness of the distances d_{13}, d_{35} and d_{57} to unity. The spectrum obtained in the open glottis region

Table 3.2: Comparison of different bandwidth windows for a natural voiced speech signal. The distances shown below are the symmetric Itakura distances for the cases mentioned in the table.

1. Without BW windowing in the closed phase (CP)
2. Without BW windowing in the open phase (OP)
3. With BW window A in the closed phase
4. With BW window A in the open phase
5. With BW window B in the closed phase
6. With BW window B in the open phase
7. With BW window C in the closed phase
8. With BW window C in the open phase

d_{ij} **is the symmetric Itakura distance between LP spectra for the cases** i **and** j**.**

$$d_{13} = 1.027 \quad d_{24} = 1.541 \quad d_{34} = 3.449$$
$$d_{35} = 1.012 \quad d_{46} = 1.010 \quad d_{56} = 3.935$$
$$d_{57} = 1.117 \quad d_{68} = 1.105 \quad d_{78} = 3.007$$

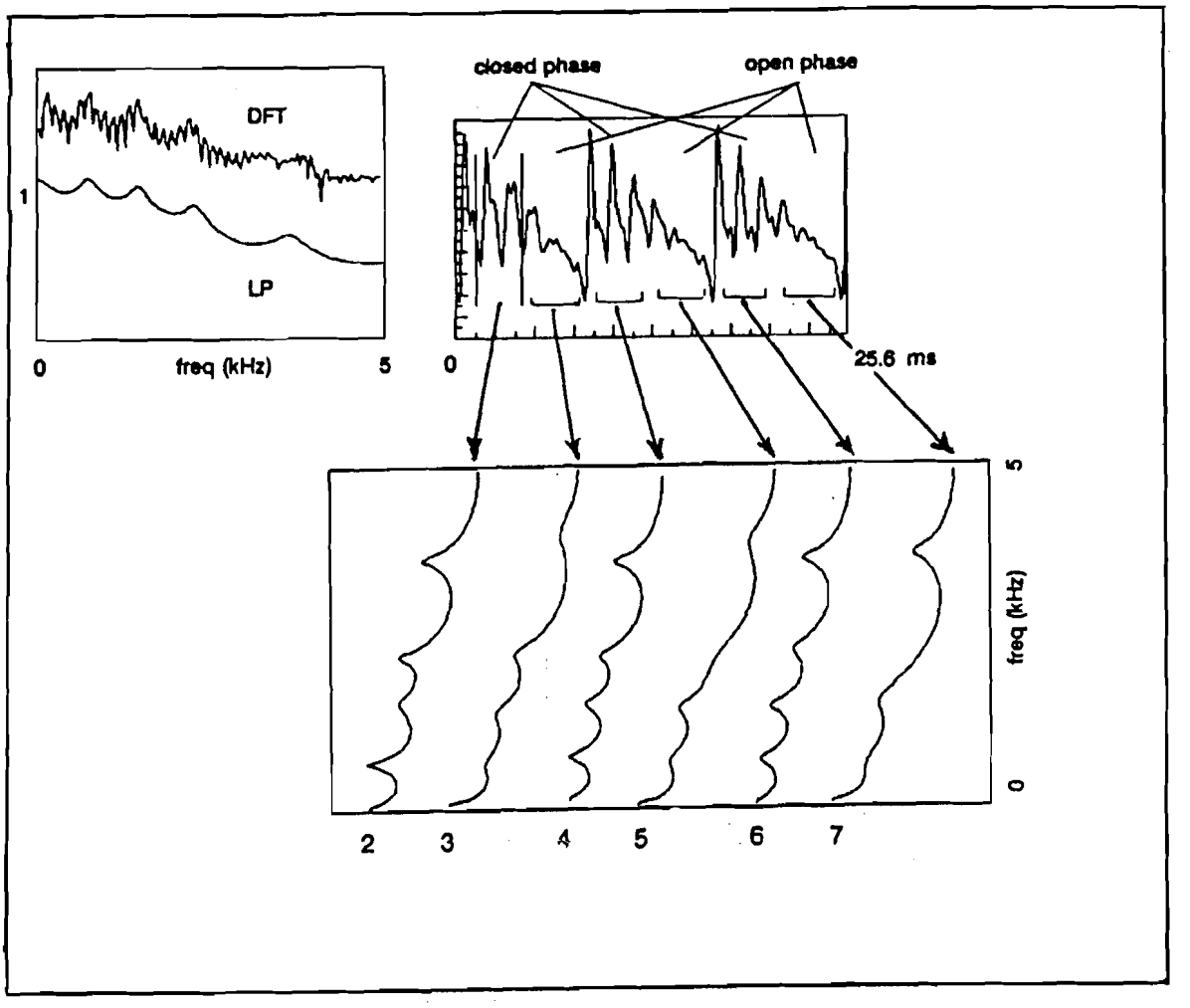
without BW windowing is dominated by the first formant (F_1) of the original LP spectrum and is different from the spectrum obtained using BW windowing for the same region. Hence, the distance d_{24} is large while the distances d_{46} and d_{68} are close to unity. The fact that both d_{46} and d_{68}, and similarly d_{35} and d_{57}, are all close to unity, even though the BW window is progressively sharpened from A to C, demonstrates that the sharpness with which the BW window is tapered is not very critical. The large distances d_{34}, d_{56} and d_{78} confirm that there is a significant change in the spectra for the closed and open glottis regions.

3.4 SHORT WINDOW ANALYSIS OF SPEECH SEGMENTS

In this section we consider analysis of several types of speech segments using source–system windowing. The speech signals analysed in this section were sampled at 10 kHz rate. In all the cases, we consider a source window of size 3 ms which includes a 0.5 ms taper on either side and a BW window of the shape shown in Fig. 3.4(g). Our

observation of the results of analysis are as follows: Analysis of the signal for nasal /m/ shows that the low, sharp first formant due to the nasal tract is not significantly influenced by the glottal opening and closure. For unvoiced speech, the sourcesystem windowing does not show significant differences when compared to the conventional LP spectrum. This is because the vocal tract system is generally steady during production of the unvoiced speech.

To demonstrate the consistency of the LP spectra derived using sourcesystem windowing, Fig. 3.5 shows the results of analysis of a natural voiced speech segment. The LP spectra are derived in the closed and open glottis regions for three successive glottal cycles. The closed phase region is identified as the 3 ms segment just after the instant of significant excitation in each glottal cycle. The open phase region is identified as the 3 ms segment just preceding the instant of significant excitation. The figure shows that the spectra in the closed and open phases of glottis are significantly different. The resonance peaks in the LP spectra for the closed glottis region are sharper than those in the original LP spectrum, and are broader in the open glottis region due to increase in the bandwidth of the resonance peaks. The Itakura distance computed between pairs of spectra are also presented in the figure. Note that among all the distances, large distances are obtained when computed between spectra in the closed and open glottis regions of each glottal cycle (d_{23}, d_{45} and d_{67}), showing that there is a significant change in the spectrum from the closed to the open glottis interval. We also note that the original LP spectrum (shown as (1) in the figure) also exhibits a large distance with the spectra in the open glottis region (d_{13}, d_{15} and d_{17}) but exhibits much smaller distances with the spectra in the closed glottis region (d_{12}, d_{14} and d_{16}). This is consistent with our expectation that the conventional LP spectrum is dominated by the vocal tract characteristics in the closed glottis interval.



SYMMETRIC ITAKURA DISTANCES (d_{ij}) BETWEEN TWO SEGMENTS i AND j :

d_{12} : 1.316	d_{13} : 2..252	d_{23} : 3.209	d_{24} : 1.082	d_{35} : 1.267
d_{14} : 1.170	d_{15} : 2.929	d_{45} : 3.125	d_{46} : 1.054	d57 : 1.954
d_{16} : 1.137	d_{17} : 2.690	d_{67} : 2.272		

Fig. 3.5: LP analysis using source-system windowing for open and closed glottis regions in each glottal cycle for three successive glottal cycles. (2), **(4)** and **(6)** are the LP spectra obtained in the closed glottis region of the three glottal cycles. (3), **(5)** and **(7)** are the corresponding spectra obtained in the open glottis region. The LP spectra for the closed glottis region show sharper resonance peaks compared to those for open glottis region.

The distance between the spectra in the closed glottis interval from one glottal cycle to the next are close to one (d_{24} and d_{46}) demonstrating the consistency of the spectral estimates obtained using the new windowing procedure. Though the distance (d_{35}) between the spectra in the open glottis region of the first and second glottal cycles is small, it is large for the second and third glottal cycles because of the strong higher formant in the open glottis region of the third glottal cycle (spectrum shown as 7).

Results of analysis of different natural vowels using the source–system windowing in the closed and open glottis regions of a glottal cycle are shown in Fig. 3.6. In the open glottis region, we observe a significant increase in the bandwidth of the first formant, and also increase in the value of the first formant (F_1) in some cases [2,11,12]. These observations are confirmed by the Itakura distances between the spectra in the open and closed regions. As in Fig. 3.4, some higher formants are lost in a few cases due to LP analysis of short data records.

For some speech segments (e.g., the vowel /u/ in high pitched female speech), it is generally difficult to identify the closed and open glottis regions, even approximately, from either the waveform or the LP residual signal. Analysis using source–system windowing brings out clearly the regions where the resonance peaks are sharp and the regions where they are damped. The differences in LP spectra enable us to identify, approximately, the closed and open glottis regions. Results of analysis of a segment of waveform of the vowel /u/ uttered by a female speaker are shown in Fig. 3.7, which clearly brings out the change in the value and bandwidth of F_1 from the closed to the open glottis region. Here again, the Itakura distances are given in the figure to indicate the spectral changes. The original LP spectrum (1) in the figure has a dominant resonance at the low frequency end since the pitch frequency and the first formant coincide (approximately). This dominant resonance is clearly not the actual

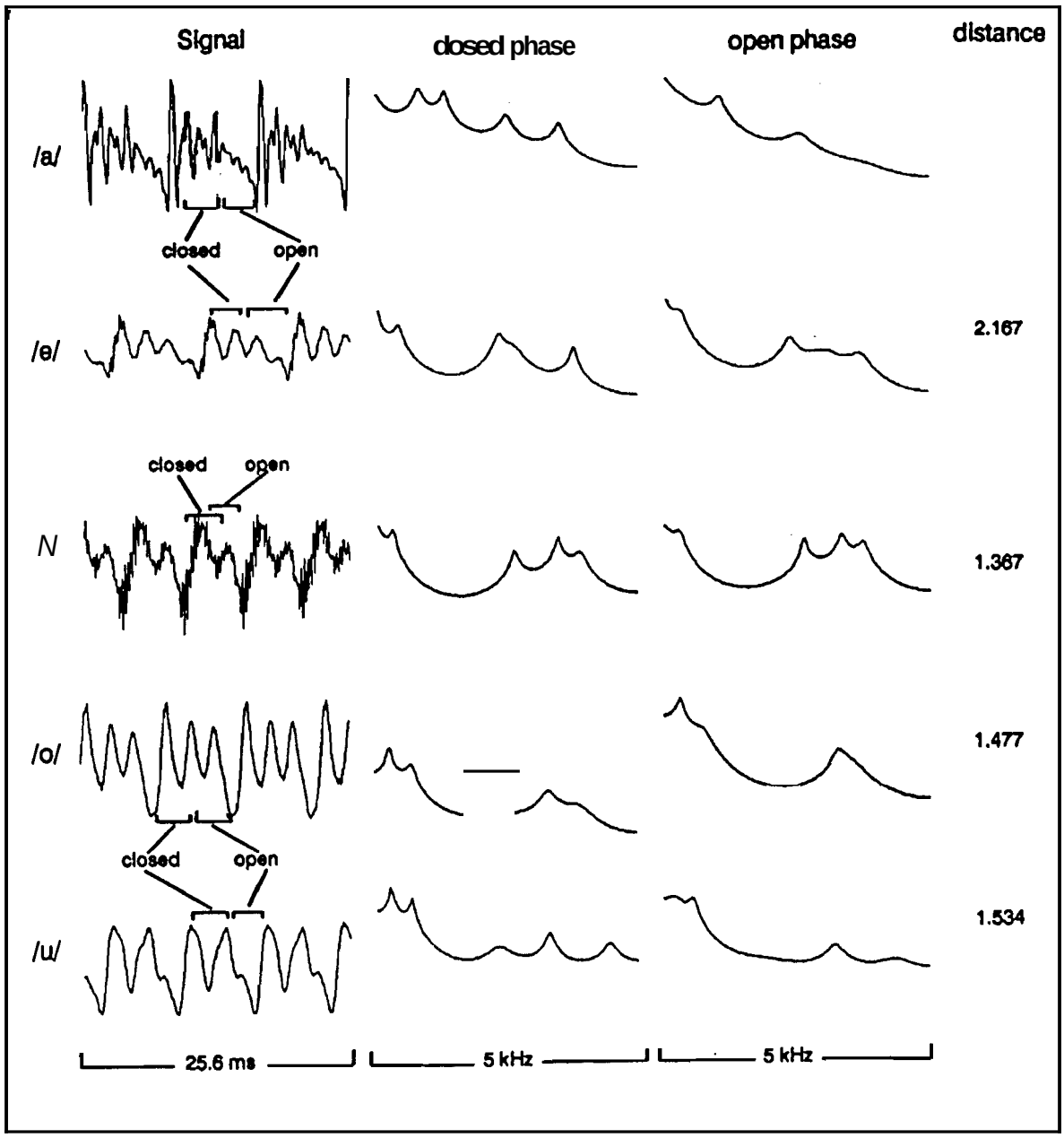
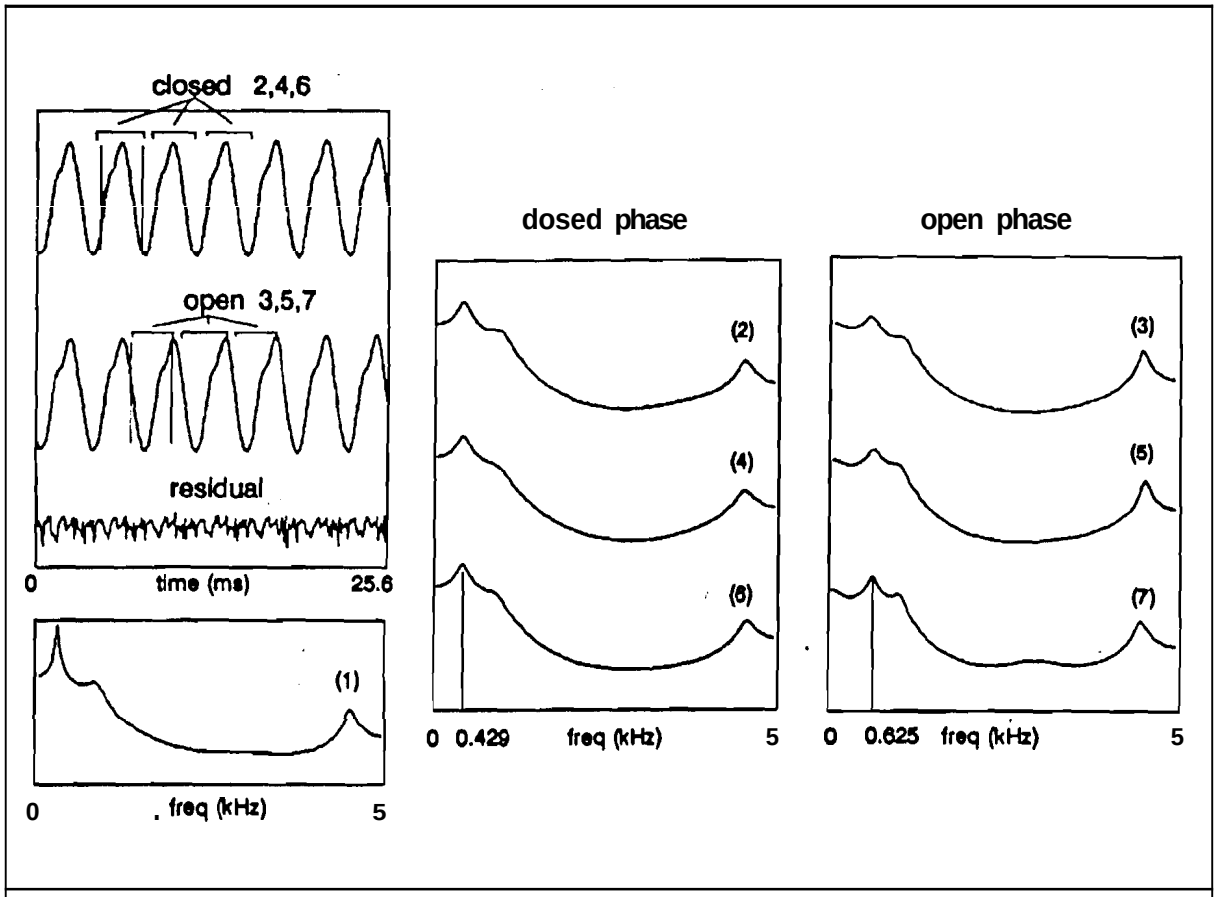


Fig. 3.6: Analysis of natural vowels using source-system windowing in the closed and open glottis regions of a glottal cycle. For each vowel, the signal waveform, LP spectra for closed and open phases, and the symmetric Itakura distance between these LP spectra are given along a row.



ITAKURA DISTANCE (d_{ij}) BETWEEN TWO SPECTRA LABELLED (i) AND (j) ABOVE :

d_{12} = 1.516	d_{13} = 2.734	d_{23} = 1.256	d_{24} = 1.043	d_{35} = 1.052
d_{14} = 1.554	d_{15} = 3.850	d_{45} = 1.448	d_{46} = 1.027	d_{57} = 1.239
d_{16} = 1.477	d_{17} = 4.726	d_{67} = 1.777		

Fig. 3.7: Analysis of high pitched female voiced speech using source–system windowing to identify approximately the closed and open glottis regions. The identified closed and open glottis regions are marked on the speech signal in the figure. The conventional LP spectrum obtained using a 25 ms window is shown as (1) and LP spectra obtained using source–system windowing are shown as 2, . . . , 7.

first formant (F_1). Hence, the conventional LP spectrum exhibits a large Itakura distance with the spectra derived using source-system windowing in the closed and open glottis regions (see distances d_{12}, d_{13}, d_{14}, d_{15}, d_{16} and d_{17}). The consistency of the estimates derived in the separate closed and open glottis regions in the three successive glottal cycles is reflected in the distances d_{24}, d_{46} and d_{35}, d_{57} respectively, which are close to one.

In CV transition regions, the characteristics of the vocal tract system exhibit rapid temporal and spectral changes. The spectra obtained using source-system windowing in one such transition for /ca/ are shown in Fig. 3.8. The LP spectra obtained in the consonant region (shown by 2,3, ...,10 in the figure) are different from those obtained in the vowel region (shown by 11 in the figure), while the conventional LP spectrum (shown by 1) exhibits the behaviour for the entire frame.

The results presented in this section show that the new method of windowing indeed helps in extracting the system characteristics in the open and closed glottis regions of voiced speech. The effects of these differences in the vocal tract system on the quality of synthetic speech is examined briefly in the next section.

3.5 EFFECT OF SOURCE-SYSTEM WINDOWING ON SYNTHESIS

Voiced parts of speech primarily dictate the quality of any synthetic speech [12], although the overall quality depends on both voiced and unvoiced speech. The quality of speech synthesised from LPC depends on modeling the excitation source for voiced speech. Modeling excitation source involves basically modeling the LP residual signal. One of the most popular models is the Liljencrants-Fant (LF) model [205] for the glottal pulse shape (see Appendix-A). Using a glottal pulse model for excitation and the same LPC system throughout the glottal cycle will not reflect the dynamics of the vocal tract system within a glottal cycle.

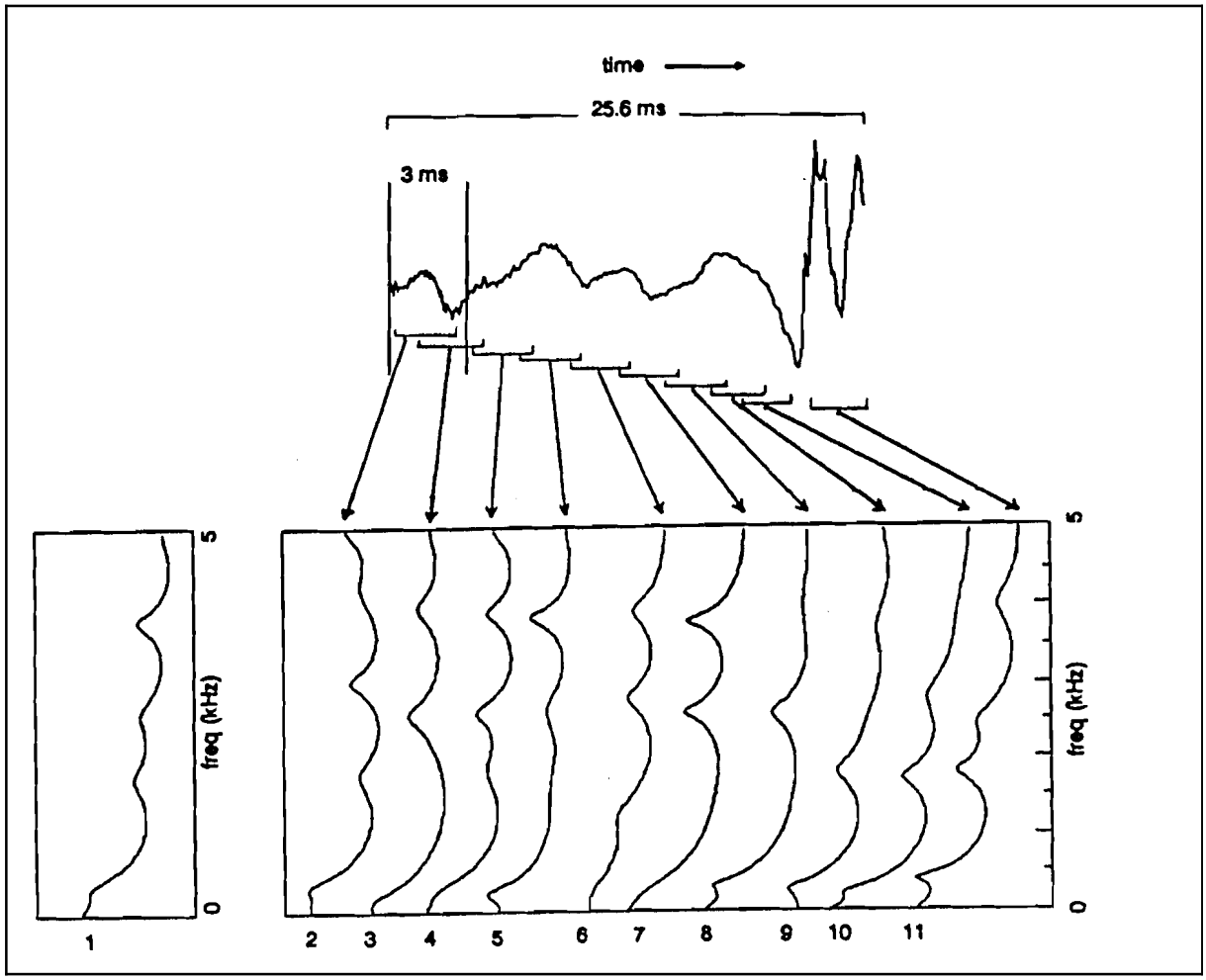


Fig. 3.8: Analysis of **a** CV transition region /ca/ using source–system windowing. The spectrum indicated as (1) is the conventional 10th order LP spectrum obtained using a **25** ms Hamming window. The LP spectra labeled $2, 3, \cdots, 11$ are obtained using source–system windowing.

In this section we compare (informally) the quality of synthetic speech for the two cases, one using the conventional LPCs and the other using separate LPCs for open and closed glottis regions. Improved naturalness was observed in the latter, since we were not only taking into account the variations of the vocal tract system within a glottal cycle but also the variations in the system from one glottal cycle to another. It is important to note that in natural speech, even two successive glottal cycles need not be alike, especially at the onset and end regions of a vowel [205]. During synthesis of speech, using the LPCs derived in the closed glottis interval and merely damping all the peaks of this all–pole system uniformly in the open glottis interval improved the naturalness significantly, compared to the conventional LPC synthesis [208]. The improvement in the naturalness of synthesised speech was obtained even though we have not incorporated the noticeable shift (increase) in F_1 observed in the open glottis interval in natural speech. This possibly indicates the significance of the bandwidth change in a glottal cycle of voiced speech in synthesising natural sounding speech.

3.6 SUMMARY

In this chapter we have shown that using suitable residual signal and bandwidth windows for source and system componerits of a speech signal, it is possible to derive the characteristics of the vocal tract system in the closed and open glottis regions within each glottal cycle. Thus source–system windowing overcomes, to some extent, the limitations of the short (10–30 ms) window analysis. This type of representation of the vocal tract system may help in generating natural sounding synthetic speech. However, the performance of this subsegmental analysis depends on the positioning of the window on the residual signal. If the system characteristics change significantly within the analysis window then it is difficult to interpret the results. In the subsegmental analysis, the positioning of the analysis window using the locations of instants

of significant excitation is examined in Chapters 7 and 8. In particular, the robustness of a group–delay function–based method for extraction of instants of significant excitation is studied. In the next chapter, we discuss a method based on the subsegmental analysis for speech enhancement.

Chapter 4

ENHANCEMENT OF NOISY SPEECH

In the previous chapter we have presented a method for subsegmental analysis of speech. In the chapters to follow, we present applications of the subsegmental analysis for enhancement of degraded speech. In this chapter, we present a method for enhancement of speech corrupted by additive random noise. The objective of the method is to selectively enhance the high SNR regions in the noisy speech in the temporal and spectral domains, without causing significant distortion in the resulting enhanced speech. This is proposed to be done at three different levels: (a) At the gross level, by identifying the regions of speech and noise in the temporal domain, (b) At the finer level, by identifying the regions of high and low SNR portions in the noisy speech, and (c) At the short–time spectrum level, by enhancing the spectral peaks over spectral valleys. The basis for the proposed method is to analyse the linear prediction (LP) residual signal in short (1–3 ms) segments to determine whether a segment belongs to a noise region or a speech region. The speech regions are emphasised relative to the noise regions to achieve enhancement.

In the next section the background to the problem is presented. In Section 4.2, we discuss the scope of study in this work. We also discuss the characteristics of noisy speech which form the basis for the proposed approach for speech enhancement. In Section 4.3, we develop a method for speech enhancement based on the characteristics of the LP residual signal. We propose enhancement at three levels, each level providing improvement of some feature of speech in the noisy signal. In Section 4.4, we discuss

application of the proposed method for different types of additive noise. We also discuss the performance and limitations of the proposed approach.

4.1 AN OVERVIEW OF SPEECH ENHANCEMENT METHODS

Speech signal collected under normal environmental conditions is usually degraded due to noise and distortion. Performance of speech systems depends critically on the effect of these environmental conditions on the parameters and features extracted from the speech signal [41, 209–214]. The quality of the recorded speech is also affected significantly due to noise and distortion. Enhancement of speech is normally required to reduce annoyance due to noise. The focus of study in this chapter is speech enhancement in additive noise.

Several approaches were studied for speech enhancement in additive noise [49, 50, 142, 215–228]. Many of these studies have focussed on enhancement based on attempts to suppress noise [49, 50, 142]. In order to suppress noise the characteristics of noise are estimated from the regions containing predominantly noise. Therefore for suppressing noise it is necessary to identify the noise regions. Subtraction of noise from noisy speech is usually performed in the spectral domain. Methods based on spectral subtraction disturb the spectral balance in speech, resulting in unpleasant distortion in the enhanced speech. Speech enhancement has also been accomplished by modifying the temporal contours of the parameters or features, like spectral band energies [26, 90]. The technique uses data–dependent filters that reduce the random fluctuations in the parameter contours caused by noise, and thus enhances the characteristics of speech. The parameters of speech are usually related to short–time spectra. Therefore modification of the temporal variations of the spectral features may sometimes introduce unnatural spectral changes which are perceived as distortion in the enhanced speech.

Methods for speech enhancement have also been developed based on extraction of

parameters from noisy speech, and synthesizing speech from these parameters [229–233]. All–pole modeling of degraded speech is one such method [234]. In the all–pole modeling, if wrong peaks are extracted, then these peaks may get enhanced. Temporal sequence of these peaks also produces discontinuities in the contours of the spectral peaks when compared with the smooth contours encountered in natural speech.

Methods of speech enhancement seem to depend generally on modification of the short–time spectral envelope. If there are errors in extracting the features of a spectral envelope, or if errors are introduced in the spectral envelope due to modification of the temporal contours of the spectral features, the resulting speech may produce unnatural audible distortion.

Several methods focussing on characteristics of speech have been proposed for enhancement of degraded speech [162, 163, 165, 217, 235–238]. Some of these methods are based on exploiting the pitch periodicity and high signal energy characteristics in **10–30** ms segments of speech [162, 163, 165, 218, 235, 238–240]. Noise samples in successive glottal cycles are uncorrelated. On the other hand, the characteristics of the vocal tract system are highly correlated due to slow movement of the articulators. These methods for enhancement of speech depend critically on the estimation of pitch from the noisy speech signal. Also, synthetic excitation signal is used for producing speech in the methods based on synthesis. Hence the quality of speech will be poor, even though the effects of noise are reduced.

Several suprasegmental parameters such as pitch contours and syllabic durations are robust features. But these features are not useful for enhancement, since for generating the enhanced speech signal one needs both the spectral envelope and excitation for each (short–time) analysis frame.

In many of the above mentioned methods, no attempt has been made to explore

the characteristics of the source signal for enhancement. The primary reason for this is that, in the source signal, such as the linear prediction residual signal, the samples are uncorrelated and hence the residual samples are more like noise than like a signal. Thus the residual signal is not expected to have any features useful for speech enhancement. We show in this work that features of the residual error signal can be exploited for enhancement of speech in the presence of additive noise.

4.2 BASIS FOR THE PROPOSED METHOD OF SPEECH ENHANCEMENT

Human beings perceive speech by capturing some features from the high signal–to–noise ratio (SNR) regions in the spectral and temporal domains, and then extrapolating the features in the low SNR regions [204]. Therefore speech enhancement should primarily aim at highlighting the high SNR regions relative to the low SNR regions. Lowering the signal levels in the low SNR regions relative to the signal levels in the high SNR regions may help in reducing the annoyance due to noise without losing the information. The relative emphasis of the features in the high SNR regions over the features in the low SNR regions should be accomplished without causing distortion in speech. Otherwise the enhancement may cause annoyance of a type different from that due to additive noise. The objective of this work is to study the enhancement produced due to modification of the characteristics of the source and system components of speech production in the signal.

4.2.1 Effects of Noise on the Speech Signal

Before we proceed to discuss our approach, we briefly review some characteristics of noisy speech. Speech signal has a large (30–60 dB) dynamic range in the temporal and spectral domains. For example, in the temporal domain some sounds have low signal energy, especially during the release of stop sounds and in the steady nasal sounds.

Speech signal energy level is also low prior to the release of a stop sound and also in some fricative sounds. Even within a glottal cycle of a voiced speech signal the energy of the signal is usually higher only in the vicinity of the major excitation of the vocal tract system, which is the instant of glottal closure in each glottal cycle [1]. This is due to damped sinusoidal nature of the impulse response of the vocal tract system. Even in the frequency domain the spectral levels of large amplitude formants are typically much higher (20–30 dB) than the low amplitude formants. The spectral envelope also decreases by 12–18 dB per octave due to glottal roll-off [11]. For a given additive noise, the SNR varies as a function of frequency in the spectral domain. Thus the SNR is different in different segments of speech in both time and frequency domains. Fig. 4.1(c) shows the SNR of a speech utterance as a function of time, where the overall SNR is 10 dB. The noisy speech signal (Fig. 4.1(b)) is generated by adding white Gaussian noise to the clean speech signal shown in Fig. 4.1(a). The SNR is computed for each frame of duration 20 ms with an overlap of 10 ms.

Typically, the correlation between noise samples is low, and speech samples are correlated. Therefore, the envelope of the speech spectrum will be less flat due to formant structure and glottal roll–off compared to the noise spectrum. Additive noise increases the spectral flatness of speech. The spectral envelope becomes more flat in the low SNR portions of the spectrum. As the noise level increases, the weaker spectral features and the low energy signal features will be progressively submerged in the noise. The proposal in this work is to identify the high SNR portions in the noisy speech signal, and enhance those portions relative to the low SNR portions, without causing significant distortion in the enhanced speech. Note that, from human perception point of view, some background noise is tolerable, but not the distortion caused by the artifacts of processing.

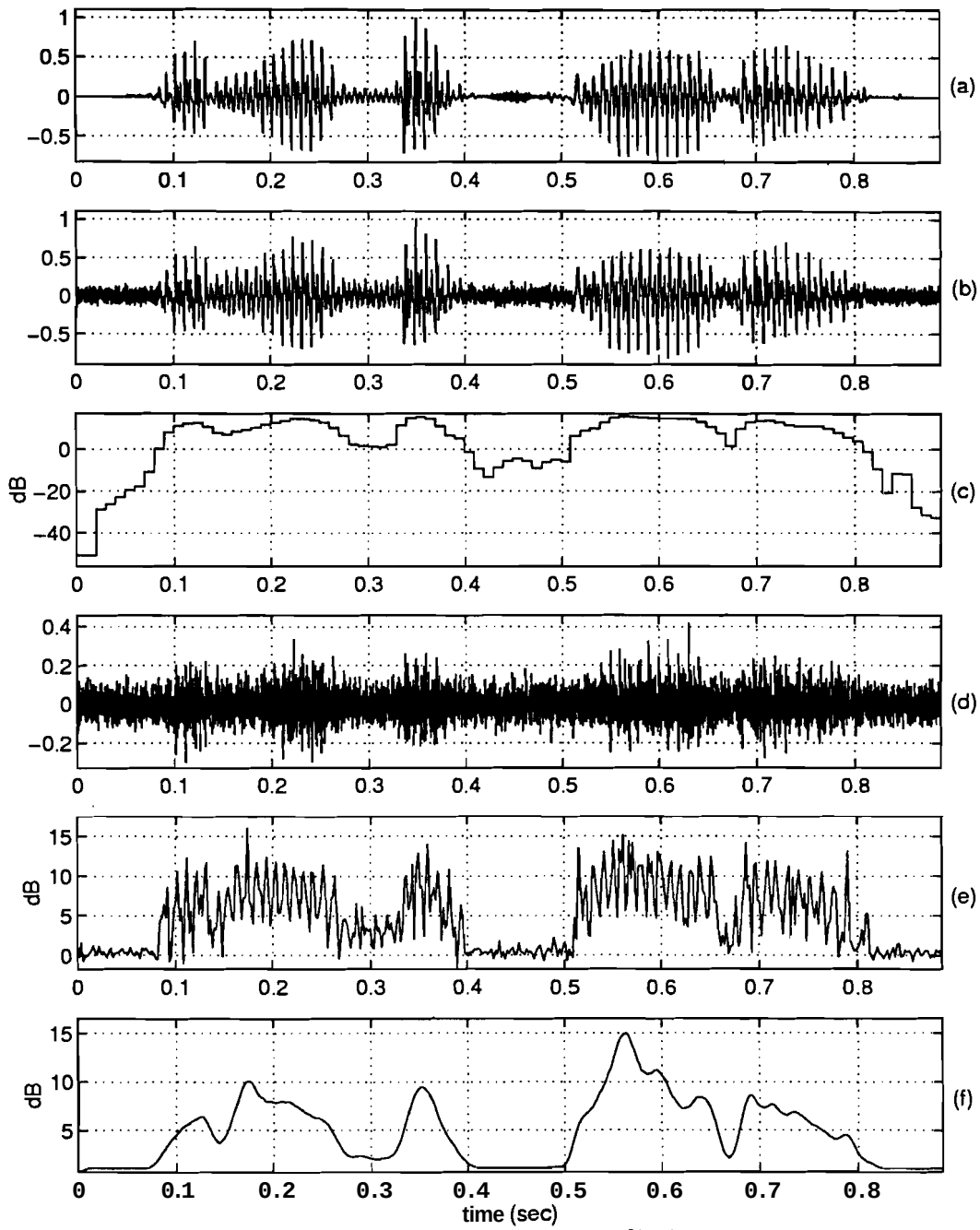


Fig. 4.1: (a) Speech signal for the utterance *"any dictionary"*. (b) Signal with an average SNR of 10 dB. (c) The SNR as a function of time. (d) The 12th order LP residual signal derived from the noisy signal in (b). (e) The ratio of energy values between (d) and (b) for 10 dB SNR case for each **2** ms frame. (f) The ratio curve in (e) smoothed using a 17–point Hamming window.

4.2.2 Approach for Speech Enhancement

In this section we present the proposed approach for speech enhancement. We attempt to emphasize the residual signal in the regions around the glottal closure in the voiced speech segments and reduce the energy levels of the residual signal in the silence regions. By exciting the time–varying all–pole filter (derived from the noisy speech) with the modified residual signal, one can produce enhanced speech without causing significant distortion.

Let y $= [y_n, y_{n+1}, \ldots, y_{n+N-1}]^T$ be a frame of N samples of the signal corrupted by additive random noise. The characteristics of the signal are assumed to be stationary within the frame. We can write y as

$$\mathbf{y} = \mathbf{s} + \mathbf{w}, \tag{4.1}$$

where

$$\mathbf{s} = [s_n, s_{n+1}, \cdots, s_{n+N-1}]^T$$

is the vector of clean signal samples and

$$\mathbf{w} = [w_n, w_{n+1}, \cdots, w_{n+N-1}]^T$$

is the vector of noise samples. Let **e** be the vector of residual error samples derived by inverse filtering the noisy signal y using a pth order linear prediction (LP) analysis. The linear prediction coefficients (LPCs) are denoted by $a_0, a_1, a_2, \ldots,$ a, with $a_0 = 1$. Assuming the initial conditions to be zero, the residual signal vector **e** may be expressed in matrix form as

$$\mathbf{e} = \mathbf{A}\mathbf{y} \tag{4.2}$$

where

$$\mathbf{A} = \begin{bmatrix} a_0 & 0 & \cdots & & 0 \\ a_1 & a_0 & \cdots & & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ a_p & a_{p-1} & \cdots & a_0 \cdots & 0 \\ \vdots & \ddots & & \ddots & \vdots \\ 0 & \cdots & a_p & \cdots a_1 & a_0 \end{bmatrix} \tag{4.3}$$

An estimate of the clean signal can be obtained by weighting the derived residual error samples appropriately and exciting the LP all–pole filter. The weighted residual error vector $\mathbf{e}_w$ can be expressed as

$$\mathbf{e}_w = \mathbf{\Gamma e} \tag{4.4}$$

where $\mathbf{\Gamma} = \text{diag}\,[\gamma(0), \gamma(1), \ldots, \gamma(\mathrm{N}-1)]$ is the diagonal N x N matrix of optimal weights to be estimated. An estimate of the clean signal is given by

$$\hat{\mathbf{s}} = \mathbf{He}, \tag{4.5}$$

where

$$\mathbf{H} = \mathbf{A}^{-1} \tag{4.6}$$

is the matrix of coefficients of truncated impulse response of the all–pole filter. The truncation effects are assumed to be negligible. Using **(4.2)** and **(4.4)**

$$\begin{aligned} \hat{\mathbf{s}} &= \mathbf{H\Gamma Ay} \\ &= \mathbf{H\Gamma As} + \mathbf{H\Gamma Aw} \end{aligned} \tag{4.7}$$

The error in reconstruction is given by

$$\begin{aligned} \epsilon &= \mathbf{s} - \hat{\mathbf{s}} \\ &= (\mathbf{I}_N - \mathbf{H\Gamma A})\mathbf{s} - \mathbf{H\Gamma Aw} \end{aligned}$$

Using (4.6) in the above equation we find

$$\boldsymbol{\epsilon} = \mathbf{H}(\mathbf{I}_N - \boldsymbol{\Gamma})\mathbf{A}\mathbf{s} - \mathbf{H}\boldsymbol{\Gamma}\mathbf{A}\mathbf{w}. \tag{4.8}$$

The energy of the reconstruction error $\boldsymbol{\epsilon}$ can be minimized with respect to the weight matrix $\boldsymbol{\Gamma}$. But this error criterion does not exploit the masking properties of the human ear [238, 241, 242]. Hence, a criterion which would be more meaningful perceptually would be the energy of filtered reconstruction error $\boldsymbol{\epsilon}_p$. The filter can be the inverse filter $A(z) = a_0 + a_1 z^{-1} + \ldots + a_p z^{-p}$ of the LP analysis. For the signal–to–noise ratio usually encountered in practice (> 10 dB) it is reasonable to assume that the inverse filter $A(z)$ exhibits valleys at approximately the formant frequencies, although its dynamic range would be low because of noise in the speech signal. Minimization of the energy of the filtered error with respect to $\boldsymbol{\Gamma}$ would allow more error in the formant regions and minimizes the error in the valley regions, which is desirable from a perceptual viewpoint. From (4.8) above, the filtered error $\boldsymbol{\epsilon}_p$ can be written as

$$\begin{aligned} \boldsymbol{\epsilon}_p &= \mathbf{A}\boldsymbol{\epsilon} \\ &= \mathbf{A}\mathbf{H}(\mathbf{I}_N - \boldsymbol{\Gamma})\mathbf{A}\mathbf{s} - \mathbf{A}\mathbf{H}\boldsymbol{\Gamma}\mathbf{A}\mathbf{w}. \end{aligned} \tag{4.9}$$

Using (4.6) in (4.9) we obtain

$$\boldsymbol{\epsilon}_p = (\mathbf{I}_N - \boldsymbol{\Gamma})\mathbf{A}\mathbf{s} - \boldsymbol{\Gamma}\mathbf{A}\mathbf{w}. \tag{4.10}$$

Let $\mathbf{e}_s = \mathbf{A}\mathbf{s}$ be the signal obtained by filtering the clean signal s using the filter $A(z)$ derived from the noisy signal y, and let $\mathrm{v} = \mathbf{A}\mathbf{w}$ be the filtered noise in the residual signal domain, then

$$\boldsymbol{\epsilon}_p = (\mathbf{I}_N - \boldsymbol{\Gamma})\mathbf{e}_s - \boldsymbol{\Gamma}\mathbf{v}. \tag{4.11}$$

Assuming that the signal $\mathbf{s}$ and noise $\mathbf{w}$ are uncorrelated, the cost function [243–245]

$$\psi(\boldsymbol{\Gamma}) = \mathcal{E}\{\|\boldsymbol{\epsilon}_p\|_2^2\} \tag{4.12}$$

is minimized to obtain the optimum weights as

$$\gamma_o(k) = \frac{\mathcal{E}\{e_s^2(k)\}}{\mathcal{E}\{e_s^2(k)\} + \mathcal{E}\{v^2(k)\}}, \qquad k = 0, 1, ..., (N-1), \tag{4.13}$$

where $e_s(k)$ and $v(k)$ are the kth components of e, and v, respectively. If we define the following ratio as an approximate measure of SNR in the residual signal domain,

$$SNR(k) = \frac{\mathcal{E}\{e_s^2(k)\}}{\mathcal{E}\{v^2(k)\}} \tag{4.14}$$

then we have

$$\gamma_o(k) = \frac{SNR(k)}{1 + SNR(k)} \tag{4.15}$$

The solution in (4.15) is clearly a time domain analogue of the optimal Wiener filter frequency response [246]. Note that in arriving at the result in (4.15), no restriction is placed on the noise samples in the vector w. The noise samples are only assumed to be uncorrelated with the signal samples in the vector s. Since it is difficult to estimate $SNR(k)$ in practice, $\gamma_o(k)$ can only be approximated as discussed in Section 4.3. Note that the optimal weight $\gamma_o(k)$ in (4.15) approaches one in the limit when $SNR(k) \gg 1$ and approaches $SNR(k)$ itself, when $SNR(k) \ll 1$. But in our method (presented in Section 4.3) the weight function used is not exactly the same as the optimal weight. Firstly, it is difficult to estimate the $SNR(k)$ in practice. Secondly, allowing the weight to assume very low values when the $SNR(k)$ is poor produces distortion in the processed speech. Hence, it is necessary to restrict the minimum value of the weight. Assuming that the noise variance in the residual signal domain is approximately constant, $SNR(k)$ is proportional to the short–time energy of the residual signal. Hence, the short–time energy values of the residual signal are used to derive the weight function at the finer (1–3 ms) level.

4.2.3 Nature of LP Residual Signal

An experiment was conducted to demonstrate the effect of processing the LP residual signal of speech and reconstructing the speech using only a part of the residual signal after the instant of glottal closure. From the clean speech, the voiced/unvoiced/silence segments and the instants of significant excitation were identified [202, 247]. The LP residual signal of noisy speech was modified retaining only the 2 ms portions of the residual signal around the instants of excitation. The modified residual signal was used to excite the time–varying all–pole filter to regenerate the speech signal. The resulting speech was significantly enhanced without causing serious distortion. This is because the high SNR segments of noisy speech were retained in the reconstructed speech. Note that the all–pole filter derived from the noisy speech may not represent the spectral features of the clean speech accurately. The coefficients of the filter were used mainly to derive the noisy residual signal by inverse filtering. Retaining the waveform bells around the glottal closure produces good quality speech as was demonstrated in PSOLA based Text–to–Speech system (TTS) [248, 249].

The LP residual signal (Fig. 4.1(d)) may be derived for the noisy speech using a frame of 20 ms duration and a frame rate of about 100 frames per second. Even in the LP residual signal of noisy speech, the SNR is a function of time or frequency. Inverse filtering reduces the correlation between samples existing in the noisy speech signal. Since the residual signal samples are less correlated, the SNR as a function of time can be studied using much smaller windows (1–3 ms) than the windows (10–30 ms) normally used in the short–time spectral analysis. The truncation effects of the analysis window are significantly reduced in the residual signal [250, 251]. For each small window of the residual signal, the energy ratio of the noisy speech signal and the corresponding portion of the residual signal gives an indication of the amount of

reduction in the correlation of the signal samples. This also gives an indication of how much the signal spectrum is flattened in the residual signal. If the signal spectrum is already flat, then the ratio of the energies of the noisy signal and the residual signal in the short (1–3 ms) window will be nearly unity. Otherwise, the ratio will be quite large. Note that for noise–like segments this ratio of the energies will be nearly unity. Thus the ratio of the energies gives an indication of the speech signal and noise regions of the signal. The ratio of the energy values for a 10 dB SNR situation computed for each 2 ms frame is shown in Fig. 4.1(e). Note that even weak signal regions are discernible in the ratio plots. The ratio can be interpreted as the inverse of spectral flatness of the noisy signal, the minimum inverse flatness being one, corresponding to the energy ratio of 0 dB.

Since the correlation between the residual signal samples is low, these samples can be manipulated to some extent without producing significant distortion in the reconstructed speech [252]. It is this manipulative capability of the residual signal we would like to exploit for enhancement of speech.

4.3 MANIPULATION OF LP RESIDUAL SIGNAL

The basic principle of our approach for speech enhancement is to identify the low SNR regions in the LP residual signal, and derive a weight function for the residual signal which will reduce the energy in the low SNR regions relative to the high SNR regions of the noisy signal. The residual signal samples are multiplied with the weight function. The modified residual signal is used to excite the time–varying all–pole filter to generate the enhanced speech. Speech enhancement is carried out at three levels: (a) at gross level, based on the overall smoothed inverse spectral flatness characteristics, (b) at finer level (1–3 ms), based on the relative energies of the residual signal between adjacent frames, and (c) at spectral level, to enhance the features in the spectrum that

could not be affected by the fine level operations.

4.3.1 Gross Temporal Level

At the gross level the regions corresponding to low and high **SNR** regions are identified from the characteristics of the LP residual signal. A weight function for the residual signal samples is derived based on the smoothed inverse spectral flatness characteristics of the noisy speech signal. The spectral flatness characteristics are derived by comparing the energy in the residual signal with the energy in the noisy speech signal in each short interval of about 2 ms.

Inverse filtering the noisy speech signal using the time–varying LP coefficients will give the residual signal. The LP residual signal for the noisy speech data is shown in Fig. 4.1(d). The ratio of the noisy speech signal energy to the residual signal energy in dB for each nonoverlapping frame of 2 ms is shown in Fig. 4.1(e). The ratio plot gives an indication of the inverse spectral flatness as a function of time. The inverse spectral flatness plot is smoothed using a 17–point Hamming window. The smoothed inverse flatness plot shown in Fig. 4.1(f) clearly indicates the low and high SNR regions. The low SNR (noisy) regions have an inverse flatness close to unity (0 dB), and the high SNR (signal) regions have larger inverse flatness values. Note that for noise–like segments the inverse flatness will be close to unity. Unvoiced segments can be distinguished from noisy segments by the higher residual signal energy value for the unvoiced region compared to the energy value in the (noisy) silence region (see Fig. 4.1(c)). A weight function is derived from the smoothed inverse flatness characteristics in such a way that the residual signal samples in the regions corresponding to low values of the inverse flatness are reduced relative to the residual signal samples in the regions corresponding to high values of the inverse flatness.

A mapping function of the type shown in Fig. 4.2 can be used to map the smoothed

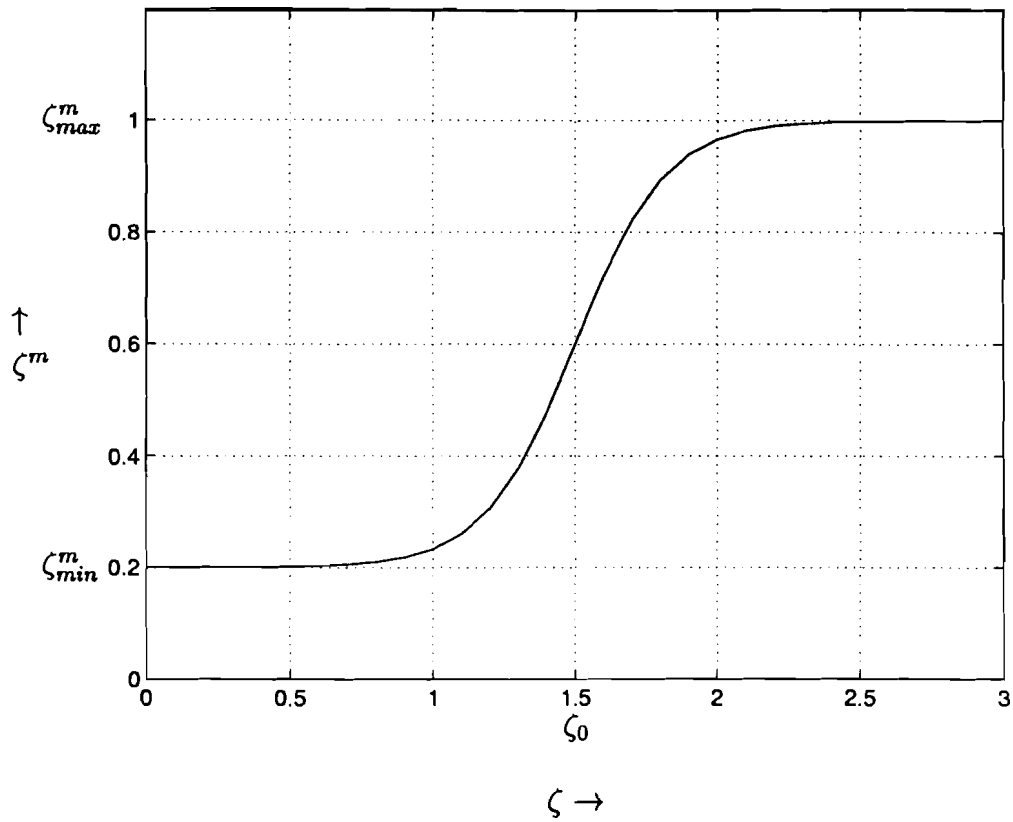


Fig. 4.2: Mapping function to generate the mapped energy ratio values (ζ^m) from the energy ratio values (ζ). The mapping function $\zeta^m = \left(\frac{\zeta^m_{max} - \zeta^m_{min}}{2}\right) \tanh(a, \pi\ (\zeta - \zeta_0)) + \left(\frac{\zeta^m_{max} + \zeta^m_{min}}{2}\right)$ is shown for a, $= 0.75$ and $\zeta_0 = 1.50$.

inverse spectral flatness values to the weight values for each short (2 ms) frame of residual signal. The mapping function is of the type $tanh(x)$. The purpose of the nonlinear mapping function is to enhance the contrast between the value of the inverse spectral flatness in the speech signal regions and its value in the background noise regions. The weight values for each frame are further smoothed using a 2 ms window to compute the running average across time. Thus we can generate a weight value for each sample of the residual signal as shown in Fig. 4.3(a). The residual signal samples are multiplied with this weight function to generate a modified residual signal.

The noisy and the enhanced signals along with their spectrograms are shown

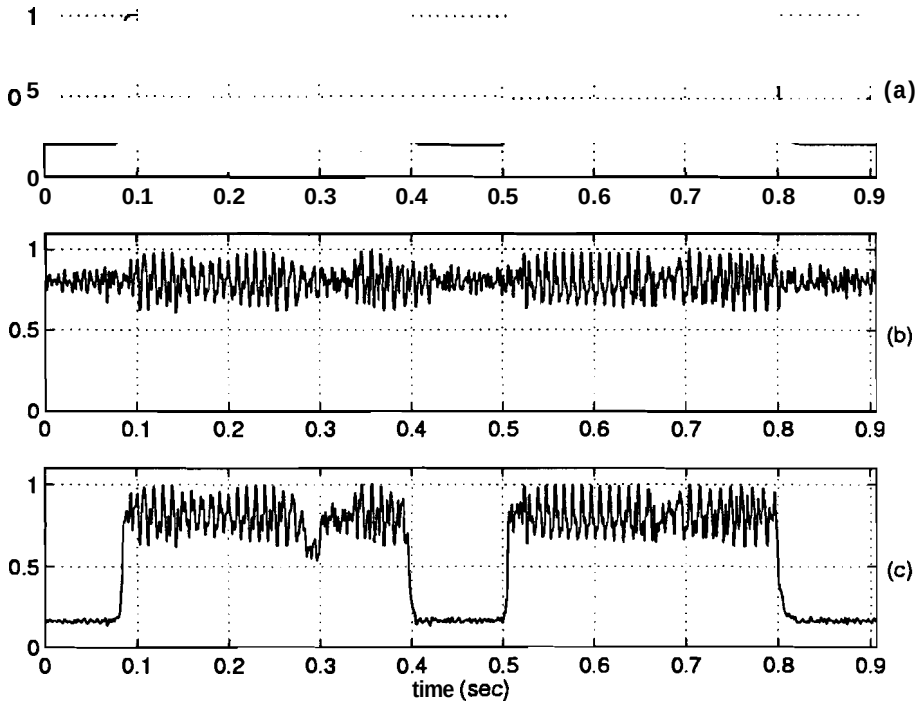


Fig. 4.3: Weight functions for the LP residual signal. (a) Gross weight function. (b) Fine weight function. (c) Final weight function.

in Fig. 4.4. The figure shows the reduction in the energy in the noisy segments relative to the speech segments. On listening, we notice a significant reduction in the annoyance due to the background noise. However, due to sudden change from low noise to the noisy speech regions, the change can be perceived in the enhanced speech. It is possible to trade between the annoyance and speech quality by adjusting the thresholds in the mapping function shown in Fig. 4.2. The more the reduction in the noise level in the low SNR regions relative to the noise in the high SNR regions, the better will be the speech quality. But then there will be more annoyance due to sudden rise in the background noise. Further improvement can be obtained by manipulating the residual signal at the finer level as discussed in the next subsection.

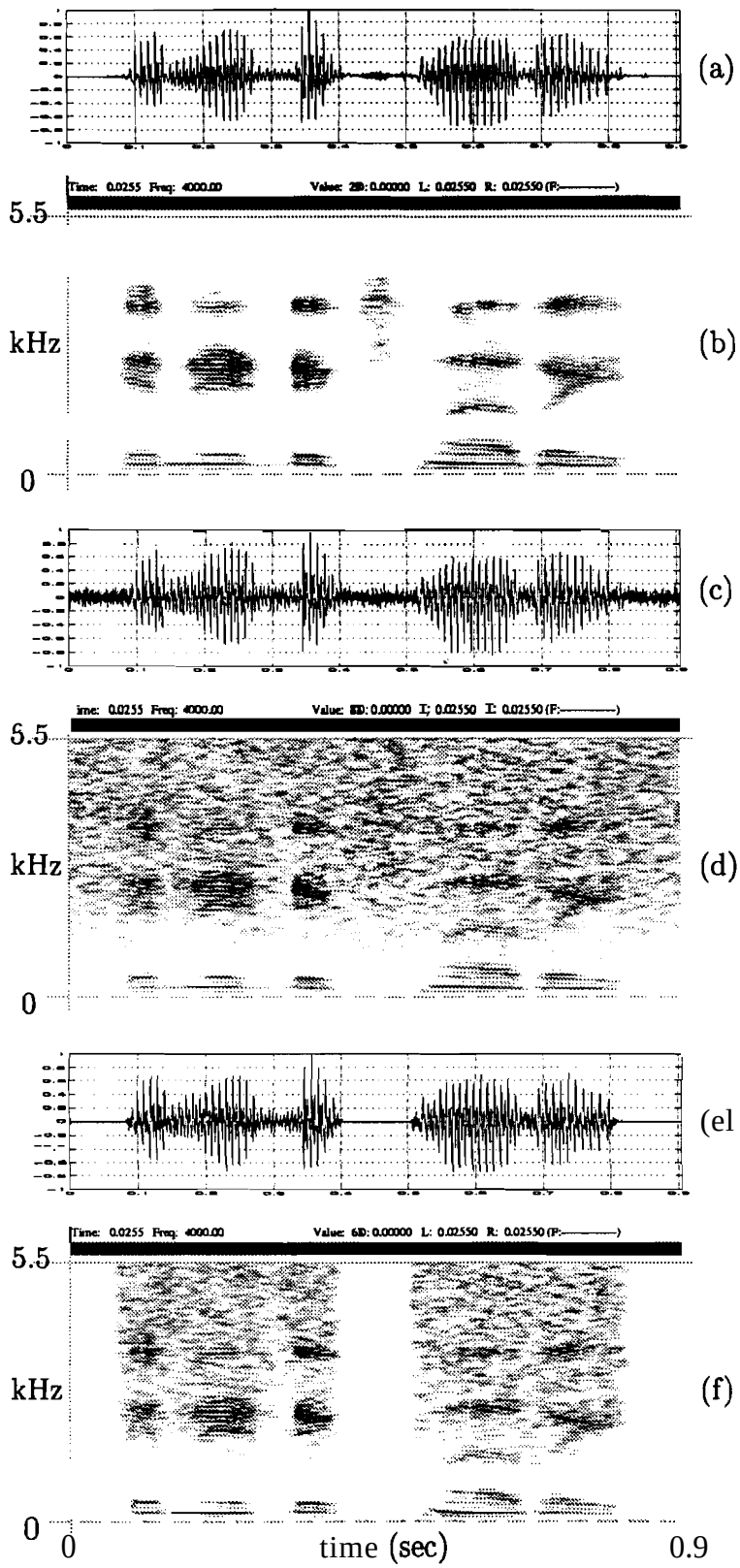


Fig. 4.4: Results of enhancement of speech degraded by additive white noise. (a) Clean speech. (c) Speech signal at 10 dB SNR. (e) Enhanced speech signal obtained using gross level weighting. (b),(d),(f) – spectrograms for the signals in (a),(c),(e), respectively.

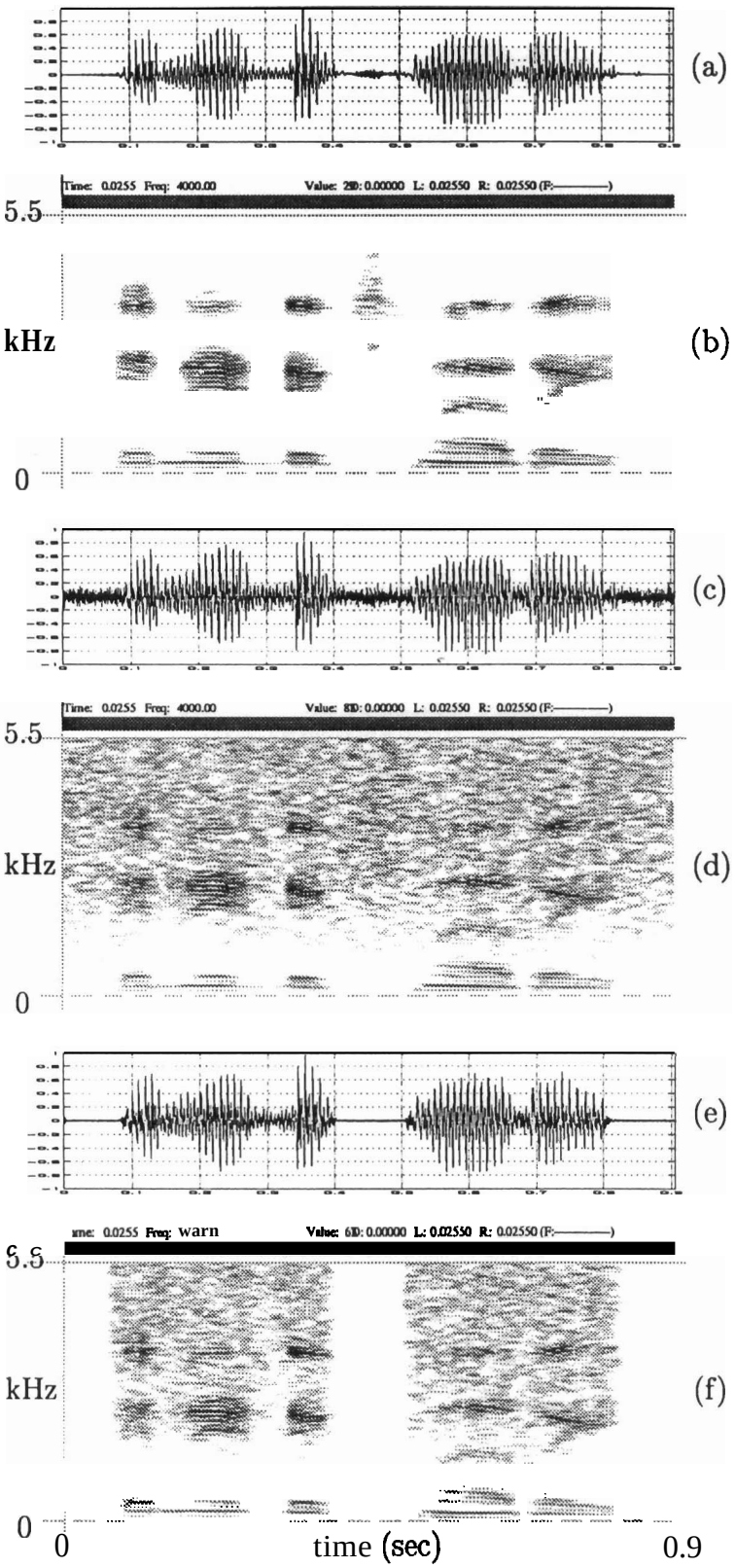


Fig. 4.4: Results of enhancement of speech degraded by additive white noise. (a) Clean speech. (c) Speech signal at 10 dB SNR. (e) Enhanced speech signal obtained using gross level weighting. (b),(d),(f) – spectrograms for the signals in (a),(c),(e), respectively.

4.3.2 Finer Temporal Level

F'rom the spectrogram in Fig. 4.4(f) we notice that the noise in the enhanced speech regions is distributed uniformly across frequency in the spectrum. This causes annoyance due to abrupt change from low noise to high noise regions in the time domain. Also the speech formant features are masked due to noise filling up the low amplitude portions in the frequency domain. Further enhancement at finer levels in the speech segments, especially in the voiced regions, may improve the quality and reduce the annoyance.

For voiced segments, if the SNR is low in some short (1–3 ms) segments, then the residual signal in those regions can be given lower weightage compared to the adjacent higher SNR segments. This is likely to happen for the regions corresponding to the open glottis portion in each glottal cycle due to damping of the formants. The fluctuations in the residual signal energy contour for short (2 ms) segments illustrate the energy differences between adjacent segments. A weight function at the fine level can be derived from the residual signal energy plot to deemphasize the segments corresponding to the valleys relative to the segments corresponding to the peaks. However for noisy speech, the residual signal is noisy and so the energy of the short segment of the residual signal may not be reliable for deriving the weight. Hence, the F'robenius norm [253] of the Toeplitz prediction matrix (see (4.16) below) constructed using the noisy speech samples in a frame of 2 ms duration is used to represent the short–time energy of the corresponding frame of LP residual signal (see Appendix–B). This approach has the advantage of exploiting the envelope information in the noisy speech

waveform. The Toeplitz prediction matrix Y is given by

$$\mathbf{Y} = \begin{bmatrix} y_{p+1} & y_p & \cdots & y_1 \\ y_{p+2} & y_{p+1} & \cdots & y_2 \\ & & \ddots & \\ \vdots & \vdots & & y_{p+1} \\ & & & \vdots \\ y_M & y_{M-1} & \cdots & y_{M-p} \end{bmatrix} \tag{4.16}$$

where $y_1, y_2, \ldots, y_M$ are the noisy speech samples in a frame of length M samples, which is 16 for 2 ms duration at 8 kHz sampling. The linear prediction order p is taken as 10. The Frobenius norm is computed for every sample. The weight function is derived using the logarithm of the ratio of the Frobenius norm of the present frame to the Frobenius norm of the frame 2 ms prior to the present frame. A mapping function of the type shown in Fig. 4.2 is used to map the log ratio values to the weight values for each sample of the signal. The objective of the mapping function is to control the relative emphasis of high SNR segments over low SNR segments in short (2 ms) intervals. The maximum change is restricted to the interval 0.2 to 1.0. The finer weight function is shown in Fig. 4.3(b). The overall weight function is obtained by multiplying the gross weight function derived from the smoothed inverse flatness plot with the fine weight function. The final weight function for the residual signal samples is shown in Fig. 4.3(c). Enhanced speech is generated by exciting the time–varying all–pole filter with this weighted residual signal. Spectrograms of the enhanced speech along with the spectrograms for noisy speech and the speech enhanced using only gross level weighting of the residual signal are shown in Fig. 4.5. From the spectrogram in Fig. 4.5(c) we observe that the spectrum of the signal is significantly enhanced in the voiced regions. The quality of speech is significantly better than in the case of the

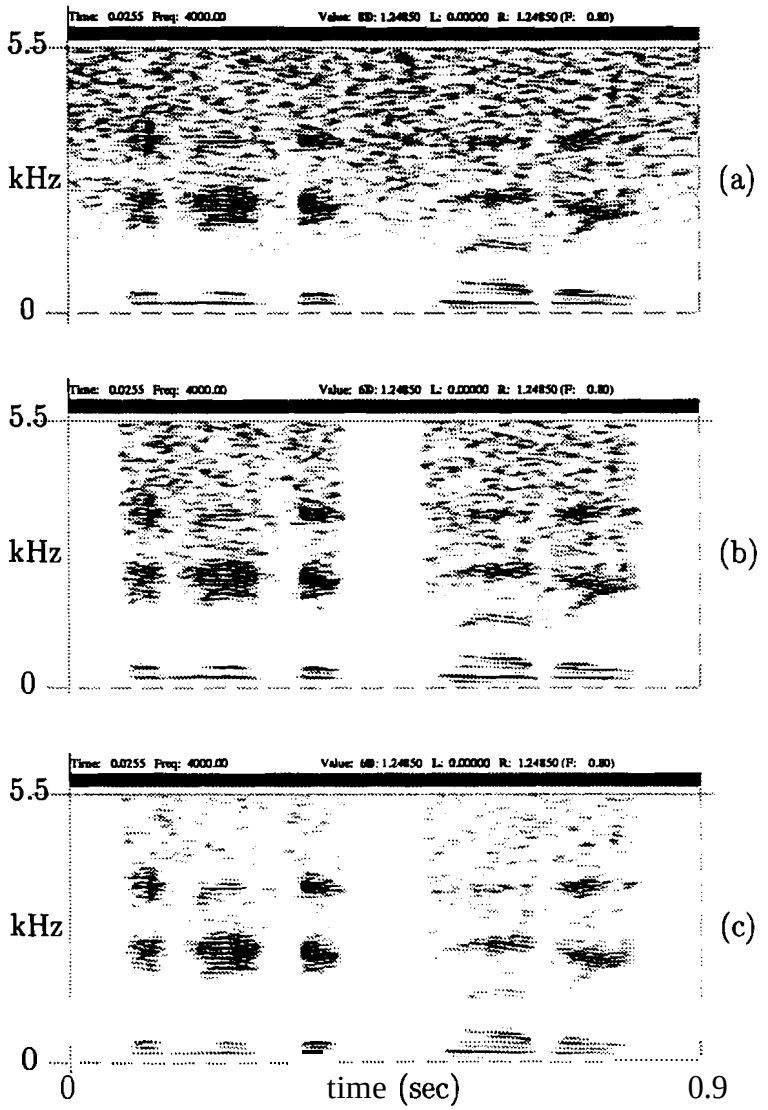


Fig. 4.5: (a) Spectrogram for 10 dB SNR speech. (b) Spectrogram for enhanced speech using gross level weighting of the residual signal. (c) Spectrogram for enhanced speech using gross and fine level weighting of the residual signal.

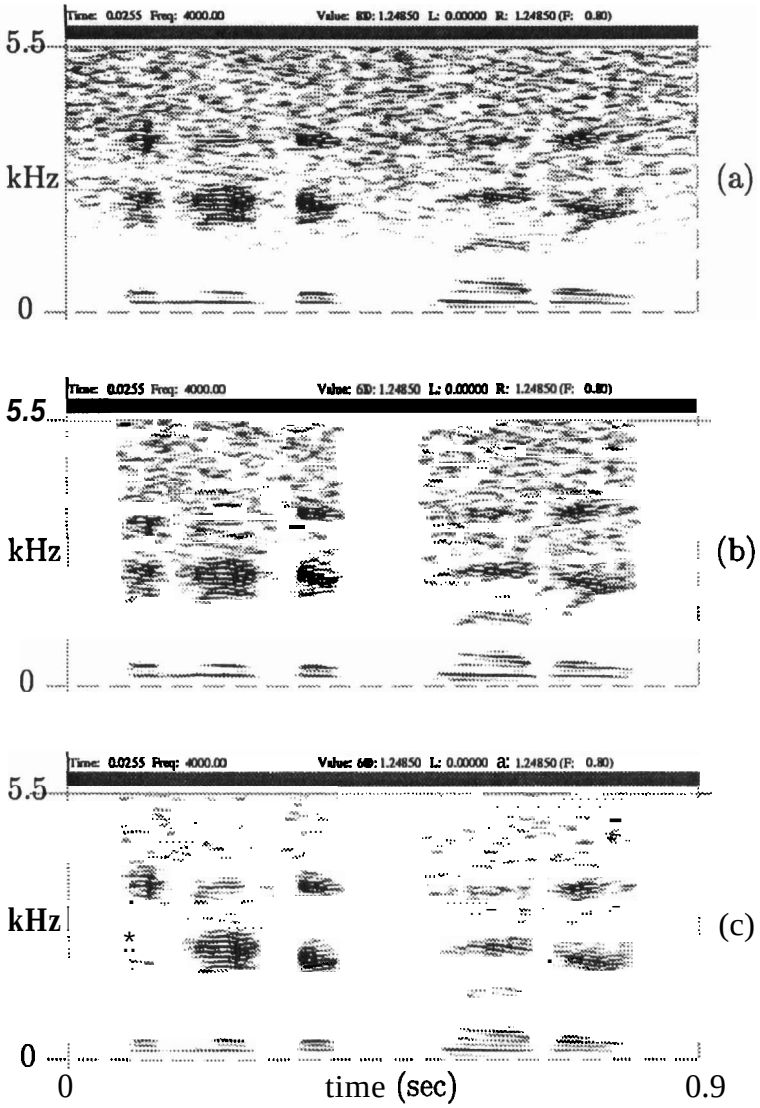


Fig. 4.5: (a) Spectrogram for 10 dB SNR speech. (b) Spectrogram for enhanced speech using gross level weighting of the residual signal. (c) Spectrogram for enhanced speech using gross **and** fine level weighting of the residual signal.

gross level modification.

4.3.3 Spectral Level

In the reconstruction of enhanced speech, even though the LP residual signal is deemphasized in the low SNR regions, the all–pole filters derived from the 20 ms segments dominate the system characteristics in the reconstructed speech signal. To improve the system characteristics at the spectral level, the LPCs for shorter (1–3 ms) segments need to be obtained from noisy speech. This will make the all–pole filter for the high SNR segments closer to the true one. For other segments the amplitude of the output signal is reduced in the reconstruction due to deemphasis of the corresponding residual signal. But unfortunately, we do not have a good method of estimating the all–pole filter for short (1–3 ms) segments.

One way to achieve spectral manipulation indirectly is to perform a low order LP analysis on the differenced (noisy) speech signal. A 7th order LP analysis is performed using 5 ms Hamming windowed segments overlapped by 2 ms. Due to the Hamming window, the effective duration of the signal used for analysis is less than 5 ms. The residual signal is computed by passing the speech signal through the inverse filter. The residual signal is manipulated as described before. The modified residual signal is used to excite the time–varying all–pole filter, updated every 2 ms, to generate the enhanced speech. The different steps in the algorithm are presented in Table–4.1.

4.4 EXPERIMENTAL RESULTS

Examples are given in this section to demonstrate the performance of the proposed method for different types of noises. The degradation is gradual and graceful as the noise level is increased. This is because the LP analysis tends to be less accurate as the SNR reduces. It is important to note that the thresholds for deriving the weight

Table 4.1: Algorithm for processing noisy speech for enhancement.

Computation of the gross weight function

- Calculate the linear prediction (LP) residual signal using a speech frame of size 20 ms, overlapping by 10 ms, Hamming window and a 10th order LP analysis by autocorrelation method. The analysis is performed on the premphasized speech signal.
- Calculate the ratio of the noisy speech signal energy and the LP residual signal energy for each nonoverlapping 2 ms frame. The ratio gives the inverse spectral flatness value for each 2 ms frame.
- Smooth the inverse spectral flatness curve using a 17–point Hamming window. The smoothed spectral flatness value is denoted by ζ_k for the kth frame.
- Obtain the output ζ_k^m of the mapping function

$$\zeta_k^m = \left(\frac{\zeta_{max}^m - \zeta_{min}^m}{2}\right) \tanh\left(\alpha_g \pi \ (\zeta_k - \zeta_0)\right) + \left(\frac{\zeta_{max}^m + \zeta_{min}^m}{2}\right)$$

  from ζ_k. (See Fig. 4.2).
- Obtain the gross weight function by repeating each mapped value ζ_k^m 16 times (2 ms at 8 kHz sampling) and smoothing it with a 2 ms mean smoothing filter. This generates a gross weight value γ_n^g for every sampling instant n.

Computation of the fine weight function

- Compute the Frobenius norm of the Toeplitz prediction matrix constructed using the noisy speech samples in each 2 ms frame, for every sampling instant n.
- Compute the logarithm of the ratio of Frobenius norms of the current frame at the nth sampling instant to the Frobenius norm of the frame 2 ms (=16 sampling instants) prior to the current frame. Normalize the log ratio w.r.t. the maximum value. Obtain the fine weight function γ_n^f by mapping the normalized log ratio using the function

$$\gamma_n^f = \left(\frac{\gamma_{max}^f - \gamma_{min}^f}{2}\right) \tanh\left(\alpha_f \pi \ y_n\right) + \left(\frac{\gamma_{max}^f + \gamma_{min}^f}{2}\right)$$

  which is similar to the function shown in Fig. 4.2. γ_n^f is the fine weight value at the nth sampling instant, y_n is the normalized log ratio of Frobenius norms at n, γ_{max}^f (= 1) is the maximum mapped value, γ_{min}^f (= 0.6) is tho minimum mapped value and α_f (=0.75) is a positive constant.

Linear prediction analysis

- Calculate the linear prediction (LP) residual signal using a speech frame of size 5 ms, overlapping by 2 ms, Hamming window and a 7th order LP analysis by autocorrelation method. The analysis is performed on the preemphasized speech signal.

Synthesis of enhanced speech

- Multiply the two weight functions γ_n^g and γ_n^f to generate the overall weight function.
- Multiply the LP residual signal obtained above using 5 ms segments of the speech signal by the overall weight function. The weighted residual signal is used to excite the time–varying all–pole filter updated every 2 ms, to generate enhanced speech.

function could be adjusted so as to obtain an acceptable trade-off between reduction in annoyance due to noise and degradation in speech quality, based on perceptual impression of the enhanced speech. However, once the listener sets the thresholds to suit his preference, they need not be adjusted again.

4.4.1 Studies on Different Types of Noises

The proposed method for speech enhancement works well even for colored additive noise. Fig. 4.6(a) shows the spectrogram of speech corrupted by noise recorded in the cockpit of an F16 aircraft [254]. The average SNR is adjusted to 10 dB. We notice from the spectrograms that the cockpit noise exhibits both broadband as well as narrowband (spectral lines at approximately 3000 and 4500 Hz) characteristics. Fig. 4.6(b) shows the spectrogram of enhanced speech. The enhancement was carried out using the algorithm proposed in the previous section in three iterations. We found that the enhancement was better when carried out in smaller steps over two or three iterations rather than in one step. In each iteration, mild enhancement can be obtained by using suitable values for the thresholds used for the mapping function in Fig. 4.2. The thresholds were chosen so as to achieve mild enhancement in each iteration and are kept constant in all the iterations.

The method was tested for real signals where the speech signal and noise were recorded simultaneously. Figs. 4.7(a), 4.7(c) and 4.7(e) show the clean, degraded and processed speech signals, respectively. The clean speech and degraded speech were collected simultaneously by two microphones at a sampling frequency of 8 kHz. One microphone was placed at a distance of 0.1m from the speaker and the other was placed 1.2m away. The speech signal shown in Fig. 4.7(a) corresponds to the sentence *"She had your dark suit in greasy wash water all year"* spoken by a male speaker and is taken from the TIMIT database [255]. It can be seen in Fig. 4.7(c) that the degraded speech

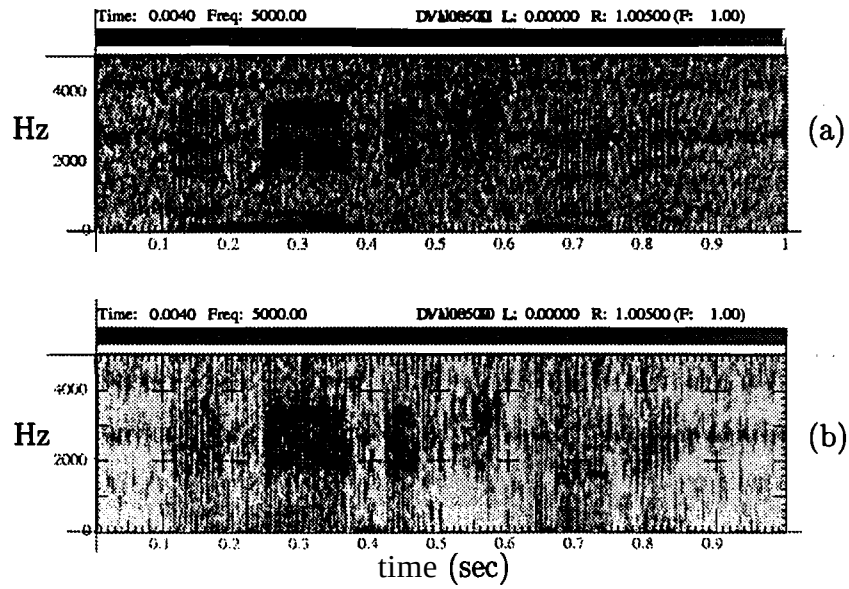


Fig. 4.6: (a) Spectrogram for 10 dB SNR speech. The speech is corrupted by aircraft cockpit noise. (b) Spectrogram for enhanced speech using spectral level manipulation besides gross and fine level weighting of the LP residual signal. The speech is enhanced using three iterations.

has small amount of room reverberation in addition to ambient (airconditioner) noise. The ambient noise has lowpass spectal characteristics and some narrowband spectral components. In fact there is ambient noise present even in the clean speech signal in Fig. 4.7(a). The speech signal in Fig. 4.7(c) was differenced before processing. The speech signal processed using the proposed algorithm and its spectrogram are shown in Figs. 4.7(e) and 4.7(f), respectively. It can be seen from the Figs. 4.7(e) and 4.7(f) that the noise level is significantly attenuated, especially in the silence regions. It is important to note that the gross weight function provides mild attenuation of the reverberation tails. Informal listening confirms that there is reduction of the annoyance due to noise in the processed speech signal, without introducing significant distortion.

The proposed method was also tested on female speech. The experimental setup for data collection was the same as that used for collection of male speech mentioned

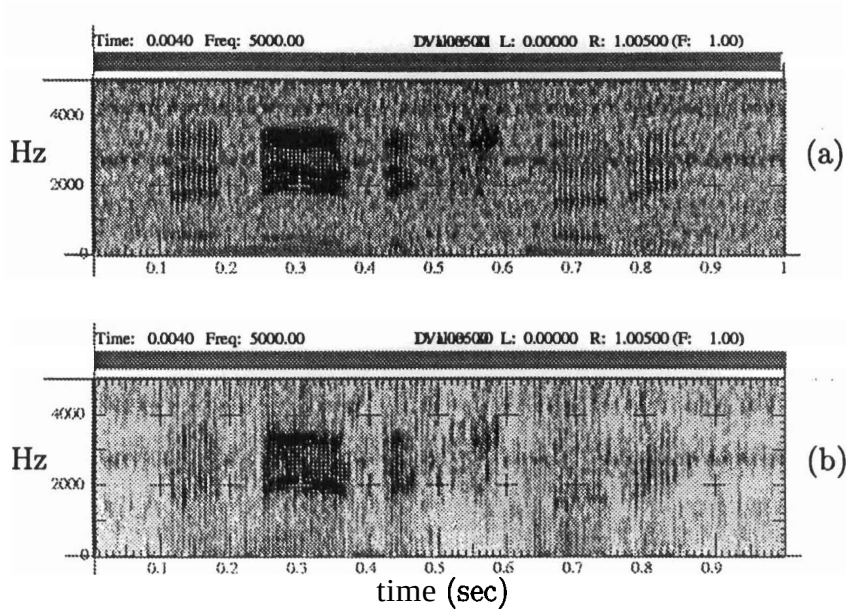


Fig. 4.6: (a) Spectrogram for 10 dB SNR speech. The speech is corrupted by aircraft cockpit noise. (b) Spectrogram for enhanced speech using spectral level manipulation besides gross and fine level weighting of the LP residual signal. The speech is enhanced using three iterations.

has small amount of room reverberation in addition to ambient (airconditioner) noise. The ambient noise has lowpass spectal characteristics and some narrowband spectral components. In fact there is ambient noise present even in the clean speech signal in Fig. 4.7(a). The speech signal in Fig. 4.7(c) was differenced before processing. The speech signal processed using the proposed algorithm and its spectrogram are shown in Figs. 4.7(e) and 4.7(f), respectively. It can be seen from the Figs. 4.7(e) and 4.7(f) that the noise level is significantly attenuated, especially in the silence regions. It is important to note that the gross weight function provides mild attenuation of the reverberation tails. Informal listening confirms that there is reduction of the annoyance due to noise in the processed speech signal, without introducing significant distortion.

The proposed method was also tested on female speech. The experimental setup for data collection was the same as that used for collection of male speech mentioned

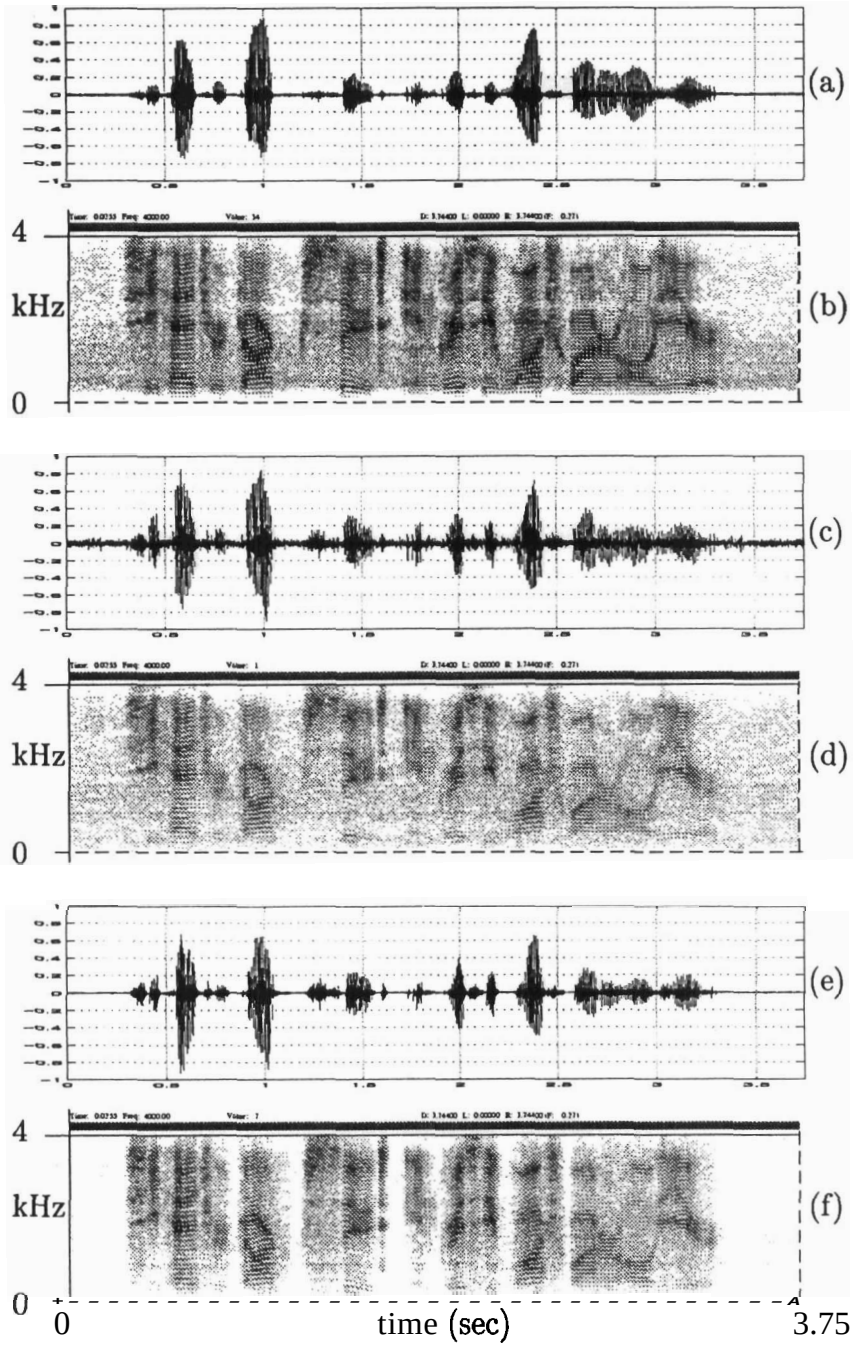


Fig. 4.7: Results of enhancement of male speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

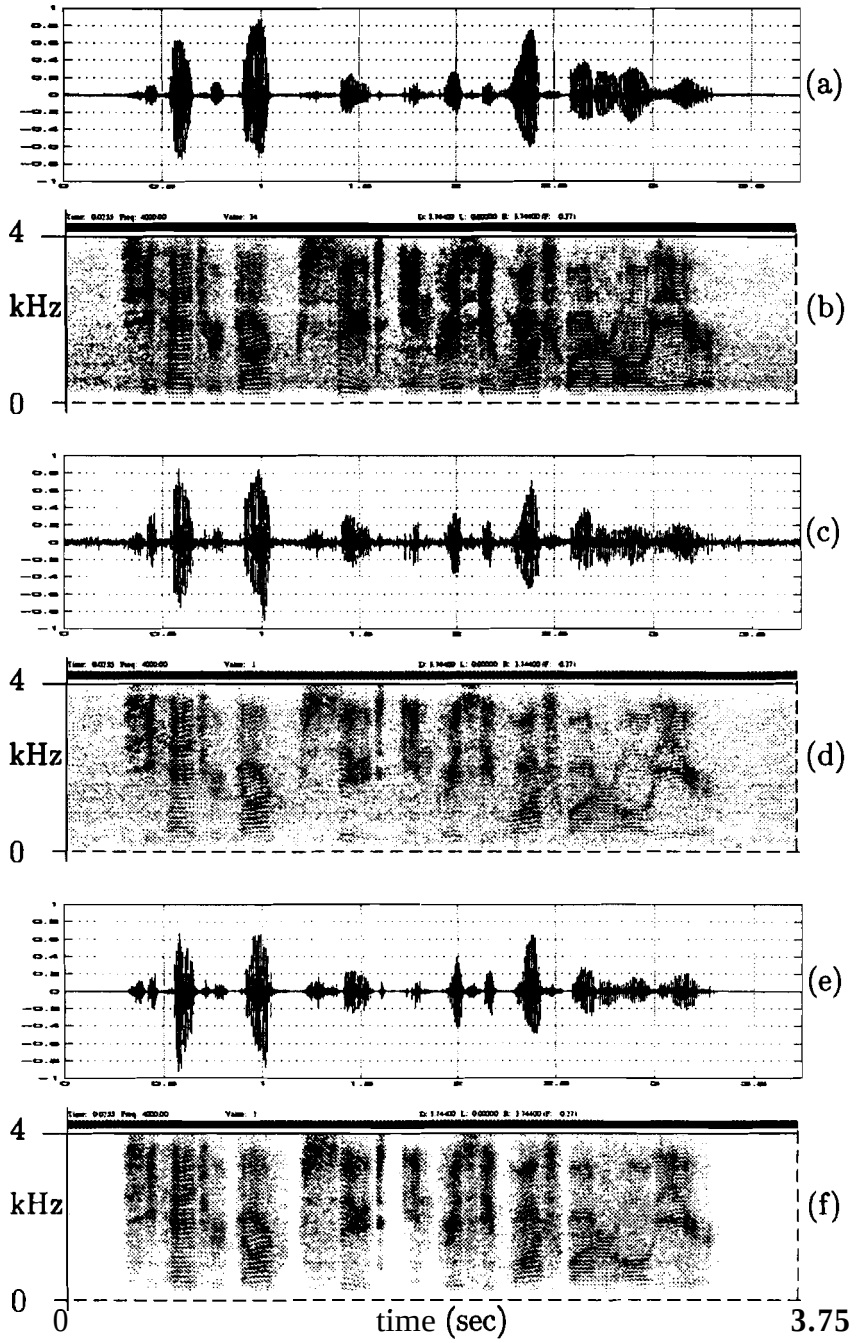


Fig. 4.7: Results of enhancement of male speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

above. Fig. 4.8(a) shows the clean speech signal corresponding to the sentence "***She had your dark suit in greasy wash water all year***" taken from the TIMIT database. The spectrograms of clean, degraded and processed speech signals are shown in Figs. 4.8(b), 4.8(d) and 4.8(f), respectively. The improvement obtained due to processing can be clearly seen in the spectrogram in Fig. 4.8(f). The dark background in the spectrogram in Fig. 4.8(d) is significantly attenuated in the spectrogram in Fig. 4.8(f), both in the silence regions as well as in the regions between the pitch harmonics. Informal listening confirms the improvement obtained due to processing. It is important to note that the same thresholds were used for the mapping functions in all the experiments.

4.4.2 Performance of the Method for Different Parameter Settings

A comparison of the performance of the proposed method for two different settings of the parameters of the mapping function is shown in Fig. 4.9 for the case of female speech. A comparison with the performance of the spectral subtraction method [142] is also given in the same figure (Fig. 4.9(c)). The speech signal used for this comparison is the same as the one shown in Fig. 4.8(c). Fig. 4.9(a) and Fig. 4.9(b) show the spectrograms of the processed speech signals for the parameter settings A and B, respectively, given in Table–4.2. The parameter settings for case A are chosen such that a mild enhancement of the noisy speech signal is obtained without introducing distortion in the processed signal. The emphasis of speech regions with respect to the background noise regions is relatively more for the parameter settings for case B compared to that for case A. But in this case mild distortion is perceived in the processed signal.

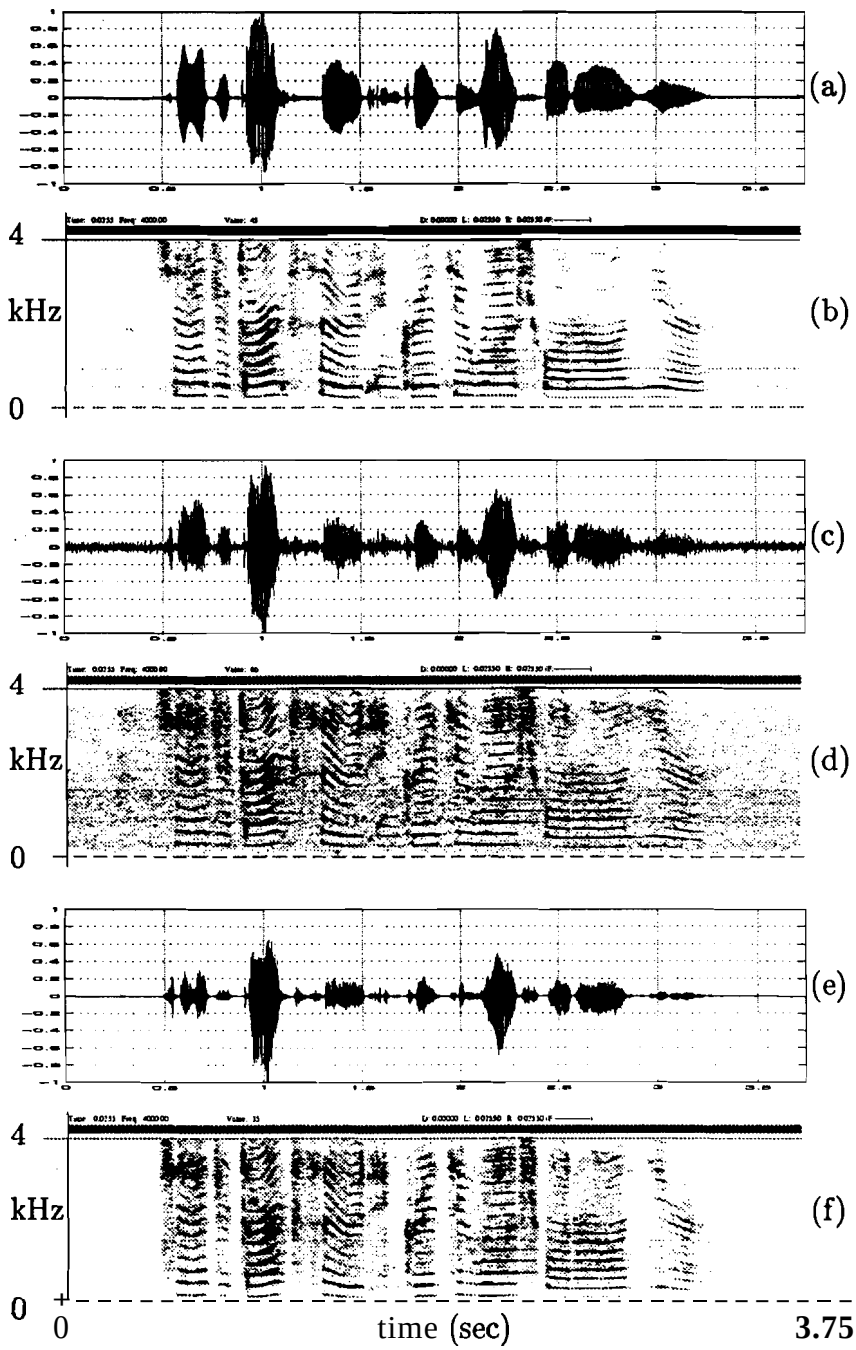


Fig. 4.8: Results of enhancement of female speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

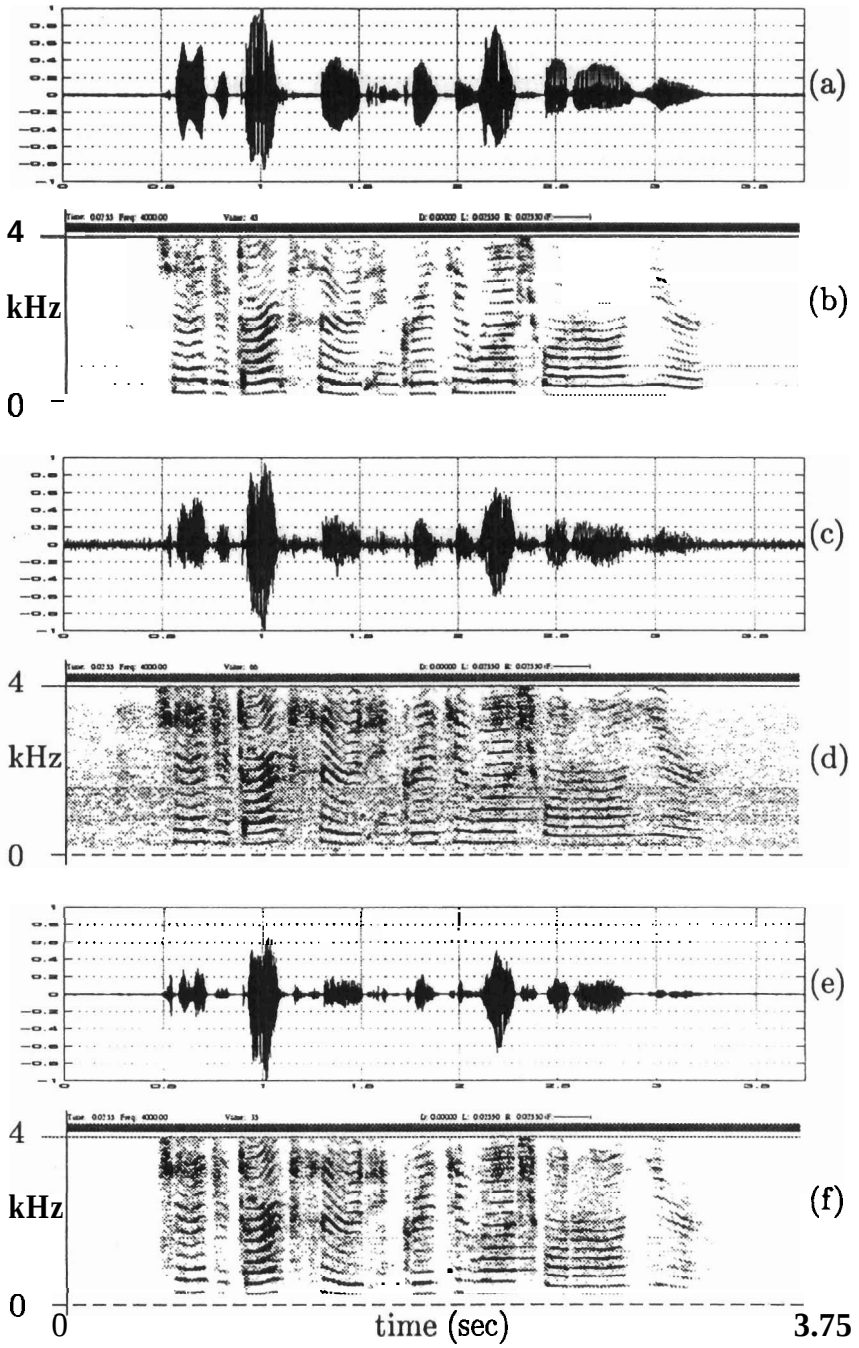


Fig. 4.8: Results of enhancement of female speech degraded by ambient noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by noise. (d) Spectrogram of speech degraded by noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

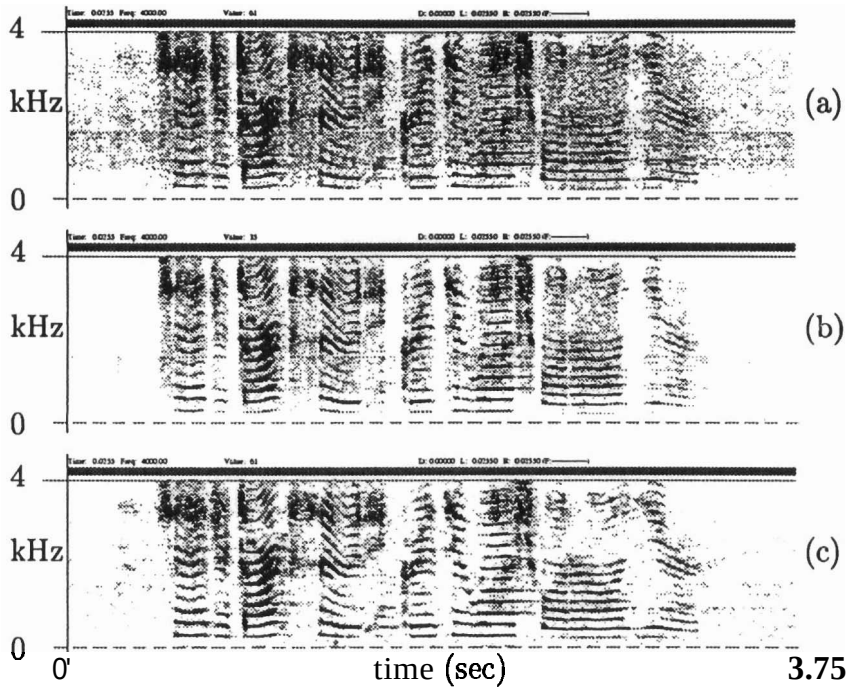


Fig. 4.9: Comparison of results of enhancement of the proposed method with spectral subtraction for female speech degraded by ambient noise. Spectrograms of speech processed using the proposed algorithm for the parameter settings of (a) case A and (b) case B in Table–4.2. (c) Spectrogram of speech processed using the spectral subtraction algorithm.

Table 4.2: Two different settings of the parameters for the mapping functions.

	ζ^m_{max}	ζ^m_{min}	α_g	ζ_0	γ^f_{max}	γ^f_{min}	α_f
Case A	1.0	0.1	1.0	1.5	1.0	0.25	0.75
Case B	1.0	0.05	2.0	2.0	1.0	0.6	0.75

Although the spectrogram in Fig. 4.9(a) does not show significant improvement when compared to the spectrogram of noisy speech in Fig. 4.8(d), the improvement can be clearly perceived while listening. The spectrogram in Fig. 4.9(b) shows a significant improvement when compared with the spectrogram of the noisy speech in Fig. 4.8(d). Note that the weighting of the residual signal at the fine level (i.e., relative emphasis

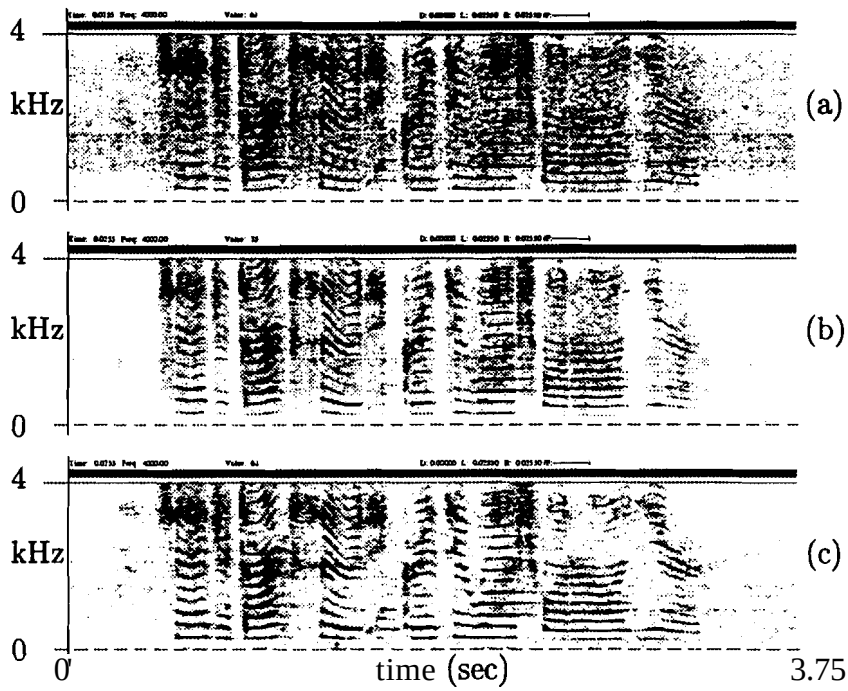


Fig. 4.9: Comparison of results of enhancement of the proposed method with spectral subtraction for female speech degraded by ambient noise. Spectrograms of speech processed using the proposed algorithm for the parameter settings of (a) case A and (b) case B in Table–4.2. (c) Spectrogram of speech processed using the spectral subtraction algorithm.

Table 4.2: Two different settings of the parameters for the mapping functions.

	ζ^m_{max}	ζ^m_{min}	α_g	ζ_0	γ^f_{max}	γ^f_{min}	α_f
Case A	1.0	0.1	1.0	1.5	1.0	0.25	0.75
Case B	1.0	0.05	2.0	2.0	1.0	0.6	0.75

Although the spectrogram in Fig. 4.9(a) does not show significant improvement when compared to the spectrogram of noisy speech in Fig. 4.8(d), the improvement can be clearly perceived while listening. The spectrogram in Fig. 4.9(b) shows a significant improvement when compared with the spectrogram of the noisy speech in Fig. 4.8(d). Note that the weighting of the residual signal at the fine level (i.e., relative emphasis

of the residual signal samples within a glottal cycle) should be mild to avoid distortion in the processed speech. In the voiced regions the spectrogram in Fig. 4.9(b) appears cleaner compared to the spectrogram in Fig. 4.9(a). In the case of speech processed using the spectral subtraction method we observe that weak spectral peaks appear randomly in the spectrogram in Fig. 4.9(c). These random spectral peaks give rise to musical noise.

4.5 SUMMARY

In this chapter we have presented a new approach for enhancement of speech based on LP residual signal. The method uses the fact that in noisy speech the SNR is a function of time and frequency. By enhancing the high SNR regions relative to the low SNR regions, the annoyance due to background noise is reduced without significantly distorting the speech. This is accomplished by identifying the low and high SNR regions based on the characteristics of the spectral flatness in short (2 ms) time frames. The spectral flatness information is derived using the ratio of energies in the LP residual signal of the speech and the noisy signal. Inverse spectral flatness characteristics are used to derive a weight function for the residual signal at gross level, and the Frobenius norm of short (2 ms) segments of the speech signal is used to derive the weight function at finer level. The two weight functions are multiplied to get the overall weight function for the residual signal. The method works since the residual signal samples are nearly uncorrelated, and hence can be manipulated without significantly affecting the quality of the speech regenerated from the modified residual signal. Since no direct manipulation in different frequency bands is involved, this method does not produce the type of distortion which the spectral subtraction and parameter smoothing methods produce.

The objective in this study is to enhance speech over background noise, and not

noise suppression or elimination. In fact even a small (3–6 dB) improvement in SNR of noisy speech may give relief to the listener. This study suggests that speech enhancement methods must aim to bring down the annoyance due to noise by mild enhancements.

The setting of various thresholds in the processing is primarily dictated by the listener's tolerance to annoyance due to noise and preference to speech quality. The various parameter values used in the processing, such as LP order, analysis frame size and thresholds of the mapping function are not critical. The choice of the parameters depends on listener's preference, as the effect of these parameters on the resulting quality of the enhanced speech is gradual and not abrupt. Another important feature is that the method does not depend on the pitch of the voice. There is no direct manipulation of the spectrum. However, a better estimation of the vocal tract system characteristics is needed to improve the enhancement at the spectral level.

The proposed method reduces the annoyance due to additive noise but is not very useful in reducing the annoyance due to reverberation. However an approach based on emphasizing the high signal energy regions relative to the low signal energy regions can be developed for enhancement of reverberant speech also [256]. This is the subject matter of the next chapter.

In our opinion the proposed approach is different from many methods available for processing degraded speech. There is scope for significant improvement by studying the effects of various parameters on the perceptual quality of the enhanced speech. Moreover, this approach may be combined with well known spectrum–based methods for speech enhancement to obtain a better quality of enhanced speech for various types of degradation.

Chapter 5

ENHANCEMENT OF REVERBERANT SPEECH

In the previous chapter we have presented a method for enhancement of speech degraded by additive random noise using subsegmental analysis. The low correlation between samples of the LP residual signal was exploited in the method. In this chapter, we propose a method for enhancement of speech degraded by reverberation in small rooms. The approach taken is similar to the one adopted in the case of enhancement of noisy speech. The method is based on analysis of short **(1–3** ms) segments of speech to enhance the regions in the speech signal having high Signal–to–Reverberant component Ratio (SRR). The short (1–3 ms) segment analysis shows that SRR is different in different segments of speech. The processing method involves identifying and manipulating the linear prediction residual signal in three different regions of the speech signal, namely, high SRR region, low SRR region and only reverberation component region. **A** weight function is derived to modify the linear prediction residual signal. The weighted residual signal samples are used to excite a time–varying all–pole filter to obtain perceptually enhanced speech. The method is robust to noise present in the recorded speech signal. Informal listening shows that the proposed method enhances the speech signal under reverberant conditions without causing significant degradation in quality.

In the next section the background to the proposed method is developed. In Section **5.2** we discuss the model of reverberant speech and some of its characteristics.

By studying the effects of degradation in short (1–3 ms) segments, we obtain clues that can be used for processing the reverberant speech. In Section 5.3 steps for processing degraded speech are discussed. In particular, the importance of processing the linear prediction (LP) residual signal is emphasized. We present some experimental results in Section 5.4. The improvement in the processed speech is demonstrated through the signal waveform, short–time spectra and spectrograms.

5.1 INTRODUCTION TO ENHANCEMENT OF REVERBERANT SPEECH

Degradations in speech are caused by additive noise and reverberation. In this chapter we consider enhancement of speech under reverberant conditions. The focus is on the degradation of speech such as in speakerphone situation. Speech from a speakerphone contains both the direct component and the reverberant component. The objective of processing is to enhance the signal in the direct component, wherever possible, so that the resulting processed speech is perceived as less reverberant and thus increasing the comfort level for listening.

Normally, degraded (additive or reverberant) speech is processed assuming that the degradation has long term stationary characteristics relative to speech. For example, for degradation due to additive noise, the noise statistics are estimated from the degraded speech and the long (100–300 ms) term noise effects are subtracted from the short (10–30 ms) time speech spectra [49, 50, 142] to reduce the effects of noise. Likewise, for reverberant speech, the reverberation effects are captured by estimating the impulse response of the room environment from long (500–1000 ms) segments of speech [52, 257–259]. The room impulse response is usually long, of the order of 200–300 ms. The reverberant speech is passed through an inverse filter for the room response to dereverberate speech. Here again the estimated long term characteristics are used to filter out its effects from the short (10–30 ms) quasistationary segments of

speech. The main problem in these approaches for processing degraded speech is that the estimates of the characteristics of the degradations may not be good enough to remove their effects in short segments of speech. This is because the level of degradation in terms of Signal-to-Noise Ratio (SNR) is different for different segments of speech. Moreover, the emphasis in many of these approaches seems to be on the degradation and not on speech. In other words, enhancement is sought to be accomplished by attempting to cancel the effects of the degrading component.

In noise suppression and dereverberation there is more emphasis on improving the overall SNR/SRR of the degraded speech. In this process most of the attention is given to improve the low SNR/SRR regions of speech. When attempting to reduce the degradation in these regions, the natural characteristics of speech are changed, causing significant distortions. This is because, all segments of degraded speech are treated equally. In order to improve the overall SNR/SRR, it is necessary to reduce the noise/reverberation in the low SNR/SRR regions, which does not produce significant enhancement perceptually.

Methods for enhancement of reverberant speech generally rely on estimating the impulse response of the inverse system for dereverberation [52]. It may not be possible to estimate this response accurately from speech in most of the situations. In some of the methods, the room response is collected separately to design the inverse system [260]. The recovery of the average envelope modulation spectrum of the original (anechoic) speech by filtering the time trajectories of spectral bands of reverberant speech has been proposed in [85, 95, 261]. Enhancement methods based on processing the speech collected by multiple microphones have also been proposed [262, 263].

There appears to be a need to look at the problem of enhancement of reverberant speech with more emphasis on the direct component of speech at the receiving

microphone. In processing, it is necessary to increase the contribution of the direct component relative to the reverberant component [256]. In such an attempt there will be more emphasis on the speech than on the degradation during the enhancement. This point of view is also reasonable, since speech is a nonstationary signal, with the energy of the signal varying over a wide (about 60 dB) dynamic range both in temporal and spectral domains. Therefore the signal–to–degradation ratio will be varying even within 10–30 ms segments of data. For short (10–30 ms) segments it is difficult to estimate the reverberant component. Moreover, the reverberant component itself will be different in different segments due to its dependence on the energy in the preceding segments of speech. That is, the reverberant component is signal dependent.

It is also essential that we specify our goal in the enhancement of degraded speech. Obviously, complete dereverberation is not a realizable task. Therefore, the emphasis should be on enhancement, but not necessarily enhancement of all segments of speech. There are segments of speech where reverberant component dominates over the direct component. For such segments there is no point in attempting to enhance the speech part. On the other hand, if regions, where the direct speech signal component is significantly higher compared to the reverberant component, could be identified, then by enhancing speech in such regions the annoyance due to reverberation could be reduced in some segments at least. Likewise the levels of the signal in the regions with higher reverberation could be reduced, if such regions could be identified. In the regions where there is only a reverberant component, such as silence regions, the levels could be reduced to very low values. Perception of the overall speech is influenced significantly by the high signal energy regions, thus giving an impression of enhancement of degraded speech. Therefore the criterion for improvement need not be based on giving equal emphasis to all the speech segments. It is better to focus on

the regions having high direct path signal component.

In this work we show that using subsegmental analysis it is indeed possible to locate the segments in the degraded speech where the direct component is higher than the reverberant component. These segments are usually much shorter than the glottal cycle. The proposed approach is different from the existing methods, as there is more emphasis on the characteristics of speech, and also the analysis segments are much shorter (1–3 ms) compared to the normal frame size (10–30 ms) used in speech analysis.

5.2 CHARACTERISTICS OF REVERBERANT SPEECH

In this section we will examine the characteristics of reverberant speech to determine clues for processing speech for enhancement. Throughout the discussion we will examine the similarities and differences in the characteristics of reverberant speech and speech corrupted by additive noise. For this purpose we consider the following models for reverberant speech and noisy speech.

$$\text{Reverberant speech:} \quad y(n) = s(n) + \sum_{k=1}^{N} \beta_k \, s(n - n_k) \tag{5.1}$$

$$\text{Noisy speech:} \quad Y(n) = s(n) + w(n) \tag{5.2}$$

where $s(n)$ is the clean speech signal, β_k is the relative amplitude of the reflection arriving after a delay of n_k samples, N is the number of such reflections, and $w(n)$ is the additive noise component. In each model the first term on the right hand side is the signal component and the second term is the component due to degradation. The main difference between these two models is that, in the case of reverberation, the degrading component is dependent on previous speech data, whereas in the case of noisy speech the degrading component is independent of speech. That is, in the reverberation the degrading component is speech–like.

The relative strength of the reverberant component over the direct component depends on the energy of the speech signal in a short segment around the current instant. This strength can be called signal–to–reverberant component ratio (SRR) at that instant. Likewise, the ratio of the signal energy to the noise energy in a short segment around the current instant is called signal–to–noise ratio (SNR) at that instant. To study the characteristics of SRR and SNR as a function of time, these ratios are computed for short (2 ms) segments of degraded speech. Due to nonstationary nature of speech, the signal energy varies with time. Fig. 1.6(a) shows the clean speech signal. The energy of the clean speech and the SRR for the reverberant speech are shown in Figs. 1.6(b) and 1.6(c), respectively. Likewise, the SNR for speech corrupted by additive noise is plotted in Fig. 1.6(d). In both cases, it is obvious that SRR and SNR vary with time since the signal energy is also a function of time. In fact, in the case of reverberant speech, both the signal energy and the energy of the degrading component are time–varying, which is not always true in the case of noise–corrupted speech. In Fig. 1.6(c) we observe that in the 300–400 ms region the SRR is very poor. This is because the direct component is small in this region, whereas there is a large reverberant tail component due to the preceding vowel. In Figs. 1.6(c) and 1.6(d) we also observe that there are finer variations (ripple) in the SRR and SNR plots. This is because of the variation of the signal energy and energy of the degrading component even within a glottal cycle.

The effects of reverberation can be seen by comparing the signal waveforms for clean and reverberant speech signals shown in Fig. 5.1. The clean speech has damped sinusoidal pattern within each glottal cycle, whereas the reverberant speech is smeared within each cycle (region AB in Fig. 5.1(b)). Smearing of the signal within each glottal cycle is more prominent when the envelope of the signal waveform is decaying as in

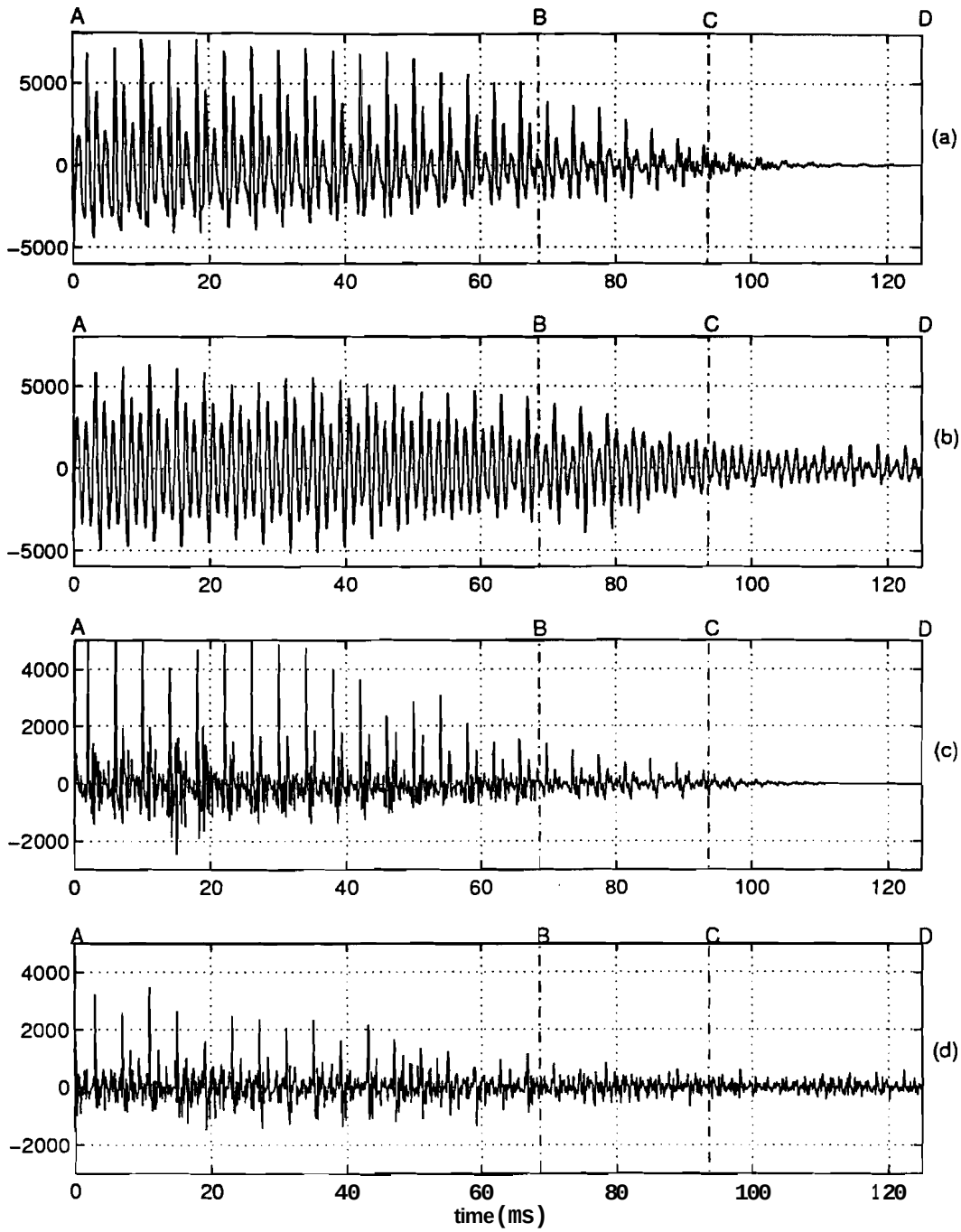


Fig. 5.1: Comparison of clean and reverberant speech signals. (a) Clean speech. (b) Signal corrupted by reverberation. (c) LP residual signal for the clean speech in (a). (d) LP residual signal for the reverberant speech in (b).

the region BC in the figure. The smearing extends for several glottal cycles due to the influence of large amplitude signal component in the region AB. Only the reverberation tail component is present in the low amplitude silence regions (CD).

Nature of the reverberant speech in the spectral domain can be observed by comparing short–time (20 ms) spectra (Fig. 5.2) for segments in each of the three regions.

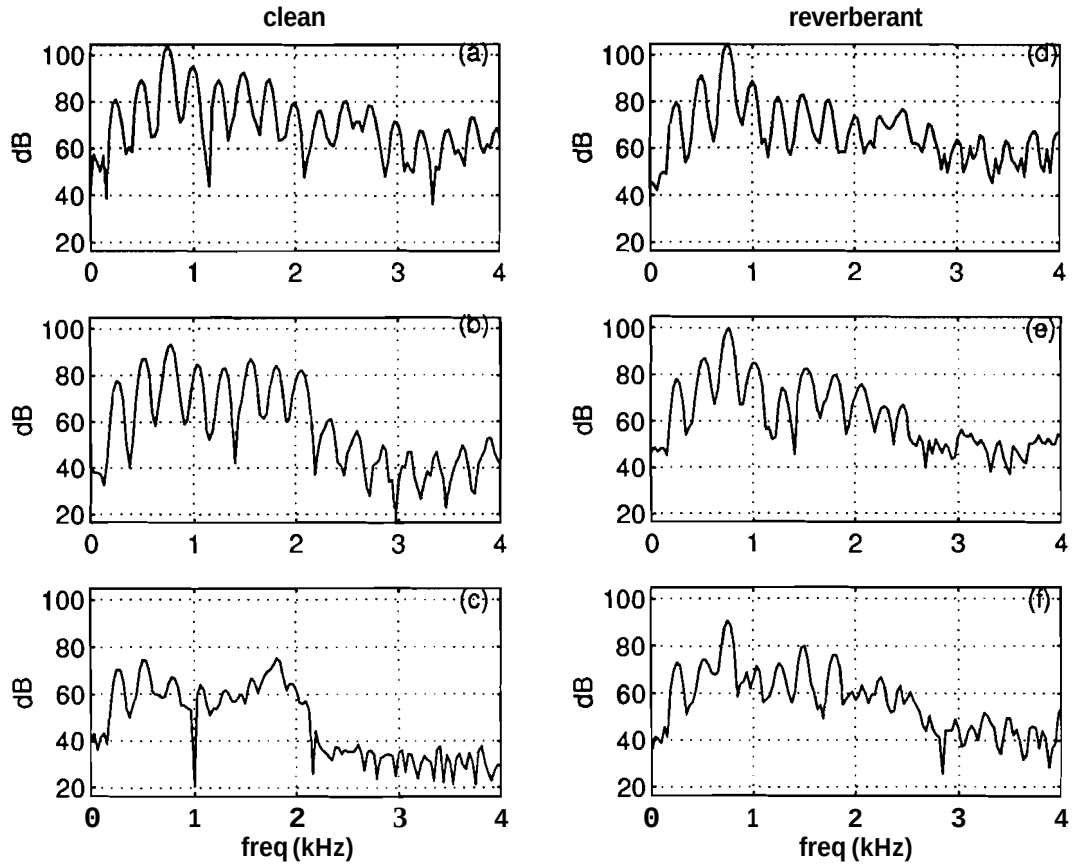


Fig. 5.2: Comparison of short–time spectra for clean and reverberant speech in different segments. (a) – (c) Short–time spectra of the clean signal in Fig. 5.1(a) in the regions AB, BC and CD, respectively. (d) – (f) Short–time spectra of the reverberant signal in Fig. 5.1(b) in the regions AB, BC and CD, respectively.

In all the three cases the dynamic range of the dominant initial portion of the spectral envelope is higher for the reverberant speech compared to that of the clean speech. Thus there is reduction in the flatness of the spectral envelope due to reverberation.

The figure also illustrates that the spectral features of the clean speech are altered significantly due to reverberation, especially for the segments in the regions BC and CD in Fig. 5.1.

Effect of reverberation can also be seen clearly in the LP residual signal waveform. Figs. 5.1(c) and 5.1(d) show the LP residual signals for clean and reverberant speech. The residual signal is computed for a segment of 2 ms at every sampling instant, using a 5th order autocorrelation LP analysis. The residual signal for reverberant speech signal clearly shows that there is a significant direct component of the signal in the reverberant speech in the region AB. This is because for the segments in the region AB the signal amplitudes at the epochs (instants of glottal closure) are higher than the signal amplitudes in the rest of the glottal cycle, like in the case of clean speech. This shows that there are segments in the reverberant speech where the direct component is significantly higher than the reverberant component. In the region BC, due to the decay of the overall signal amplitudes, the reverberation effects of the preceding speech dominate over the direct component. In the region CD the residual signal is mainly due to reverberation.

Comparing the residual signals for clean and reverberant speech signals, the effects of reverberation can be seen within each glottal cycle since the residual signal is much higher in between two epochs when the reverberant component dominates. Whenever the direct component of speech is higher than the reverberant component, the LP residual signal at the epochs has significant energy around the instants of glottal closure. Figs. 5.1(c) and 5.1(d) show that there are regions where the direct component is dominant. We need to identify such regions so that the signals in those regions can be processed to enhance the direct component over the reverberant component. Note that there is no clear evidence of the direct component in the region BC, and there is

only reverberant component in the region **CD.** So the signals in the regions **BC** and **CD** need to be attenuated relative to the signal in the region **AB.** Within the region **AB** the signal around the instants of glottal closure need to be enhanced compared to the signal in the rest of the glottal cycle.

First of all it is necessary to identify these three different regions in the reverberant speech. For this purpose let us observe some more characteristics of the reverberant speech. Fig. **5.3** shows the normalized error η (defined in 2.20) of clean and reverberant

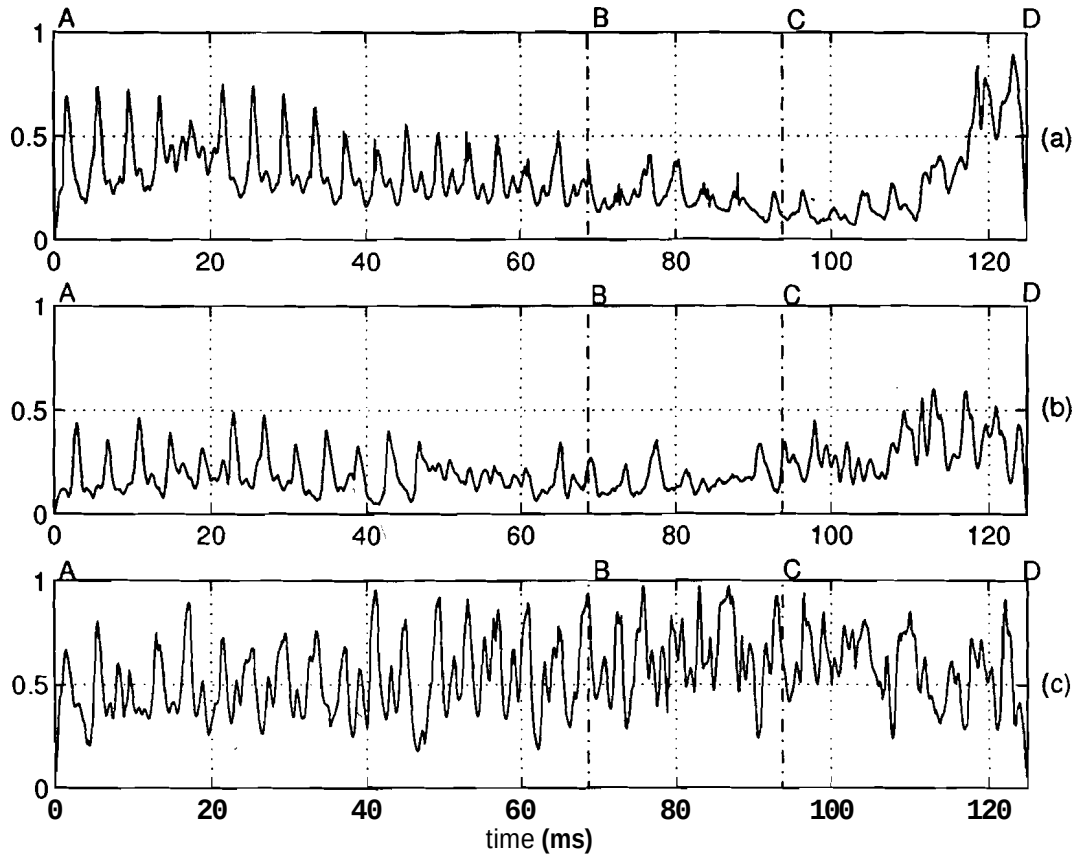


Fig. 5.3: Comparison of normalized prediction error for (a) clean, (b) reverberant and (c) noisy speech (average SNR = 10 dB).

speech, computed at every sampling instant using a 5th order autocorrelation LP analysis using a frame size of 2 ms. The normalized errors for both the clean and reverberant speech are similar in the high SRR regions. However the normalized error

for the reverberant speech is generally lower than for the clean speech. This is due to the multiplicative effect of the frequency response of the room on the speech spectrum. Multiplication of two spectra produces larger dynamic range and hence reduces the spectral flatness.

In contrast, the speech corrupted by additive noise has higher spectral flatness compared to the clean speech. Thus the normalized error for the additive noise case is higher than for the clean speech as shown in Fig. **5.3.** Although the LP residual signal for noisy and reverberant speech look similar, their spectral flatness characteristics are distinct. Reverberation decreases the spectral flatness of speech whereas additive noise increases the spectral flatness. In fact, the increase in spectral flatness for additive noise was exploited for developing a method for enhancement of noisy speech [264].

A closer examination of the normalized error plot within each glottal cycle shows that the error is maximum just before glottal closure. This is because the speech signal amplitude is low in this region. The points of maximum η within each glottal cycle can be identified in the high SRR regions such as AB in Fig. **5.3.** It is difficult to see the distinction between open and closed glottis regions in the low SRR regions such as BC. The normalized error in the purely reverberant region (CD) does not show any periodic peaks.

The above study of the characteristics of reverberant speech suggests that we need to address the following issues for enhancement:

(a) Which domain to process, temporal or spectral ? Which signal to manipulate, original or residual ?

(b) How to identify the high SRR regions in short **(2** ms) segments as well as in the long segments such as AB, BC and CD ?

(c) How to process the signal in each of these regions so that the SRR is increased

at the fine level (2 ms) within a glottal cycle, and at the gross level (> 20 msec segments) as in the regions AB, BC and CD ?

(d) How to increase the spectral flatness to the levels of clean speech signal by increasing the normalized error in each segment of speech ?

(e) How to measure the enhancement realized by a processing method ?

In the next section we discuss some approaches to deal with each of these issues, and present a method for processing reverberant speech for enhancement. The important point to be noted is that for enhancement of degraded speech, different segments need to be processed differently according to the characteristics of speech in the temporal and short–time spectral domains.

5.3 PROCESSING REVERBERANT SPEECH USING LP RESIDUAL SIGNAL FOR ENHANCEMENT

For processing reverberant speech for enhancement we propose manipulation of the LP residual signal in short (2 ms) and in longer (20 ms) segments in a selected manner. The manipulation basically involves weighting the residual signal samples appropriately. Manipulation of the residual signal is more appropriate than the manipulation of speech signal, especially for short (2 ms) segments, as the residual signal samples are generally less correlated than the speech samples. On the other hand, for manipulation of the speech signal directly, the choice of the size and shape of the window may affect the results significantly. It is interesting to note that any distortion caused by processing the residual signal is smoothed out by the all–pole filter used for synthesis.

LP residual signal is computed by performing the LP analysis on short (2 ms) segments of speech data around every sampling instant. Differenced speech signal samples are used to perform the LP analysis. The LP residual signal is obtained by

inverse filtering the speech signal using the LPCs. The decorrelation achieved by the inverse filtering is useful to modify the residual signal.

As mentioned earlier, processing of the LP residual signal involves determination of suitable weight function for the residual signal. The weight function is derived for modifying the residual signal both at the fine (within glottal cycle) level and at the gross level. To derive the weight function we need to identify the different SRR regions at the fine and gross levels from the reverberant speech signal. That is, we need to determine the three types of regions such as AB, BC and CD shown in Fig. 5.1, and also the regions around the instants of glottal closure in AB. These regions can be identified using the properties of the LP residual signal for reverberant speech. The regions at the gross level are determined using the statistics of the LP residual signal. In the high SRR regions the entropy of the distribution of the samples in the LP residual signal is low compared to the entropy in the low SRR regions. This is because the LP residual signal samples exhibit a Gaussian–like probability density function in the reverberant tail regions, and hence the entropy is high. In the high SRR regions, especially in the voiced regions, the peaks in the LP residual signal due to strong excitations of the vocal tract system produce a skewed density function, and hence the resulting entropy is low. To compute the entropy, the probability density function of the samples in each of the 20 ms segments of the LP residual signal is estimated. A longer (20 ms) segment is used to obtain a good estimate of the histograms of the samples and hence their probability density function. The entropy H_k for the kth frame is given by the following expression [265] :

$$H_k = -\sum_{i=1}^{M} p_i \log(p_i) \tag{5.3}$$

where p_i is the estimated probability for the ith bin of the histogram, and M is the number of bins in the histogram. The number of bins (M) can be chosen to be in

the range 5 – 20, making sure that there are enough LP residual signal samples per bin. We have chosen a value of M = 7. This ensures that there are, on an average, about 20 samples per bin in each 20 ms frame. The entropy is computed for a 20 ms frame at every 10 ms. Figs. 5.4(a) and 5.4(b) show the clean and reverberant speech signals, respectively. Figs. 5.4(c) and 5.4(d) show the skewness and kurtosis [266, 267] computed for a 20 ms frame of the LP residual signal at every 10 ms. Fig. 5.4(e) shows the entropy function. It is clear from the figure that both the skewness and kurtosis are high in the regions where the direct component of the signal is strong and so the corresponding entropy is low. The skewness and kurtosis assume values close to zero in the silence and reverberation tail regions because the shape of the estimated probability density function is Gaussian–like [266]. Therefore the entropy in these regions is high as shown in Fig. 5.4(e).

The entropy function is smoothed by repeating each entropy value in Fig. 5.4(e) 80 times (corresponding to 10 ms at 8 kHz sampling rate), and smoothing the resulting function using a 600–point mean smoothing filter. From the smoothed entropy function (Fig. 5.5(a)) a gross weight function (Fig. 5.5(b)) is derived using the nonlinear mapping function shown in Fig. 5.6. The objective of the nonlinear mapping function is to enhance the contrast between the strong direct speech component and the reverberant component. The values of H_0 and γ_{min}^{g} in the mapping function in Fig. 5.6 can be varied to derive a suitable mapping function, although the setting of these thresholds is not critical.

Note that the entropy function is preferable to the skewness and kurtosis functions, for deriving the gross weight function. This is because the entropy function detects even weak speech regions (both voiced and unvoiced) while the skewness and the kurtosis functions were found to be sensitive to only strong voiced regions. This can

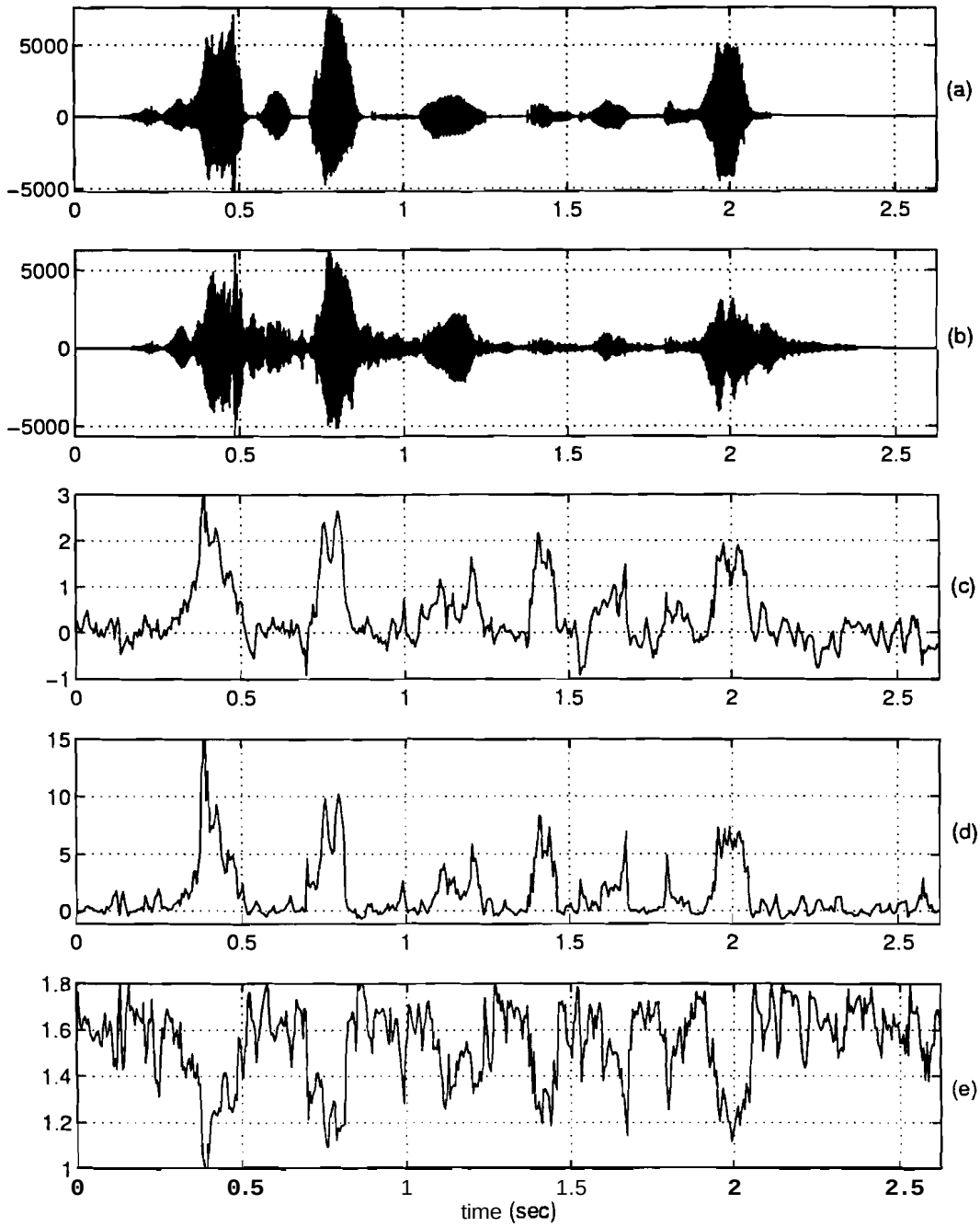


Fig. 5.4: Characteristics of LP residual signal for reverberant speech. (a) Clean speech signal. (b) Reverberant speech signal. (c) Skewness. (d) Kurtosis. (e) Entropy function.

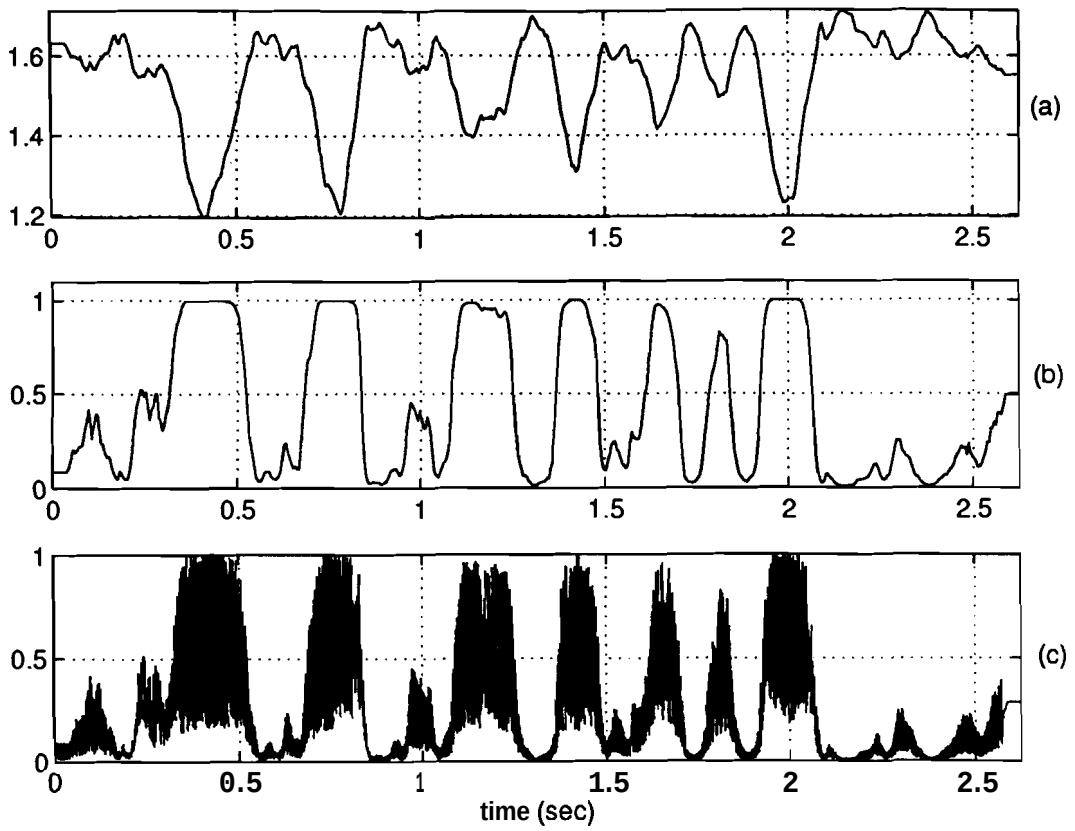


Fig. 5.5: Various stages in the derivation of the weight function for the LP residual signal. (a) Smoothed entropy function. (b) Gross weight function. (c) Overall weight function.

be observed for the segment of the signal at the beginning of the utterance shown in Fig. 5.4(a), which corresponds to the fricative /sh/ as in *she*. The entropy function in Fig. 5.4(e) shows a prominent dip at $t = 0.25$ s, while the skewness and kurtosis functions in Figs. 5.4(c) and 5.4(d), respectively, do not exhibit a peak in that region.

From the gross weight function (Fig. 5.5(b)) the three different types of SRR regions can be identified. The regions of rising and high values of the weight function correspond to the high SRR regions (like region AB in Fig. 5.1). The falling portions correspond to the low SRR regions (like region BC in Fig. 5.1). The low weight function regions correspond to the reverberant component regions (like region CD in Fig. 5.1).

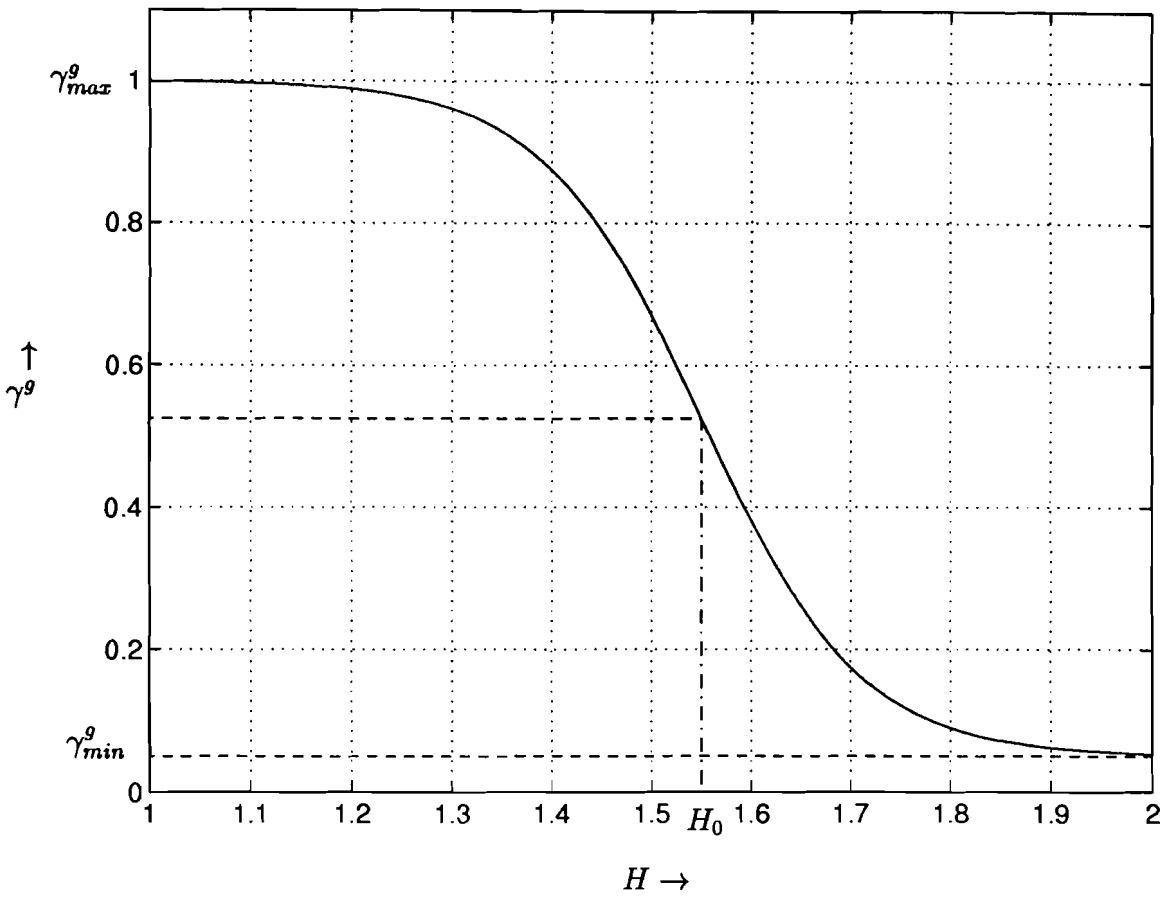


Fig. 5.6: Mapping function to generate the weight values from the entropy values. The mapping function $\gamma^g = \left(\frac{\gamma^g_{max}-\gamma^g_{min}}{2}\right) \tanh\left(-\alpha_g \pi \ (\mathrm{H} - H_0)\right) + \left(\frac{\gamma^g_{max}+\gamma^g_{min}}{2}\right)$ is shown for $\alpha_g = 1.5$, $H_0 = 1.55$ and $\gamma^g_{min} = 0.05$.

To derive the fine weight function, the normalized error (q) is computed at each sampling instant using a frame size of 2 ms and a 5th order LP analysis. The normalized error is shown in Fig. 5.7(c) for a segment of 80 ms of speech shown in Fig. 5.7(a). The peaks in the error function generally correspond to the region around the glottal excitation points, at which the LP residual signal (Fig. 5.7(b)) also has large amplitudes. Note that the normalized LP error shows the characteristic peaks in the initial 50 ms segment because of the strong direct component. These peaks are not prominent in the latter 30 ms segment because of the stronger reverberant component. **A**

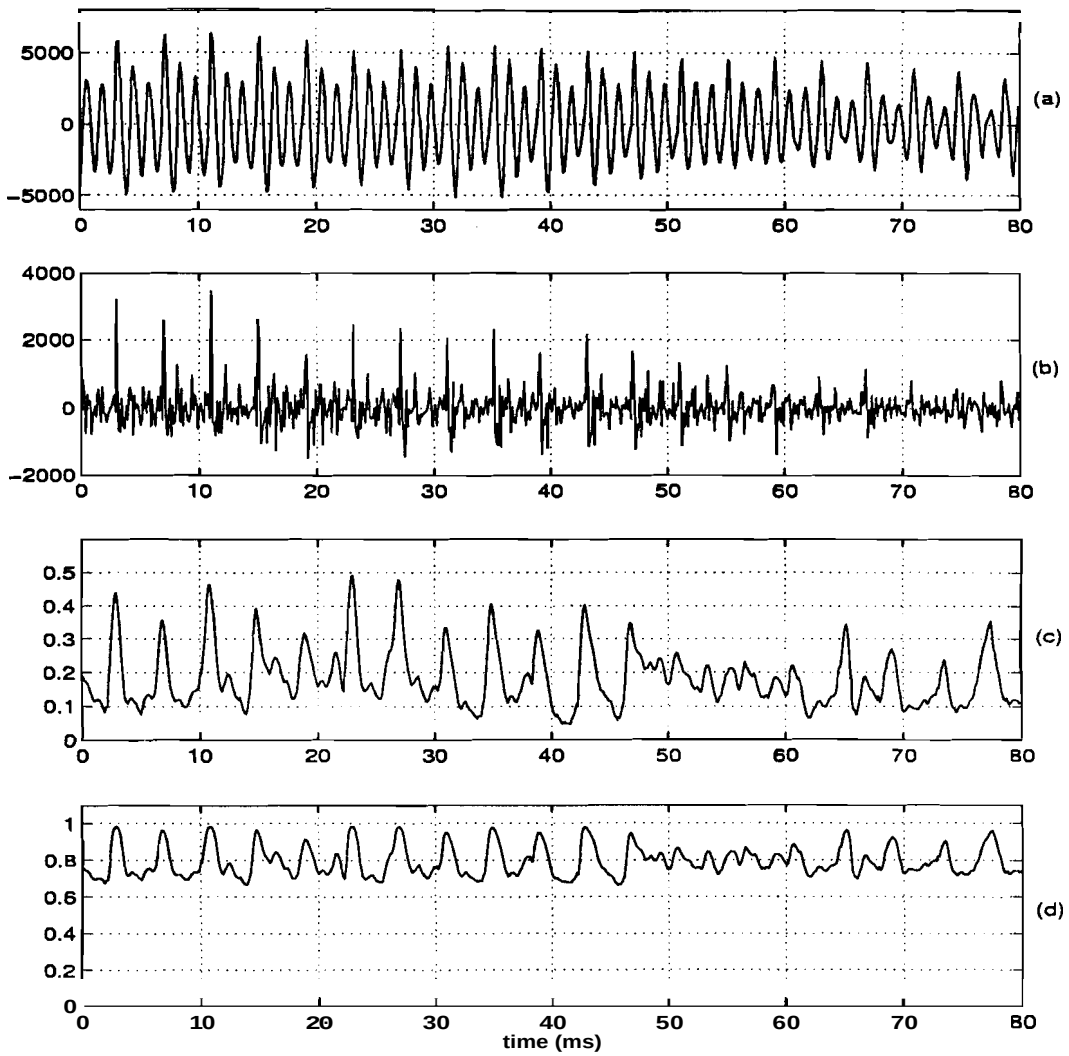


Fig. 5.7: Derivation of the fine weight function. (a) Segment of reverberant speech. (b) LP residual signal. (c) Normalized prediction error. (d) Fine weight function.

second weight function, which we refer to as the fine weight function, is derived from the normalized error by removing the global trend in the normalized error function and then mapping it using the following function:

$$\gamma_n^f = \left(\frac{\gamma_{max}^f - \gamma_{min}^f}{2}\right) \tanh(\alpha_f \, \pi \, \eta_n) + \left(\frac{\gamma_{max}^f + \gamma_{min}^f}{2}\right) \tag{5.4}$$

where γ_n^f is the weight value at the sampling instant n, γ_{max}^f (= 1) is the maximum weight value, γ_{min}^f is the minimum weight value, α_f (=1.5) is a positive constant which decides the slope of the weight function and η_n is the detrended normalized error value

at the sampling instant n. The fine weight function for the segment of the signal in Fig. 5.7(a) is shown in Fig. 5.7(d). The fine weight function provides relative weighting of short segments within a glottal cycle in the high SRR regions. The overall weight function (Fig. 5.5(c)) is obtained by multiplying the gross weight function with the fine weight function. The overall weight function and the LP residual signal are multiplied to derive a modified residual signal. The modified residual signal is used to excite the 5th order all–pole filter to obtain enhanced speech. The filter is updated at every sampling instant.

A comparison of the clean speech waveform and reverberant speech waveform in the voiced regions shows that within a glottal cycle the reverberant speech waveform does not decay as rapidly as the clean speech waveform. This can be seen by a comparison of Figs. 5.8(a) and 5.8(b). Despite the deemphasis of low SRR regions within a glottal cycle by the fine level weight function, the decay of the envelope within a glottal cycle is not restored in the processed speech waveform. Hence, there is a need to increase the flatness by manipulating the spectrum. One way of doing this is to modify the filter coefficients to $a_{kn}\,\xi_n^{-k}$ for $k = 1, 2, \ldots,$ p, where p is the order of the all–pole filter, $\xi_n < 1$ and a_{kn} is the kth LPC at the sampling instant n. The damping factor ξ_n at each sampling instant is varied according to the value of the fine weight function. The value of ξ_n is restricted to the range 0.9–1.0. The modification of LPCs will enable the roots of the all–pole filter move closer to the origin in the z–plane. Due to dependence of ξ_n on the fine weight function, the proposed modification of LPCs is equivalent to damping the resonances of the vocal tract system towards the end of the glottal cycle. The enhanced speech waveform obtained using gross and fine level weighting and spectral damping is shown in Fig. 5.8(c). Within each glottal cycle, the enhanced speech waceform has a rate of decay which is more than that of the reverberant speech

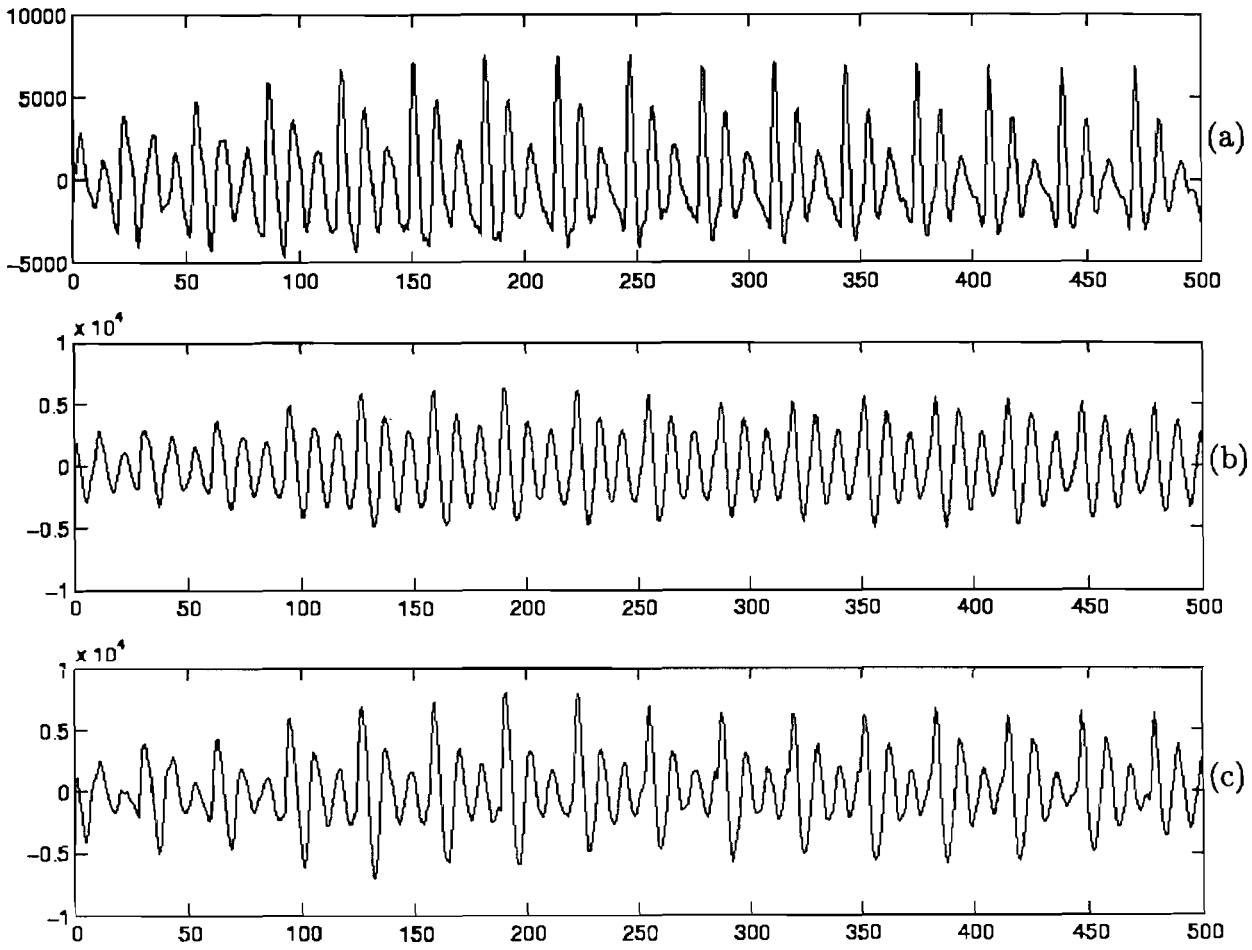


Fig. 5.8: Effect of damping the LPCs on the enhanced speech during synthesis. (a) Clean speech waveform. (b) Reverberant speech waveform. (c) Enhanced speech waveform.

waveform. The algorithm for processing reverberant speech for enhancement is given in Table–5.1.

5.4 EXPERIMENTAL RESULTS

In this section the performance of the proposed method is examined for processing speech data collected under reverberant conditions. For this purpose the speech data was collected at a sampling frequency of 8 kHz in a normal office room, with the microphone placed about 1.5 m away from the speaker. Speech data was also collected simultaneously close to the speaker to obtain a clean speech signal for comparison. The speech signal corresponds to the utterance "***She had your dark suit in greasy wash water all year***" spoken by a male speaker and is taken from the TIMIT database. The preemphasized speech signal was processed using the algorithm given in Table–5.1. The signal waveform and its spectrogram are given in Fig. 5.9 for the clean speech, reverberant speech, and the processed speech. From the spectrograms it is evident that the effects of reverberation are significantly reduced. Perceptually also the processed signal sounds less reverberant than the unprocessed one. The results for different values of the parameters used in the algorithm show that the parameter settings are not very critical. They merely provide a tradeoff between quality and enhancement in the processed signal. Fig. 5.10 shows the short–time (20 ms) spectra for a voiced segment of speech for clean, reverberant and processed speech. The reduction in the dynamic range of the spectra after processing can be seen clearly, especially around the formant regions. Thus the spectral flatness of the clean speech is restored to some extent. For enhancement of noisy speech, on the other hand, one attempts to lower the spectral flatness by increasing the spectral dynamic range [264].

The performance of the method was tested for female voice also. The resulting signal waveforms and spectrograms are shown in Fig. 5.11. The signal corresponds to

Table 5.1: Algorithm for processing reverberant speech for enhancement.

Computation of the gross weight function

- Calculate the linear prediction (LP) residual signal using a speech frame of size 20 ms, Hamming window and a 10th order LP analysis by autocorrelation method.
- Block the LP residual signal into 20 ms frames with 10 ms overlap. Compute an M–bin (M = 7) histogram of the samples in each frame of the LP residual signal.

a Compute the entropy $H_k = -\sum_{i=1}^{M} p_i \log(p_i)$ for the kth frame, where p_i is the estimated probability in the 4th bin of the histogram.

- Compute a smoothed entropy function H_n^s by repeating each entropy value H_k 80 times (corresponds to a frame shift of 10 ms at 8 kHz sampling) and smoothing it with a 600-point mean smoothing filter. This generates a smoothed entropy value at every sampling instant.
- Compute the gross weight function by mapping the smoothed entropy values to weight values using the function

$$\gamma_n^g = \left(\frac{\gamma_{max}^g - \gamma_{min}^g}{2}\right) \tanh\left(-\alpha_g \pi \ (H_n^s - H_0)\right) + \left(\frac{\gamma_{max}^g + \gamma_{min}^g}{2}\right)$$

where γ_n^g is the weight value for sampling instant n, γ_{max}^g (= 1) is the maximum weight value, γ_{min}^g (= 0.05) is the minimum weight value (denoted in Fig. 5.6), α_g (=1.5) is a positive constant which decides the slope of the weight function, H_0 (=1.55) is the entropy value about which the tanh function is anti-symmetric and H_n^s is the smoothed entropy at the sampling instant n.

Computation of the fine weight function

- Calculate the normalized LP error for every sample of the differenced speech signal using a frame of duration 2 ms and 5th order LP analysis using the autocorrelation method.

a Remove the trend in the normalized LP error by smoothing it with a 10 ms Hamming window and subtracting the smoothed function from the normalized LP error. The resulting detrended error function η_n is mapped using the nonlinear function

$$\gamma_n^f = \left(\frac{\gamma_{max}^f - \gamma_{min}^f}{2}\right) \tanh(\alpha_f \pi \eta_n) + \left(\frac{\gamma_{max}^f + \gamma_{min}^f}{2}\right)$$

where γ_n^f is the weight value for sampling instant n, γ_{max}^f (= 1) is the maximum weight value, γ_{min}^f (= 0.6) is the minimum weight value, α_f (=1.5) is a positive constant which decides the slope of the weight furiction and η_n is the detrended error value at the sampling instant n.

Synthesis of enhanced speech

a Compute the overall weight function by multiplying the gross and fine weight functions.

a LP residual signal is derived for every sample using 2 ms frames. The residual signal is multiplied with the overall weight function. The weighted residual signal is passed through the time–varying LP all–pole filter to obtain enhanced speech. At each sampling instant the LPCs are given by $a_{kn}\xi_n^{-k}$, where a_{kn} is the kth LPC at instant n. The damping factor ξ_n, restricted to the range 0.9–1.0, is derived using a linear map of the fine weight function.

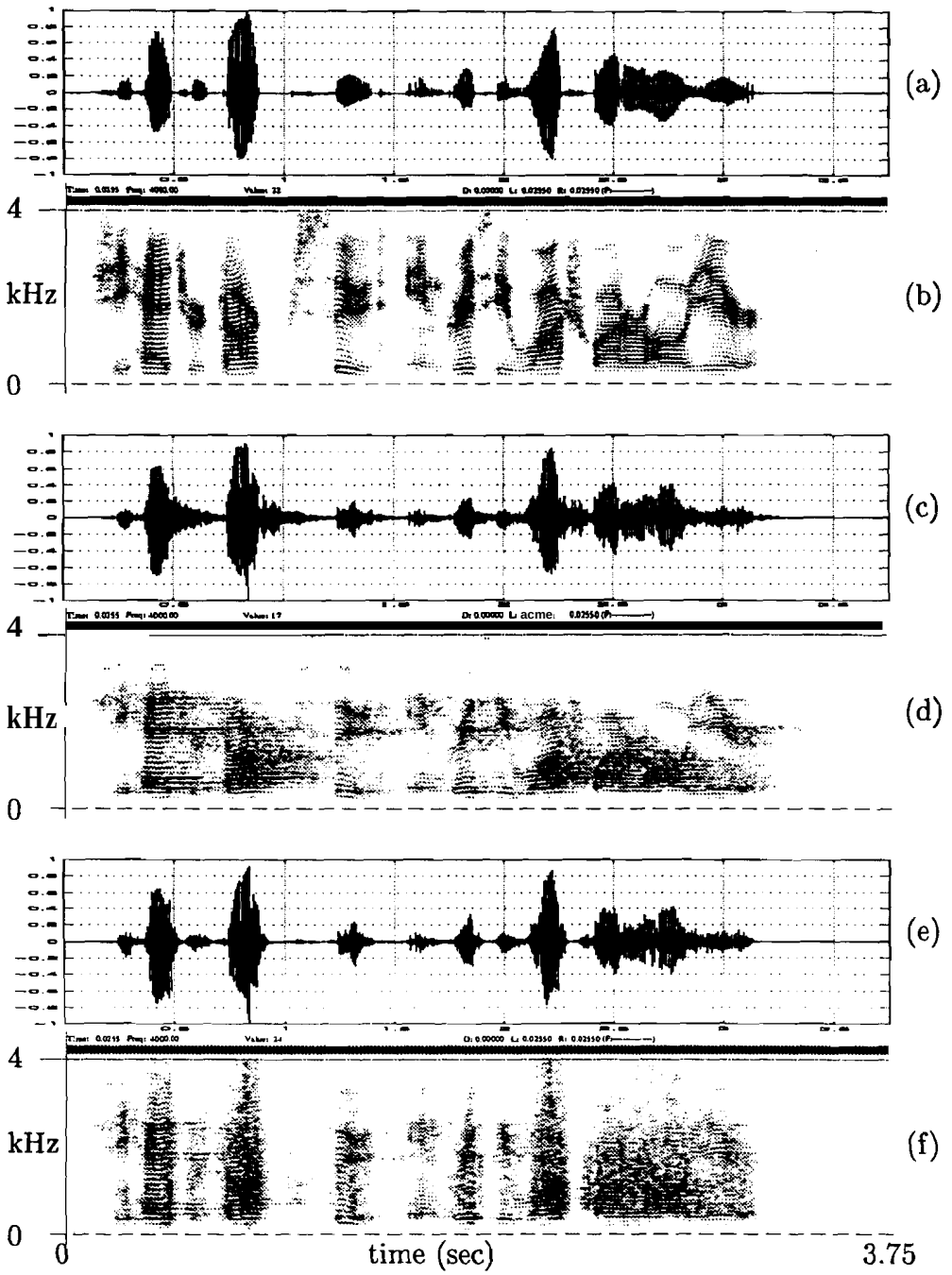


Fig. 5.9: Results of enhancement of reverberant speech of a male voice. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by reverberation. (d) Spectrogram of speech degraded by reverberation. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

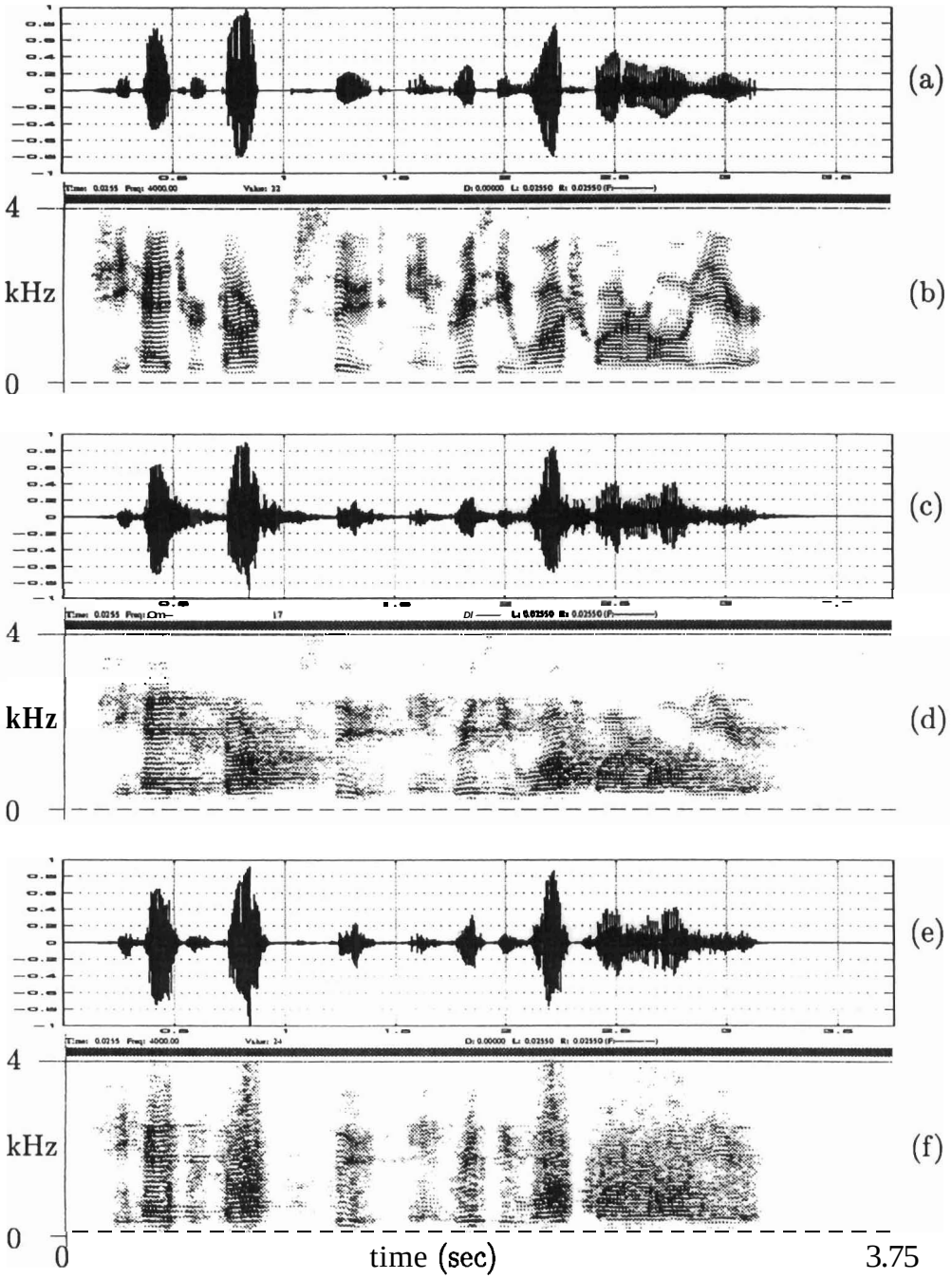


Fig. 5.9: Results of enhancement of reverberant speech of a male voice. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by reverberation. (d) Spectrogram of speech degraded by reverberation. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

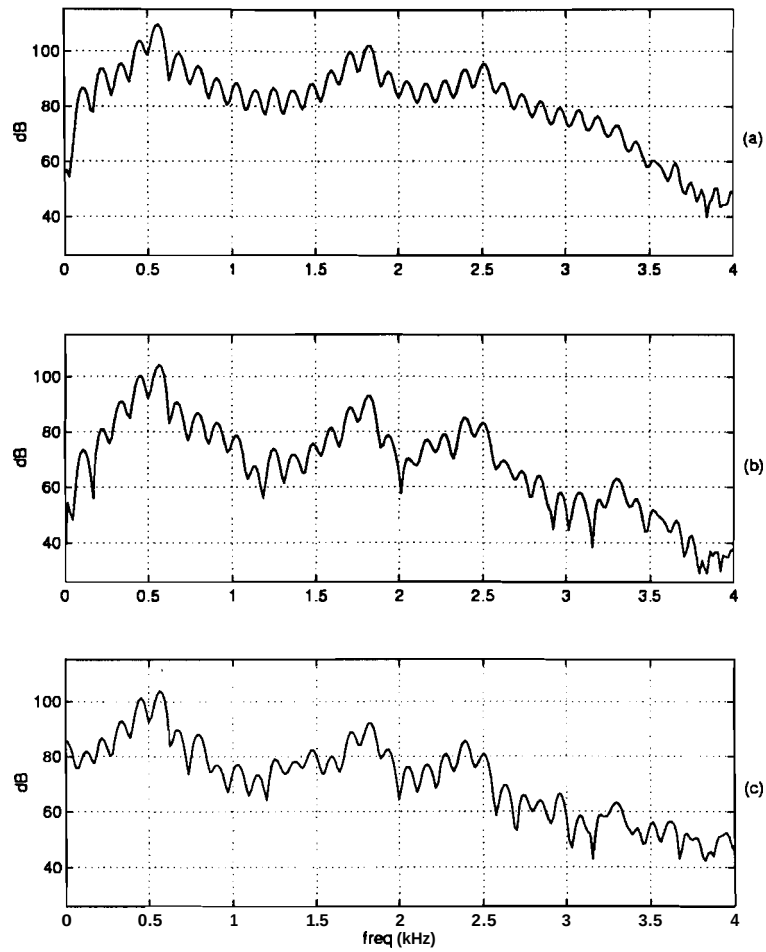


Fig. 5.10: Short–time spectra of a segment of speech for (a) clean speech signal (b) reverberant speech signal (c) processed speech signal.

the sentence *"She had your dark suit in greasy wash"* taken from the TIMIT database. Perceptual listening confirms the improvement in quality of the processed speech for this case also.

In a practical speakerphone–like situation, in addition to degradation due to reverberation, there will be ambient noise also. Fig. 5.12 shows this situation. The reverberant speech signal in Fig. 5.11(c) is corrupted by additive random noise so that the overall SNR is 20 dB. The noise added reverberant speech is shown in Fig. 5.12(c). The processed speech signal is shown in Fig. 5.12(e). The spectrograms for Figs. 5.12(c)

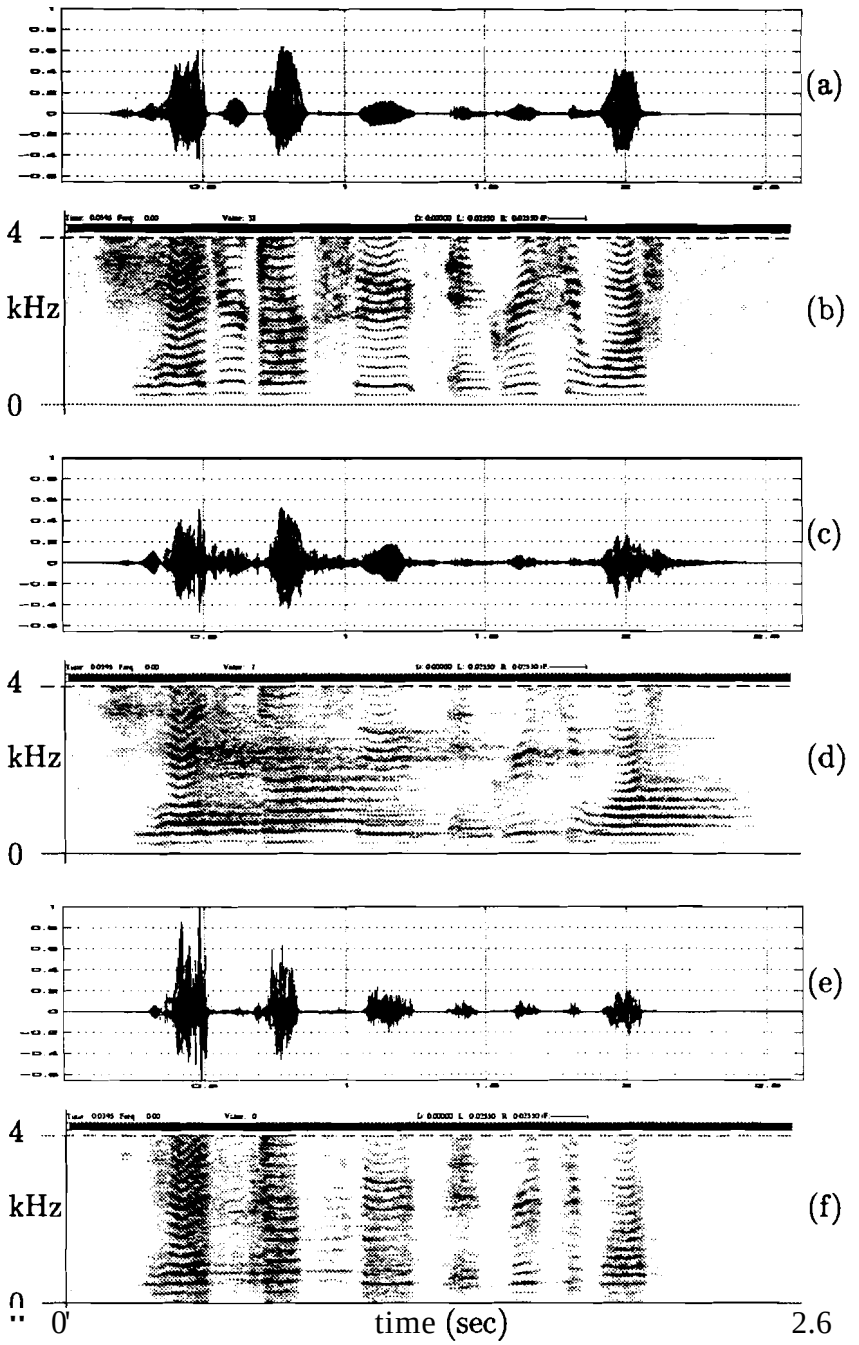


Fig. 5.11: Results of enhancement of reverberant speech of a female voice. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by reverberation. (d) Spectrogram of speech degraded by reverberation. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

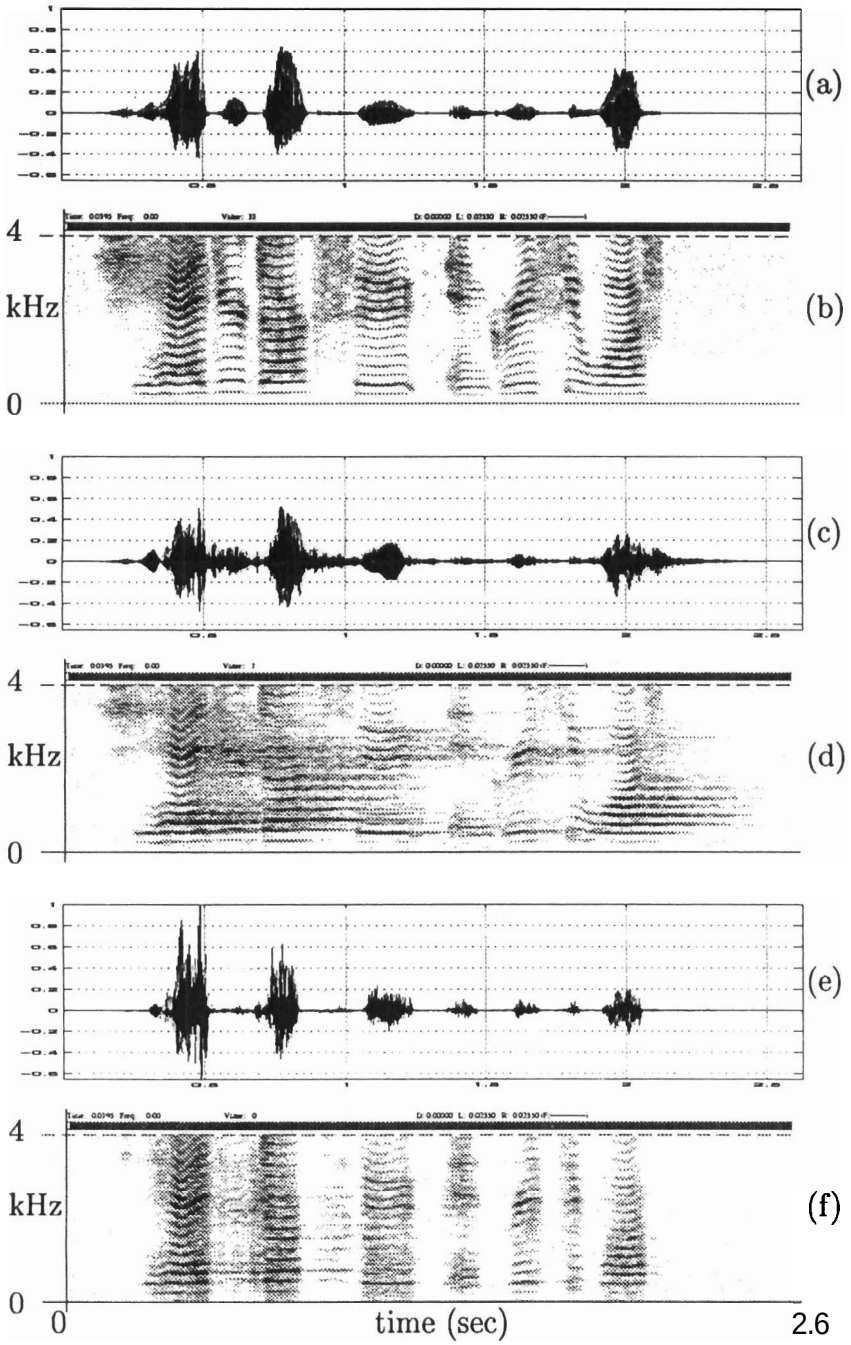


Fig. 5.11: Results of enhancement of reverberant speech of a female voice. (a) Clean speech. (b) Spectrogram of clean speech. ***(c)*** Speech degraded by reverberation. (d) Spectrograrn of speech degraded by reverberation. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

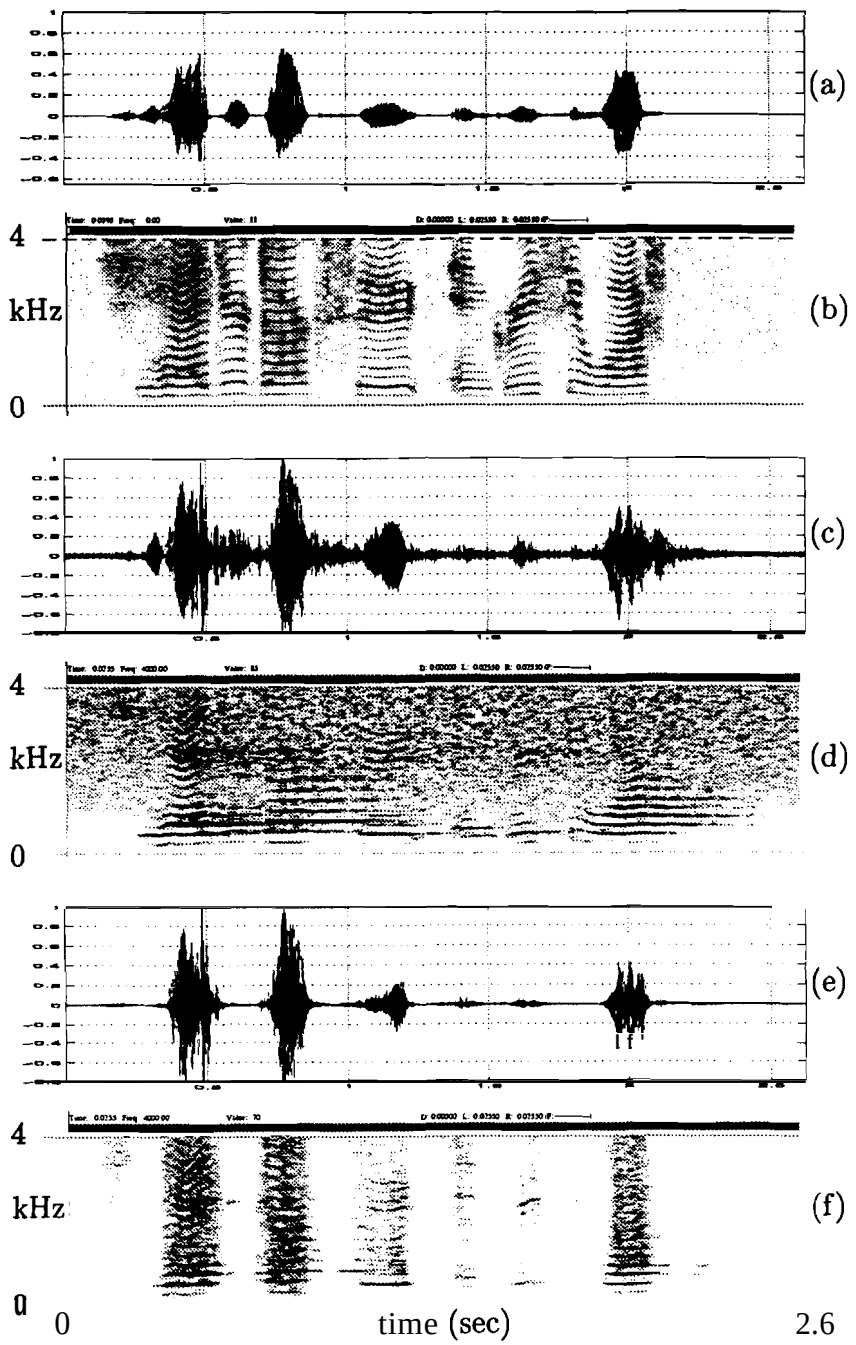


Fig. 5.12: Results of enhancement of speech degraded by reverberation and noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by reverberation and noise (SNR = 20 dB). (d) Spectrogram of speech degraded by reverberation and noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

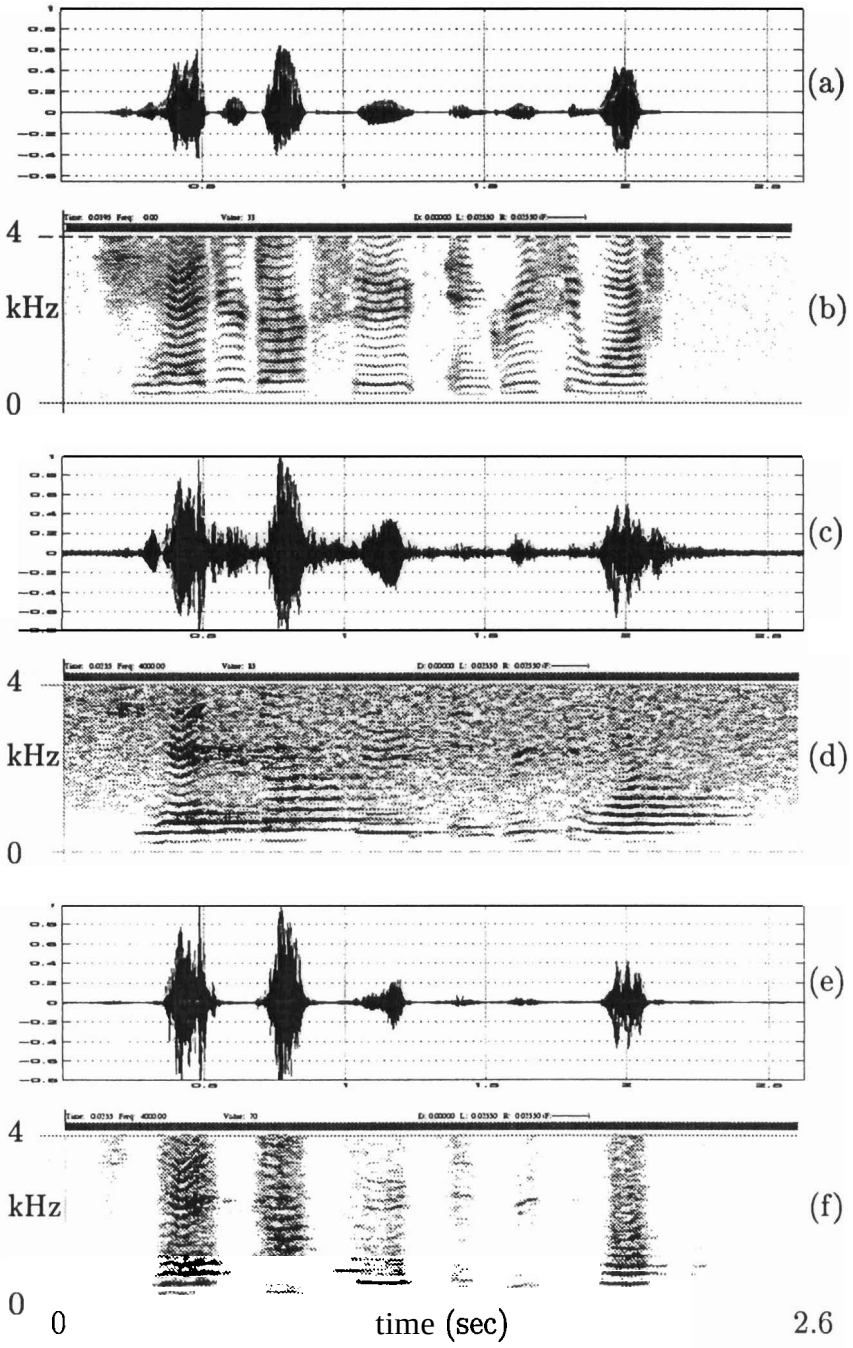


Fig. 5.12: Results of enhancement of speech degraded by reverberation and noise. (a) Clean speech. (b) Spectrogram of clean speech. (c) Speech degraded by reverberation and noise (SNR = 20 dB). (d) Spectrogram of speech degraded by reverberation and noise. (e) Speech processed using the proposed algorithm. (f) Spectrogram of processed speech.

and 5.12(e) are shown in Figs. 5.12(d) and 5.12(f), respectively. The improvement can be clearly seen in the spectrogram in Fig. 5.12(f). We observe that in the silence regions the noise level as well as the reverberation are significantly reduced. This is because the noise increases the randomness in the LP residual signal, more so in the silence regions and hence increases the entropy. Hence, in the silence regions both the reverberation tails and the noise increase the entropy. Thus the gross weight function will have small values in the silence regions. We also observe from the processed signal in Fig. 5.12(e) and the spectrogram in Fig. 5.12(f) that the weak signal segments are severely attenuated, which produces some distortion in the processed speech signal.

There will be some reduction in quality when the proposed algorithm is applied to clean speech. But this reduction in quality is offset by the advantage due to enhancement obtained in processing degraded speech.

5.5 SUMMARY

In this chapter we have proposed a method for processing reverberant speech. The proposed method is based on the knowledge that the speech signal energy fluctuates over a large dynamic range in short segments (2 ms). Thus the SRR varies significantly over different segments of speech. By identifying the high SRR regions, and enhancing such regions at gross level and at fine (within glottal cycle) level one can achieve enhancement of reverberant speech. The processing was done by weighting the LP residual signal. The weight function was derived using the characteristics of the reverberant speech in different regions. The resulting signal shows reduction in the perceived reverberation without significantly affecting the quality. By adjusting the parameters used for obtaining the weight function, the comfort level in the processed signal can be traded with the distortion caused by the manipulation. Thus processing the LP residual signal provides an alternative approach for enhancement of reverberant

speech. A uniform approach for processing reverberant speech as in [85, 95, 261] may not be satisfactory, since the reverberation affects the speech differently in different segments due to nonstationary nature of the speech signal.

The key ideas presented in this chapter are : (a) the need to process different regions of reverberant speech differently, (b) the advantage of manipulating the residual signal samples for enhancement and (c) ability to tune the processing depending on the level of tolerance of distortion vs the desired level of comfort. It is interesting to note that only regions of high SRR need to be processed for enhancement, whereas the low SRR and the reverberant tail regions should be deemphasized to obtain perceptually significant enhancement.

In the next chapter we consider some practical issues that arise in the implementation of the proposed methods for speech enhancement.

Chapter 6

PRACTICAL ISSUES IN THE IMPLEMENTATION OF SPEECH ENHANCEMENT METHODS

In the previous chapters, we have proposed methods for enhancement of speech corrupted by additive noise and room reverberation, based on processing **1–3** ms segments of the speech signal. We have tacitly assumed that we know a priori whether the speech signal is degraded or not, and if degraded, the type of degradation (additive noise or reverberation). However, in practical conditions, we do not know whether the speech signal to be processed is clean or degraded. If it is degraded, it is necessary to know the type of degradation and the level of degradation so that suitable method for enhancement can be employed. In the previous chapters we have also mentioned the limitations of the proposed methods. Firstly, when a clean speech signal is processed using these methods some distortion may result due to the processing. Secondly, when the SNR/SRR is poor ($\leq$ 5 dB), there may not be a significant improvement due to processing since linear prediction analysis performs poorly under such conditions. Therefore, it is preferable not to process the speech signal when there is no degradation or when the SNR/SRR is poor. Thirdly, it is also necessary to identify the type of degradation, i.e., additive noise/reverberation, so that the corresponding enhancement method could be used for processing the speech. Thus there is a need to identify the nature and level of degradation in the speech signal for processing the signal for enhancement. This is the theme of the present chapter.

The modulation spectrum of the speech signal has been proposed as a cue for identifying the type of degradation in [66]. A method based on the histograms of temporal contours of each frequency component in the short–time spectrum of speech has also been proposed [91]. The method based on the modulation spectrum requires several minutes of speech for reliable estimation of the modulation spectrum. In this chapter we propose a method based on the normalized prediction error of speech for identification of type and level of degradation. The proposed method is based on the studies made on normalized prediciton error in the previous chapter (see Section 5.2 and Fig. **5.3).** The method typically requires only **3** s of speech for processing.

This chapter is organized as follows. In Section 6.1 the basis for the proposed method for identification of the type and level of degradation is presented. In Section **6.2** the characteristics of the normalized prediction error of speech degraded under practical conditions is illustrated. In Section **6.3** the parameter settings for the methods for speech enhancement proposed in this thesis are given for different levels of degradation. The results presented in this chapter are summarized in Section **6.4.**

6.1 NATURE OF THE NORMALIZED PREDICTION ERROR

In the case of noisy speech, generally speech samples are correlated and the noise samples are less correlated. Additive noise reduces the correlation between the samples of the speech signal. As the noise level increases, the low amplitude segments will progressively submerge in noise. The normalized prediction error (η) for a noisy signal is always greater than that of **a** clean signal for a given prediction order p. The prediction error rises with increasing noise level. This can be shown as follows. Recalling the notation in equation (5.2), let the noisy speech signal for two different noise levels

be represented as

$$y_1(n) = s(n) + w_1(n) \tag{6.1}$$

$$y_2(n) = s(n) + w_2(n) \tag{6.2}$$

where $s(n)$ is the clean speech signal and $w_1(n), w_2(n)$ are the additive noise components, such that the SNR of $y_2(n)$ is lower compared to that of $y_1(n)$. Therefore, the energy of $w_2(n)$ is more than that of $w_1(n)$. The expression in (2.20) for the normalized prediction errors of y_1(n),$y_2(n)$ can be rewritten as

$$\eta_{y1} = 1 - \frac{\mathbf{r}_{yy1}^T \mathbf{R}_{yy1}^{-1} \mathbf{r}_{yy1}}{r_{yy1}(0)} \tag{6.3}$$

$$\eta_{y2} = 1 - \frac{\mathbf{r}_{yy2}^T \mathbf{R}_{yy2}^{-1} \mathbf{r}_{yy2}}{r_{yy2}(0)} \tag{6.4}$$

where $\mathbf{R}_{yy1}, \mathbf{R}_{yy2}, \mathbf{r}_{yy1}$ and $\mathbf{r}_{yy2}$ are similar to the quantities defined in (2.13) arid (2.14). Assuming that the speech signal and noise are uncorrelated, from (6.1) and (6.2) we have

$$\mathbf{r}_{yy1} = \mathbf{r}_{ss} + \mathbf{r}_{ww1} \tag{6.5}$$

$$\mathbf{r}_{yy2} = \mathbf{r}_{ss} + \mathbf{r}_{ww2} \tag{6.6}$$

$$\mathbf{R}_{yy1} = \mathbf{R}_{ss} + \mathrm{R\,w\,w\,l} \tag{6.7}$$

$$\mathbf{R}_{yy2} = \mathbf{R}_{ss} + \mathbf{R}_{ww2} \tag{6.8}$$

where $\mathbf{R}_{ww1}$, $\mathbf{R}_{ww2}$, $\mathbf{r}_{ww1}$ and $\mathbf{r}_{ww2}$ are similar to $\mathbf{R}_{yy1}$ and $\mathbf{r}_{yy1}$. Further assuming that $w_1(n)$ and $w_2(n)$ are white arid Gaussian distributed with variances σ_{w1}^2 and σ_{w2}^2, respectively, $\mathbf{r}_{ww1}$ and $\mathbf{r}_{ww2}$ are identically zero.

$$\mathbf{r}_{ww1} = \mathbf{0} \tag{6.9}$$

$$\mathbf{r}_{ww2} = \mathbf{0} \tag{6.10}$$

The autocorrelation matrices for $w_1(n)$ and $w_2(n)$ reduce to a diagonal form.

$$\mathbf{R}_{ww1} = \sigma_{w1}^2 \mathbf{I}_p \tag{6.11}$$

$$\mathbf{R}_{ww2} = \sigma_{w2}^2 \mathbf{I}_p \tag{6.12}$$

Using expressions (6.5) – (6.8) and (6.9) – (6.12) in (6.3) and (6.4), we have

$$\eta_{y1} = 1 - \frac{\mathbf{r}_{ss}^T \left(\mathbf{R}_{ss} + \sigma_{w1}^2 \mathbf{I}_p\right)^{-1} \mathbf{r}_{ss}}{r_{ss}(0) + \sigma_{w1}^2} \tag{6.13}$$

$$\eta_{y2} = 1 - \frac{\mathbf{r}_{ss}^T \left(\mathbf{R}_{ss} + \sigma_{w2}^2 \mathbf{I}_p\right)^{-1} \mathbf{r}_{ss}}{r_{ss}(0) + \sigma_{w2}^2} \tag{6.14}$$

Assuming $\mathbf{R}_{ss}$ is full rank and since it is Toeplitz symmetric, it is also positive definite [268, 269]. Therefore, $\mathbf{R}_{ss}$ admits an eigen–decomposition

$$\mathbf{R}_{ss} = \mathbf{U}_{ss} \mathbf{\Lambda}_{ss} \mathbf{U}_{ss}^T \tag{6.15}$$

where

$$\mathbf{\Lambda}_{ss} = \text{diag}[\lambda_1 \ \lambda_2 \ \cdots \ \lambda_p] \tag{6.16}$$

is a diagonal matrix of p distinct positive eigenvalues, and $\mathbf{U}_{ss}$ is a unitary matrix having the eigenvectors of $\mathbf{R}_{ss}$ as its columns. Since $\mathbf{U}_{ss}$ is unitary, the inverse of $\mathbf{R}_{ss}$ is given by

$$\mathbf{R}_{ss}^{-1} = \mathbf{U}_{ss} \mathbf{\Lambda}_{ss}^{-1} \mathbf{U}_{ss}^T \tag{6.17}$$

where

$$\mathbf{\Lambda}_{ss}^{-1} = \text{diag}\left[\frac{1}{\lambda_1} \ \frac{1}{\lambda_2} \ \cdots \ \frac{1}{\lambda_p}\right] \tag{6.18}$$

Using (6.17) and noting that $\mathbf{U}_{ss}$ is unitary, we have [270]

$$\begin{aligned} \left(\mathbf{R}_{ss} + \sigma_{w1}^2 \mathbf{I}_p\right)^{-1} &= \left(\mathbf{U}_{ss} \mathbf{\Lambda}_{ss} \mathbf{U}_{ss}^T + \sigma_{w1}^2 \mathbf{I}_p\right)^{-1} \\ &= \left(\mathbf{U}_{ss} \left[\mathbf{\Lambda}_{ss} + \sigma_{w1}^2 \mathbf{I}_p\right] \mathbf{U}_{ss}^T\right)^{-1} \\ &= \mathbf{U}_{ss} \left[\mathbf{\Lambda}_{ss} + \sigma_{w1}^2 \mathbf{I}_p\right]^{-1} \mathbf{U}_{ss}^T \end{aligned} \tag{6.19}$$

Similarly,

$$\left(\mathbf{R}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1} \;=\; \mathbf{U}_{ss}\, \left[\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right]^{-1}\, \mathbf{U}_{ss}^T \tag{6.20}$$

where

$$\left[\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right]^{-1} \;=\; \text{diag}\left[\frac{1}{\lambda_1 + \sigma_{w1}^2}\;\; \frac{1}{\lambda_2 + \sigma_{w1}^2} \cdots \frac{1}{\lambda_p + \sigma_{w1}^2}\right] \tag{6.21}$$

$$\left[\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right]^{-1} \;=\; \text{diag}\left[\frac{1}{\lambda_1 + \sigma_{w2}^2}\;\; \frac{1}{\lambda_2 + \sigma_{w2}^2} \cdots \frac{1}{\lambda_p + \sigma_{w2}^2}\right] \tag{6.22}$$

Using (6.19) and (6.20) in (6.13) and (6.14), we have

$$\eta_{y1} \;=\; 1 \;-\; \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T\; \left(\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right)^{-1}\; \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{r_{ss}(0) + \sigma_{w1}^2} \tag{6.23}$$

$$\eta_{y2} \;=\; 1 \;-\; \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T\; \left(\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1}\; \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{r_{ss}(0) + \sigma_{w2}^2} \tag{6.24}$$

Subtracting (6.23) from (6.24), we have

$$\begin{aligned} \eta_{y2} \;-\; \eta_{y1} \;=\; & \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T\; \left(\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right)^{-1}\; \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{r_{ss}(0) + \sigma_{w1}^2} \\ & - \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T\; \left(\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1}\; \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{r_{ss}(0) + \sigma_{w2}^2} \end{aligned} \tag{6.25}$$

After some algebraic manipulations, (6.25) can be written as

$$\begin{aligned} \eta_{y2} \;-\; \eta_{y1} \;=\; & r_{ss}(0)\, \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T \left[\left(\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right)^{-1} - \left(\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1}\right] \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{\left(r_{ss}(0) + \sigma_{w1}^2\right)\; \left(r_{ss}(0) + \sigma_{w2}^2\right)} \\ & + \frac{\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)^T \left[\sigma_{w2}^2 \left(\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right)^{-1} - \sigma_{w1}^2 \left(\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1}\right] \left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)}{\left(r_{ss}(0) + \sigma_{w1}^2\right)\; \left(r_{ss}(0) + \sigma_{w2}^2\right)} \end{aligned} \tag{6.26}$$

In the above expression (6.26) for $\eta_{y2} - \eta_{y1}$, the denominator terms are always positive. The numerator terms are quadratic forms in $\left(\mathbf{U}_{ss}^T \mathbf{r}_{ss}\right)$. The diagonal matrix $\left[\left(\mathbf{\Lambda}_{ss} + \sigma_{w1}^2\, \mathbf{I}_p\right)^{-1} - \left(\mathbf{\Lambda}_{ss} + \sigma_{w2}^2\, \mathbf{I}_p\right)^{-1}\right]$ can be written as

$$\text{diag}\left[\frac{\sigma_{w2}^2 - \sigma_{w1}^2}{\left(\lambda_1 + \sigma_{w1}^2\right)\left(\lambda_1 + \sigma_{w2}^2\right)} \;\cdots\; \frac{\sigma_{w2}^2 - \sigma_{w1}^2}{\left(\lambda_p + \sigma_{w1}^2\right)\left(\lambda_p + \sigma_{w2}^2\right)}\right] \tag{6.27}$$

Similarly, the diagonal matrix $\left[\sigma_{w2}^2 (\mathbf{\Lambda}_{ss} + \sigma_{w1}^2 \mathbf{I}_p)^{-1} - \sigma_{w1}^2 (\mathbf{\Lambda}_{ss} + \sigma_{w2}^2 \mathbf{I}_p)^{-1}\right]$ can be written as

$$\text{diag}\left[\frac{(\sigma_{w2}^2 - \sigma_{w1}^2)(\lambda_1 + \sigma_{w1}^2 + \sigma_{w2}^2)}{(\lambda_1 + \sigma_{w1}^2)(\lambda_1 + \sigma_{w2}^2)} \cdots \frac{(\sigma_{w2}^2 - \sigma_{w1}^2)(\lambda_p + \sigma_{w1}^2 + \sigma_{w2}^2)}{(\lambda_p + \sigma_{w1}^2)(\lambda_p + \sigma_{w2}^2)}\right] \quad (6.28)$$

Since the energy in $w_2(n)$ has been assumed to be more than that of $w_1(n)$ $(\sigma_{w2}^2 > \sigma_{w1}^2)$, the diagonal matrices *(6.27)* and *(6.28)* are positive definite. Therefore the quadratic forms in the numerator terms of *(6.26)* are always positive. Hence,

$$\eta_{y2} > \eta_{y1} \qquad \text{when} \qquad a_{,,}^2 > \sigma_{w1}^2$$

By following the steps in the analysis presented above, it is easy to show that the normalized prediction error of noisy speech $y_1(n)$ is higher than that of clean speech $s(n)$, i.e.,

$$\eta_{y2} > \eta_{y1} > \eta_s \quad (6.29)$$

where η_s is the normalized prediction error of clean speech signal $s(n)$. If the additive noise component $w(n)$ is colored, the above analysis is still valid, provided that a noise whitening transformation [49] is applied to the noisy speech signal $y(n)$.

Reverberant speech can be interpreted as the convolution of the room impulse response and the anechoic speech signal $s(n)$. Therefore, there is a multiplicative effect of the frequency response of the room on the speech spectrum. Multiplication of the two spectra produces larger dynamic range and hence reduces the spectral flatness. Therefore, the normalized prediction error for the reverberant speech is generally lower than for the clean speech. This is also the observation arrived at by studying real reverberant data. However, the reduction of the normalized prediction error for reverberant speech cannot easily be shown analytically. The degrading component in *(5.1)* (see Chapter *5)* is dependent on the signal as well as on the characteristics of

the room. Therefore, the expression for normalized error in the case of reverberant speech becomes involved. In the following section, we present studies made on some real reverberant data.

6.2 STUDIES ON SPEECH DEGRADED IN PRACTICAL CONDITIONS

The normalized prediction error is relatively high in case of noisy speech and low in case of reverberant speech, compared to that of clean speech. This can be seen in Fig. 5.3 of Chapter 5. In Table–6.1 the variation of the averaged normalized prediction error for different types and levels of degradation is given.

Table 6.1: Variation of the averaged normalized prediction error depending upon the type and level of degradation.

Case	Type of degradation	Level of degradation	Range of $\bar{\eta}$
1	additive noise	5 dB	0.50 – 0.49
2	additive noise	10 dB	0.45 – 0.42
3	additive noise	20 dB	0.39 – 0.32
4	additive noise	30 dB	0.28 – 0.25
5	clean		0.26 – 0.22
6	reverberation	0.6 m	0.21 – 0.14
7	reverberation	1.2 m	0.19 – 0.10
8	reverberation	1.5 m	0.15 – 0.08
9	reverberation	2.1 m	0.15 – 0.07

The averaged normalized prediction error $(\bar{\eta})$ was computed for four different speakers (two male and two female) for the cases of clean, noisy and reverberant speech. In Fig. 6.1 $\bar{\eta}$ is shown for the four different speakers mentioned above. The figure shows $\bar{\eta}$ for the types and levels of degradation (cases 1–9) in Table–6.1. The two dotted horizontal lines in Fig. 6.1 indicate the range of $\bar{\eta}$ for clean speech. The sentences used for the study were taken from the TIMIT database. The duration of the sentences is about 3 s each. The averaged normalized prediction error was obtained by computing

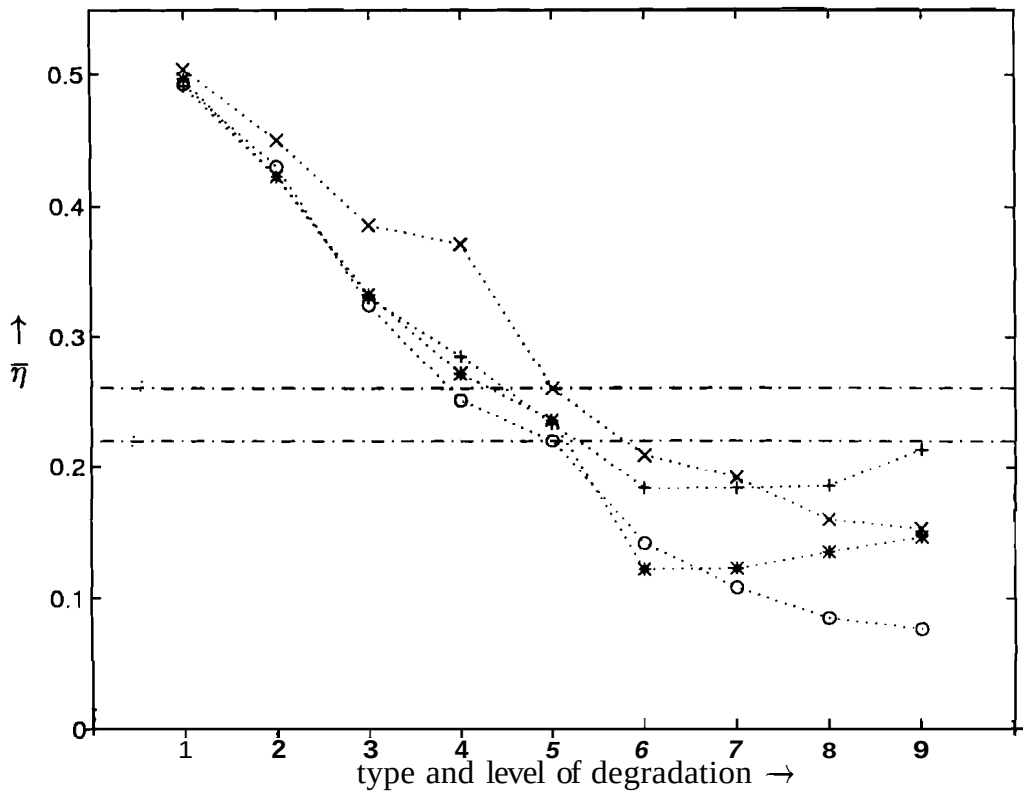


Fig. 6.1: The averaged normalized prediction error $\bar{\eta}$ plotted as a function of type and level of degradation for four different speakers, 2 male and 2 female.

the mean of the prediction errors obtained in 20 ms frames using a 12th order autocorrelation LP analysis. The frames were overlapped by 10 ms. The speech signal was preemphasized before processing. The variance (σ_{η}^2) of the normalized prediction error for each of the four speakers for the different types and levels of degradation is shown in Fig. 6.2. It can be seen from the figure that the variance also exhibits a downward trend from left to right, but is not as consistent as $\bar{\eta}$.

The noisy speech data was generated by adding the noise generated by an airconditioner to the clean speech signal such that an average SNR of 5 dB, 10 dB, 20 dB and 30 dB are obtained. The reverberant speech data was obtained by playing out the clean speech from a loudspeaker in a quiet room and recording the signal with a micro-

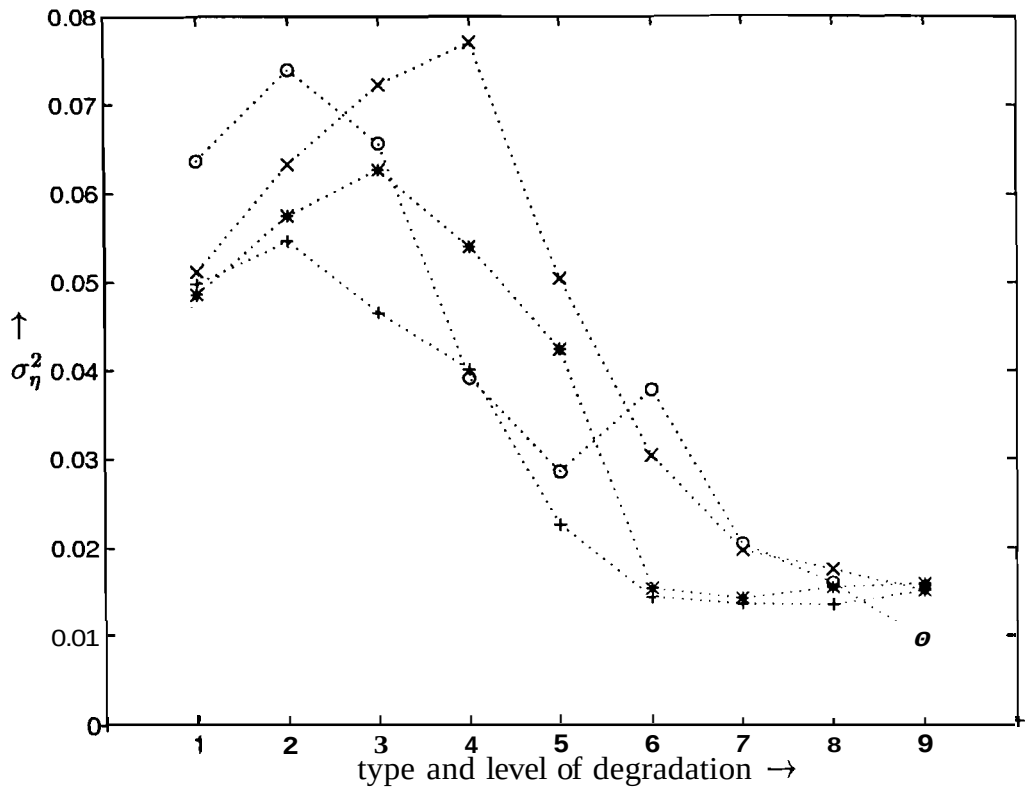


Fig. 6.2: The variance of normalized prediction error σ_η^2 plotted as a function of type and level of degradation for four different speakers, 2 male and 2 female.

phone. The microphone was placed at four different distances from the speaker: 0.6m, 1.2m, 1.5m and 2.1m. Thus the criterion used for judging the level of reverberation is the distance of the microphone from the speaker. We observe in Fig. 6.1 that the curves for all the four speakers are similar. Table–6.1 shows the range of the averaged normalized prediction error $(\bar{\eta})$ for each of the cases. This is used for identifying the level of degradation and is dealt with in the next section.

6.3 PARAMETER SETTINGS FOR DIFFERENT CONDITIONS OF DEGRADATION

From Table–6.1 the type of degradation (noise or reverberation) can be identified by computing the averaged normalized prediction error from about 3 s of data and

comparing it with the thresholds of 0.22 and 0.26. If it is less than 0.22, the source of degradation is reverberation and if it is greater than 0.26, it is additive random noise. At an average SNR of 30 dB, the noise is barely perceptible and needs no processing. In the methods for speech enhancement proposed in this thesis, the setting of the parameters has to be changed for the more severe levels of degradation, i.e., 1, 2 and 8. For the case 9, the reverberation swamps the direct path speech signal. Therefore any processing is unwarranted as the processing may distort the speech further. The parameter settings for the different cases 1, 2, 3 and 4 for processing noisy speech are given in Table–6.2. The parameter settings for cases 6, 7 and 8 for processing reverberant speech are given in Table–6.3. These settings of the parameters were arrived at by conducting several informal listening tests and choosing those values which provided reduction in the annoyance due to degradation without introducing perceptible distortion. The steps in the proposed method for identifying the type and level of degradation are given in Table–6.4.

Table 6.2: Different settings of the parameters for the mapping functions depending upon the level of additive noise.

Case	SNR (dB)	ζ^m_{max}	ζ^m_{min}	α_g	ζ_0	γ^f_{max}	γ^f_{min}	α_f
1	5	1.0	0.3	0.75	1.5	1.0	0.6	0.75
2	10	1.0	0.1	0.75	1.5	1.0	0.1	0.75
3	20	1.0	0.3	0.75	1.5	1.0	0.1	0.75
4	30	1.0	0.5	0.75	1.5	1.0	0.6	0.75

Table 6.3: Different settings of the parameters for the mapping functions depending upon the level of reverberation.

Case	distance from microphone	γ_{max}^{g}	γ_{min}^{g}	α_g	H_0	γ_{max}^{f}	γ_{min}^{f}	α_f
6	0.6 m	1.0	0.001	1.5	1.55	1.0	0.4	2.0
7	1.2 m	1.0	0.010	1.5	1.65	1.0	0.4	2.0
8	1.5 m	1.0	0.100	1.5	1.65	1.0	0.4	2.0

Table 6.4: Algorithm for determining the type and level of degradation.

- Compute the normalized prediction error using 12th order autocorrelation LP analysis in each 20 ms Hamming windowed frame. The frames are overlapped by 10 ms. The speech signal is preemphasized before processing.
- Obtain the averaged normalized prediction error ($\overline{\eta}$) by computing the mean of the errors obtained in all the frames.
- Identify the type of degradation using the decision logic given below:
  - If $\overline{\eta} > 0.26$, then the degradation is additive noise. Apply the algorithm for enhancement of noisy speech (Table-4.1).
  - If $\overline{\eta} < 0.22$, then the degradation is reverberation. Apply the algorithm for enhancement of reverberant speech. (Table–5.1).
- For the computed value of $\overline{\eta}$, perform a table-lookup in Table-6.1 to determine the level of degradation.
- Depending on the level of degradation choose the appropriate parameter settings from Tables–6.2 and 6.3.
- Perform enhancement using the appropriate method of enhancement.

6.4 SUMMARY

In this chapter we have considered some practical issues in the implementation of the methods proposed in this thesis for speech enhancement. The averaged normalized prediction error obtained over an interval of about 3 s of speech has been proposed as a criterion for identifying whether the source of degradation is additive random noise

or reverberation. The averaged normalized prediction error was also used to estimate the level of degradation. One of the limitations of this method is that when noise and reverberation are simultaneously present, the criterion based on normalized prediction error is not useful. It is also not useful when the degradation is due to the speech of a competing speaker. The degradation due to the speech of a competing speaker too reduces the averaged normalized error to a value below that of clean speech. But the amount of reduction is dependent on the characteristics of the speech of the competing speaker (e.g., voiced, unvoiced).

Chapter 7

ROBUSTNESS OF GROUP–DELAY–BASED METHOD FOR EXTRACTION OF INSTANTS OF SIGNIFICANT EXCITATION

In the previous chapters, we have seen the effectiveness of subsegmental analysis of speech for capturing rapid changes in the cllaracteristics of the vocal tract system and for speech enhancement. In subsegmental analysis, the positioriing of the short (1–3 ms) analysis window is crucial for deriving reliable estimates of the vocal tract characteristics from the speech signal. In this chapter, we investigate the robustness of a group–delay–based method advanced earlier for deriving the instants of significant excitation and suggest some refinements to the method. The instants of significant excitation correspond to the instants of glottal closure for voiced speech. The method uses the properties of the global phase characteristics of minimum phase signals. Robustness of the method against noise and distortion is due to the fact that the average phase characteristics of a signal is determined mainly by the strength of the excitation impulse. The strength of excitation is determined by the energy of the residual error signal around the instant of excitation. We propose a measure for the strength of the excitation based on Frobenius norm of the preemphasized signal. The robustness of the group–delay–based method is illustrated for speech under different types of degradations and for speech from different speakers.

The organization of the chapter is as follows. In Section 7.2, the modified group–

delay–based method for the extraction of the instants of significant excitation is briefly reviewed. Some refinements of the method are also discussed. Since robustness of the method is due to the strength of the excitation, we discuss in Section 7.3 the need for a measure for the strength of excitation, and propose a measure based on the Frobenius norm of the prediction matrix of differenced speech signal. In Section 7.4, the robustness of the group–delay–based method is discussed for speech signals corrupted by additive noise arid reverberation. In Section 7.5, we study the performance of the method for different types of speech data with natural degradations.

7.1 IMPORTANCE OF INSTANTS OF SIGNIFICANT EXCITATION

Speech is produced as a result of excitation of a time–varying vocal tract system. In the case of voiced speech the excitation is due to the quasiperiodic airflow resulting from the opening and closing of the glottis in each glottal cycle. Within a glottal cycle, the vocal tract system is strongly excited around the instant of glottal closure. We refer to this instant as an *instant of significant excitation* in this chapter. Strong excitations such as at the release of unvoiced or voiced stops can also be considered as instants of significant excitation.

Instants of significant excitation are useful in several situations, for example, for accurate analysis and synthesis of speech [7, 10, 249]. For noisy speech, knowledge of the instants of significant excitation helps in performing robust spectrum analysis. This is because a short (1–3 ms) segment in the voiced speech signal immediately after the instant of significant excitation usually corresponds to a high signal–to–noise ratio (SNR) portion of the speech within a glottal cycle [1]. Hence, analysis of these short segments may yield better estimates of the characteristics of the vocal tract system.

Determination of the instants of significant excitation is difficult even for clean speech. In the case of strong voicing, due to sharp glottal closure in the voiced speech,

the instant of significant excitation can be perceived even in the presence of noise. But in the case of voiced sounds where the glottal closure is gradual, the instant of glottal closure is difficult to perceive or identify, especially if the signal is corrupted by noise. Reliable identification of the instant of significant excitation depends on the strength of the excitation.

Several methods have been proposed in the literature for determining the instants of significant excitation [1, 192, 271–274]. Most of them depend on either the short-time energy of the speech signal or on the linear prediction (LP) residual signal. These methods are based on block-data processing, and hence there is some ambiguity in the locations of the instants. Moreover, the performance of these methods generally deteriorates when the speech signal is corrupted by noise and distortion.

In [202] a method was proposed for the extraction of the instants of significant excitation. The method is based on the fact that the average value of the group-delay function of a signal within an analysis frame corresponds to the location of the significant excitation within the frame. An improved method based on the computation of the group-delay function directly from the speech signal was proposed in [203]. In this work, we propose further refinements of the method and then discuss the robustness of the group-delay-based method. Even though it was mentioned in [202] that the method would be sensitive to additive noise, the studies in this chapter show that the group-delay-based method is indeed robust against additive random noise and channel distortions. This is because it is the strength of the excitation that determines the robustness of the method against noise.

7.2 DETERMINATION OF INSTANTS OF SIGNIFICANT EXCITATION

In this section, we briefly present the groupdelay–based method proposed in [202, 203] for determining the instants of significant excitation from speech signals, and propose some refinements to the method. The method is based on the global phase characteristics of minimum phase signals. Since the average groupdelay of a minimum phase system is zero [275], the average slope of the phase spectrum of the impulse response of the system corresponds to the location of the excitation impulse within the analysis frame [202]. In practice, the computed phase spectrum or the group–delay function depends on the window function used for analysis. To reduce the effects of the window function on the estimated group–delay function, it is preferable to compute the group–delay function from the LP residual signal. The residual signal is also preferable because some characteristics of the glottal source can be seen better in the residual error signal than in the speech signal. The average slope of the phase spectrum of the speech signal is the same for the residual signal also, because the inverse filter of the LP analysis is a minimum phase system [40]. The residual signal is derived by inverse filtering the speech signal, and the inverse filter is obtained using LP analysis. For LP analysis, a frame size of about 25 ms for every 10 ms may be chosen [202, 203]. The instants of significant excitation can be derived from the LP residual signal as follows [203]. Around each sampling instant a 10 ms segment of the LP residual signal is considered and the group–delay function is computed using the formula [276]

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)} \tag{7.1}$$

where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(\omega) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the windowed residual signal $x(n)$ and $y(n) = nx(n)$, respectively. The group–delay function is smoothed using a 3–point median filter [277, 278] to remove any discon-

tinuities in the group–delay function. The negative of the average of the smoothed group–delay function is called *phase slope*. The phase slope value is computed at each sampling instant to obtain the *phase slope function*. If the instant of significant excitation within a frame is at the midpoint of the frame, then the phase slope is zero. Therefore the positive zero–crossings of the phase slope function correspond to the instants of significant excitation. Short–time (1–3 ms) energy of the LP residual signal around the instant can be used to represent the strength of excitation associated with the instant [202,203]. Figs. 7.1(a)–(d) show a segment of speech signal, the LP residual signal, the phase slope function and the extracted instants with estimated strengths, respectively. The speech signal shown corresponds to the utterance */dz ua/*, where */dz/* is a voiced palatal affricate as in ***Julie.***

Sometimes the LP residual signal may contain some spurious impulses which may result in wrong estimation of the instants of significant excitation, as can be seen in Fig. 7.1(d), where the strengths are computed using the short–time energy of the residual signal centered around the estimated instants of significant excitation. The effect of these spurious impulses can be reduced by enhancing the region around the instants of significant excitation relative to other regions in the LP residual signal. This can be accomplished by deriving a weight function for the LP residual signal. The different steps involved in deriving such a weight function are illustrated for a signal segment in the transition region of a CV */dz a/* (see Fig. 7.2(a)). For clarity only a 100 ms segment of the signal has been chosen for illustration. The weight function (shown in Fig. 7.2(d)) is derived by smoothing the LP residual signal with a Hamming window of duration 0.75 ms (8 samples at 11 kHz sampling rate). The smoothing reduces the noise fluctuations in the residual signal. The smoothed residual signal is shown in Fig. 7.2(c). The short–time energy of the smoothed residual signal is

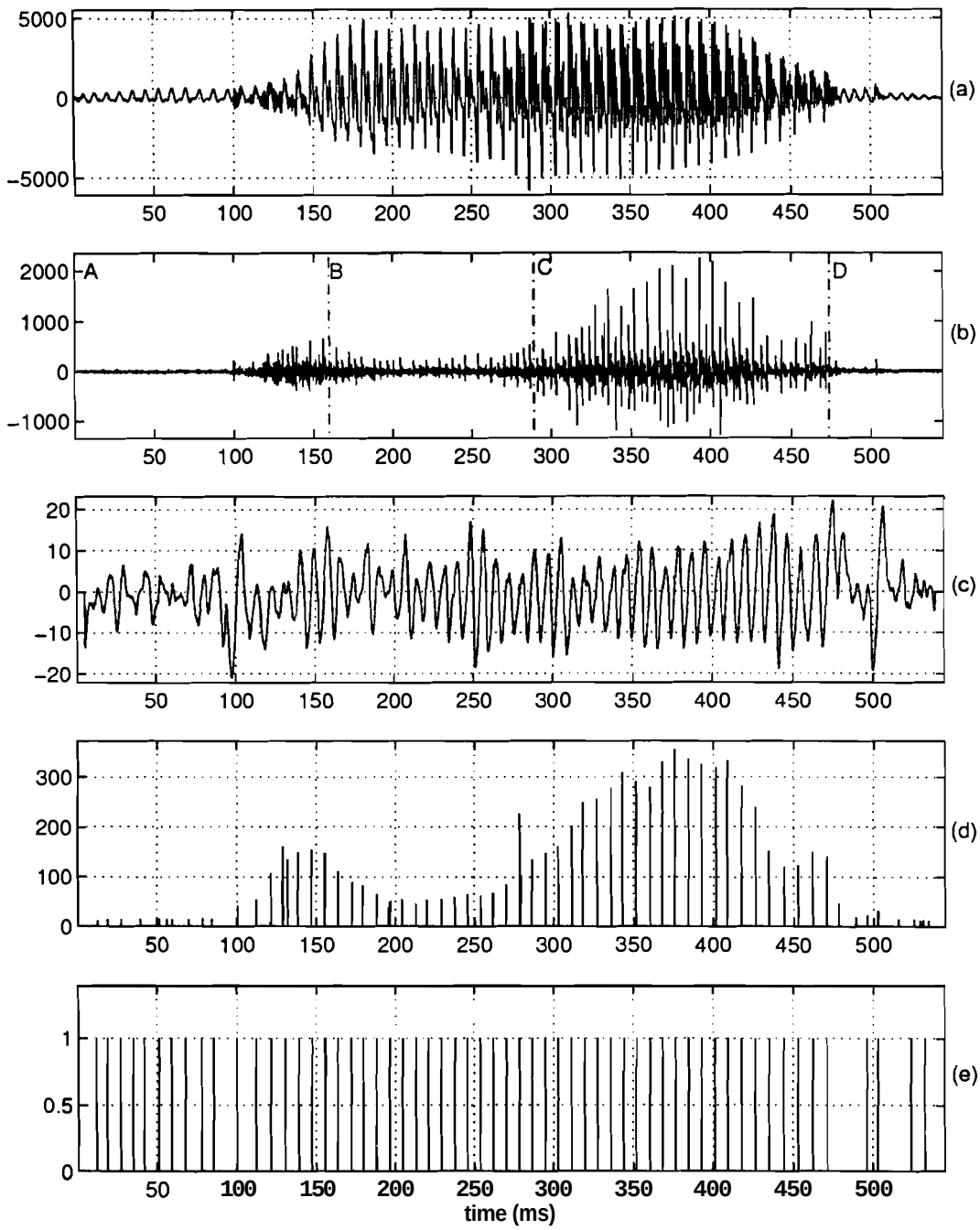


Fig. 7.1: (a) Clean speech for the utterance */dz ua/*. (b) LP residual signal derived from the signal in (a). (c) Phase slope function. (d) Instants of significant excitation, weighted by their strengths, derived from the signal in (a). (e) Instants of significant excitation, derived from the signal in (a) using the proposed algorithm.

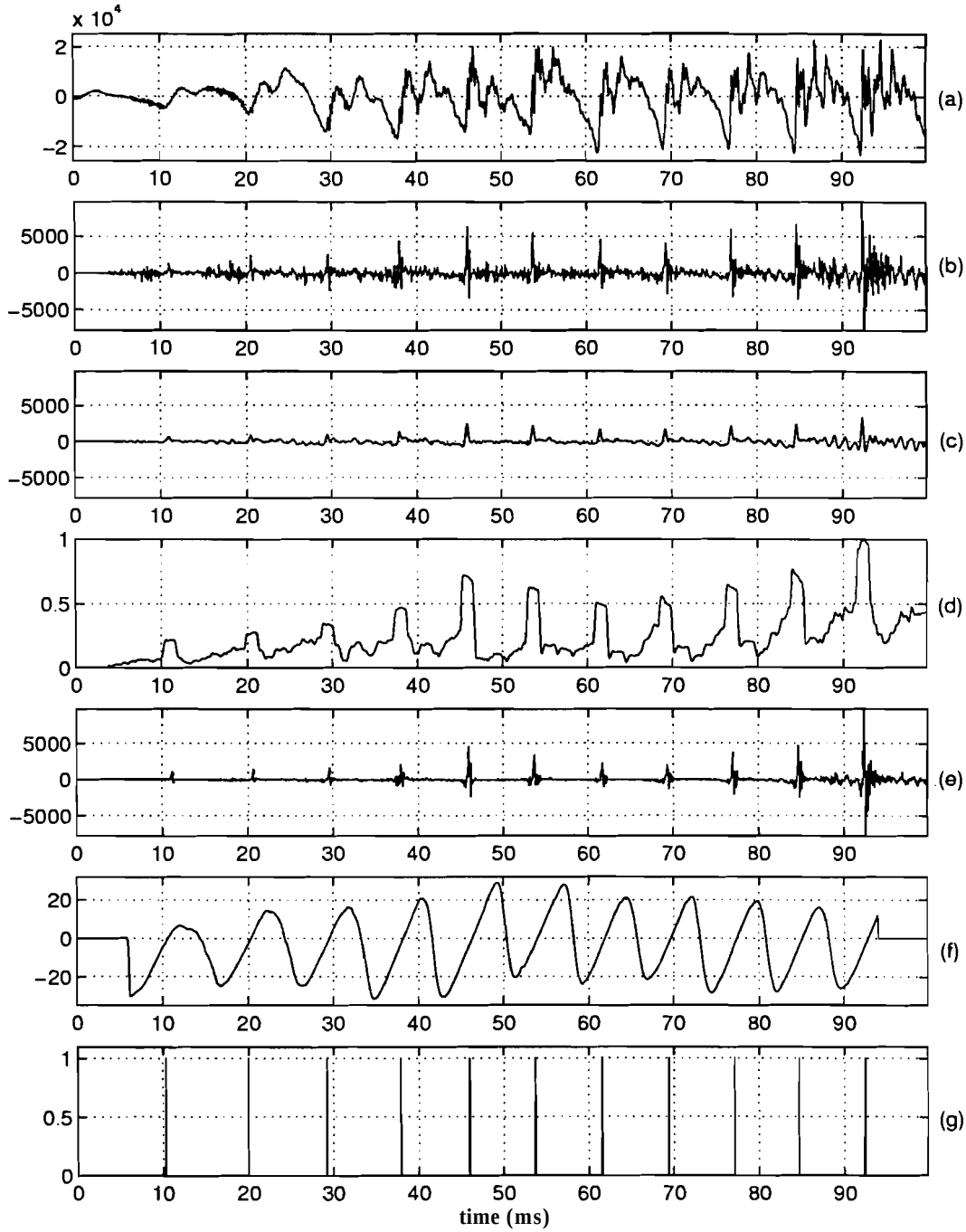


Fig. 7.2: (a) Speech signal in the transition region of the utterance /*dz a*/. (b) LP residual signal derived from the signal in (a). (c) Smoothed residual signal. (d) Normalized short−time energy function of the signal in (c). (e) Weighted residual signal. (f) Phase slope function. (g) Instants of significant excitation indicating the positive zero−crossings of the phase slope function.

computed at every sample using a frame size of 1.4 ms (15 samples at 11 kHz sampling rate). The short-time energy curve will have large amplitudes around the instants of significant excitation. The short-time energy is normalized to a maximum value of one (see Fig. 7.2(d)) and is used as a weight function for the residual signal to enhance the regions in the residual signal around the significant excitations. The weighted residual signal shown in Fig. 7.2(e) is used to derive the instants of significant excitation. By comparing Figs. 7.2(b) and 7.2(e) we observe that weighting the residual signal has indeed enhanced the excitation peaks in the residual signal with respect to the low amplitude regions. The phase slope function smoothed using a 5-point Hamming window, is shown in Fig. 7.2(f). Positive zero-crossings of the smoothed phase slope function are identified as the instants of significant excitation (see Fig. 7.2(g)). For the utterance */dz ua/* in Fig. 7.1, the plot of the instants derived after the above mentioned refinements is shown in Fig. 7.1(e). Some of the errors in the estimation of instants in Fig. 7.1(d) are corrected in Fig. 7.1(e). The different steps in the algorithm for the computation of the instants of significant excitation are given in Table-7.1.

7.3 MEASURE OF STRENGTH OF EXCITATION

Reliability of the extracted instants depends on the strength of excitation around the instants. In [202, 203] the short-time energy of the LP residual signal was used to represent the strength of excitation at each instant. In some cases it is difficult to use the short-time energy around the instant as a measure of the strength, especially when the residual signal is noisy, as in the region BC in Fig. 7.1(b). Moreover, the derived residual signal energy depends on the effectiveness of the LP analysis for these segments.

We propose an alternative rneasure for the strength of excitation, which is based on the use of F'robenius norm. In [273] the F'robenius norm of a signal prediction

Table 7.1: Algorithm for determination of instants of significant excitation.

- Calculate the linear prediction residual signal using a frame of size 25 ms, Hamming window and a 10th order linear prediction analysis by autocorrelation method. The frame is shifted successively by 10 ms.
- Smooth the residual signal with a 0.75 ms Hamming window. Compute the short–time energy of the smoothed residual signal for every sample, over a frame of duration 1.4 ms. Normalize the short–time energy function to a maximum value of one. Multiply the residual signal obtained from the speech signal with the normalized short–time energy function, to obtain the ***weighted residual signal.***
- Select a frame size between one to two periods of a glottal cycle, apply a Hamming window and compute the groupdelay of the weighted residual signal at each sampling instant using the formula

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{X_R^2(\omega) + X_I^2(\omega)}$$

  where $X(\omega) = X_R(\omega) + jX_I(\omega)$ and $Y(w) = Y_R(\omega) + jY_I(\omega)$ are the Fourier transforms of the weighted residual signal $x(n)$ and $y(n) = nx(n)$, respectively.
- Smooth the groupdelay function computed at each sampling instant using a 3–point median filter.
- Compute the average of the smoothed groupdelay function.
- The negative of the average of the groupdelay function obtained at each sampling instant is plotted with time to obtain the ***phase slope function.*** Smooth the phase slope function using a Hamming window of length 5 samples.
- The positive zero–crossings in the smoothed phase slope function are identified as the instants of significant excitation.

matrix, formed by using the samples in a frame of about 3 ms, was proposed to locate the instants of glottal closure. The Frobenius norm was computed at each sampling instant. The locations of the peaks in the plot of the Frobenius norm as a function of time were considered as the desired instants. In this section, we propose that the Frobenius norm [279] of the signal prediction matrix [273] formed by using the samples in a 3 ms frame of differenced speech centered around the identified instant of significant excitation can be used to represent the strength of excitation at that instant.

Consider a frame of the differenced speech signal with M samples, s_1, s_2,...,s_M. Assuming a linear prediction order of p, the following prediction error vector can be formed

$$\mathbf{e} = \mathbf{S}\mathbf{a}_a \tag{7.2}$$

where S is the Toeplitz signal prediction matrix of dimension (M – p) x (p + 1)

$$\mathbf{S} = \begin{bmatrix} s_{p+1} & s_p & \cdots & s_1 \\ s_{p+2} & s_{p+1} & \cdots & s_2 \\ & & \ddots & \vdots \\ \vdots & \vdots & & s_{p+1} \\ & & & \vdots \\ s_M & s_{M-1} & \cdots & s_{M-p} \end{bmatrix} \tag{7.3}$$

and $\mathbf{a}_a$ is the augmented vector of LPCs $[1\ a_1\ a_2\ \ldots\ a_p]^T$. Assuming s_n, $n = 1,..,M$ are the samples of a signal at the output of an all–pole system excited by a periodic impulse train, there is a linear dependence between the column vectors of S, when the instant of excitation is not included in the analysis frame [273]. The error vector is then zero. But when the instant of excitation is included, the norm of the error vector goes up. The amplitudes of signal samples in the signal prediction matrix also go up,

because of the excitation. Thus the Fkobenius norm of the signal prediction matrix, computed as the square root of the sum of all squared elements of the matrix, also goes up. The square of the Euclidean norm of **e,** which is a measure of the energy (strength) of excitation, is given by

$$\begin{aligned} \|\mathbf{e}\|_2^2 &= \|\mathbf{S}\mathbf{a}_a\|_2^2 \\ &\leq \|\mathbf{S}\|_F^2 . \|\mathbf{a}_a\|_2^2 \end{aligned} \tag{7.4}$$

where $\|\mathbf{S}\|_F$ is the Frobenius norm of S. The ratio $\frac{\|\mathbf{e}\|_2^2}{\|\mathbf{a}_a\|_2^2}$ is upper bounded by $\|\mathbf{S}\|_F^2$. Ignoring the variation in $\|\mathbf{a}_a\|_2^2$ compared to $\|\mathbf{S}\|_F^2$, we can use $\frac{\|\mathbf{e}\|_2^2}{\|\mathbf{a}_a\|_2^2}$ as a measure of the strength of excitation. Computing the Euclidean norm of e from (7.2), we get

$$\frac{\|\mathbf{e}\|_2^2}{\|\mathbf{a}_a\|_2^2} = \frac{\mathbf{a}_a^T (\mathbf{S}^T \mathbf{S}) \mathbf{a}_a}{\mathbf{a}_a^T \mathbf{a}_a} \tag{7.5}$$

$$= \rho(\mathbf{a}_a) \tag{7.6}$$

where $\rho(\mathbf{a}_a)$ is the Rayleigh quotient of $(S^T S)$ [279]. It is shown in Appendix–C (see (C.8)) that

$$\sigma_{p+1}^2 \leq \rho(\mathbf{a}_a) \; \mathbf{I} \; \sigma_1^2 \tag{7.7}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{p+1} > 0$ are the singular values of S, and are also the eigenvalues of $S^T S$. It is also known that the square of the Fkobenius norm is the sum of squared singular values [253]. So we have the inequality

$$\sigma_{p+1}^2 \leq \frac{1}{(p+1)} \|\mathbf{S}\|_F^2 \leq \sigma_1^2 \tag{7.8}$$

since $(p+1)^{-1}\|\mathbf{S}\|_F^2$ is the arithmetic mean of squared singular values. It is known that all the singular values rise in magnitude when there is an excitation within the analysis frame and fall when there is no excitation [273]. Therefore, both $\rho(a_a)$ in (7.7) and $(p+1)^{-1}\|\mathbf{S}\|_F^2$ in (7.8) track these changes. Therefore $(p+1)^{-1}\|\mathbf{S}\|_F^2$ can be used

as a measure of the strength of excitation. We note that though this is a measure of energy of the residual signal, it is computed directly from the speech signal.

It is to be noted that since the square of the F'robenius norm of the signal prediction matrix is the sum of squares of all samples in the matrix, it is nothing but the short–time energy of the speech signal computed from the weighted samples of the speech signal.

To illustrate the need for a measure for the strength of excitation, let us consider the differentiated glottal pulses (Fig. 7.3(a)) generated using the LF–model [205]. All the parameters of the model are kept constant except the time constant of the return phase (T_a) and the instant of peak positive excitation (T_i) (see Appendix–A). To vary the rate of closure, the time constant of the return phase is increased from 0.05 ms to 1.5 ms from left to right. The amplitudes of the pulses are progressively scaled up (from left to right) so that all the pulses have equal negative peak amplitudes. These differentiated glottal pulses are used to excite an all–pole model to obtain a synthetic voiced sound shown in Fig. 7.3(c). It should be noted that, in the first 40 ms of the speech signal, the signal components due to higher formants can be clearly seen. This is due to the sharp closing phase, which results in a relatively flat magnitude spectrum of the excitation pulses. This feature is not seen in the latter portion of the signal in Fig. 7.3(c) due to gradual closing phase. The second derivative of the glottal pulse and the 12th order LP residual signal are shown in Figs. 7.3(b) and 7.3(d), respectively. F'rom these figures it is evident that the amplitudes of the excitation impulses are higher for the glottal pulses with sharper closure. The strength of excitation is higher for sharper closure, although the amplitude and energy of the speech signal in Fig. 7.3(c) is nearly the same throughout for all the glottal pulse shapes. It should be noted in Fig. 7.3(a) that the energy concentration is higher for

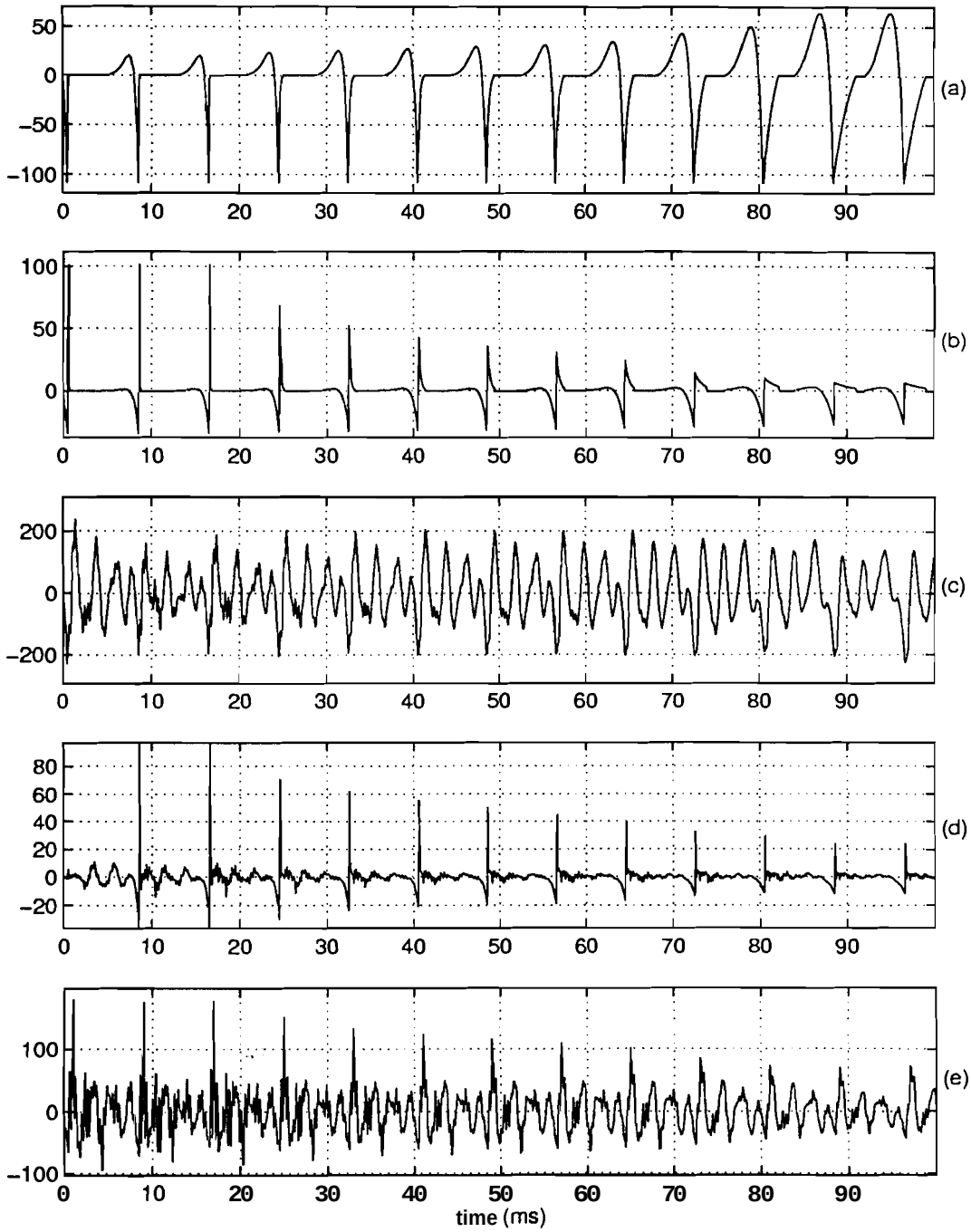


Fig. 7.3: (a) Differentiated glottal pulses. (b) Second derivative of glottal pulses. (c) Synthetic signal. (d) Residual signal derived from the signal in (c). (e) The signal in (c) after preemphasis.

the pulses in the initial portion than in the latter portion of the signal.

If we consider the differenced signal of Fig. 7.3(c), as shown in Fig. 7.3(e), we notice that the strength of excitation is also evident in the differenced signal. It can also be seen by considering a difference operation $(1 - z^{-1})$ on the z-transform of the signal, $S(z) = E(z)H(z)$, where $E(z)$ corresponds to the differentiated glottal pulse excitation, and $H(z)$ corresponds to the vocal tract system. We have

$$(1 - z^{-1})\, S(z) \;=\; (1 - z^{-1})\, \mathrm{E(z)}\, \mathrm{H(z)} \tag{7.9}$$

Thus the differenced signal can be viewed as a signal that results due to the excitation of the vocal tract system with the second derivative of the glottal pulse. The second derivative of the glottal pulse in Fig. 7.3(b) and the differenced signal in Fig. 7.3(e) both show the characteristics of the strength of excitation. These figures suggest that the Frobenius norm of the differenced signal can be used as a measure of the strength of excitation around the instant of significant excitation.

7.4 ROBUSTNESS OF THE GROUP–DELAY–BASED METHOD

In this section we shall examine the robustness of the group–delay–based method for two types of degradations, namely, additive random noise and echo/reverberation.

7.4.1 Robustness Against Additive Noise

Let us consider an excitation signal $x(n)$ consisting of an impulse of amplitude A at time n = L and a zero–mean additive white Gaussian noise $v(n)$.

$$x(n) = \mathcal{A}\, \delta(n - L) + v(n)\,, \qquad n = 0, 1, ..., N - 1 \tag{7.10}$$

The Fourier transform of $x(n)$ is

$$X(\omega) = \mathcal{A}\, \exp(-j\omega L) + V(\omega) \tag{7.11}$$

where

$$V(\omega) = \sum_{n=0}^{N-1} v(n) \exp(-j\omega n)$$
$$= |V(\omega)| \exp(j\phi_v(\omega)) \quad (7.12)$$

$|V(\omega)|$ and $\phi_v(\omega)$ are random variables corresponding to the magnitude and phase of $V(\omega)$, respectively. Without loss of generality, the phase spectrum $\phi_v(\omega)$ can be assumed to have a uniform probability density function over the range $[-n,\pi]$ [280]. Let $|X(w)|$ and $\phi_x(\omega)$ be the magnitude and phase of $X(\omega)$, respectively. Then

$$\log[|X(w)|] + j\,\phi_x(\omega) = \log(\mathcal{A}) - j\omega L + \log[1 + \frac{|V(\omega)|}{\mathcal{A}} \exp(j(\phi_v(\omega) + \omega L))] \quad (7.13)$$

It is shown in Appendix–D (see (D.4)) that

$$\frac{\mathcal{E}[|V(\omega)|]}{\mathcal{A}} < 10^{-\frac{E_s}{20}} \quad (7.14)$$

where $\mathcal{E}$ denotes ensemble average and E_s is the excitation signal–to-noise ratio, defined as the logarithm of the ratio of average excitation signal power per sample $(\frac{\mathcal{A}^2}{N})$ to the average noise power per sample (σ_v^2)

$$E_s = 10 \log_{10}(\frac{\mathcal{A}^2}{N\sigma_v^2}) \text{ dB} \quad (7.15)$$

For $E_s = 0$ dB, the upper bound on the expected value of the magnitude of $[\frac{|V(\omega)|}{\mathcal{A}} \exp(j(\phi_v(\omega) + \omega L))]$ is one. If the Fourier transform in (7.11) is evaluated using an N–point Discrete Fourier Transform (DFT), the magnitude of the DFT $|V(\omega_k)|$ can be shown to be less than A with 99% confidence when $E_s \geq 6.6$ dB (see (D.7) in Appendix–D). Expanding the third term on the right hand side of (7.13) by Taylor series expansion, the phase term of (7.13) can be approximated as

$$\phi_x(\omega) = -\omega L + \frac{|V(\omega)|}{\mathcal{A}} \sin(\phi_v(\omega) + \omega L) \quad (7.16)$$

The group–delay function $(\tau_x(\omega))$ is given by

$$\tau_x(\omega) = -\frac{d\phi_x(\omega)}{d\omega} \tag{7.17}$$

Hence, the average value of the group–delay function is given by

$$\begin{aligned} \bar{\tau}_x &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \tau_x(\omega)\, d\omega \\ &= -\frac{1}{2\pi}[\phi_x(\pi) - \phi_x(-\pi)] \end{aligned} \tag{7.18}$$

Substituting *(7.16)* in *(7.18)* and noting that $\phi_x(\omega)$ is an odd function of *w* and that the second term in *(7.16)* vanishes at π, we have

$$\bar{\tau}_x = L \tag{7.19}$$

i.e., the average value of the group–delay function gives the location of the impulse.

In practice, the group–delay function is computed at discrete frequencies, and hence the computed average deviates from *(7.19)*. Random fluctuations and spikes appear in the group–delay function [281]. These spikes may bias the mean value of the group–delay function. Therefore, it is preferable to use median smoothing of the computed group–delay function before computing the average.

So far we have considered an excitation signal corrupted by additive noise. Let us now consider a noisy speech signal $y(n)$

$$y(n) = s(n) + w(n) \tag{7.20}$$

where $s(n)$ is the speech signal and $w(n)$ is the additive white noise. To derive the instants of significant excitation, let us consider the LP residual signal. The frequency response of the inverse filter obtained from the LP analysis is given by

$$\begin{aligned} A(\omega) &= |A(\omega)| \exp(j\phi_A(\omega)) \\ &= \sum_{k=0}^{p} a_k \exp(-j\omega k) \end{aligned} \tag{7.21}$$

where $a_0 = 1$ and $a_1, ..., a_p$ are the LPCs. The residual error signal obtained after inverse filtering is given by

$$x(n) = e(n) + v(n) \tag{7.22}$$

where $e(n)$ is the component at the output of the inverse filter due to the speech signal $s(n)$ and $v(n)$ is the coloured noise due to filtering of the white noise $w(n)$. Note that even though the speech signal is assumed to be the output of an all–pole system, the noisy signal $y(n)$ corresponds to a pole–zero system [268]. The power spectrum of the coloured noise component $v(n)$ is given by

$$\begin{aligned} P_v(\omega) &= \frac{1}{N} \mathcal{E}[|\, V(\omega)\,|^2] \\ &= |\, A(\omega)\,|^2\, \sigma_w^2 \end{aligned} \tag{7.23}$$

The second moment $\mathcal{E}[|\, V(\omega)\,|^2]$ depends on the frequency w. Let us consider the worst case situation, i.e., the maximum value of $\frac{1}{N} \mathcal{E}[|\, V(\omega)\,|^2]$. Let

$$\begin{aligned} \sigma_{v\max}^2 &= \max_{\omega} \frac{1}{N} \mathcal{E}[|\, V(\omega)\,|^2] \\ &= A_{\max}^2\, \sigma_w^2 \end{aligned} \tag{7.24}$$

where $A_{\max}$ is the maximum value of $|\, A(\omega)\,|$ given by

$$\begin{aligned} A_{\max} &= \max_{\omega} |\, A(\omega)\,| \\ &= \max_{\omega} \left|\, 1 + \sum_{k=1}^{p} a_k \exp(-j\omega k)\,\right| \end{aligned} \tag{7.25}$$

In the expression for E_s in (7.15), the σ_v^2 is replaced by σ_{vmax}^2. Assuming that $A_{\max} > 1$, the effective E_s for the residual signal is reduced.

The above analysis is valid even when the speech is corrupted by additive coloured random noise, except that $A_{\max}$ now also depends on the maximum value of the power spectrum of the coloured noise.

The robustness of estimation of the instant of excitation depends on the excitation signal–to–noise ratio (E,). For a constant additive noise, E_s will decrease as the strength of the excitation decreases. This is illustrated in Fig. 7.4 for a noisy case of the synthetic signal generated by exciting an all–pole filter with the differentiated

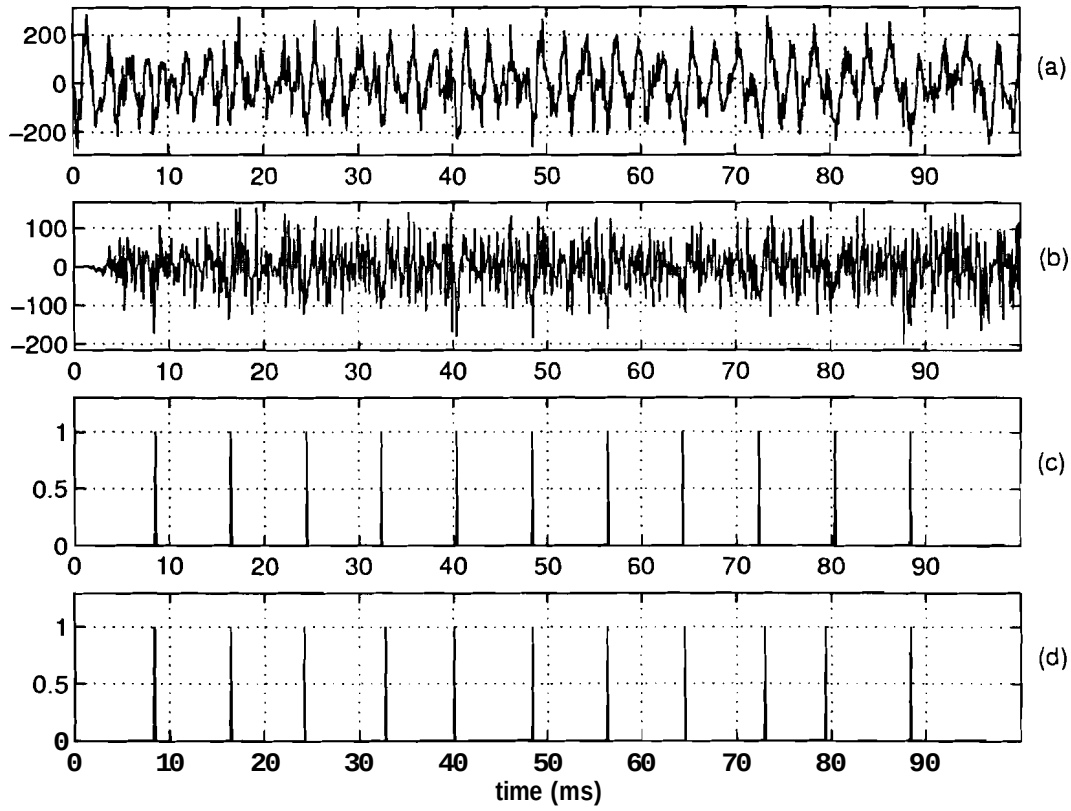


Fig. 7.4: (a) Synthetic speech of Fig. 7.3(c) at an average SNR of 5 dB. (b) LP residual signal derived from the signal in (a). (c) The true locations of the instants of significant excitation. (d) The instants of significant excitation derived from the noisy signal in (a).

glottal pulses of Fig. 7.3(a). The overall SNR of the speech signal is 5 dB. Note that the periodicity cannot be immediately seen from the noise corrupted speech signal. Since it is a synthetic case, the strength of excitation can be approximated to the amplitude of the second derivative of the glottal pulse shown in Fig. 7.3(b). Fig. 7.4(c) shows the

actual instants of significant excitation. Fig. 7.4(d) shows the instants of significant excitation estimated from the noisy speech signal. The figure shows that the accuracy of the extracted instants depends on the excitation signal–to–noise ratio. Reliability of the extracted instants decreases with decrease in the excitation signal–to–noise ratio, as can be seen from the deviation of the instants in Fig. 7.4(d) relative to the instants in Fig. 7.4(c). The excitation signal–to–noise ratio (E_s) is defined as the ratio of the square of the amplitude of the impulse and the noise power. Note that even though the average SNR of the speech signal is nearly constant i.e., 5 dB, the excitation signal–to–noise ratio is decreasing from left to right on the time scale.

7.4.2 Robustness Against Echo and Reverberation

Let us consider the following reverberant signal $x(n)$ for an impulse of strength A and delayed by n = L samples.

$$x(n) = \mathcal{A}\,\delta(n-L) + \beta\,\mathcal{A}\,\delta(n-L-D) + \beta^2\,\mathcal{A}\,\delta(n-L-2D) + \ldots \tag{7.26}$$

where β is the attenuation factor ($0 < \beta < 1$) and D is the delay due to reverberation. The Fourier transformation of (7.26) yields

$$|\,X(\omega)\,|\;\exp(j\phi_x(\omega)) = \frac{\mathcal{A}\,\exp(-j\omega L)}{(1-\beta\,\exp(-j\omega D))} \tag{7.27}$$

where $|\,X(\omega)\,|$ and $\phi_x(\omega)$ are the magnitude and phase of the Fourier transform of $x(n)$, respectively. Taking natural logarithm on both sides of (7.27), we get [282]

$$\log[|\,\mathrm{X}(\omega)\,|] + j\phi_x(\omega) = \log(\mathcal{A}) - j\omega L - \log\,[1-\beta\,\exp(-j\omega D)] \tag{7.28}$$

Neglecting the higher order terms in the Taylor series expansion of the last term above, the phase component is given by

$$\phi_x(\omega) = -\omega L - \beta\sin(\omega D) \tag{7.29}$$

The group–delay is

$$\tau_x(\mathrm{w}) = \mathrm{L} + \beta D \cos(\mathrm{wD}) \tag{7.30}$$

The mean value of the group–delay $\tau_x(\omega)$ is L. For a single echo, the term $-\log\ [1-\beta\ \exp(-j\omega D)]$ in (7.28) can be replaced by $\log[1+\beta\ \exp(-j\omega D)]$. The expression for the phase is the same as in (7.29) and hence the group–delay for the case of echo is the same as in (7.30).

It should be noted that the above analysis is valid only under mild echo and reverberant conditions ($\beta \ll 1$). We have also assumed that the signal characteristics are stationary. Due to nonstationarity of speech signals, the model of reverberation in (7.26) may not be valid for reverberation in large rooms ($T_{60} > 400$ ms).

7.4.3 Robustness due to Weighting of the LP Residual Signal

In this subsection we show that suitable weighting of the LP residual signal improves the robustness of the algorithm for extraction of the instants of significant excitation. This is because the excitation signal–to–noise ratio E_s can be improved by weighting, as shown below.

Consider the impulse–in–noise sequence $x(n)$ in (7.10). Let $\gamma(\mathrm{n}), \mathrm{n} = -\frac{N-1}{2}, .., 0, .., \frac{N-1}{2}$ be a positive window function such that $\gamma(0) > \gamma(n)$, $\mathrm{n} \neq 0$. Let

$$\begin{aligned} x_\gamma(n) &= x(n)\ \gamma(n-L) \\ &= \mathcal{A}\ \gamma(0)\ \delta(n-L) + \gamma(n-L)\ v(n)\ , \qquad n = 0, 1, ..., (N-1) \end{aligned} \tag{7.31}$$

be the weighted excitation signal, such that the impulse at $\mathrm{n} = \mathrm{L}$ is given the maximal weight of $\gamma(0)$. By following the steps in the analysis presented in Section 7.4.1, we have

$$\phi_{x\gamma}(\omega) = -\omega L + \frac{|\ V_\gamma(\omega)\ |}{\mathcal{A}\ \gamma(0)}\ \sin(\phi_{v\gamma}(\omega) + \omega L) \tag{7.32}$$

where

$$V_\gamma(\omega) = \sum_{n=0}^{N-1} v(n)\ \gamma(n-L)\ \exp(-j\omega n)$$
$$= |\ V_\gamma(\omega)\ |\ \exp(j\phi_{v\gamma}(\omega)) \tag{7.33}$$

and $\phi_{x\gamma}(\omega)$ is the phase of the Fourier transform of the weighted excitation sequence $x_\gamma(n)$. The above approximation in (7.32) is justified provided that

$$\frac{\mathcal{E}\,[|\ V_\gamma(\omega)\ |]}{\mathcal{A}\ \gamma(0)} < 1$$

Assuming that $\{v(n)\}$ are zero mean Gaussian random variables with variance σ_v^2, we have from (7.33),

$$\mathcal{E}\left[|\ V_\gamma(\omega)\ |^2\right] = \sigma_v^2 \sum_{n=0}^{N-1} \gamma^2(n-L)$$
$$= \gamma^2(0)\ \sigma_v^2\ S_\gamma \tag{7.34}$$

where

$$S_\gamma = \sum_{n=0}^{N-1} \left[\frac{\gamma(n-L)}{\gamma(0)}\right]^2 \tag{7.35}$$

Following the steps in the analysis presented in Appendix–D, we define the weighted excitation signal–to–noise ratio as

$$E_w = 10\ \log_{,,}\ \left(\frac{\mathcal{A}^2\ \gamma^2(0)}{\mathcal{E}\,[|\ V_\gamma(\omega)\ |^2]}\right)\ \text{dB}$$
$$= 10\ \log_{10}\left(\frac{\mathcal{A}^2}{N\sigma_v^2}\ \frac{N}{S_\gamma}\right)\ \text{dB} \tag{7.36}$$

Using (7.15), we get

$$E_w = E_s + 10\ \log_{10}\left(\frac{N}{S_\gamma}\right) \tag{7.37}$$

Note that for the case without weighting of the LP residual signal, $\gamma(n) = 1$. Therefore from (7.35) and (7.37), $E_w = E_s$. For any other wiridow function with a broad peak

around the location of the impulse i.e., n = L, S_γ < N. Thus there is some gain in the excitation signal–to–noise ratio. For the limiting case of a weight function with a narrow peak at n = L, the gain in the excitation signal–to–noise ratio tends to 10 $\log_{10}(N)$.

7.5 PERFORMANCE EVALUATION OF THE GROUP–DELAY–BASED METHOD

In this section we consider some examples of speech data under actual conditions of degradation, and examine the performance of the group–delay–based method for extraction of the instants of significant excitation. Since we do not have a method for estimating the SNR of the strength of excitation (E_s) for signals with natural degradations, the results can only be interpreted from our a priori knowledge of the characteristics of the excitation for different categories of sounds. Wherever appropriate, the Frobenius norm of the differenced speech signal can be used as a measure of the strength of excitation.

Fig. 7.5 shows the performance of the algorithm for noise and telephone channel degradations for the segment of speech given in Fig. 7.1(a). The strengths of excitation at the extracted instants computed using Frobenius norm are shown in Fig. 7.5(b). For this signal, the strength of excitation is lower for the segment /u/ in the region BC compared to the region CD. The noisy speech signal in Fig. 7.5(c) corresponds to the same speech as in Fig. 7.5(a), but recorded by a microphone placed 50 cm away from the speaker. The signal in the region AB is affected by the additive noise more than the signal in the region CD due to lower signal amplitudes. Hence the instants extracted for the signal in region AB are not reliable. Most of the extracted instants (Fig. 7.5(d)) for the signal in the region BC are correct, eventhough in Fig. 7.5(c) there appears to be no visible periodicity in the signal in the region BC. From Figs. 7.5(b)

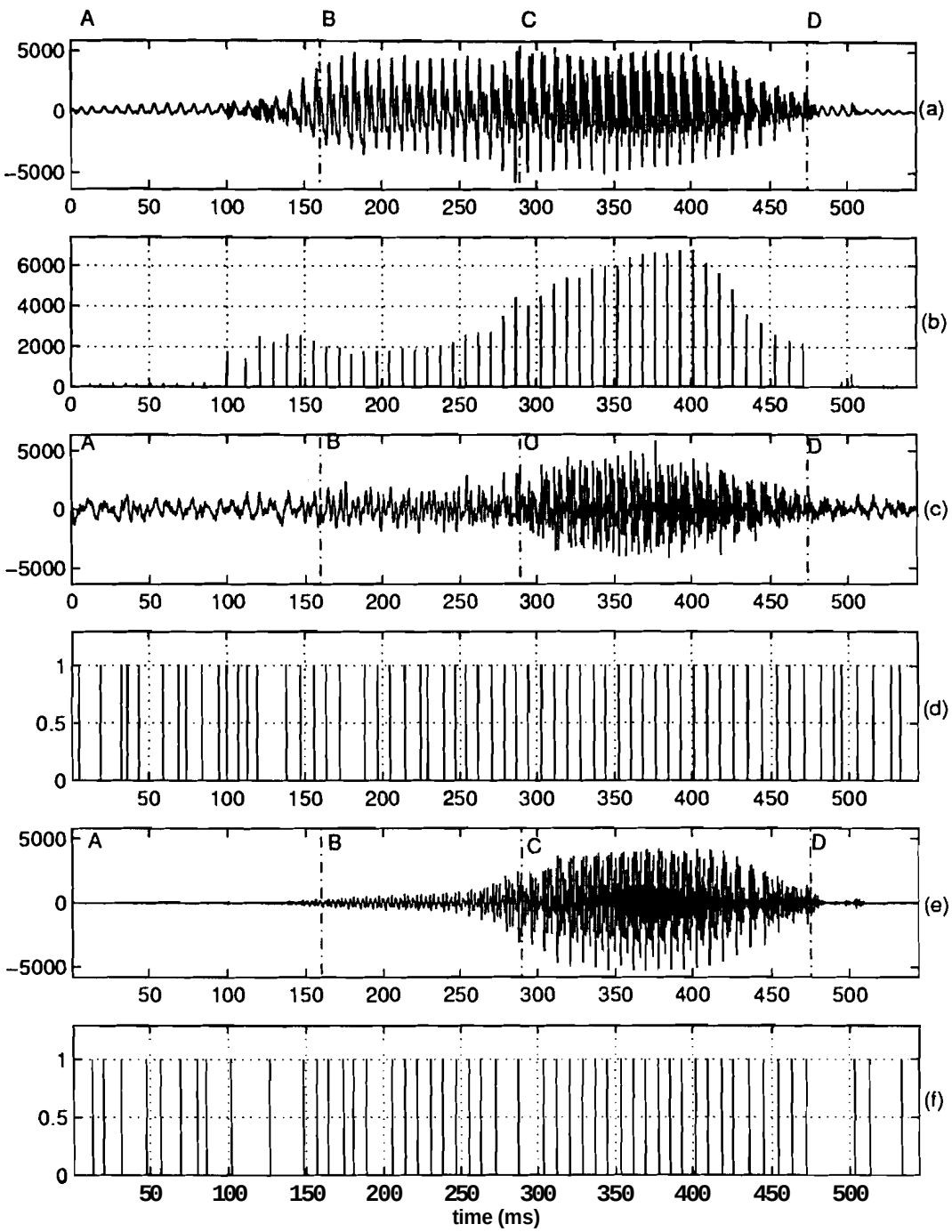


Fig. 7.5: (a) Clean speech for the utterance */dz ua/*. (b) Strengths of excitation based on the Frobenius norm. (c) Speech degraded by ambient noise. (d) Instants of significant excitation derived from the signal in (c). (e) Telephone speech. (f) Instants of significant excitation derived from the signal in (e).

and 7.5(d), it can be seen that the instants are correctly extracted for the signal in the region CD. The results are similar for the case of telephone speech shown in Figs. 7.5(e) and 7.5(f). In the telephone speech shown in Fig. 7.5(e), the signal in the region AB is lost and it is significantly attenuated in the region BC. This is because the low first formant of the vowel /u/, is severely attenuated due to the bandpass nature of the telephone channel characteristics. The errors in the region AB are due to low levels of the signal itself in that region. It is important to note that although the signal level is high in the region BC for the clean speech, the strength of excitation is low for the instants in that region. Hence, the extracted instants in this region are more prone to errors compared to the extracted instants in the region CD.

A systematic investigation was carried out to study the accuracy of the extracted instants for synthetic and natural vowels. Histograms of the spread of the errors are shown in Figs. 7.6 and 7.7 for five synthetic and natural vowels (/a/, /e/, /i/, /o/ and /u/), respectively, for an overall SNR of 10 dB. All the synthetic vowels are generated by the same LF–model–based differentiated glottal pulses. The length of each pulse was chosen to be 80 samples. In the case of the natural vowels, the glottal cycle duration varied from 9 ms for vowel /a/ to 7 ms for vowel /u/. In Fig. 7.6 the histogram for each synthetic vowel is obtained by computing the histogram of deviations of the estimated instants of significant excitation from the true locations for 50 realizations of noise. There are 10 glottal cycles in the signal for each vowel and hence we get 500 such deviations for each vowel. In Fig. 7.7 the deviations are obtained by subtracting the estimated locations from the locations extracted from the clean speech signal. Larger spread of the histograms indicates more deviation of the extracted instants from the true locations of the instants. The errors are typically rnore for the close vowels /u/ and /i/ than for the open vowels /a/, /e/ and /o/. For the synthetic case shown in

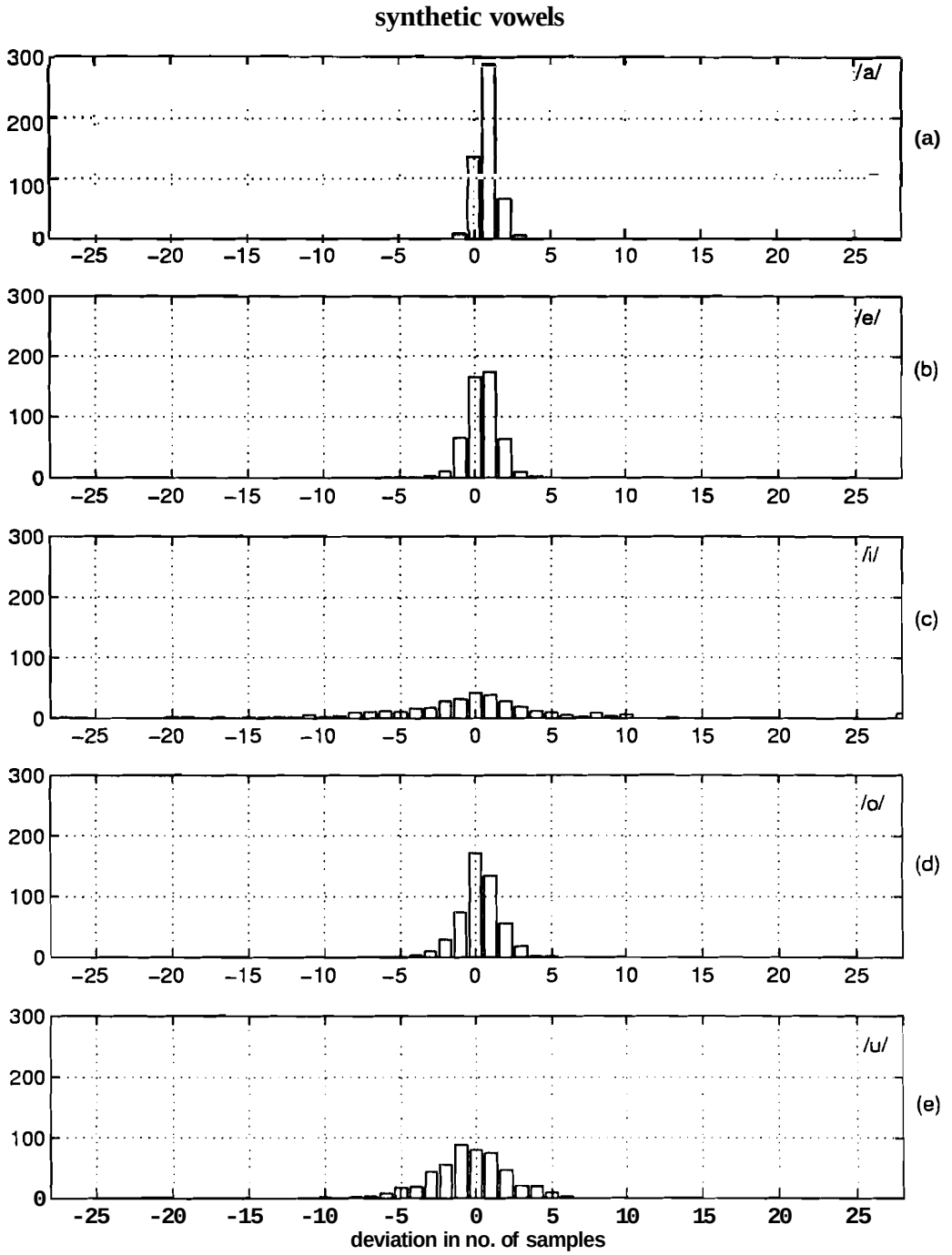


Fig. 7.6: Histogram of errors in the estimated instants for five synthetic vowels for SNR = 10 dB. (a) /*a*/ (b) /*e*/ (c) /*i*/ (d) /*o*/ (e) /*u*/.

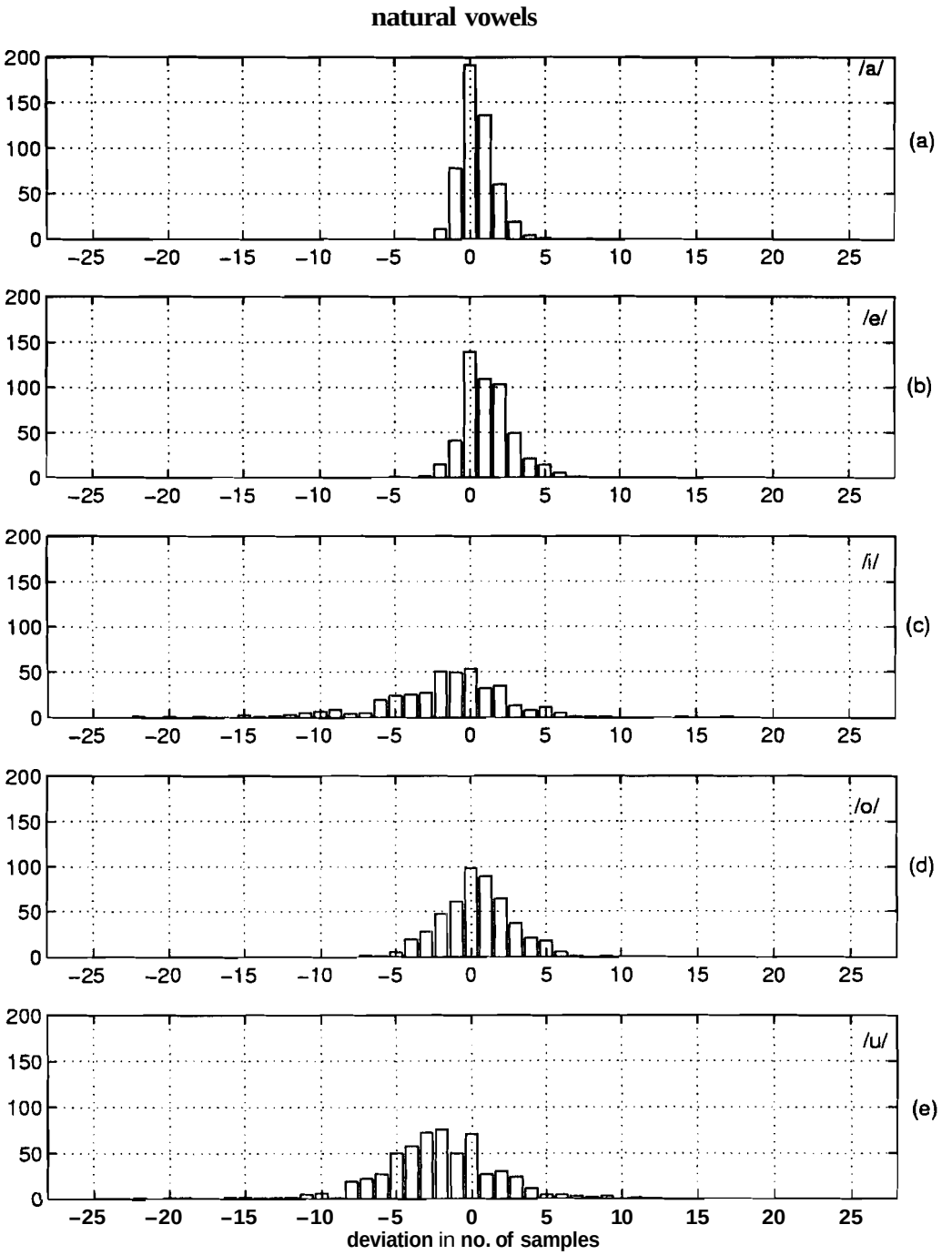


Fig. 7.7: Histogram of errors in the estimated instants for five natural vowels for SNR = 10 dB. (a) /*a*/ (b) /*e*/ (c) /*i*/ (d) /*o*/ (e) /*u*/.

Fig. 7.6, all the instants have the same strength and hence the spread of errors is less compared to the case of natural vowels. It is important to note that the variation in the spread of the errors for different vowels is also due to the artifacts of the LP analysis. For the synthetic case shown in Fig. 7.6, the spread is more for the close vowels /u/ and /i/, despite the excitation strength being the same for all the five vowels, because of the dominance of the first formant in the LP analysis of the noise corrupted signals for these close vowels. This is also true in the case of natural vowels shown in Fig. 7.7. There is a systematic bias in the estimated locations of the instants of excitation for the case of synthetic vowels. The bias is about 2 samples for the average glottal cycle length of 80 samples. That is, the bias is about 3%. The bias may have been caused due to weighting the LP residual signal before computing the instants of excitation. The weight function depends on the nature of the voiced sound, and the extent of degradation caused by noise. That is why the bias is positive in some cases and negative in some other cases.

Errors in the extracted instants were also studied for utterances taken from the standard NTIMIT [283, 284] data for male and female speech. Since the TIMIT [255] data was available for reference, the spread was estimated using the deviations of the extracted instants for the NTIMIT data from those for the TIMIT data. The TIMIT and NTIMIT data taken for study were lowpass filtered and downsampled to 8 kHz before processing. The TIMIT and NTIMIT data were first time–aligned before the deviations were computed. The histograms of errors for one male voice and one female voice are shown in Figs. 7.8 and 7.9. The data for the male voice corresponds to the file: */test/dr2/mrndm2/sa1.wav* in the TIMIT/NTIMIT database. The data for the female voice corresponds to the file: */test/dr5/fjcs0/sa1.wav*. The instants of significant excitation were extracted only from the voiced regions, which were identi-

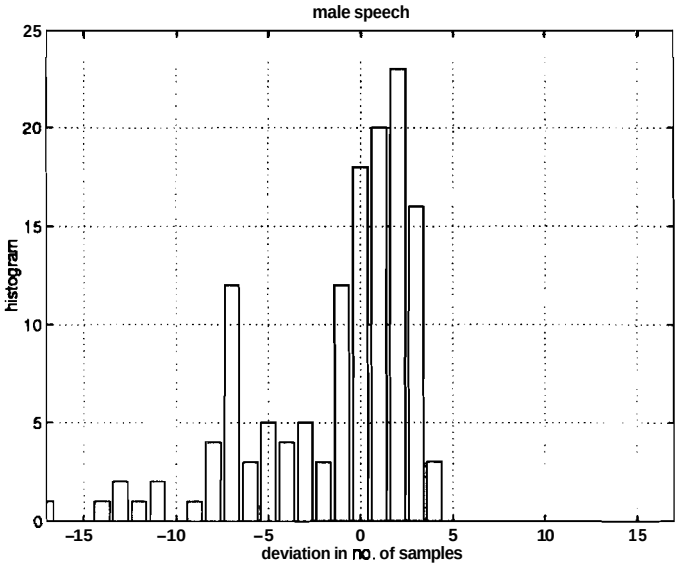


Fig. 7.8: Histogram of errors for the utterance *"She had your dark suit in greasy wash water all year"* uttered by a male speaker.

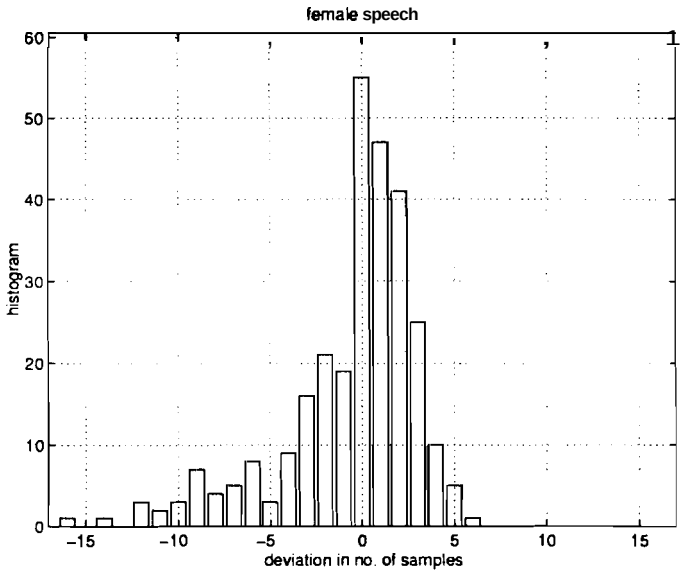


Fig. 7.9: Histogram of errors for the utterance *"She had your dark suit in greasy wash water all year"* uttered by a female speaker.

fied using the phonetic transcription files provided with the TIMIT database. From Figs. 7.8 and 7.9, it can be seen that there are more values of deviation in the histogram of deviations for female speech than for the male speech. This is because the average pitch of the female speaker is about 210 Hz and that of the male speaker is about 100 Hz. So there are more glottal cycles in the utterance of the female speaker than for the male speaker. The spread of errors is more for these utterances compared to the errors for the vowels in Fig. 7.7, because the SNR is different for different segments in this case, whereas for vowels it was constant. The speech SNR varies over a range of 20–50 dB for the utterances taken from the TIMIT data and over a range of 5–40 dB for the utterances taken from the NTIMIT data for both male and female voices. The SNR for different segments was computed as the ratio of the energy of the signal samples to the energy of the noise samples in the silence regions. The bias and spread of the errors in Figs. 7.8 and 7.9 can be attributed not only to the variations of SNR for different segments, but also to the weight function used on the LP residual signal before computing the instants of excitation.

7.6 SUMMARY

In this chapter, we have demonstrated that the group–delay–based method proposed in [202,203] is robust against degradations in speech due to additive noise and channel distortions. The robustness is due to the fact that the energy of the signal is concentrated around the instant of significant excitation, which for voiced speech corresponds to the instant around glottal closure. We have discussed the importance of the strength of excitation, which cannot be directly inferred from the speech signal. We have shown that the errors in the extracted instants are small for many practical signals such as in the NTIMIT speech data.

In the next chapter we present some applications of the instants of significant

excitation for processing degraded speech. In particular, we propose a method for enhancement of speech degraded by speech of a competing speaker.

Chapter 8

IMPORTANCE OF INSTANTS OF SIGNIFICANT EXCITATION

In Chapters 4 and 5 we have proposed methods for enhancement of speech degraded by additive random noise and reverberation, respectively. A source of degradation which is more difficult to handle is that due to speech of a competing speaker (popularly known as Cocktail party effect) [92]. This case is difficult for enhancement because the degrading signal too has the spectral characteristics of speech, and hence it is difficult to distinguish it from the desired signal. It is difficult to track the individual voices, even for a human listener, when the pitch of the two competing speakers is approximately the same. In this chapter we propose a method, based on the instants of significant excitation and subsegrnental analysis, for enhancement of speech degraded by speech of a competing speaker. We also study two other applications of instants of significant excitation, namely, enhancement of noisy speech using comb filtering and pitch synchronous subsegrnental analysis of noisy speech. The chapter is organized as follows. In Section 8.1, we propose a method for enhancement of speech degraded by speech of a competing speaker. In Section 8.2, we study the adaptive comb filtering technique of noisy speech using instants of significant excitation. In Section 8.3, we present some results obtained using pitch synchronous subsegmerital analysis of noisy speech. The analysis is performed in short (1–3 ms) high SNR segments of the signal identified using the instants of significant excitation. A surnmary of the results presented in this chapter is given in Section 8.4.

8.1 ENHANCEMENT OF SPEECH DEGRADED BY SPEECH OF A COMPETING SPEAKER

In this section, we propose a new method for enhancement of speech degraded by speech of a competing speaker. The method uses the characteristics of the speech signal and the linear prediction residual signal in the short (1–3 ms) segments immediately after the instants of significant excitation of the vocal tract. These characteristics are used to classify each instant of significant excitation as belonging to one speaker or the other, or to a spurious instant. The speech of each speaker is enhanced with respect to the speech of the other by performing relative emphasis of the speech signal in the short 3 ms interval around each instant of significant excitation of the speaker. The relative emphasis is achieved by weighting the linear prediction residual signal. The weighted linear prediction residual signal is used to excite the time–varying all–pole filter to obtain enhanced speech.

8.1.1 Overview of Previous Methods for Speaker Separation

Several approaches have been proposed in the literature to process speech degraded by speech of competing speakers [285–293]. Most of the methods proposed so far are based on separating the speakers in the short–time spectral domain by retaining the pitch harmonics of the desired speech and suppressing the other frequency components [286–289, 291, 294, 295]. Clearly, this requires accurate estimation of pitch of the desired speech. The estimation of pitch of the desired speech is difficult in the regions where the degrading component also happens to be a voiced region. Methods for processing degraded speech collected simultaneously by multiple microphones have also been suggested [285, 292]. These methods model the degraded speech recorded at each microphone as a linear weighted sum of the individual signals. The weights corresponding to each of the microphone signals are represerited in the form of a

matrix called the *transfer-matrix.* To achieve separation of the different speech signals these methods attempt to invert the transfer matrix. In [290], one such transfer matrix–based approach has been proposed. The method assumes that there are two independent signals present in the degraded signal and two separate degraded inputs are available for processing. The separation of the two signals is achieved in two stages. The first stage consists of two linear predictors which whiten the input signals. The coefficients of the linear predictors are estimted using the LMS algorithm. The second stage consists of decoupling filters that are estimated by imposing the constraint of statistical independence on the outputs. A transfer matrix–based approach has also been proposed in [292]. The transfer matrix is estimated by minimizing the mutual information between the signals.

In this work, we consider the case of mixture of two voices. We propose a method for enhancement of speech of each of the two speakers with respect to the speech of the other speaker. Enhancement of a speaker's voice is achieved by performing relative emphasis of the short segments of the signal immediately after the instants of significant excitation corresponding to that speaker, with respect to the other segments of the signal. The relative emphasis is achieved by giving a larger weight to the LP residual signal samples in the region around the instants of significant excitation and a lower weight to the samples in the other regions. The LP residual signal is derived from the degraded speech signal. The weighted residual signal is passed through the time–varying all–pole filter to obtain enhanced speech. Clearly, it is necessary to know the instants of significant excitation corresponding to the speech of each of the speakers, to perform weighting. However, instants of significant excitation extracted from the degraded speech signal will have instants corresponding to both the speakers as well as some spurious instants. The instants of significant excitation corresponding

to the two speakers are obtained using the properties of the residual signal in short (1–3 ms) segments. This is because, the characteristics of the vocal tract and the voice source of a speaker are better preserved in the short (1–3 ms) segments of the signal immediately after the instants of significant excitation corresponding to the speech of that speaker.

8.1.2 Proposed Method for Speech Enhancement

8.1.2.1 Basis for the proposed method

The following experiment was conducted to demonstrate the efficacy of the instants of significant excitation for enhancement of speech degraded by speech of a competing speaker. Firstly, the instants of significant excitation corresponding to the clean speech of each of the two speakers were extracted. Two weight functions were derived for the two speakers. The weight function for a speaker was generated by convolving an (asymmetric) normalized Gaussian-bell-shaped curve with the chain of instants of significant excitation for the utterance of that speaker. Next, the degraded speech was generated by adding the speech signals corresponding to the utterances of the two speakers. The LP residual signal was computed from the degraded speech signal using 20 ms analysis windows and overlapped by 10 ms. The LP residual signal was multiplied by the two weight functions to generate two weighted LP residual signals corresponding to the two speakers. The two weighted LP residual signals were used to excite the time-varying all–pole filter derived from the degraded speech to generate enhanced speech signals corresponding to the two speakers. It was observed that the spectrograms of the enhanced speech signals clearly showed the horizontal striations due to the pitch periodicity. These striations were smeared in the spectrogram of the degraded speech signal. Informal listening tests also confirmed that there was a significant enhancement of the speech of each of the speakers with respect to that of

the other speaker.

Clearly, the above experiment assumes that the instants of significant excitation corresponding to the speech of each speaker are available to us. However, the locations of the instants of significant excitation corresponding to the speech of each speaker have to be estimated from the degraded speech signal. This is discussed in the following subsections.

8.1.2.2 Issues in the proposed method

The issues that arise in the proposed method are:

- Identification of instants of significant excitation for determining the short (1–3 ms) high energy regions corresponding to each speaker.
- Classification of extracted instants into the two speaker classes.
- Weighting the LP residual signal to enhance the characteristics of the desired speaker.

The instants of significant excitation are extracted for the degraded speech using the method given in Table–7.1. The duration of the analysis window for computation of the group–delay function was chosen as 4 ms to capture closely–spaced instants also. The sequence of instants so derived includes the instants of significant excitation corresponding to the voices of both the speakers as well as due to secondary excitations in the speech signal.

In the degraded speech, the important characteristics of a speaker's voice are preserved in the short (1–3 ms) interval immediately after each instant of significant excitation in the voiced regions. This is because of the relatively higher energy of the desired signal in these short segments. Therefore the characteristics of the speech signal as well as the LP residual signal in these short segments can be used for classification

of the instants corresponding to each of the speakers' voices.

8.1.2.3 Classification of instants

The gross spectral features corresponding to the voices of both the speakers are captured by the **LP** spectral envelope. The order of **LP** analysis is 14–16. Inverse filtering removes the gross spectral envelope. Therefore, in the **LP** residual signal the characteristics of the speaker are reflected in the short (1–3 ms) segments immediately after the instants of significant excitation.

The normalized crosscorrelation coefficient c_{ij} [276] would show how much a 2 ms segment of the residual signal resembles another 2 ms segment. The normalized crosscorrelation coefficient is computed between the 2 ms segment immediately after every instant of significant excitation i and a similar 2 ms segment for the four neighbouring instants of significant excitation $(j = 1, \ldots, 4)$ succeeding the instant i. The normalized crosscorrelation coefficient is given by

$$c_{ij} = \frac{\sum_{n=1}^{M} x_i(n)\, x_j(n)}{\left[\sum_{n=1}^{M} x_i^2(n)\right]^{\frac{1}{2}} \left[\sum_{n=1}^{M} x_j^2(n)\right]^{\frac{1}{2}}} \tag{8.1}$$

where $x_i(n)$ are the samples of the residual signal in the 2 ms segment immediately after the ith instant of significant excitation, x_j(n) are the samples of the residual signal after the jth neighbouring instant of significant excitation and M (=16) is the number of samples in a 2 ms segment. To allow for an error of a few samples in the estimation of the locations of instants, a search is performed for the maximum crosscorrelation value by shifting the 2 ms segment after each instant either side by 5 samples. Among the four neighbouring instants of excitation $(j = 1, \ldots, 4)$ the one which yields the highest crosscorrelation and also satisfies the pitch constraint is chosen as the next instant in the chain, after the instant i. The pitch constraint that is imposed on the separation

between the instants of a chain is that, the reciprocal of the separation between the instant i and its jth neighbour should lie in the range 70–330 Hz. In every voiced region we may obtain several such chains of instants of excitation. For a given chain, let the separation between the ith instant and the $(i+1)$th instant be denoted by L_i (in number of samples). For each chain the mean L_i is obtained (denoted by $\overline{L}_i$). To select the chain (or two chains) corresponding to the speaker (two speakers) among all the chains, only those chains are considered whose maximum absolute deviation of L_i about the mean separation $\overline{L}_i$ is less than 10% of $\overline{L}_i$. Among the considered chains, the two chains which give the least deviation are chosen. Depending on the $\overline{L}_i$ of the speakers in the previous voiced region, the chain (or chains) chosen in the present voiced region is associated with one of the two speakers. Thus we obtain two chains of instants for the entire utterance corresponding to the two speakers. A weight function is obtained by placing a normalized Gaussian–bell–shaped curve centred at each instant in the chain of instants corresponding to each of the two speakers. The LP residual signal is multiplied with the two weight functions to generate two weighted residual signals. The weighted residual signals are used to excite the time–varying all–pole filter to generate enhanced speech signals corresponding to the two speakers. The different steps in the proposed method are given in Table–8.1.

The method is illustrated on a segment of speech where the voiced regions of the two speakers overlap. Both the speakers are male speakers. A 100 ms segment of the clean speech corresponding to one of the speakers (speaker # 1) is shown Fig. 8.1(a). The instants of significant excitation for this 100 ms segment are shown in Fig. 8.1(b). A 100 ms segment of a voiced region in the speech of the competing speaker is shown in Fig. 8.1(c). The instants of significant excitation for the signal in Fig. 8.1(c) are shown in Fig. 8.1(d). The degraded speech signal obtained by adding the signal

Table 8.1: Algorithm for enhancement of speech degraded by speech of a competing speaker.

Identification of voiced regions

- Compute the mean μ_E and the standard deviation σ_E of the log-energy values computed in 20 ms frames.
- Identify a frame as voiced if the log-energy in that frame is greater than $\mu_E - \sigma_E$. Else identify the frame as nonvoiced.

Computation of the weight function

- In each voiced region of the degraded signal identify the instants of significant excitation using the method in Table-7.1.
- For each instant i, compute the normalized crosscorrelation coefficient between the 2 ms segment in the residual signal immediately after the instant of significant excitation and a similar 2 ms segment corresponding to the succeeding four neighbouring instants. Of the four, choose the instant which gives the maximum crosscorrelation, subject to the constraint that the separation between the two instants (L_i) lies within the bounds 3 ms and 14 ms (corresponding to a pitch of 330 Hz and 70 Hz, respectively). Obtain all such chains of instants for each voiced region.
- Compute the mean of the separation L_i between every pair of successive instants in each chain. The mean value is denoted by $\overline{L}_i$.
- Among all the chains choose only those chains whose maximum absolute deviation of separations L_i about the mean separation $\overline{L}_i$ is less than 10% of the mean of the separation between the instants $(\overline{L}_i)$. From all the chains so chosen, identify the two chains which have the least deviation about the mean, as belonging to the two speakers. A chain is classified as belonging to one of the two speakers based on the mean separation $(\overline{L}_i)$ between the instants of the two speakers in the previous voiced region.
- Derive the weight functions γ_{n1} and γ_{n2} for the two speakers by convolving a normalized Gaussian-bell-shaped curve with the chain of instants corresponding to the voice of each speaker.

Synthesis of enhanced speech

- Multiply the LP residual signal derived from the degraded speech with the two weight functions γ_{n1} and γ_{n2} to derive two weighted residual signals e_{n1} and e_{n2} corresponding to the two speakers. The LP residual signal is derived by performing 14th order LP analysis on 20 ms Hamming windowed segments of the degraded speech signal. The analysis frames are overlapped by 10 ms.
- The two weighted residual signals e_{n1} and e_{n2} are used to excite the time-varying all-pole filter to obtain enhanced speech corresponding to each speaker.

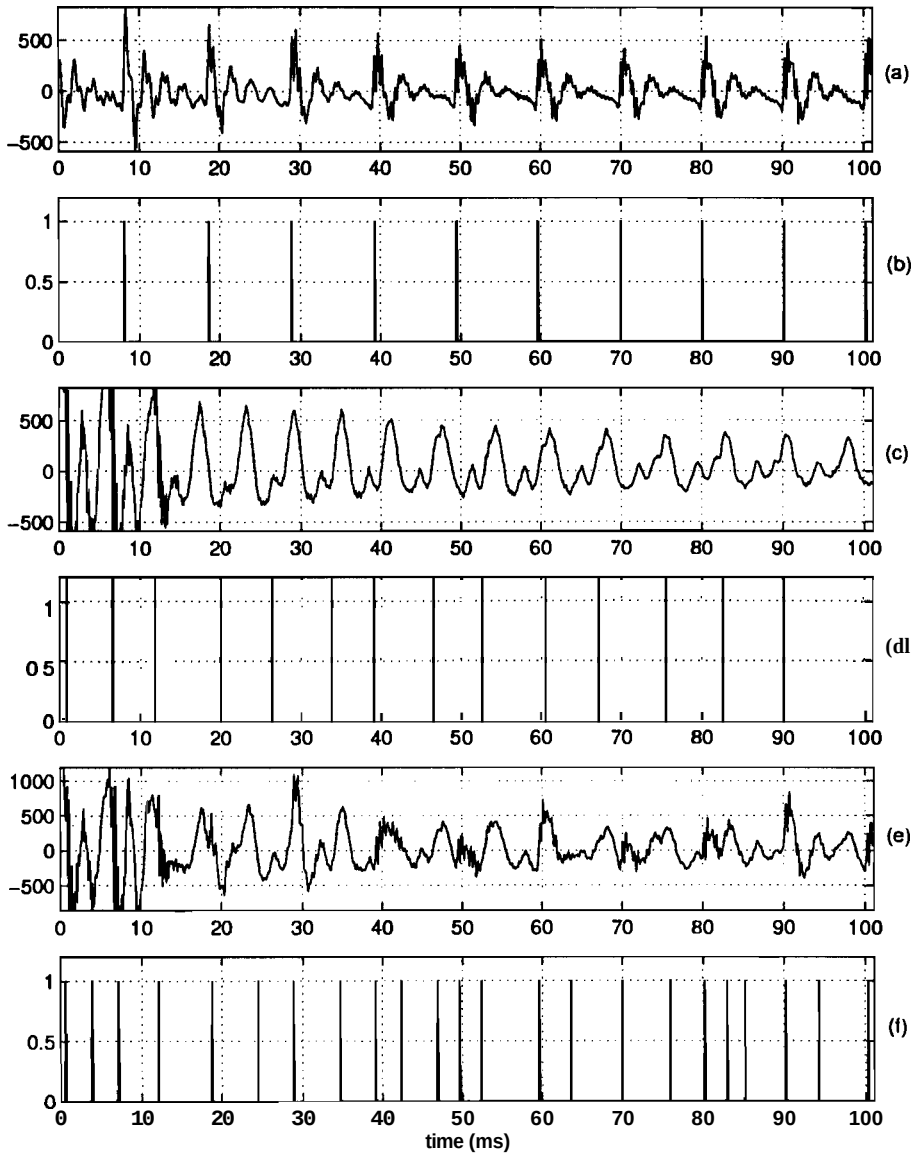


Fig. 8.1: Instants of significant excitation for clean speech, competing speaker's speech and degraded speech. (a) Clean speech signal, (c) a segment of the speech of the competing speaker, (e) degraded speech signal. (b),(d),(f) – instants of significant excitation for the signals in (a),(c),(e), respectively.

in Figs. 8.1(a) and 8.1(c) is shown in Fig. 8.1(e). Note that the periodicity due to the pitch of either of the two speakers cannot be made out from the degraded signal in Fig. 8.1(e). The instants of significant excitation derived from the degraded signal in Fig. 8.1(e) are shown in Fig. 8.1(f). By comparing Figs. 8.1(b), 8.1(d) and 8.1(f), we observe that the sequence of instants in Fig. 8.1(f) has instants corresponding to the voices of both the speakers as well as instants due to some secondary excitations in the signal. The objective now is to separate the chain of instants–of-excitation corresponding to each speaker from the sequence in Fig. 8.1(f). The normalized crosscorrelation values (c_{ij}) computed between each of the 22 instants of significant excitation in Fig. 8.1(f) and the 4 immediate neighbours to the right of each of these instants are given in Table–8.2. Each column in the table corresponds to each of the four neighbours ($j = 1, \cdots, 4$) of the ith instant, $i = 1, \cdots, 22$. The maximum crosscorrelation value in each row is shown in boldface print and is also pointed to by an arrow. Similarly, the separation l_{ij} (in number of sampling periods) between the ith instant of significant excitation and its four neighbouring instants of significant excitation are given in Table–8.3.

Table 8.2: Normalized crosscorrelation values c_{ij} computed between the 2 ms segment of the residual signal immediately following each instant i and its four succeeding neighbours $j = 1,\ldots,4$. The highest crosscorrelation coefficient corresponding to each instant i is shown in boldface print and is pointed to by an arrow.

Instant No.	Neighbour#1	Neighbour#2	Neighbour#3	Neighbour#4
i	c_{i1}	c_{i2}	c_{i3}	c_{i4}
1	0.3236	0.1986	**0.9004** ←	0.5057
2	0.2246	0.1089	0.2229	**0.5570** ←
3	0.2931	0.3991	**0.5080** ←	0.1633
4	**0.6441** ←	0.2695	0.4363	0.5325
5	0.3906	**0.6528** ←	0.4887	0.4768
6	0.5532	0.5545	0.4047	**0.6701** ←
7	0.4259	**0.8490** ←	0.4768	0.5137
8	**0.5630** ←	0.4874	0.4706	0.5425
9	0.4917	0.3664	**0.6502** ←	0.4727
10	0.4196	0.3164	**0.4617** ←	0.2443
11	0.4601	0.5518	**0.5904** ←	0.4535
12	0.4315	0.4304	0.5057	**0.5063** ←
13	**0.4937** ←	0.1385	0.2689	0.3657
14	0.2899	0.7302	0.3051	**0.7973** ←
15	**0.6010** ←	0.5512	0.4859	0.4414
16	0.6121	**0.6270** ←	0.4076	0.3990
17	0.4789	**0.6882** ←	0.3004	0.4039
18	0.2377	0.3879	**0.9076** ←	0.2567
19	0.3582	**0.6814** ←	0.2536	0.5780
20	0.3107	**0.5927** ←	0.2756	0.3316
21	0.4124	0.3065	**0.7005** ←	0.2978
22	0.1674	**0.6107** ←	0.5797	0.4495

Table 8.3: Separation in number of samples of the ith instant from its four neighbours $j = 1, \cdots, 4$, denoted by l_{ij}. The separation from the neighbour which yields the highest crosscorrelation coefficient is shown in boldface print and pointed to by an arrow.

Instant No.	Neighbour#1	Neighbour#2	Neighbour#3	Neighbour#4
i	l_{i1}	l_{i2}	l_{i3}	l_{i4}
1	26	52	**92** ←	145
2	26	66	119	**165** ←
3	40	93	**139** ←	174
4	**53** ←	99	134	181
5	46	**81** ←	128	163
6	35	82	117	**143** ←
7	47	**82** ←	108	144
8	**35** ←	61	97	119
9	26	62	**84** ←	106
10	36	58	**80** ←	138
11	22	44	**102** ←	134
12	22	80	112	**163** ←
13	**58** ←	90	141	189
14	32	83	131	**165** ←
15	**51** ←	99	133	155
16	48	**82** ←	104	122
17	34	**56** ←	74	114
18	22	40	**80** ←	113
19	18	**58** ←	91	139
20	40	**73** ←	121	161
21	33	81	**121** ←	152
22	48	88 ←	119	144

The separation from the neighbour which yields the highest crosscorrelation coefficient is shown in boldface print and also pointed to by an arrow. We now proceed as follows: Starting from the first instant, its third neighbour i.e., the 4th instant, gives the highest crosscorrelation value. The separation of the third neighbouring instant from the first instant is $l_{13} = 92$ samples. Since l_{13} satisfies the pitch constraint that $40 < l_{13} < 112$ samples, where 40 samples correspond to 5 ms and 112 sam-

ples correspond to 14 ms, respectively, at a sampling rate of 8 kHz, the 4th instant is chosen as the instant of significant excitation succeeding the first in this chain. If the pitch constraint mentioned above is not met by the neighbour with the highest corsscorrelation then the neighbour with the next highest crosscorrelation is considered, and so on. Similarly, the next instant in the chain, after the 4th, is chosen. This procedure is continued till the end of the voiced region is reached. The first chain of instants–of–excitation is thus obtained. Similarly, we now start from the 2nd instant and form another chain. The third chain starts from the 3rd instant. The 4th instant belongs to the chain of the first instant. Hence, it cannot be the starting point of a new chain. After obtaining all such possible chains, the chain with the least absolute deviation of the separation between instants of the same chain (i.e., L_i) about the mean separation $\bar{L}_i$ of that chain is chosen. For the example waveform considered in Fig. 8.1, the chain so chosen is given in Fig. 8.2(c). We observe that the chosen instants–of–excitation correspond to the actual instants of significant excitation for the signal in Fig. 8.1(a) (speaker # 1), which are reproduced in Fig. 8.2(a) for convenience. The instants of significant excitation obtained for the degraded speech signal, which are shown in Fig. 8.1(f), are also reproduced here in Fig. 8.2(b) for comparison. The weight function γ_{n1} derived for enhancing the speech of speaker # 1 is shown in 8.2(d). The weight function is derived by convolving a normalized (asymmetric) Gaussian–bell–shaped curve with the chain of instants in Fig. 8.1(c). The maximum and minimum values of the weight function are controlled by mapping the convolved sequence using a nonlinear function of the tanh type shown in Fig. 4.2.

In the signal segment shown in Fig. 8.1(a) the strengths of excitation corresponding to the voice of speaker # 2 are low. Therefore, the characteristics of the LP residual signal derived from the degraded signal are dominated by the voice characteristics of

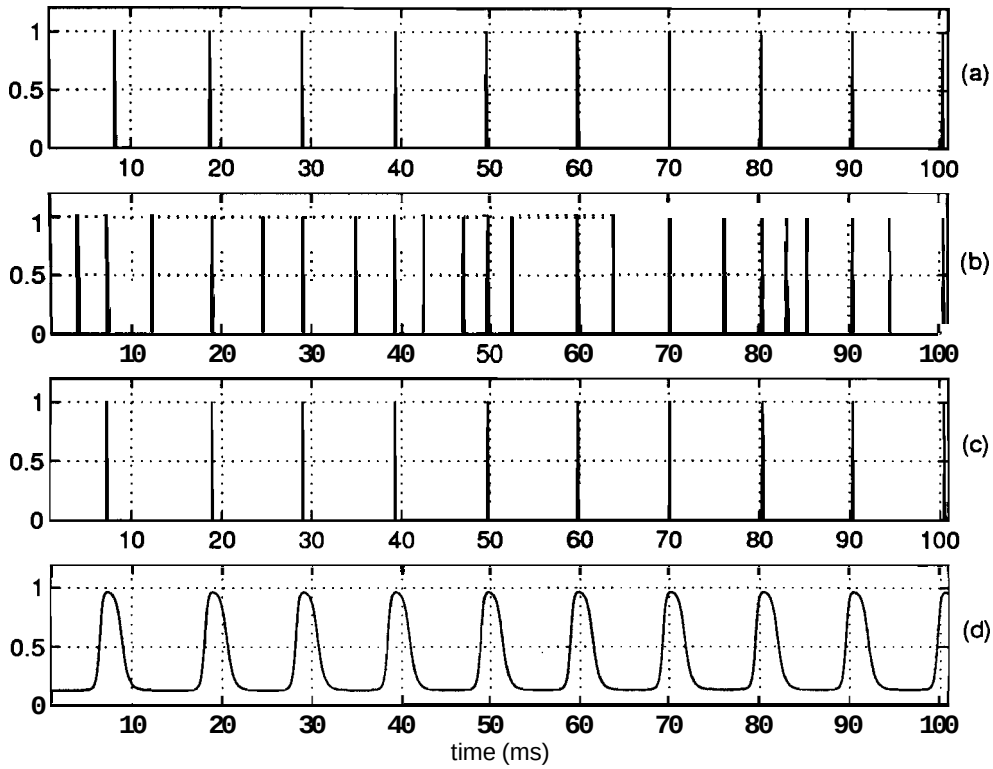


Fig. 8.2: (a) Instants of significant excitation for the clean signal in Fig. 8.1(a). (b) Instants of significant excitation for the degraded signal in Fig. 8.1(e). (c) Instants of significant excitation for speaker#1 separated from the mixture in (b). (d) Weight function derived for enhancing the speech of speaker # 1.

speaker # 1. Hence, the crosscorrelation–based approach does not yield good estimates of the locations of the instants of significant excitation corresponding to the voice of speaker # 2.

In the next section we present some results of enhancement of speech degraded by speech of a competing speaker.

8.1.3 Experimental Studies

The data for the study in this section was obtained by adding the speech of two male speakers. The amplitude of the speech signal corresponding to each speaker was scaled so that the total energy of the two utterances is approximately the same. The duration

of each utterance is about 3 s. The utterance of one of the speakers (# 1) corresponds to the sentence "***Any dictionary will give atleast any ...***". The utterance of the other speaker (# 2) corresponds to the sentence "***She had your dark suit in greasy wash water all yea***+' taken from the TIMIT database. The average pitch frequencies of the two speakers are 130 Hz and 110 Hz. The results of enhancement are shown in Fig. 8.3. In the figure, the panels labelled (a), (b) and (c) show the spectrograms of the clean speech signal, the degraded speech signal and the processed speech signal, respectively, for speaker # 1. By comparing the spectrograms (b) and (c), we observe that a there is significant attenuation of the formant structure of the competing speaker in the regions around 1.25 s and 2 s. In the spectrogram (b), in the region around 1.6 s, the horizontal striations due to the pitch are smeared due to the speech of the competing speaker. These horizontal striations are enhanced in spectrogram (c). However, the enhancement cannot be seen in other regions, for example in the region between 2.5 s and 3.0 s. This is because the instants of significant excitation of speaker # 1 have not been selected correctly in this region. Thus there is a need for **a** measure to supplement the crosscorrelation–based measure to select the instants of significant excitation corresponding to the two speakers reliably. Secondly, due to the domination of the instants of significant excitation of speaker # 1 over those of speaker # 2 due to higher strengths, the crosscorrelation–based method does not identify the instants of significant excitation of speaker # 2 reliably. Therefore there was no perceptible enhancement in case of the speech of speaker # 2.

8.2 COMB FILTERING OF NOISY SPEECH USING INSTANTS OF SIGNIFICANT EXCITATION

The pitch periodicity in voiced speech has been exploited for enhancement of noisy speech. One of the methods based on pitch periodicity is ***adaptive*** comb *filtering* [162,

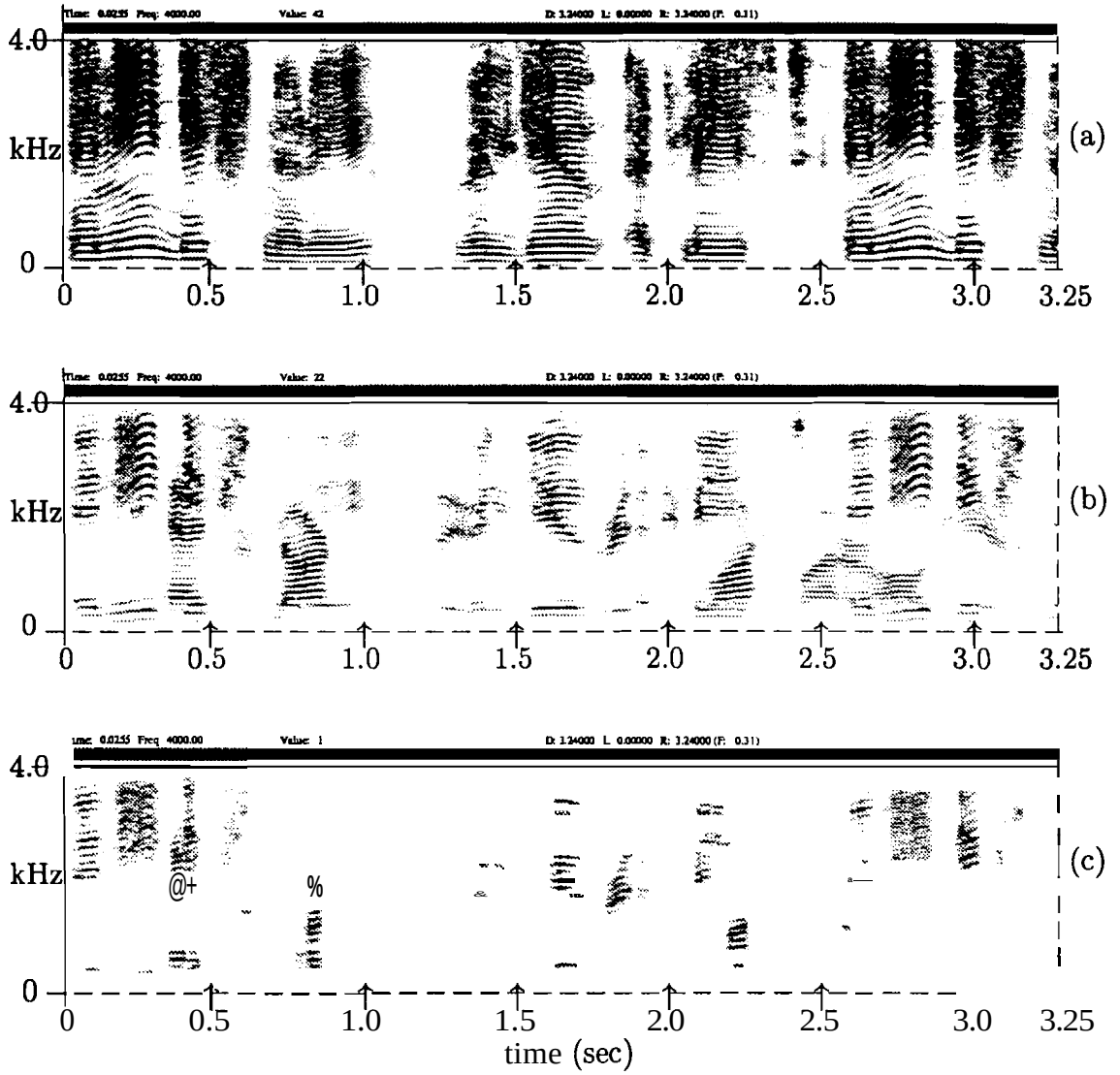


Fig. 8.3: Comparison of spectrograms before, and after processing speech degraded by speech of a competing speaker. (a) Spectrogram for clean speech of speaker # 1. (b) Spectrogram for degraded speech. (c) Spectrogram for processed speech of speaker # 1.

238,2961. This method is based on the fact that in voiced speech the spectral energy resides only at the pitch harmonics while the spectral distribution of noise may span most of the frequency range. Therefore, the annoyance due to noise in the speech could be reduced by attenuating the spectral energy of noisy speech between the pitch harmonics. This can be achieved by using a filter whose frequency response exhibits peaks at the pitch harmonics and valleys in between, and hence it resembles the teeth of a *comb*. The comb filter is a finite impulse response (FIR) filter with the filter taps spaced a glottal cycle length apart. To obtain a linear phase response for the filter, often the taps are made symmetric with respect to the midpoint. Clearly, comb filtering requires an accurate estimation of the glottal cycle length in the voiced regions of the noisy signal to be filtered. In the traditional method of comb filtering, a separate pitch estimation algorithm such as *Simplified Inverse Filter Tracking* (SIFT) [297] or the *maximum likelihood pitch estimation* algorithm [298] is used to estimate the pitch from noisy speech. We have seen in Chapter 7 that the instants of significant excitation can be reliably estimated from noisy speech using the group–delay–based method. In this section we use the instants of significant excitation to construct the impulse response of the comb filter.

8.2.1 Frequency Response of a Comb Filter

Let a comb filter with 2M + 1 taps have an impulse response $h_c(n)$ given by

$$\begin{aligned} h_c(n) &= \sum_{k=-M}^{M} c_k\, \delta(n - kL) \\ &= c_0\, \delta(n) + \sum_{k=1}^{M} [c_k\, \delta(n - kL) + c_{-k}\, \delta(n + kL)] \end{aligned} \tag{8.2}$$

where c_k, k $= -M \cdots$ M are the non–zero taps of the filter. To preserve the phase of the speech signal, a comb filter with a zero–phase response is preferable. Therefore, we choose a filter with a symmetric impulse response i.e., $c_k = c_{-k}$. Equation (8.2)

can now be written as

$$h_c(n) = c_0\,\delta(n) + \sum_{k=1}^{M} c_k\,[\delta(n-kL) + \delta(n+kL)] \tag{8.3}$$

The Fourier transform of (8.3) is

$$H_c(\omega) = c_0 + \sum_{k=1}^{M} 2c_k\,\cos(\omega kL) \tag{8.4}$$

which is a real function of w and hence has a zero–phase response. For M = **1,** the response of a 3–tap filter from (8.4) can be written as

$$H_c(\omega) = c_0 + 2c_1\,\cos(\omega L) \tag{8.5}$$

which has maxima at $\omega_l^{max} = \left(\frac{2\pi}{L}\right)1,\ l = 0,1,2,\cdots$. The maximum value of the magnitude response $|\,H_c(\omega)\,|$ is $c_0 + 2c_1$. The minima of $|\,H_c(\omega)\,|$ occur at $\omega_l^{min} = \frac{(2l+1)\pi}{L},\ l = 0,1,2,\cdots$. The minimum value of the magnitude response is $c_0 - 2c_1$, assuming that $c_0 > 2c_1$. To achieve an attenuation of **A** dB in the regions midway between pitch harmonics and a gain of 0 dB at the pitch harmonics, the coefficients of a third order filter are to be chosen as

$$c_0 = \frac{1 + 10^{-\left(\frac{A}{20}\right)}}{2} \tag{8.6}$$

$$c_1 = \frac{1 - 10^{-\left(\frac{A}{20}\right)}}{4} \tag{8.7}$$

In Chapter 7, we have seen the robustness of the group–delay–based method for extraction of instants of significant excitation to additive random noise and channel distortion. Since the instants of significant excitation are spaced a glottal cycle length apart they could be used as the taps of the comb filter. The number of taps and the shaping window for the impulse response may be chosen based on the desired attenuation between the pitch harmonics. Fig. 8.4(a) shows the frequency response of a 3–tap comb filter assuming a pitch frequency of **100** Hz and a sampling rate of **11.025** kHz.

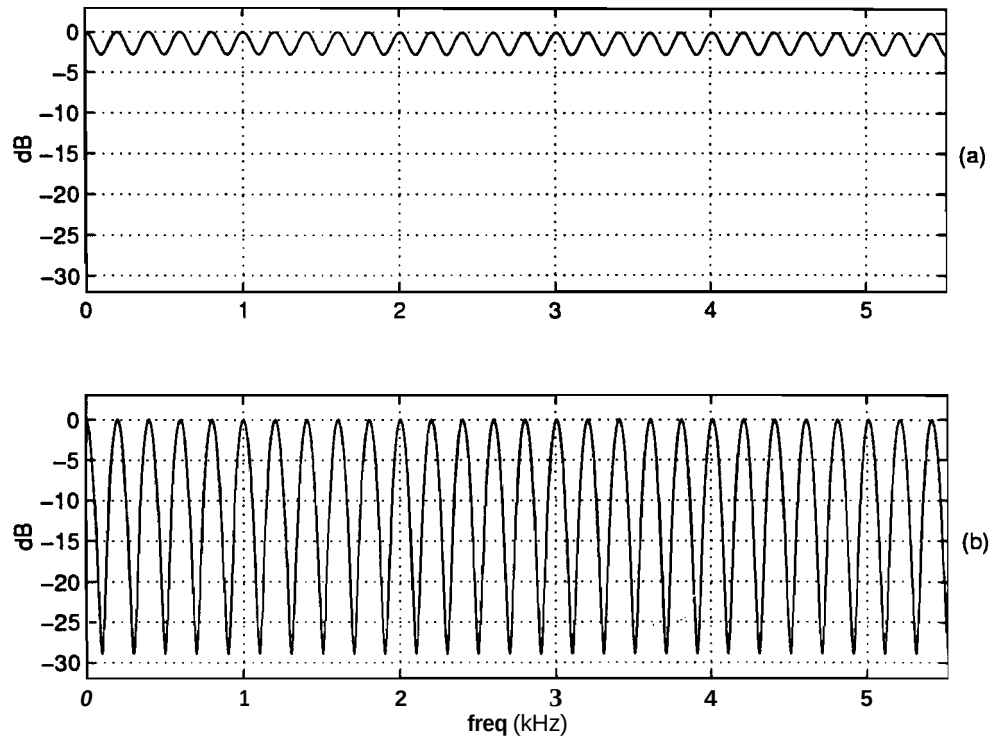


Fig. 8.4: Comparison of frequency responses of (a) 3–tap and (b) 5–tap comb filters.

Fig. 8.4(b) shows the frequency response of a 5–tap filter. A Hamming window has been used for shaping the impulse responses of both the filters. The peaks in the frequency response correspond to harmonics of the pitch frequeny of 100 Hz. The 3–tap filter provides an attenuation of 3 dB between the pitch harmonics while the 5–tap filter provides an attenuation of 30 dB.

8.2.2 Experimental Results

Fig. 8.5(a) shows a voiced speech segment corresponding to the diphthong /*ai*/ as in *five*. The duration of the utterance is about 1.1 s. The sampling frequency is 11.025 kHz. Fig. 8.5(c) shows a noise corrupted version of the same segment at an average SNR of 10 dB. Fig. 8.5(e) shows the noisy signal after comb filtering. A 3–tap comb filter was used. The taps were chosen using (8.6) and (8.7) so as to yield an

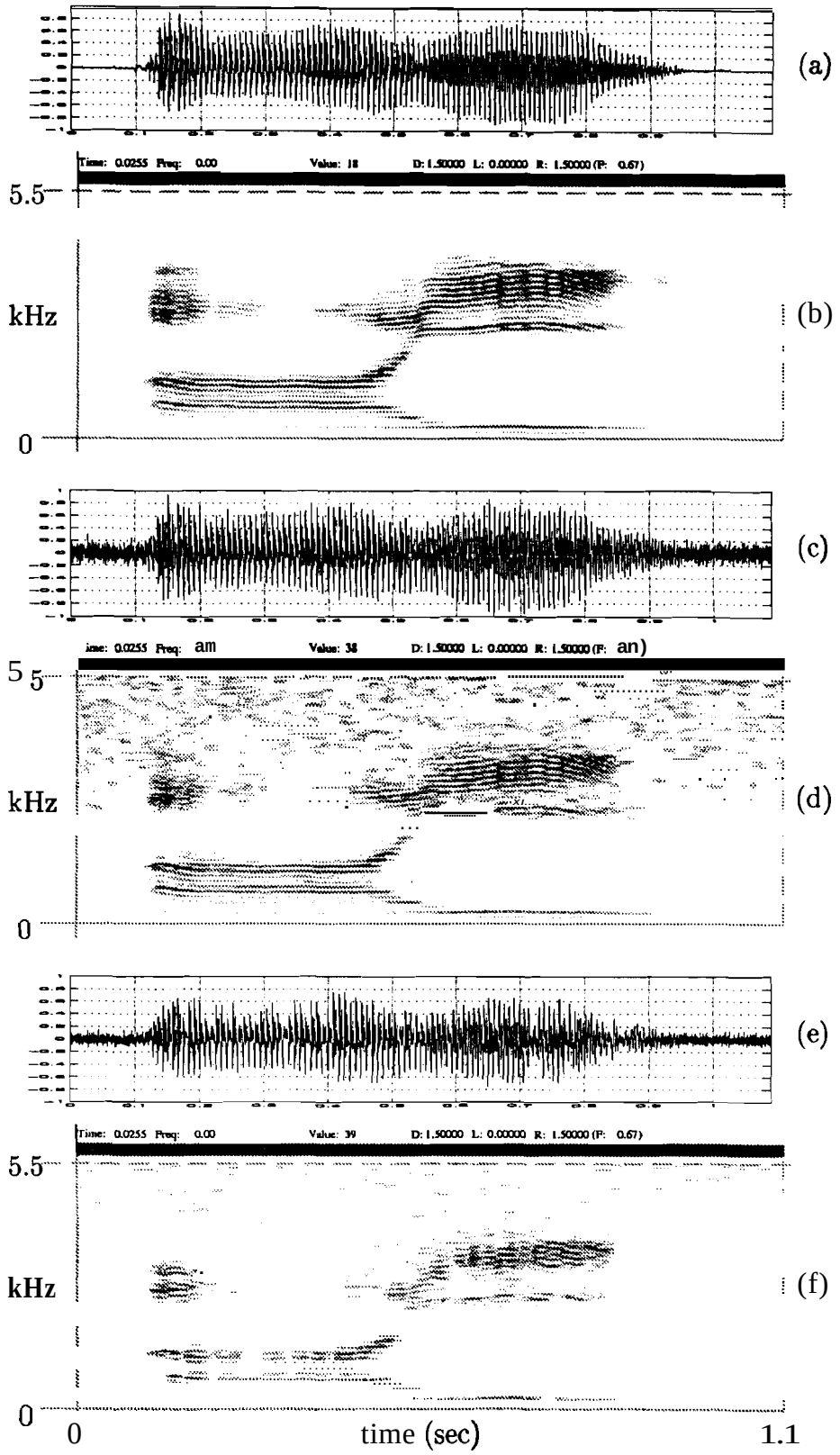


Fig. 8.5: Results of enhancement of speech degraded by additive white noise. (a) Clean speech. (c) Speech signal at 10 dB SNR. (e) Enhanced speech signal obtained using comb filtering. (b),(d),(f) – spectrograms for the signals in (a),(c),(e), respectively.

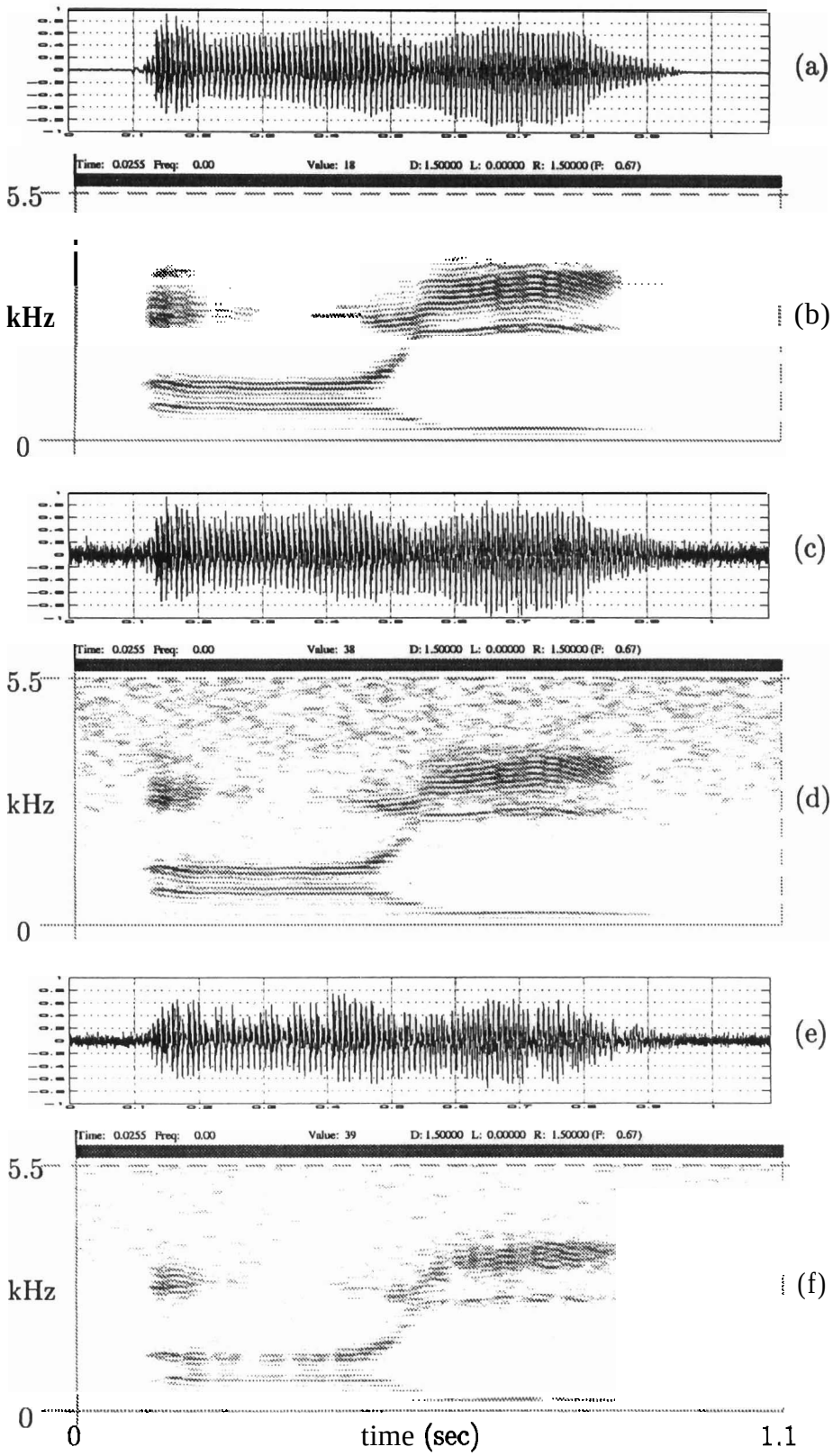


Fig. 8.5: Results of enhancement of speech degraded by additive white noise. (a) Clean speech. ***(c)*** Speech signal at 10 dB SNR. (e) Enhanced speech signal obtained using comb filtering. (b),(d),(f) – spectrograms for the signals in (a),(c),(e), respectively.

attenuation of 50 dB in the regions between the pitch harmonics. The tap spacing L was chosen as the distance between the current instant of significant excitation and the next. The tap spacing is updated at every instant of significant excitation. For a comb filter with more than three taps, distortion may be introduced into the processed speech due to severe attenuation of spectral amplitude in the regions between the pitch harmonics. It is also reported in [162] that as the number of taps of the comb filter are increased the intelligibility of processed speech reduces. The Figs. 8.5(b), 8.5(d) and 8.5(f) show the spectrograms corresponding to the Figs. 8.5(a), 8.5(c) and 8.5(e), respectively. From a comparison of Figs. 8.5(d) and 8.5(f), we observe that there is significant attenuation of noise power in the higher formant regions where the SNR is poor. Informal listening tests confirm that the annoyance due to noise is reduced. To illustrate the improvement due to comb filtering, a voiced segment was chosen above for processing. However, the above method can be used in general for enhancement of noisy speech. Due to lack of harmonic structure in the unvoiced regions of the speech signal to be processed, there is not likely to be any improvement in the unvoiced regions.

8.3 ANALYSIS OF DEGRADED SPEECH USING INSTANTS OF SIGNIFICANT EXCITATION

One of the effects of noise contamination of a signal is the reduction in the spectral dynamic range of the signal [268]. Hence, in the case of noisy speech signals, the higher formant peaks in the spectrum are lost. It is also well known that the conventional method of LP analysis is sensitive to noise [299]. Moreover, noise introduces spurious peaks into the spectrum of the estimated all–pole model [98]. In this section, we show that pitch synchronous spectral analysis is more effective for noisy speech.

Pitch synchronous analysis of noisy speech is known to improve the results of

analysis [19]. One important advantage of pitch synchronous analysis is that the variability introduced into the results of analysis due to arbitrary placement of the analysis window is now avoided. Moreover, pitch synchronous analysis facilitates the tracking of formant changes within a glottal cycle and also from one glottal cycle to another [10, 23, 25].

The three panels in Fig. 8.6 show the results of pitch synchronous weighted-covariance analysis for the cases of clean speech in Fig. 8.7(a), noisy speech in Fig. 8.7(c) and telephone speech in Fig. 8.7(e), respectively. Fig. 8.7 is same as Fig. 7.5 and is reproduced here for convenience. The instants of significantexcitation are determined using the algorithm given in Table-7.1. The LP spectra, plotted as mesh plots in Fig. 8.6 are obtained by performing a 12th order weighted-covariance analysis. The conventional covariance analysis performed using a 3 ms analysis window placed 3-4 samples after the instant of significant excitation (to avoid transient effects) gives spurious peaks. Hence, to determine the LPCs corresponding to the 3 ms duration immediately after a given instant of significant excitation, the covariance coefficients computed for the present instant, the previous instant and the next instant are averaged. In voiced speech, this short region after the instant of significant excitation is not only the closed glottis region but also, usually, has a high SNR. Choosing a segment for analysis in this region gives reliable estimates of the vocal tract characteristics. The LPCs are determined by solving the normal equations using the averaged covariance coefficients. The LP analysis performed using averaged covariance coefficients is equivalent to the weighted-covariance analysis, if the weight function is chosen such that it takes a value of unity in the three desired 3 ms regions and zero otherwise. The three desired 3 ms regions are chosen to be 4-5 samples after each of the three successive instants of significant excitation. The mesh plot of LP spectra in the top

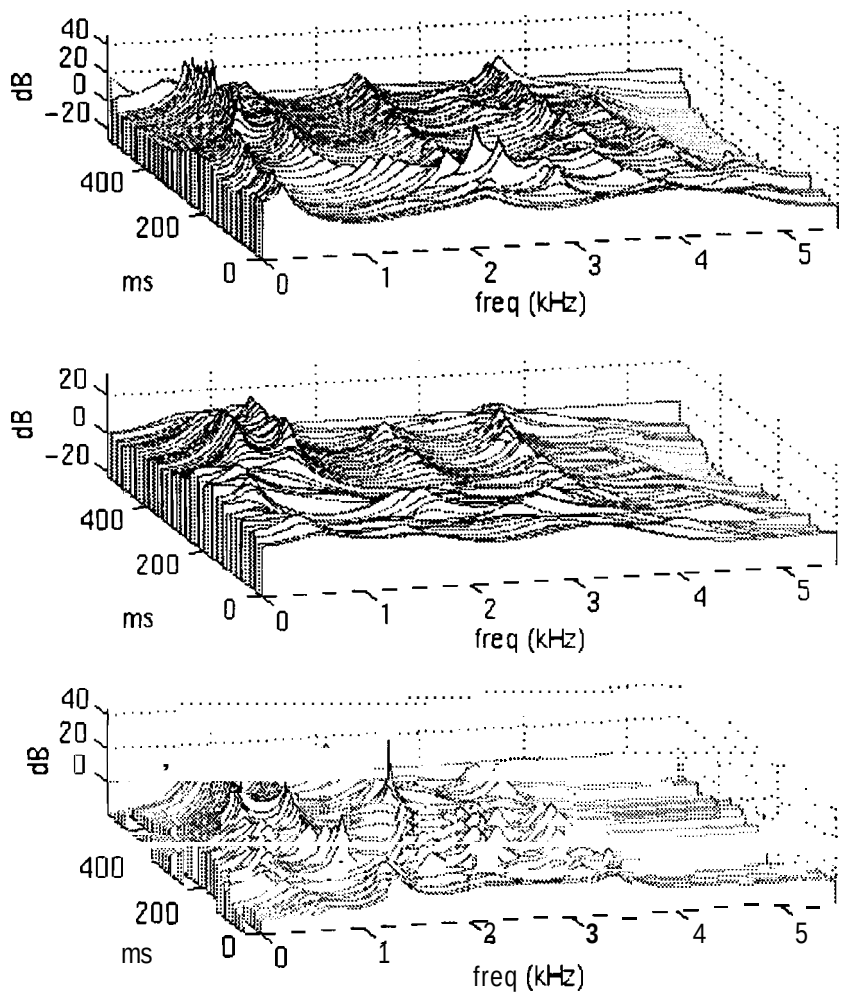


Fig. 8.6: Pitch–synchronous weighted-covariance LP analysis of the utterance /*dz ua*/ simultaneously sampled with a mic 10 cm away from the lips, with another mic 45 cm away under noisy conditions and on a telephone channel. Top panel: Mesh plot of power spectra for clean speech (Sampling rate: 11 **kHz).** Middle panel: Mesh plot of power spectra for noisy speech (Sampling rate: 11 **kHz).** Bottom panel: Mesh plot of power spectra for telephone speech (Sampling rate: 8 **kHz).**

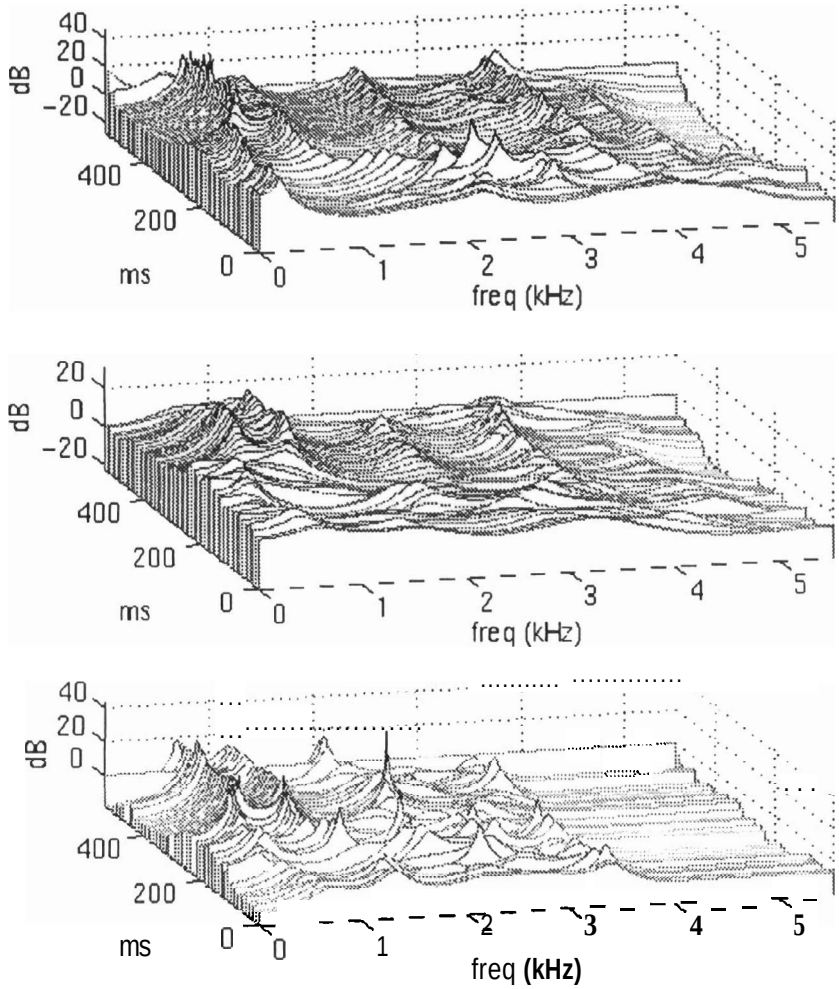


Fig. 8.6: Pitch-synchronous weighted–covariance LP analysis of the utterance /*dz ua*/ simultaneously sampled with a mic 10 cm away from the lips, with another mic 45 cm away under noisy conditions and on a telephone channel. Top panel: Mesh plot of power spectra for clean speech (Sampling rate: **11 kHz).** Middle panel: Mesh plot of power spectra for noisy speech (Sampling rate: **11 kHz).** Bottom panel: Mesh plot of power spectra for telephone speech (Sampling rate: 8 **kHz).**

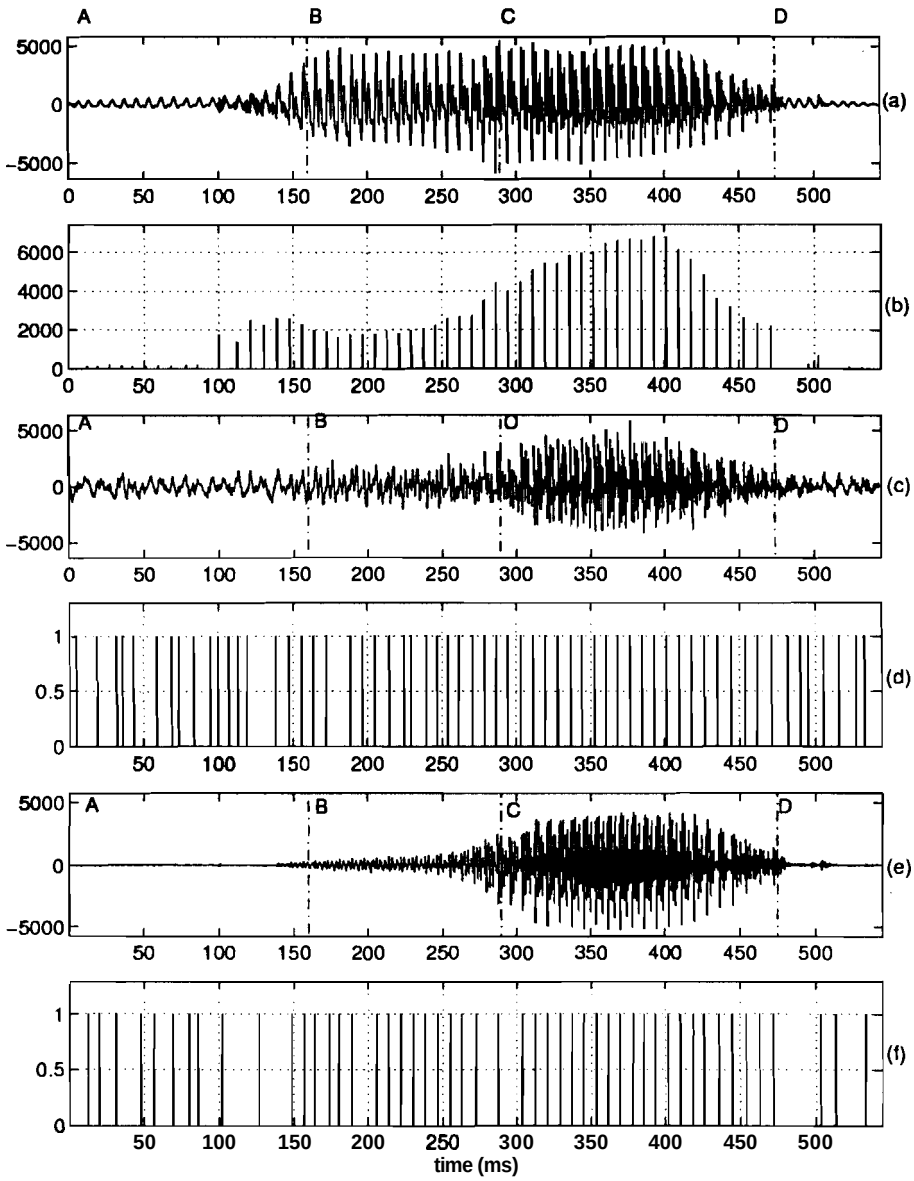


Fig. 8.7: (a) Clean speech for the utterance */dz ua/*. (b) Strengths of excitation based on the Frobenius norm. (c) Speech degraded by ambient noise. (d) Instants of significant excitation derived from the signal in (c). (e) Telephone speech. (f) Instants of significant excitation derived from the signal in (e).

panel of Fig. 8.6, which corresponds to clean speech, clearly shows the spectral transition from the vowel /u/ to the vowel /a/ and the first five formants of the vowel /a/. The mesh plot in the middle panel shows similar behaviour despite the fact that the signal in Fig. 8.7(c) has a poor SNR. The mesh plot in the bottom panel too shows the transition from /u/ to /a/ and the first four formants in vowel /a/ clearly, for the case of telephone speech. Since, the bandwidth of the telephone channel is only about 4 kHz, the order of LP analysis was reduced to 10 for telephone speech. Note that the fifth formant in vowel /a/ is lost due to the telephone channel.

8.4 SUMMARY

In this chapter we have considered the importance of instants of significant excitation and subsegmental analysis in some applications. We have discussed the usefulness of instants of significant excitation and subsegmental analysis for enhancing speech degraded by speech of a competing speaker. The proposed method is different from the traditional frequency domain–based methods, and also the methods which attempt to separate the voices. The proposed method attempts to identify the regions where the desired speech signal is stronger compared to the speech of the competing speaker and performs relative emphasis of these regions. The main problem in the proposed method for enhancement is the identification of the instants of significant excitation corresponding to the voice of each of the speakers from the instants of significant excitation extracted from the degraded speech signal. We have proposed a crosscorrelation–based method for separating the instants of excitation. However, if the strengths of excitation of one of the speakers are low in a particular segment, then the instants of significant excitation corresponding to the voice of that speaker cannot be obtained in that segment.

We have also shown that the instants of significant excitation can be used to

perform adaptive comb filtering of noisy speech. In the traditional method of comb filtering, a separate pitch estimation algorithm such as SIFT is used to estimate the pitch from noisy speech. The instants of significant excitation can also be used to identify short **(1–3** ms) high SNR segments in the voiced regions of noisy speech for accurate analysis.

Chapter 9

SUMMARY AND CONCLUSIONS

9.1 SUMMARY OF THE WORK

In this thesis we have studied the issues that arise in processing short (1–3 ms) segments of speech. Speech signal is produced as a result of excitation of the time–varying vocal tract system. Therefore, the speech signal is processed in short segments to capture the time–varying characteristics of the speech production mechanism. Conventional short–time spectrum–based methods have a poor frequency resolution when short (1–3 ms) segments of the signal are analysed. Therefore, in this thesis we have used the properties of the LP residual signal for processing short (1–3 ms) segments of speech. The speech signal collected under normal environmental conditions is usually degraded due to noise, reverberation and channel (e.g., telephone) distortion. Performance of automatic speech systems depends critically on the environmental conditions. The perceptual quality is also affected significantly due to degradations in the speech signal. Enhancement of speech is generally attempted to reduce annoyance due to degradation.

Methods for processing speech can be broadly classified into three categories:

- Suprasegmental level (100--300 ms of signal for analysis)
- Segmental level (10–30 ms of signal for analysis)
- Subsegmental level (1–3 ms of signal for analysis)

Methods for processing speech at the suprasegmental level are guided mainly by perception. Methods for processing speech at the segmental level (10–30 ms) are dictated

more by signal processing considerations, such as window effects and time–frequency resolution, rather than by the characteristics of the signal. The methods proposed to process speech at the subsegrnental level are primarily guided by the characteristics of the speech signal and the speech production mechanism.

In the suprasegmental and segmental analyses of speech, the temporal changes in the characteristics of the vocal tract and its excitation are smeared. For example, in consonant–to–vowel (CV) and vowel–to–consonant (VC) transitions, rapid changes occur in the vocal tract shape and excitation characteristics, in a duration of about 50–100 ms. Hence, an analysis window of duration 10–30 ms may not provide adequate temporal resolution of the changes in the vocal tract characteristics. The quasistationary assumption over a 10–30 ms duration is not valid even for steady voiced regions in the speech signal. This is because the vocal tract system changes due to coupling and decoupling of the trachea during open and closed phases of the glottal excitation, respectively. If short (1–3 ms) segments of speech signal are considered pitch synchronously for analysis, then one may capture the variations in similar segments in successive glottal cycles. The focus of the work presented in this thesis is on subsegmental analysis of speech and its application to speech enhancement.

To reduce the effects of the small (1–3 ms) size of the window on the results of analysis, we have proposed a new windowing procedure called *source–system windowing.* The method is based on the fact that in LP analysis, the characteristics of the vocal tract system not captured in the LPCs appear in the LP residual signal. Selecting a short (1–3 ms) segment in the residual signal would generate a signal whose characteristics will be similar to the characteristics of the speech signal corresponding to the selected window.

As an application of the subsegmental analysis we have proposed methods for

enhancement of speech corrupted by additive random noise and reverberation in small rooms. The proposed methods for enhancement primarily aim at emphasising the high SNR/SRR regions relative to the low SNR/SRR regions in the speech signal to reduce the annoyance due to the degradation. However, the relative emphasis should be accomplished without causing distortion in speech. The main objective of the work has been to address the signal processing issues involved in the enhancement of degraded speech. Therefore, formal intelligibility tests and listener ratings on the processed speech have not been carried out.

In practical conditions, we do not know a priori (i) whether a given speech signal is degraded or not, (ii) whether the degradation is due to noise or reverberation and (iii) the level of degradation. Therefore, we proposed a method based on the averaged normalized prediction error to identify the type and level of degradation. The averaged normalized prediction error is higher for a signal degraded by additive random noise, compared to that for the clean speech. When the speech signal is corrupted by reverberation, the averaged normalized prediction error is lower compared to that for the clean speech.

In the subsegrnental analysis, both the size and location of the short segment are crucial for accurate analysis of changes of the vocal tract system. Moreover, when speech is corrupted by additive noise, the short (1–3 ms) segments of the signal immediately after the instants of significant excitation usually have a high SNR. A method based on group-delay function was proposed in [202, 203] for determining the instants of significant excitation from speech signals. We have investigated the robustness of the group-delay-based method for determination of iristants of significant excitation. We proposed some refinements to the method motivated by the analysis. We have illustrated the robustness of the method on speech degraded by noise and telephone

channel effects.

To illustrate the importance of the instants of significant excitation and subsegmental analysis we have proposed a method for enhancement of speech corrupted by speech of a competing speaker.

9.2 MAJOR CONTRIBUTIONS OF THE WORK

The most important contribution of the research reported in this thesis is that it presents an attempt to process short (1–3 ms) segments of the speech signal for analysis and enhancement. Processing 1–3 ms segments of the speech signal poses time–frequency resolution problems. Therefore, the low correlation between the samples of the LP residual signal was exploited to perform the analysis and enhancement. The major contributions of this thesis are:

- A method for enhancement of noisy speech was proposed which performs relative emphasis of high SNR segments of the speech signal with respect to the low SNR segments.
- A method for enhancement of reverberant speech was proposed for relative emphasis of high SRR regions with respect to low SRR regions. Both the methods for speech enhancement are based on the fact that there is a low correlation between the samples of the LP residual signal. Therefore, the LP residual signal can be manipulated to some extent without introducing distortion into the processed speech signal.
- For speech data collected in practical situations, we do not know a priori whether the speech signal is clean or degraded. If degraded, whether the source of degradation is noise or reverberation. A method based on the averaged normalized prediction error was proposed to identify the type and level of degradation.

- The robustness of the group–delay–based method for extraction of the instants of significant excitation was illustrated for degradations in practical conditions such as additive broadband noise, telephone channel distortion and reverberation in small–rooms.
- A method for enhancement of speech corrupted by speech of a competing speaker was proposed. The method identifies the instants of significant excitation corresponding to the voice of each speaker. Short (1–3 ms) segments of the LP residual signal around each instant of significant excitation are emphasised and are used as excitation for synthesis of enhanced speech.
- The concept of source-system windowing for speech signals was proposed for analysing short (1–3 ms) segments of the signal.

9.3 DISCUSSION ON THE PROPOSED METHODS

The work presented in this thesis is based on the linear prediction analysis of speech and the linear prediction residual signal.

The methods proposed in Chapters 4 and 5 for enhancement of speech depend upon the results of LP analysis in the high SNR/SRR segments of the speech signal. However, when the SNR/SRR reduces to low (< 10 dB) levels, LP analysis performs poorly.

The criterion based on the averaged normalized prediction error for identification of the type and level of degradation given in Chapter 6 cannot be used when the additive noise and reverberation are simultaneously present in the degraded speech. This is because noise and reverberation have opposite effects on the normalized prediction error, and hence their effects cancel each other.

The group–delay–based method for determination of instants of significant exci-

tation given in Chapter 7 involves computation of a 256 or 512-point FFT for every sample of the residual signal to obtain the group-delay function. The computational cost could be reduced significantly by estimating the instants only in the voiced regions. Secondly, the group-delay function need not be computed for every sample. It could be computed once every 3 or 4 samples and whenever there is a zero-crossing, that region can be closely investigated [300]. With the help of a pitch-tracker the group-delay computation need be performed only for 4–5 successive samples in every glottal cycle, after detecting one of the instants of significant excitation. We have also seen in Figs. 7.6–7.9 that there is a 2–3% bias in the estimation of the location of instants of significant excitation for noisy speech. The bias may be negligible for most practical purposes, especially for large (> 8 ms) glottal cycle durations.

The results of analysis obtained using the source-system windowing procedure presented in Chapter 3 are meaningful only when the window is placed in relatively steady regions in a glottal cycle, e.g., the closed or open glottis region. If the system characteristics change significantly within the analysis window, then it is difficult to interpret the results. Secondly, in the case of some cases of voiced speech a closed phase may not exist at all.

9.4 DIRECTIONS FOR FUTURE RESEARCH

The methods proposed in this thesis for enhancement of speech under noisy and reverberant conditions do not explicitly take into consideration the unvoiced regions for enhancement. It was observed in some of our studies that weak unvoiced segments, which are often submerged by noise, are attenuated as a result of processing. This is perceived as distortion in the processed speech. Therefore, there is a need for a robust method which can reliably detect voiced and unvoiced regions in degraded speech. The method for voiced/unvoiced detection can be combined with the proposed methods for

enhancement for improved performance.

The method proposed for identification of the type and level of degradation based on the averaged normalized error may not be reliable when the speech signal is corrupted by both additive noise and reverberation simultaneously. Therefore, there is a need to explore a different criterion to supplement the averaged normalized prediction error–based criterion.

The group–delay–based method for extraction of the instants of significant excitation from noisy speech performs poorly when the SNR is poor (< 0 dB). This is because the LP analysis performs poorly under such conditions, and hence the residual signal reduces to a random noise-like waveform. Therefore it is necessary to enhance the speech signal without disturbing its phase characteristics before the group–delay–based method can be applied for extraction of the instants of significant excitation.

The method proposed for processing speech corrupted by speech of a competing speaker is based on the instants of significant excitation. The instants of significant excitation corresponding to the speech of each of the speakers were identified using the crosscorrelation between the **2** ms segments of the LP residual signal immediately after the instants of significant excitation. However, we found that the crosscorrelation–based criterion is not always reliable to identify the instants of significant excitation. The characteristics of the speech signal in the short **(1–3** ms) segments immediately after the instants of significant excitation need to be combined with the crosscorrelation–based criterion to arrive at a more reliable method for identifying the instants of significant excitation corresponding to the speech of each of the speakers from the degraded speech signal. The Itakura distance could be used to measure the similarity between two speech segments of duration 1–3 ms.

APPENDIX–A
LF–MODEL FOR DIFFERENTIATED GLOTTAL PULSE

Several methods have been proposed in literature for modeling the glottal pulse and its derivative [205, 301–303]. In this appendix, the Liljencrants–Fant (LF) model for the differentiated glottal pulse [205] is briefly discussed. In speech production, the lip radiation has a high pass filtering effect which can be approximately considered as a differentiation operation in the time domain. Generally, in modeling the speech production mechanism, the lip radiation effect is combined with the glottal source signal (volume velocity in cm^3/sec) and is considered the effective voice source. The effective voice source is therefore the derivative of the glottal source signal (in cm^3/sec^2). Let $u_g(t)$ denote the glottal volume velocity. Then the LF model for differentiated glottal pulse, denoted by $u'_g(t)$, is given by [205]

$$\begin{aligned} u'_g(t) &= E_0 \exp(\alpha t) \sin(2\pi f_g t), \qquad 0 \le t \le T_e \\ &= -\frac{E_e}{\epsilon T_a} \{\exp[-\epsilon(t - \mathrm{T},)] - \exp[-\epsilon(T_c - T_e)]\}, \quad T_e \le t \le T_c \\ &= 0, \qquad T_c \le t \le T_o \end{aligned} \tag{A.1}$$

where the different parameters governing the model are indicated in Fig. A.1(c). Fig. A.1(a) shows the glottal volume velocity waveform derived by the integration of the differentiated pulse in Fig. A.1(c). From Fig. A.1(c) and equation (A.l) above, we observe that

$$f_g = \frac{1}{2T_p} \tag{A.2}$$

Since the value of $u'_g(t)$ given by both the expressions in (A.l) should be same at

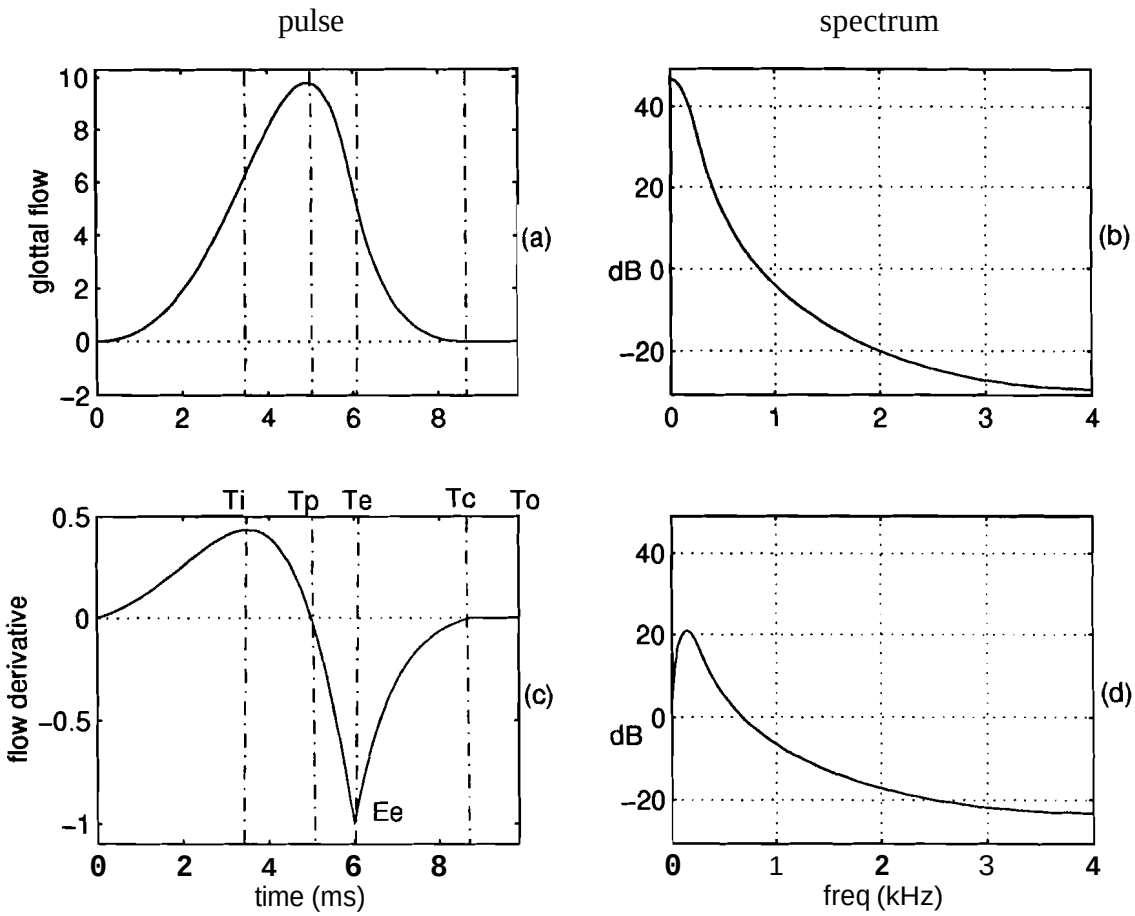


Fig. A.1: Liljencrants–Fant model for differentiated glottal pulse. (a) Glottal volume velocity $u_g(t)$ (integral of the pulse in (c)). (b) Spectrum of the signal in (a). (c) Differentiated glottal pulse $u'_g(t)$ (LF model). (d) Spectrum of the signal in (c).

t $= T_e$, we have

$$E_0 \;=\; -\frac{E_e}{\epsilon T_a}\left[\frac{1-\exp\left[-\epsilon\left(T_c-T_e\right)\right]}{\exp(\alpha T_e)\,\sin(2\pi f_g T_e)}\right] \tag{A.3}$$

Since the glottal pulse $u_g(t)$ starts at the value 0 and returns to 0, the differentiated glottal pulse should be area balanced i.e.,

$$\int_0^{T_o} u'_g(t)\,dt \;=\; 0 \tag{A.4}$$

To determine the time instant T_i at which the differentiated glottal pulse reaches a

maximum, its derivative is set to zero:

$$\frac{du'_g(t)}{dt}\,|_{t=T_i} \;=\; 0 \tag{A.5}$$

Solving (A.5) above for T_i, we have

$$T_i \;=\; T_p\left[1 - \frac{1}{\pi}\tan^{-1}\left(\frac{\pi}{\alpha T_p}\right)\right] \tag{A.6}$$

The maximum value of the flow derivative (E_i) at the instant T_i is given by

$$\begin{aligned} E_i \;&=\; u'_g(t)\,|_{t=T_i} \\ &=\; E_0\,\exp(\alpha T_i)\frac{\pi}{\sqrt{\pi^2 + \alpha^2\,T_p{}^2}} \end{aligned} \tag{A.7}$$

When the relations given above are combined, the equations involving the model parameters are nonlinear. Therefore, for generating a pulse, some parameters were chosen and the others were determined using the Newton–Raphson iterative method.

The return phase time constant (T_a) determines the roll–off of the glottal pulse spectrum. Figs. A.1(b) and A.1(d) show the magnitude spectra of the glottal pulse and its derivative, respectively. When T_a is reduced, the return phase is made sharper and the roll–off of the glottal pulse spectrum reduces and vice–versa.

APPENDIX–B
FROBENIUS NORM OF SIGNAL PREDICTION MATRIX IN THE PRESENCE OF NOISE

In the case of noisy speech, the signal prediction matrix

$$\mathbf{S} = \begin{bmatrix} s_{p+1} & s_p & \cdots & s_1 \\ s_{p+2} & s_{p+1} & \cdots & s_2 \\ & & \ddots & \vdots \\ \vdots & \vdots & & s_{p+1} \\ & & & \vdots \\ s_M & s_{M-1} & \cdots & s_{M-p} \end{bmatrix} \tag{B.1}$$

is perturbed by a matrix of noise samples $\mathbf{W}$ to yield the noisy signal matrix

$$\mathbf{Y} = \mathbf{S} + \mathbf{W} \tag{B.2}$$

For mild perturbation, the singular value decomposition (SVD) of $\mathbf{Y}$ is robust [279]. When the noise is white and Guassian distributed, all the squared singular values of $\mathbf{S}$ are augmented by the noise variance $N\sigma_w^2$ [150]. However, in practice, the speech signal is preemphasized before processing. So, even if the noise corrupting the speech signal is white, the preemphasized noise cannot be assumed to be white. However, even for coloured noise, the squared singular values of the matrix of preemphasized speech samples $\mathbf{S}$ are augmented by an amount $\|\mathbf{W}\|$ [150]. For a reasonably good SNR ($\geq$ 5 dB), the $\|\mathbf{W}\|$ is small compared to $\|\mathbf{S}\|_F$ and thus the variations due to $\|\mathbf{S}\|_F$ are preserved in $\|\mathbf{Y}\|_F$. Therefore, $\|\mathbf{Y}\|_F^2$ could be used as a measure of energy of the residual signal. Note that the square of the Frobenius norrn of a matrix is the sum of squared singular values of the matrix.

APPENDIX–C
BOUNDS ON THE RAYLEIGH QUOTIENT

Let the Singular Value Decomposition (SVD) [253] of S be

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{C.1}$$

where the columns of $\mathbf{U}_{(M-p)\times(M-p)}$ and $\mathbf{V}_{(p+1)\times(p+1)}$ are the left and right singular vectors of S, respectively. $\boldsymbol{\Sigma}_{(M-p)\times(p+1)}$ is the matrix of singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{p+1} > 0$. Therefore

$$\mathbf{S}^T\mathbf{S} = \mathrm{V}\ (\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})\ \mathbf{V}^T \tag{C.2}$$

So of ... σ_{p+1}^2 are the eigenvalues of (S^TS) and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{p+1}\}$, the columns of V, are its eigenvectors. The Rayleigh quotient of (S^TS) is defined as [279]

$$\rho(\mathbf{a}_a) = \frac{\mathbf{a}_a^T(\mathbf{S}^T\mathbf{S})\mathbf{a}_a}{\mathbf{a}_a^T\mathbf{a}_a} \tag{C.3}$$

where $\mathbf{a}_a \in R^{p+1}$. Assuming that the eigenvalues of (S^TS) are all distinct, its eigenvectors form an orthonormal basis in R^{p+1}. Hence, $\mathbf{a}_a$ can be expressed as

$$\mathbf{a}_a = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_{p+1}\mathbf{v}_{p+1} \tag{C.4}$$

where $c_1, c_2, \ldots, c_{p+1}$ are the components of $\mathbf{a}_a$ w.r.t. the basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{p+1}\}$. Premultiplying both sides of (C.4) by (S^TS) and noting that o_i^2 and $\mathbf{v}_i$, $i = 1 \ldots p+1$, are its eigenvalues and eigenvectors, respectively, we have

$$(\mathbf{S}^T\mathbf{S})\mathbf{a}_a = c_1\sigma_1^2\mathbf{v}_1 + c_2\sigma_2^2\mathbf{v}_2 + \ldots + c_{p+1}\sigma_{p+1}^2\mathbf{v}_{p+1} \tag{C.5}$$

Premultiplying (C.5) by $\mathbf{a}_a^T$ and noting that the eigenvectors form an orthonormal set, we have

$$\mathbf{a}_a^T(\mathbf{S}^T\mathbf{S})\mathbf{a}_a = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots + c_{p+1}^2\sigma_{p+1}^2 \tag{C.6}$$

From (C.3), (C.4) and (C.6), we have

$$\rho(\mathbf{a}_a) = \frac{(c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots + c_{p+1}^2\sigma_{p+1}^2)}{(c_1^2 + c_2^2 + \ldots + c_{p+1}^2)} \tag{C.7}$$

From (C.7), it is clear that

$$\sigma_{p+1}^2 \leq \rho(\mathbf{a}_a) \leq \sigma_1^2 \tag{C.8}$$

i.e., the Rayleigh quotient is bounded by the extreme eigenvalues of (S^TS).

APPENDIX–D
EXCITATION SIGNAL–TO–NOISE RATIO

For the zero–mean Gaussian distributed random variables $v(n)$, the Fourier transform $V(\omega)$ is a complex zero–mean Gaussian random variable [268]. Therefore we have

$$\mathcal{E}[|\, V(\omega)\,|^2] = N\sigma_v^2 \tag{D.1}$$

Since the square of the mean is always less than the second moment, i.e.,

$$(\mathcal{E}[|\, V(\omega)\,|])^2 < \mathcal{E}[|\, V(\omega)\,|^2] \tag{D.2}$$

we have

$$\mathcal{E}[|\, V(\omega)\,|] < (N\sigma_v^2)^{\frac{1}{2}} \tag{D.3}$$

Hence

$$\frac{\mathcal{E}[|\, V(\omega)\,|]}{\mathcal{A}} < 10^{-\frac{E_s}{20}} \tag{D.4}$$

where E_s is the excitation signal–to–noise ratio:

$$E_s = 10\ \log_{10}(\frac{\mathcal{A}^2}{N\sigma_v^2})\ \text{dB} \tag{D.5}$$

Let us consider an N–point DFT of the sequence given in (7.10), computed at $\omega_k =$, k $= 0, \ldots,$ N $- 1$. It can be shown [304] that the real and imaginary parts of the DFT of $v(n)$, $V_R(\omega_k)$ and $V_I(\omega_k)$, are (real) independent identically distributed (i.i.d.) Gaussian random variables for k $= 1, 2, \ldots, (\frac{N}{2} - 1)$. Therefore, the vectors $\mathbf{v}_R = [V_R(\omega_1)\ V_R(\omega_2) \ldots V_R(\omega_{\frac{N}{2}-1})]^T$ and $\mathbf{v}_I = [V_I(\omega_1)\ V_I(\omega_2) \ldots V_I(\omega_{\frac{N}{2}-1})]^T$ are $\sim \mathcal{N}(0, \frac{N\sigma_v^2}{2}\mathbf{I})$. Under these conditions the magnitude of the DFT of $v(n)$, $|\, V(\omega_k)\,| = [V_R{}^2(\omega_k) + V_I{}^2(\omega_k)]^{\frac{1}{2}}$, is Rayleigh distributed [305]. Since we have the knowledge of both the mean and variance of $|\, V(\omega_k)\,|$, we get

$$\begin{aligned} \frac{\mathcal{E}[|\, V(\omega_k)\,|]}{\mathcal{A}} &= \frac{\sqrt{\pi}}{2}\, 10^{-\frac{E_s}{20}} \\ &\approx (0.9)\ 10^{-\frac{E_s}{20}} \end{aligned} \tag{D.6}$$

which is indeed close to the upper bound $10^{-\frac{E_s}{20}}$ given in (D.4) above. From the cumulative distribution function of a Rayleigh distribution [305], we may write

$$\begin{aligned} P[|V(\omega_k)| < \mathcal{A}] &= 1 - \exp(-\frac{\mathcal{A}^2}{N\sigma_v^2}) \\ &= 1 - \exp(-10^{\frac{E_s}{10}}) \end{aligned} \tag{D.7}$$

where $P[|V(\omega_k)| < \mathrm{A}]$ is the probability that $|V(\omega_k)|$ is less than A. From (D.7), we note that $|V(\omega_k)| < \mathrm{A}$ with more than 99% confidence, when $E_s \geq 6.6$ dB.

Bibliography

[1] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–27, pp. 309–319, Aug. 1979.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among male and female talkers," J. Awust. *Soc.* Amer., vol. 87, Feb. 1990.

[3] D. O' Shaughnessy, Speech Communication–Human and Machine. New York: Addison–Wesley, 1987.

[4] M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic properties of stop consonants," J. Awust. *Soc.* Amer., vol. 29, pp. 107–116, 1957.

[5] M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic properties of stop consonants," in Readings in Clinical Spectrography of Speech (R. J. Baken and R. G. Daniloff, eds.), pp. 216–225, San Diego, California: Singular Publishing Group and KAY Elemetrics Corp., 1991.

[6] J. M. Pickett, "Consonant features, glides and stops," in Readings in Clinical Spectrograph~of Speech (R. J. Baken and R. G. Daniloff, eds.), pp. 96–112, San Diego, California: Singular Publishing Group and KAY Elemetrics Corp., 1991.

[7] K. S. Nathan, Y.-T. Lee, and H. F. Silverman, "A time–varying analysis method for rapid transitions in speech," IEEE *Trans.* Signal Processing, vol. 39, pp. 815–824, Apr. 1991.

[8] K. S. Nathan and H. F. Silverman, "Tirne–varyingfeature selection and classification of unvoiced stop consonants," IEEE *Trans.* Speech, Audio Processing, vol. 2, pp. 395–405, July 1994.

[9] B. Cranes and L. Boves, "The effects of glottal termination impedance on the formants of speech signals," in Signal Processing *II: Theories* and Applications (H. W. Schussler, ed.), Amsterdam, The Netherlands: Elsevier Science, 1983.

[10] A. K. Krishnamurthy, "Glottal source estimation using a sum–of–exponentials model," IEEE *Trans.* Signal Processing, vol. 40, pp. 682–686, Mar. 1992.

[11] G. Fant, "Some problems in voice source analysis," Speech Comm., vol. 13, pp. 7–22, Oct. 1993.

[12] D. G. Cliilders and C. F. Wong, "Measuring and modeling vocal source–tract interaction," IEEE *Trans.* Biomedical Engineering, vol. 41, pp. 663–671, July 1994.

[13] M. V. Mathews, J. E. Miller, and E. E. David, "Pitch synchronous analysis of voiced sounds," J. *Acoust.* Soc. Amer., vol. 33, pp. 179–186, 1961.

[14] L. R. Rabiner, B. S. Atal, and M. R. Sambur, "LPC prediction error analysis of its variation with position of the analysis frame," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–25, pp. 434–442, 1977.

[15] J. D. Markel and A. H. Gray, *Linear* Prediction of Speech. Berlin: Springer, 1976.

[16] G. S. Kang and S. S. Everett, "Improvement of the LPC analysis," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 1, (Boston, Massachusetts), pp. 89–92, Apr. 1983.

[17] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida, and O. Kakusho, "Analysis of speech signals of short pitch period by a sample–selective linear prediction," IEEE *Trans.* Awust., Speech, Signal *Processing*, vol. ASSP–35, pp. 1233–1240, Sept. 1987.

[18] B. Yegnanarayana, "On timing in time–frequency analysis of speech signals," Sadhana, vol. 21, pp. 5–20, Feb. 1996.

[19] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal–tract system characteristics from speech signals," IEEE *Trans.* Speech, Audio Processing, vol. 6, pp. 313–327, July 1998.

[20] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE *Trans.* Awust., Speech, Signal Processing, vol. 27, pp. 350–355, Aug. 1979.

[21] A. K. Krishnamurthy and D. G. Childers, "Two–channel speech analysis," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. 34, no. 4, pp. 730–743, 1986.

[22] D. G. Childers and C. K. Lee, "Voice quality factors: Analysis, synthesis and perception," J. Acoust. *Soc.* Amer., vol. 90, no. 5, pp. 2394–2410, 1991.

[23] S. Parthasarathy and D. W. Tufts, "Excitation synchronous modeling of voiced speech," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–35, pp. 1241–1249, Sept. 1987.

[24] O. Ghitza, "Robustness against noise : The role of timing–synchrony measurement," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 4, (Dallas, Texas), pp. 2372–2375, Apr. 1987.

[25] O. Ghitza, "Auditory nerve representation criteria for speech analysis/synthesis," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–35, June 1987.

[26] H. Hermansky and N. Morgan, "RASTA Processing of speech," IEEE *Trans.* Speech, Audio Processing, vol. 2, pp. 578–589, Oct. 1994.

[27] D. O'Shaughnessy, Modeling Fundamental Frequency *and* its Relationship to Syntax, Semantics and Phonetics. PhD dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1976.

[28] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J.* Acoust. *Soc.* Jap., vol. (E)5, pp. 233–242, 1985.

[29] C. Gussenhoven and A. C. M. Rietweld, "Voice fundamental frequency declination in Dutch: Testing three hypotheses," *J.* Phonetics, pp. 355–369, 1988.

[30] J. t'Hart, R. Collier, and A. Cohen, A *Perceptual* Study of Intonation: An Ezperimental–Phonetic Approach to Speech Melody. Cambridge: Cambridge University Press, 1990.

[31] A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Significance of prosodic knowledge in a text-to-speech system for Hindi," in *Proc.* XII Int. Cong. Phonetic Sciences, (Aix-en-Provence, France), pp. 494497, 1991.

[32] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization for continuous speech in Hindi based on F_0 patterns," Speech Comm., vol. 18, pp. 2146, Jan. 1996.

[33] I. Lehiste, Suprasegmentals. Cambridge, Massachusetts: M. I. T. Press, 1970.

[34] S. R. Rajesh Kumar, "Significance of durational knowledge for a text–to–speech system in an Indian language," Master's thesis, Indian Institute of Technology, Madras, 1990.

[35] K. Samudravijaya, S. K. Singh, and P. V. S. Rao, "Pre–recognition measures of speaking rate," Speech Comm., vol. 24, no. 1, pp. 73–84, 1998.

[36] L. R. Rabiner and R. W. Schaefer, Digital Processing of Speech Signals. Englewood Cliffs, New Jersey: Prentice Hall, 1978.

[37] F. J. Harris, "On the use of windows for harmonic analysis with discrete Fourier transform," Proc. IEEE, vol. 66, pp. 51–83, 1978.

[38] J. W. Adams, "A new optimal window," IEEE *Trans.* Signal Processing, vol. 39, pp. 1753–1769, Aug. 1991.

[39] J. Le Roux and J. Menez, "A cost minimization approach for optimal window design in spectral analysis of sampled signals," IEEE *Trans.* Signal Processing, vol. 40, pp. 996–999, Apr. 1992.

[40] J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, vol. 63, pp. 561–580, Apr. 1975.

[41] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, Discrete–Time Processing of Speech Signals. New York: Macmillan, 1993.

[42] S. L. Marple Jr., Digital Spectral Analysis with Applications. Englewood Cliffs, New Jersey: Prentice Hall, 1987.

[43] B. Yegnanarayana and P. Satyanarayana Murthy, "On windowing speech data for analysis," in Proceedings of 111 ICAPRDT, (ISI, Calcutta, India), pp. 334–345, Dec. 1993.

[44] B. Yegnanarayana, P. Satyanarayana Murthy, and J. H. Eggen, "Source-System windowing for speech analysis," in Annual Progress Report, *no.28*, (Institute for Perception Research, Eindhoven, The Netherlands), pp. 53–58, 1993.

[45] P. Satyanarayana Murthy, "Analysis of Short Segments of Speech," M.Tech. project rep., Indian Institute of Technology, Madras, Department of Electrical Engineering, Jan. 1994.

[46] E. N. Pinson, "Pitch synchronous time–domain estimation of formant frequencies and bandwidths," J. Acoust. *Soc.* Amer., vol. 35, pp. 1264–1273, Aug. 1963.

[47] K. Steiglitz and B. Dickinson, "The use of time–domain selection for improved linear prediction," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–25, pp. 34–39, Feb. 1977.

[48] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in Proceedings of IEEE Int. Conf. Awwt., Speech, and Signal Processing, (Washington, D.C.), pp. 200–203, Apr. 1979.

[49] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 251–266, July 1995.

[50] R. L. Bouquin–Jeannes, "Enhancement of noisy speech signals: Application to mobile radio communications," Speech Comm., vol. 18, pp. 3–19, Jan. 1996.

[51] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE *Trans.* Awust., Speech, Signal Processing, vol. 36, pp. 145–152, Feb. 1988.

[52] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum–based deconvolution for speech dereverberation," IEEE *Trans.* Speech, Audio Processing, vol. 4, pp. 392–396, Sept. 1996.

[53] G. Fant, The Awustic *Theory* of Speech Production. The Hague, The Netherlands: Mouton, second ed., 1970.

[54] J. L. Flanagan, Speech Analysis Synthesis and Perception. New York: Springer-Verlag, second ed., 1972.

[55] R. N. J. Veldhuis, I. J. M. Bogaert, and N. J. C. Lous, "Two mass models for speech synthesis," in Proceedings of *EUROSPEECH'95*, (Madrid, Spain), pp. 1853–1856, 1995.

[56] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acustica*, vol. 49, pp. 179–182, 1981.

[57] A. Busala, "Fundamental considerations in the design of a voice-switched speakerphone," Bell Syst. Tech. *J.*, vol. 39, p. 265, 1960.

[58] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," Speech Comm., vol. 16, pp. 153–164, 1995.

[59] J. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," BBN Rep. 2304, Bolt, Beranek and Newman, Inc., Cambridge, Massachusetts, Aug. 1972.

[60] A. N. Chasaide and C. Gobl, "Linguistic and paralinguistic variation in the voice source," in Proceedings of Int. Conf. Spoken Language Processing, vol. 1, (Kobe, Japan), pp. 85–88, Nov. 1990.

[61] H. Strik and L. Boves, "On the relation between voice source parameters and prosodic features in connected speech," Speech Comm., vol. 11, pp. 167–174, 1992.

[62] G. Fant, "The voice source in connected speech," Speech Comm., vol. 22, pp. 125–139, Aug. 1997.

[63] D. H. Klatt, "Speech processing strategies based on auditory models," in The Representation of Speech in the Peripheral Auditory System (R. Carlson and B. Granstrom, eds.), pp. 181–196, New York: Elsevier/North Holland, 1982.

[64] H. Hermansky, "Should recognizers have ears?," in Proceedings of ESCA *Tutorial* and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, (France), pp. 1–10, 1997.

[65] E. Zwicker, "Scaling," in Handbook of Sensory Physiology (Keidel and Neff, eds.), vol. V.3, pp. 401–448, Berlin: Springer Verlag, 1975.

[66] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Awust. *Soc.* Amer., vol. 77, pp. 1069–1077, Mar. 1985.

[67] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Comput. Speech Language, vol. 1, p. 109, Dec. 1986.

[68] O. Ghitza, "Temporal nonplace information in the auditory–nerve firing patterns as front-end for speech recognition in noisy environment," J. Phonetics, vol. 16, pp. 109–123, Jan. 1988.

[69] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in Advances in Speech Signal Processing (S. Furui and M. M. Sondhi, eds.), ch. 15, pp. 453–485, New York: Marcel Dekker, 1992.

[70] O. Ghitza, "Auditory neural feedback as a basis for speech processing," in Proceedings of IEEE Int. Conf. Awust., Speech, and *Signal* Processing, vol. 1, (New York City), pp. 91–94, Apr. 1988.

[71] S. Sandhu and O. Ghitza, "A comparative study of Mel cepstra and EIH for phone classification under adverse conditions," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, vol. 1, (Detroit, Michigan), pp. 409–412, May 1995.

[72] H. Hermansky, "Auditory model for parametrization of speech in real–life environment based on reintegration of temporal derivative of auditory spectrum," File Folder ST 04-01, U S West Advanced Technologies, Boulder, Colorado, Oct. 1990.

[73] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory–like analysis of speech (RASTA–PLP)," in Proceedings of *EUROSPEECH'91*, (Genova, Italy), pp. 1367–1371, 1991.

[74] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "(RASTA–PLP) speech analysis technique," in Proceedings of IEEE Int. Conf. *Acoust.*, Speech, and Signal Processing, vol. 1, (San Francisco, California), pp. 121–124, Mar. 1992.

[75] H. Fletcher, "Loudness, masking and their relationship to the hearing process and the problem of noise measurement," J. Acoust. *Soc.* Amer., vol. 9, pp. 275–293, Apr. 1938.

[76] R. V. Cox, B. G. Haskell, Y. LeCun, B. Shahraray, and L. Rabiner, "Scanning the technology: On the applications of multimedia processing to communications," *Proc.* IEEE, vol. 86, pp. 755–824, May 1998.

[77] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, vol. 1, (Tampa, Florida), pp. 509–512, 1985.

[78] N. Morgan and H. Hermansky, "RASTA extensions: Robustness to additive and convolutional noise," in Proceedings of Workshop Speech Processing Adverse Environments, (Cannes, France), Nov. 1992.

[79] C. Avendaiio, S. van Vuuren, and H. Hermansky, "Data based filter design for RASTA–like channel normalization in ASR," in Proceedings of Int. Conf. Spoken Language Processing, (Philadelphia), pp. 2087–2090, Oct. 1996.

[80] C. Avendaiio and H. Hermansky, "On the effects of short–term spectrum smoothing in channel normalization," IEEE *Trans.* Speech, Audio Processing, vol. 5, pp. 372–374, July 1997.

[81] K. Young, S. Sackin, and P. Howell, "The effects of noise on speech : A consideration for automatic speech processing," tech. rep., University College, Gower St., London, Apr. 1992.

[82] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–26, p. 180, 1978 (abstract).

[83] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica, vol. 28, p. 66, 1973.

[84] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. Part I: General room acoustics," Acustica, vol. 46, pp. 60–72, 1980.

[85] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Paris, France), pp. 156–159, Apr. 1982.

[86] J. B. Allen, "Short–term spectral analysis, synthesis and modification by discrete Fourier transform," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–25, pp. 235–238, June 1977.

[87] H. Hermansky, N. Morgan, and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. II, (Minneapolis, Minnesota), pp. 83–86, Apr. 1993.

[88] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," IEEE *Trans.* Awust., Speech, Signal Processing, vol. 30, pp. 679–681, Aug. 1982.

[89] H. Hermansky, E. A. Wan, and C. Avendaiio, "Speech enhancement based on temporal processing," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 1, (Detroit, Michigan), pp. 405–408, May 1995.

[90] C. Avendaño, H. Hermansky, M. Vis, and A. Bayya, "Adaptive speech enhancement using frequency–specific SNR estimates," in *Proceedings* of 111 IEEE Workshop on *Interactive* Voice Technology for Telecommunications Applications, (Basking Ridge, New Jersey), pp. 65–68, Sept. 1996.

[91] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR estimation and speech enhancement," Tech. Rep. TR–93–012, International Computer Science Institute, Berkeley, CA, 1993.

[92] D. A. Berkley and O. M. M. Mitchell, "Seeking the ideal in Hands–Free telephony," Bell *Laboratories* Record, vol. 52, pp. 318–325, Nov. 1974.

[93] O. M. M. Mitchell and D. A. Berkley, "Reduction of long–time reverberation by a center-clipping process," J. Awust. *Soc.* Amer., vol. 47, p. 84, 1970 (abstract).

[94] O. M. M. Mitchell and D. A. Berkley, "A full-duplex echo suppressor using center–clipping," Bell *Syst.* Tech. J., vol. 40, pp. 1619–1630, 1971.

[95] H. Hirsch, "Automatic speech recognition in rooms," in Signal Processing IV: *Theories* and Applications (J. L. Lacome, A. Chehilian, N. Martin, and J. Malbos, eds.), B. V.: Elsevier Science Publishers, 1988.

[96] C. Avendaño and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in Proceedings of Int. Conf. Spoken Language Processing, (Philadelphia), pp. 889–892, Oct. 1996.

[97] J. Mourjopoulos and J. K. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in Proceedings of IEEE Int. Conf. Acowt., Speech, and Signal Processing, vol. 3, (Boston, Massachusetts), pp. 1144–1147, Apr. 1983.

[98] S. M. Kay and S. L. Marple Jr., "Spectrum analysis – A modern perspective," Proc. IEEE, vol. 69, pp. 1380–1419, Nov. 1981.

[99] E. A. Robinson, "A historical perspective of spectrum estimation," Proc. IEEE, vol. 70, pp. 885–906, Sept. 1982.

[100] L. Cohen, "Time-frequency distributions–A review," Proc. IEEE, vol. 77, pp. 941–981, July 1989.

[101] F. Hlawatsch and G. F. Boudreaux–Bartels, "Linear and quadratic time–frequency signal representations," IEEE Signal Processing Magazine, pp. 21–67, Apr. 1992.

[102] M. T. Heideman, D. H. Johnson, and C. S. Burrus, "Gauss and the history of the fast Fourier transform," IEEE ASSP Magazine, pp. 14–21, Oct. 1984.

[103] R. N. Czerwinski and D. L. Jones, "Adaptive short–time Fourier analysis," IEEE Signal Processing Lett., vol. 4, pp. 4245, Feb. 1997.

[104] O. Rioul and M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing Magazine, pp. 14–38, Oct. 1991.

[105] B. Yegnanarayana, "Formant extraction from linear prediction phase spectra," J. Acoust. *Soc.* Amer., vol. 63, pp. 1638–1640, May 1978.

[106] B. Yegnanarayana and D. Raj Reddy, "A distance measure based on the derivative of linear prediction phase spectrum," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, (Washington, D.C.), pp. 744–747, Apr. 1979.

[107] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," Speech Comm., vol. 10, pp. 209–221, Aug. 1991.

[108] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," Signal Processing, vol. 22, pp. 259–267, Mar. 1991.

[109] R. J. McAulay and T. F. Quatieri, "Phase modeling and its application to sinusoidal transform coding," in Proceedings of IEEE Int. Conf. Awwt., Speech, and Signal Processing, (Tokyo, Japan), Apr. 1986.

[110] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," IEEE *Trans.* Awwt., Speech, Signal Processing, vol. 34, pp. 744–754, Aug. 1986.

[111] R. J. McAulay and T. F. Quatieri, "Multirate sinusoidal transform coding at rates from 2.4 kbps," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, 1987.

[112] R. J. McAulay and T. F. Quatieri, "Computationally efficient sine-wave synthesis and its application to sinusoidal transform coding," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, (New York), pp. 370–373, Apr. 1988.

[113] A. H. Gray and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. 24, pp. 207–217, 1974.

[114] G. P. Wadsworth, E. A. Robinson, J. G. Bryan, and P. M. Hurley, "Detection of reflections on seismic records by linear operators," Gwphys., vol. 18, pp. 539–586, 1953.

[115] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc.* IEEE Conf. Communication and Processing, pp. 360–361, 1967.

[116] T. Kailath, "A view of three decades of linear filtering theory," IEEE *Trans.* Inform. Theory, vol. IT–20, pp. 146–181, Mar. 1974.

[117] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," IEEE *Trans. Inform.* Thwry, vol. IT–18, pp. 725–730, Nov. 1972.

[118] R. M. Gray, "Toeplitz and circulant matrices: A review," tech. rep., Dept. of Electrical Engineering, Stanford University, California 94305, Apr. 1993.

[119] S. U. H. Qureshi, "Adaptive equalization," Proc. IEEE, vol. 73, pp. 1349–1387, Sept. 1985.

[120] A. C. Kot, D. W. Tufts, and R. J. Vaccaro, "Analysis of linear prediction by matrix approximation," IEEE *Trans.* Signal Processing, vol. 41, pp. 3174–3177, Nov. 1993.

[121] J. S. Erkelens and P. M. T. Broersen, "Bias propagation in the autocorrelation method of linear prediction," IEEE *Trans.* Speech, Audio Processing, vol. 5, pp. 116–119, Mar. 1997.

[122] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," IEEE *Trans.* Awust., Speech, Signal Processing, vol. ASSP–24, pp. 488494, Dec. 1976.

[123] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," IEEE *Trans. Acoust.,* Speech, Signal Processing, vol. ASSP–27, pp. 478–485, Oct. 1979.

[124] B. Porat, Digital Processing of Random Signals: *Theory* and Methods, ch. 9, p. 272. Englewood Cliffs, New Jersey: Prentice Hall, 1994.

[125] J. A. Cadzow, B. Baseghi, and T. Hsu, "Singular–value decomposition approach to time series modelling," IEE Proc., vol. 130, Part F, pp. 202–210, Apr. 1983.

[126] M. A. Rahman and K.-B. Yu, "Total least squares approach for frequency estimation using linear prediction," IEEE *Trans.* Awwt., Speech, Signal Processing, vol. 35, pp. 1440–1454, Oct. 1987.

[127] Y. Hua and T. K. Sarkar, "On the total least squares linear prediction method for frequency estimation," IEEE *Trans.* Awust., Speech, Signal Processing, vol. 38, pp. 2186–2189, Dec. 1990.

[128] T. J. Abatzoglou, J. M. Mendel, and G. A. Harada, "The constrained total least squares technique and its application to harmonic superresolution," IEEE *Trans.* Signal Processing, vol. 39, pp. 1070–1087, May 1991.

[129] A. Papoulis, "Maximum entropy and spectral estimation: A review," IEEE *Trans.* Acowt., Speech, Signal Processing, vol. ASSP–29, pp. 1176–1186, Dec. 1981.

[130] M. R. Schroeder, "Linear prediction, entropy and signal analysis," IEEE ASSP Magazine, pp. 3–11, July 1984.

[131] A. Van Den Bos, "Alternative interpretation of maximum entropy spectral analysis," IEEE *Trans. Inform.* Theory, vol. IT–17, pp. 493494, July 1971.

[132] H. Hermansky, B. A. Hanson, H. J. Wakita, and H. Fujisaki, "Linear predictive modeling of speech in modified spectral domains," Tech. Rep. 1, Speech Technology Laboratory, Division of Panasonic Technologies, Inc., State Street, Santa Barbara, California, Nov. 1987.

[133] H. Hermansky, B. A. Hanson, and H. J. Wakita, "Low–dimensional representation of vowels based on all–pole modeling in the psychophysical domain," Tech. Rep. 1, Speech Technology Laboratory, Division of Panasonic Technologies, Inc., State Street, Santa Barbara, California, Nov. 1987.

[134] H. Hermansky, "Perceptual linear predictive analysis of speech," J. Acoust. *Soc.* Amer., vol. 87, pp. 1738–1752, Apr. 1990.

[135] H. Hermansky and L. A. Cox Jr., "Perceptual linear predictive analysis–resynthesis technique," in Proc. 2nd European Conference on Speech *Communication* and Technology, (Genova, Italy), Sept. 1991.

[136] I. B. Thomas and R. J. Niederjohn, "The intelligibility of filtered–clipped speech in noise," J. Audio Eng. *Soc.*, vol. 18, pp. 299–303, June 1970.

[137] R. J. Niederjohn and J. H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high–pass filtering followed by rapid amplitude compression," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–24, pp. 277–282, Aug. 1976.

[138] L. M. Arslan and J. H. L. Hansen, "Speech enhancement for crosstalk interference," IEEE Signal Processing Lett., vol. 4, pp. 92–95, Apr. 1997.

[139] M. R. Weiss et al, "Processing speech signals to attenuate interference," in presented at IEEE Symp. Speech Recognition, Apr. 1974.

[140] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Final Rep. NSC–FR/4023, Nicolet Scientific Corp., Dec. 1974.

[141] in Speech Enhancement (J. S. Lim, ed.), New Jersey: Prentice Hall, 1983.

[142] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–27, pp. 113–120, Apr. 1979.

[143] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE *Trans. Acoust.*, Speech, Signal Processing, vol. ASSP–26, pp. 471–472, Oct. 1978.

[144] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, pp. 1586–1604, Dec. 1979.

[145] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. 28, pp. 137–145, Apr. 1980.

[146] C. E. Mokbel and G. F. A. Chollet, "Automatic word recognition in cars," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 346–356, Sept. 1995.

[147] T. Gulzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral–subtraction speech enhancement," Signal Processing, vol. 64, pp. 5–19, Jan. 1998.

[148] G. Whipple, "Low residual noise speech enhancement utilizing time–frequency filtering," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Adelaide, Australia), pp. 15–18, Apr. 1994.

[149] K. Samudravijaya, "Knowledge based spectral subtraction," in Proceedings of Int. Conf. Knowledge Based Computer Systems, (Mumbai, India), pp. 237–246, 1998.

[150] A. J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace–based signal analysis using singular value decomposition," Proc. IEEE, vol. 81, pp. 1277–1308, Sept. 1993.

[151] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad–band noise in speech by truncated QSVD," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 439–448, Nov. 1995.

[152] I. Dologlou and G. Carayannis, "Physical interpretation of signal reconstruction from reduced rank matrices," IEEE *Trans.* Signal Processing, vol. 39, pp. 1681–1682, July 1991.

[153] R. L. Bouquin–Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross–spectral estimator," IEEE *Trans.* Speech, Audio Processing, vol. 5, pp. 484–487, Sept. 1997.

[154] R. L. Bouquin–Jeannes and G. Faucon, "Maximum likelihood noise cancellation with spectral constraints," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 2, (Toronto, Canada), pp. 941–944, May 1991.

[155] R. L. Bouquin–Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," Speech Comm., vol. 16, pp. 245–254, 1995.

[156] F. Ehrmann, R. L. Bouquin–Jeannes, and G. Faucon, "Optimization of a two–sensor noise reduction technique," IEEE Signal Processing Lett., vol. 2, pp. 108–110, June 1995.

[157] S. F. Boll and D. C. Pulsipher, "Suppressionof acoustic noise in speech by two microphone adaptive noise cancellation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP–28, pp. 752–753, 1980.

[158] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–286, Feb. 1989.

[159] L. R. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[160] Y. Ephraim, "On minimum mean square error speech enhancement," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, (Toronto, Canada), pp. 997–1000, May 1991.

[161] F. Xie and D. Van Compernolle, "Speech enhancement by spectral magnitude estimation – A unifying approach," Speech Comm., vol. 19, pp. 89–104, Aug. 1996.

[162] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," IEEE Rans. Acoust., Speech, Signal Processing, vol. ASSP–26, pp. 354–358, Aug. 1978.

[163] M. R. Sambur, "Adaptive noise canceling for speech signals," IEEE Trans. Awust., Speech, Signal Processing, vol. ASSP–26, pp. 419423, Oct. 1978.

[164] B. Widrow, J. R. Glover Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong Jr., and R. C. Goodlin, "Adaptive noise cariceling: Principles and applications," Proc. IEEE, vol. 63, pp. 1692–1716, Dec. 1975.

[165] D. Malah and R. V. Cox, "A generalized comb filtering technique for speech enhancement," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, (Paris, fiance), pp. 160–163, May 1982.

[166] D. L. Donoho, "De–noising by soft–thresholding," tech. rep., Dept. of Statistics, Stanford University, Apr. 1992. presented at the Symposium on Wavelet Theory, Vanderbilt University.

[167] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," tech. rep., Dept. of Statistics, Stanford University, Dec. 1994.

[168] D. L. Donoho, "De–noising by soft–thresholding," IEEE Trans. Inform. Theory, vol. 41, pp. 613–627, May 1995.

[169] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete Wavelet transform," IEEE Signal Processing Lett., vol. 3, pp. 10–12, Jan. 1996.

[170] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 433–438, Nov. 1995.

[171] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal *Processing*, vol. 1, (New York City), pp. 553–556, Apr. 1988.

[172] S. Haykin, Neuml Networks. New York: Macmillan, 1994.

[173] B. Yegnanarayana, Artificial Neuml Networks. Connaught Circus, New Delhi: Prentice–Hall India, 1999.

[174] S. Tamura, "An analysis of a noise reduction neural network," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, vol. 3, (Glasgow, Scotland), pp. 2001–2004, May 1989.

[175] S. Tamura and M. Nakamura, "Improvements to the noise reduction neural network," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 2, (Albuquerque, New Mexico, USA), pp. 825–828, Apr. 1990.

[176] M. J. Anitha and B. Yegnanarayana, "Neural network models for processing degraded speech," in Proc. National Seminar on Artificial Neuml Networks and Cognitive Systems, (Cochin, India), Sept. 1998.

[177] H. Bourlard and Y. Kamp, "Auto–association by multilayer perceptrons and singular value decomposition," Biol. *Cybernet.*, vol. 59, pp. 291–294, 1988.

[178] S. Y. Kung, Digital Neuml Netowrks. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[179] X. Shen, L. Deng, and A. Yasmin, "H, filtering for speech enhancement," in Proceedings of Int. Conf. Spoken Language Processing, (Philadelphia), Oct. 1996.

[180] B. Hassibi and T. Kailath, "H^∞ adaptive filtering," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 2, (Detroit, Michigan), pp. 949–952, May 1995.

[181] L. V. Atkinson and P. J. Harley, An *Introduction* to Numerical Methods with Pascal, ch. 6, p. 201. London: Addison–Wesley, 1983.

[182] C. R. Rao, Linear Statistical Inference und its Applications, section 5b.6, p. 341. New Delhi, India: Wiley Eastern, second ed., June 1989.

[183] R. W. Schafer, "Echo removal by generalized linear filtering," NEREM Record, pp. 118–119, 1967.

[184] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," J. Acoust. *Soc.* Amer., vol. 66, pp. 165–169, July 1979.

[185] J. Flanagan and R. C. Lummis, "Signal processing to reduce multipath distortion in small rooms," J. Acoust. *Soc.* Amer., vol. 47, pp. 1475–1481, June 1970.

[186] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal processing technique to remove room reverberation from speech signals," J. Acoust. Soc. Amer., vol. 62, pp. 912–915, Oct. 1977.

[187] P. J. Bloom and G. D. Cain, "Evaluation of two–input speech dereverberation techniques," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Paris, France), pp. 164–167, May 1982.

[188] K. Farrell, R. J. Mammone, and J. L. Flanagan, "Beamforming microphone arrays for speech enhancement," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, (San Francisco), pp. 285–288, Mar. 1992.

[189] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode, "Adaptive antenna systems," Proc. IEEE, vol. 55, pp. 2143–2159, Dec. 1967.

[190] Q. G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," Speech Comm., vol. 18, pp. 317–334, 1996.

[191] in Signal Processing: Special Issue on Acoustic Echo and Noise Control (E. Hansler, ed.), vol. 64, The Netherlands: Elsevier Science B.V., Jan. 1998.

[192] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," IEEE Trans. Awust., Speech, Signal Processing, vol. ASSP–23, pp. 562–570, Dec. 1975.

[193] C. L. Nikias and P. D. Scott, "The covariance least–squares algorithm for spectral estimation of processes of short data length," IEEE Trans. Geosci. Remote Sensing, vol. GE–21, pp. 180–190, Apr. 1983.

[194] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," IEEE Trans. Acoust., Speech, Signal Processing, vol. 50, pp. 637–655, 1971.

[195] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole–zero modeling in noise," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP–30, pp. 833–840, Dec. 1982.

[196] R. Kumaresan, D. W. Tufts, and L. L. Scharf, "A Prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," Proc. IEEE, vol. 72, Feb. 1984.

[197] S. Parthasarathy and D. W. Tufts, "Maximum–likelihood estimation of the parameters of exponentially damped sinusoids," Proc. IEEE, vol. 73, pp. 1528–1530, Oct. 1985.

[198] R. Kumaresan and A. K. Shaw, "An algorithm for pole–zero modeling and spectral analysis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP–34, pp. 637–640, June 1986.

[199] Y. Li, K. J. R. Liu, and J. Razavilar, "A parameter estimation scheme for damped sinusoidal signals based on low–rank Hankel approximation," IEEE Trans. Signal Processing, vol. 45, pp. 481–486, Feb. 1997.

[200] Y.-T. Lee and H. F. Silverman, "On a general time–varying model for speech signals," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (New York), pp. 95–98, Apr. 1988.

[201] E. W. Kamen, "The poles and zeros of a linear time–varying system," Linear Algebm Appl., vol. 98, pp. 263–290, 1988.

[202] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay functions," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 325–333, Sept. 1995.

[203] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Detroit, Michigan), pp. 776–779, May 1995.

[204] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," J. Awust. *Soc.* Amer., vol. 68, pp. 18–21, 1980.

[205] G. Fant, "Glottal flow: Models and interaction," J. Phonetics, vol. 14, pp. 393–399, Oct.–Dec. 1986.

[206] A. A. Giordano and F. M. Hsu, Least Squaw Estimation with Applications to Digital Signal Processing. New York: John Wiley, 1985.

[207] P. E. Papamichalis, Practical Approaches to Speech Coding. Englewood Cliffs, New Jersey: Prentice Hall, 1987.

[208] D. R. Allen and W. J. Strong, "A model for the synthesis of natural sounding vowels," J. Awust. *Soc.* Amer., vol. 78, pp. 58–69, July 1985.

[209] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," IEEE Trans. Speech, Audio Processing, vol. 2, pp. 245–257, Apr. 1994.

[210] B.-H. Juang, "Recent developments in robust speech recognition," in Modern Methods of Speech Processing (R. P. Ramachandran and R. Mammone, eds.), p. 232, Boston: Kluwer Academic, 1995.

[211] J. C. Junqua and J. P. Haton, Robustness in Automatic Speech Recognition: *Fundamentals* and Applications. Boston: Kluwer Academic, 1996.

[212] B.-H. Juang, "Speech recognition in adverse environments," Comput. Speech Language, vol. 5, pp. 275–294, 1991.

[213] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Comm., vol. 16, pp. 261–291, Apr. 1995.

[214] S. V. Vaseghi and B. P. Milner, "Noise compensation methods for Hidden Markov Model speech recognition in adverse environments," IEEE Trans. Speech, Audio Processing, vol. 5, pp. 11–21, Jan. 1997.

[215] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to automatic speech recognition," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signul Processing, (New York), pp. 561–564, Apr. 1988.

[216] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhance ment and coding," IEEE *Trans.* Signal Processing, vol. 39, pp. 1732–1742, Aug. 1991.

[217] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," IEEE *Trans.* Signal Processing, vol. 39, pp. 1943–1954, Sept. 1991.

[218] H. Kobatake, K. Gyoutoku, and S. Li, "Enhancement of noisy speech by maximum likelihood estimation," in Proceedings of IEEE Int. Conf. *Acoust.*, Speech, and Signal Processing, vol. 2, (Toronto, Canada), pp. 973–976, May 1991.

[219] Y. Ephraim, "Speech enhancement using state dependent dynamical system model," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (San Francisco), pp. 289–292, Mar. 1992.

[220] Y. Ephraim, "Statistical model based speech enhancement systems," Proc. IEEE, vol. 80, pp. 1526–1555, Oct. 1992.

[221] J. H. L. Hansen and L. M. Arslan, "Markov model–based phoneme class partitioning for improved constrained iterative speech enhancement," IEEE mans. Speech, Audio Processing, vol. 3, pp. 98–104, Jan. 1994.

[222] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints on an auditory spectrum," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 22–34, Jan. 1995.

[223] K. Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," IEEE Signal Processing Lett., vol. 3, pp. 196–199, July 1996.

[224] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," IEEE *Trans.* Speech, Audio Processing, vol. 4, pp. 383–389, Sept. 1996.

[225] J. H. L. Hansen and B. L. Pellom, "Text–directed speech enhancement employing phone class parsing and feature map constrained vector quantization," Speech Comm., vol. 21, pp. 169–189, 1997.

[226] J. Huang and Y. Zhao, "Energy–constrained signal subspace method for speech enhancement and recognition," IEEE Signal Processing Lett., vol. 4, pp. 283–285, Oct. 1997.

[227] K. Y. Lee, B.-G. Lee, and S. Ann, "Adaptive filtering for speech enhancement in colored noise," IEEE Signal Processing Lett., vol. 4, pp. 277–279, Oct. 1997.

[228] K. Y. Lee, S. McLaughlin, and K. Shirai, "Speech enhancement based on neural predictive Hidden Markov Model," Signal Processing, vol. 65, pp. 373–381, Mar. 1998.

[229] B. Yegnanarayana, "Processing noisy speech using group delay functions," in Proceedings of Int. Workshop on Speech Tech. for Man–Machine Interaction, (TIFR, Bombay, India), pp. 145–157, Dec. 1990.

[230] B. Yegnanarayana, H. A. Murthy, and V. R. Ramachandran, "Speech enhancement using group–delay functions," in Proceedings of Int. Conf. Spoken Language Processing, vol. 1, (Kobe, Japan), pp. 301–304, Nov. 1990.

[231] B. Yegnanarayana, H. A. Murthy, and V. R. Ramachandran, "Processing of noisy speech using modified group delay functions," in Proceedings of IEEE Int. Conf. Awwt., Speech, and Signal Processing, (Toronto, Canada), pp. 945–948, May 1991.

[232] H. A. Murthy, Algorithms for Processing Fourier *Transform* Phase of Signals. PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, Dec. 1991.

[233] B. Yegnanarayana and V. R. Ramachandran, "Group delay processing of speech signals," in Proceedings of ESCA Workshop on Comparing Speech Signal Representation, (Sheffield, England), pp. 411418, Apr. 1992.

[234] J. S. Lim and A. V. Oppenheim, "All–pole modeling of degraded speech," IEEE *Trans.* Acowt., Speech, Signal Processing, vol. ASSP–26, pp. 197–210, June 1978.

[235] R. H. Frazier, "An adaptive filtering approach toward speech enhancement," Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1975.

[236] Y. M. Perlmutter, "Evaluation of a speech enhancement system," Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1976.

[237] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a speech enhancement system," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, pp. 212–215, May 1977.

[238] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 59–71, Jan. 1995.

[239] A. Erell and M. Weintraub, "Pitch–aided spectral estimation for noise-robust speech recognition," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 2, (Toronto, Canada), pp. 909–912, May 1991.

[240] A. Erell and M. Weintraub, "Estimation of noise-corrupted speech DFT–spectrum using the pitch period," IEEE *Trans.* Speech, Audio Processing, vol. 2, pp. 1–8, Jan. 1994.

[241] P. Mermelstein, "Threshold of degradation for frequency-distributed band–limited noise in continuous speech," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. 72, pp. 1368–1373, Nov. 1982.

[242] N. S. Jayant and P. Noll, Digital Coding of *Waveforms* - Principles and Applications. New Jersey: Prentice Hall, 1984.

[243] H. Kanai and K. Kido, "Estimation of input pulse locations from the response of an all–pole transfer system using tapered rank reduction," IEEE *Trans.* Signal Processing, vol. 39, pp. 148–159, Jan. 1991.

[244] H. Kanai, K. Ikikame, and N. Chubachi, "A tapered SVD without rank determination for estimation of multipulse input time series from noisy output," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Adelaide, Australia), pp. IV–45–IV–48, Apr. 1994.

[245] H. Kanai and N. Chubachi, "Accurate estimation of AR model by tapered SVD without rank determination," in Proceedings of IEEE Int. Conf. Acoust., Speech, and *Signal* Processing, (Adelaide, Australia), pp. IV–473–IV–476, Apr. 1994.

[246] S. Haykin, Adaptive Filter *Theory*. New Jersey: Prentice-Hall, 1991.

[247] B. Yegnanarayana and R. Teunen, "Prosodic manipulation of speech using knowledge of instants of significant excitation," Tech. Rep. 1029, Institute for Perception Research, Eindhoven, The Netherlands, Dec. 1994.

[248] F. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveform concantenation," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, vol. 3, (Tokyo, Japan), pp. 38.5.1-38.5.4, Apr. 1986.

[249] C. Hamon, E. Moulines, and F. J. Charpentier, "A diphone synthesis system based on time domain prosodic modifications of speech," in Proceedings of IEEE Int. Conf. Awwt., Speech, and Signal Processing, (Glasgow, Scotland), pp. 238-241, May 1989.

[250] B. Yegnanarayana, C. d'Allesandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," Tech. Rep. NDL-95-01, Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur (LIMSI), CNRS, Orsay, France, July 1995.

[251] B. Yegnanarayana, C. d'Allesandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," IEEE *Trans*. Speech, Audio Processing, vol. 6, pp. 1-11, Jan. 1998.

[252] B. Yegnanarayana and P. Satyanarayana Murthy, "Source-System windowing for speech analysis and synthesis," IEEE *Trans*. Speech, Audio Processing, vol. 4, pp. 133-137, Mar. 1996.

[253] S. J. Leon, Linear Algebra with Applications. New York: Macmillan, 1990.

[254] Website, "http://spib.rice.edu/spib/select_noise.html," IEEE Signal Processing Information Base, 1997.

[255] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in Proceedings of DARPA Workshop on Speech Recognition, pp. 93-99, Feb. 1986.

[256] B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendaño, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 1, (Seattle, Washington), pp. 405-408, May 1998.

[257] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in Proceedings of IEEE Int. Conf. Acoust., Speech, und Signal Processing, vol. 2, (Toronto, Canada), pp. 977-980, May 1991.

[258] M. Tohyama, R. H. Lyon, and T. Koike, "Pulse waveform recovery in a reverberant condition," J. Acoust. *Soc*. Amer., vol. 91, pp. 2805-2812, May 1992.

[259] A. P. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in Proceedings of IEEE Int. *Conf*. Acoust., Speech, and Signal Processing, (Adelaide, Australia), pp. 9-13, Apr. 1994.

[260] M. Tohyarna, H. Suzuki, and Y. Ando, The Nature and Technology of Acoustic Space. London: Academic Press, 1995.

[261] C. Avendaño, Temporal Processing of Speech in a Time–Feature Space. PhD dissertation, Oregon Graduate Institute of Science and Technology, Department of Electrical Engg., Portland, Oregon, Apr. 1997.

[262] D. Giulani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation," in Proceedings of Int. Conj. *Spoken Language* Processing, Oct. 1996.

[263] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands–free speech recognition," Speech Comm., vol. 25, pp. 75–95, Aug. 1998.

[264] B. Yegnanarayana, C. Avendaño, H. Hermansky, and P. Satyanarayana Murthy, "Processing linear prediction residual for speech enhancement," in Proceedings of EU-*ROSPEECH'97*, (Patras, Greece), pp. 1399–1402, Sept. 1997.

[265] A. Papoulis, Probability, Random Variables, and Stochastic Processes. New York: McGraw–Hill, third ed., 1991.

[266] W. H. Press, S. A. Teukolsky, W. T. Vellerling, and B. P. Flannery, Numerical Recipes in *C*. New Delhi: Cambridge University Press, 1992.

[267] G. J. Borse, Numerical Methods with *MATLAB*, ch. 13, p. 292. Boston: International Thomson, 1997.

[268] S. M. Kay, Modem Spectral Estimation - *Theory* and Application. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[269] P. M. Clarkson, Optimal and Adaptive Signal Processing, ch. 3, p. 131. Boca Raton: CRC Press, 1993.

[270] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature–based approach," IEEE Signal Processing Magazine, vol. 13, pp. 58–71, Sept. 1996.

[271] H. W. Strube, "Determination of the instant of glottal closure," J. Acoust. *Soc*. Amer., vol. 56, no. 5, pp. 1625–1629, 1974.

[272] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," IEEE *Trans*. Awust., Speech, Signal Processing, vol. ASSP–37, pp. 1805–1814, Dec. 1989.

[273] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," IEEE *Trans*. Speech, Audio Processing, vol. 2, pp. 258–265, Apr. 1994.

[274] C. J. Bleakley, "Improved automatic estimation of the glottal closure instant and period," IEEE *Trans*. Speech, Audio Processing, vol. 44, p. 87, Dec. 1996 (abstract).

[275] E. A. Robinson, T. S. Durrani, and L. *G*. Peardon, Geophysical Signal Processing. Englewood Cliffs, New Jersey: Prentice Hall, 1986.

[276] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[277] X. Li, W. Fang, and Q. Tian, "Error criteria analysis and robust data fusion," in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 4, (Adelaide, Australia), pp. 37–41, Apr. 1994.

[278] V. Koivunen, N. Himayat, and S. A. Kassam, "Nonlinear filtering techniques for multivariate images–Design and robustness characterization," Signal Processing, vol. 57, pp. 81–91, 1997.

[279] G. H. Golub and C. F. Van Loan, *Matrix* Computations. Baltimore, Maryland: The Johns Hopkins University Press, 1983.

[280] X. Li and N. M. Bilgutay, "Wiener filter realization for target detection using group delay statistics," IEEE *Trans.* Signal Processing, vol. 41, pp. 2067–2074, June 1993.

[281] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," IEEE *Trans.* Signal Processing, vol. 40, pp. 2281–2289, Sept. 1992.

[282] R. C. Kemeriat and D. G. Childers, "Signal detection and extraction by cepstrum techniques," IEEE *Trans.* Inform. *Theory*, vol. IT–18, pp. 745–759, Nov. 1972.

[283] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in Proceedings of IEEE Int. Conf. *Acoust.,* Speech, and Signal Processing, vol. 1, (Albuquerque, New Mexico; USA), pp. 109–112, Apr. 1990.

[284] C. Jankowski, "The NTIMIT speech database," printed documentation which accompanies the NTIMIT CD–ROM, Nynex Science and Technology Center, White Plains, New York, Jan. 1991.

[285] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal processing for a cocktail party effect," J. Awust. *Soc.* Amer., vol. 50, pp. 656–660, Aug. 1971.

[286] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," J. Awust. *Soc.* Amer., vol. 60, pp. 911–918, Oct. 1976.

[287] B. A. Hanson and D. Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in Proceedings of IEEE Int. Conf. *Acoust.,* Speech, and Signal Processing, vol. 2, (San Diego, California), pp. 18.A.5.1–18.A.5.4, Mar. 1984.

[288] B. A. Hanson and D. Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," Tech. Rep. 1, Speech Technology Laboratory, Division of Panasonic Technologies, Inc., State Street, Santa Barbara, California, Nov. 1987.

[289] K. Min, D. Chien, S. Li, and C. Jones, "Automated two speaker separation system," in Proceedings of IEEE Int. Conf. Acoust., Speech, *and* Signal Processing, (New York City), pp. 537–540, Apr. 1988.

[290] M. Najar, M. A. Lagunas, and I. Bonet, "Blind wideband source separation," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, (Adelaide, Australia), pp. IV65–IV68, Apr. 1994.

[291] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," IEEE *Trans.* Speech, Audio Processing, vol. 5, pp. 407–424, Sept. 1997.

[292] T. Taniguchi, S. Kajita, K. Takeda, and F. Itakura, "Applying blind signal separation to the recognition of overlapped speech," in Proceedings of E*UROSPEECH'97*, (Patras, Greece), pp. 1103–1106, Sept. 1997.

[293] Y. Cao, S. Sridharan, and M. Moody, "Multichannel speech separation by eigendecomposition and its application to co–talker interference removal," IEEE *Trans. Speech, Audio Processing*, vol. 5, pp. 209–219, May 1997.

[294] B. A. Hanson, D. Y. Wong, and B.-H. Juang, 'Speech enhancement with harmonic synthesis," in *Proceedings* of IEEE *Int.* Conf. *Acoust.,* Speech, and Signal Processing, vol. 3, (Boston, Massachusetts), pp. 1122–1125, Apr. 1983.

[295] B. A. Hanson, D. Y. Wong, and B.-H. Juang, "Speech enhancement with harmonic synthesis," Tech. Rep. 1, Speech Technology Laboratory, Division of Panasonic Technologies, Inc., State Street, Santa Barbara, California, Nov. 1987.

[296] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in Speech Enhancement (J. S. Lim, ed.), (Appeared in Proc. ICASSP, April 1976) 1–10, pp. 85–87, New Jersey: Prentice Hall, 1983.

[297] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE *Trans.* Audio Electroacoustics, vol. 20, pp. 367–378, Dec. 1972.

[298] J. D. Wise, J. R. Caprio, and T.'W. Parks, "Maximum likelihood pitch estimation," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–24, pp. 418–423, Oct. 1976.

[299] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," IEEE *Trans.* Speech, Audio Processing, vol. 3, pp. 117–125, Mar. 1995.

[300] R. Teunen, "Efficient implementation of the group–delay–based algorithm for identification of the instants of significant excitation," Personal wmmunication, 1996.

(301) H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proceedings* of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 3, (Tokyo, Japan), p ~1605–1608, Apr. 1986.

[302] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE *Trans.* Acoust., Speech, Signal Processing, vol. ASSP–34, pp. 28–42, Feb. 1986.

[303] D. G. Childers, "Glottal source modelling for voice conversion," Speech *Comm.*, vol. 16, pp. 127–138, 1995.

(304) S. M. Kay, *Fundamentals* of Statistical Signal Processing - Estimation Theory. New Jersey: Prentice Hall, 1993.

[305] J. G. Proakis, Digital *Communications.* Singapore: McGraw–Hill, 1989.

LIST OF PAPERS SUBMITTED ON THE BASIS OF THIS THESIS

REFEREED JOURNALS

- B. Yegnanarayana and P. Satyanarayana Murthy, "Source-System windowing for speech analysis and synthesis," IEEE *Trans.* Speech, Audio Processing, vol. 4, pp. 133–137, Mar. 1996.
- P. Satyanarayana Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," accepted for publication in IEEE *Trans.* Speech, Audio Processing, 1998.
- B. Yegnanarayana, C. Avendaiio, H. Hermansky, and P. Satyanarayana Murthy, "Speech enhancement using linear prediction residual," Speech Comm., vol. 27, Feb. 1999.
- B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," submitted to IEEE *Trans.* Speech, Audio Processing, 1998.
- "Practical issues in the implementation of speech enhancement methods," manuscript under preparation.

PRESENTATIONS IN CONFERENCES

- B. Yegnanarayana and P. Satyanarayana Murthy, "New directions in speech processing – A Review," Proc. National Conf. Commn., IIT Madras, pp. 169–174, Feb. 1997.
- B. Yegnanarayana, C. Avendaiio, H. Hermansky, and P. Satyanarayana Murthy, "Processing linear prediction residual for speech enhancement," in Proceedings of EU-*ROSPEECH'97*, (Patras, Greece), pp. 1399–1402, Sept. 1997.
- B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendaiio, and H. Hermansky, "Enhancement of reverberant speech using LP residual," in Proceedings of IEEE Int. Conf. Awust., Speech, and Signal Processing, vol. 1, (Seattle, Washington), pp. 405408, May 1998.

ke:yu:ra:ṇi na bhu:shayanti purusham

ha:ra: na chandro:dzwala:h

na sna:nam na vile:panam na kusumam

na alankrita:h mu:rdhaja:h

va:ṇye:ka: samalankaro:thi purusham

ya: samskruta: dha:ryete:

kshi:yanthe: akhilabhu:shaṇa:ni sathatham

va:kbhu:shaṇam bhu:shaṇam

- Bhartruhari,

Bhartruhari Subha:shita:ni

"Neither armlets, nor necklaces (resplendant like the moon), nor sandal paste, nor decorated hair beautify a person. A person's refined speech is his real ornament. All other ornaments perish".