Speech Analysis and Quality Enhancement Using Higher Order Cumulants

By

Elias J. Nemer, M.Eng, M.B.A.

A thesis submitted to the Faculty of Graduate Studies and
Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Ottawa-Carleton Institute for Electrical and Computer Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
August 1999

Abstract

This thesis presents robust methods for speech analysis and enhancement based on newly established properties of the higher order cumulants (HOC) of speech signals. In the exploratory part of this work, it is shown that the HOC of speech are non-zero and may be expressed in terms of speech parameters, such as energy, harmonic amplitudes and frequencies. These properties are established assuming a sinusoidal model for speech, and considering two specific domains, namely a subband representation and the linear predictive coding (LPC) residual.

The issues pertaining to the bias and variance of the HOC estimators are examined and in the case of a sinusoid in white Gaussian noise, these entities are quantified in terms of the process variance. An algorithm is proposed for computing the 3^{rd}-order cumulant with a reduced number of multiplications, and a scheduling algorithm is proposed to map a set of DSP operations on a configurable multi-unit architecture. General properties relating 2^{nd} and higher order statistics (HOS) in the frequency domain are derived, such as the recovery of the Fourier magnitude spectrum from the bispectrum.

The application part of this work exploits the HOC properties thus established and the limitations identified to build two algorithms, the first for quality enhancement and the second for voice activity detection. The algorithm for speech enhancement uses subband domain optimal filters based on a minimum mean square error criteria (MMSE) to recover the speech signal from the noisy observation. The key idea is to use the 4^{th}-order cumulant of the noisy speech to estimate the parameters required for the filters, namely the 2^{nd}-order statistics of the speech and noise as well as the probability of speech presence. The algorithm proposed for voice activity detection (VAD) combines HOS metrics and SNR measures to classify frames as speech or noise, using the LPC residual. A voicing condition for speech frames is derived based on the relation between the skewness and kurtosis of voiced speech. In addition, the variance of the HOS estimators is used to yield a likelihood measure for noise frames.

The two algorithms developed demonstrate that, in spite of the practical limitations of using these cumulants and the approximate nature of the speech model assumed, effective application of HOC is possible. By making use of only HOC measures, the performance of these algorithms is shown to be comparable, even better in some respects, to the current standards. As this is the first iteration of this type of approach, it clearly demonstrates the promising potential of HOC in yielding algorithms that would surpass the current state of the art.

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisors Professor Rafik Goubran and Professor Samy Mahmoud for their guidance and advice throughout the course of this study. I am also grateful to the members of my committee, particularly to Professor Douglas O'Shaughnessy and Professor Victor Aitken for their thorough review and detailed comments on this work.

My colleagues at Nortel Networks and my fellow graduate students at Carleton and McGill have been very helpful, and I thank them all for sharing their knowledge, insights, and criticism. I am particularly thankful to Dr. Stéphane Coulombe for diligently reviewing my papers and for his excellent mentoring; to Dr. Majid Foodeei for valuable criticism on my approaches; to Dr. Robert Dijkerman for proofreading the math derivations; to Mr. Mohamed Zad-Issa, Mr. James Loo and Ms. Madeleine Saikaly for insightful discussions on speech coding and modeling.

I am grateful to the management at Nortel for the special arrangement to make this study/work project possible, particularly to Mr. Rafi Rabipour, Mr. Tony Addona and Mr. Duncan Bees.

Finally, I am indebted to family and good friends, whose unwavering support was a great encouragement during this endeavor; to my sisters, my dad, my cousins, and to David, Jean-Pierre, Sandra, Michael, Sasha, Joseph, Kathy, Derek, Steve, Pauline, Elie and Peter.

In memory of my mother

Happy is the man who finds wisdom
and the man who gains understanding.

Proverbs 3:13

Contents

CHAPTER 1 Introduction 1

1.1 Motivation and Rationale 1

1.2 Goal of this Thesis 4

1.3 Structure of this Thesis 6

1.4 Contributions of this Thesis 8

CHAPTER 2 Background and Literature Review 11

2.1 An Analytical Model For Speech 11

2.2 Speech Quality Enhancement Techniques 13

2.2.1 Wiener Filtering and Related Methods 14

2.2.1.1 Infinite smoothing and frequency-domain filters 15

2.2.1.2 White observation and causal filters 18

2.2.1.3 Some reported systems 18

2.2.2 ML and MMSE Spectral Estimation 21

2.2.3 Kalman Filters 24

2.2.4 Comb Filtering 29

2.2.4.1 Comb filters to reinforce F_0 29

2.2.4.2 Comb filters to cancel interfering harmonics 33

2.2.5 Wavelet-based Denoising 34

2.2.6 Psychoacoustic-based Methods 35

2.3 The Use of HOS for Speech Analysis 36

CHAPTER 3 Higher Order Statistics: Definition & General Derivations 38

3.1 Definitions and Notation 38

3.1.1 Time Domain Definitions 38

3.1.1.1 Zero-mean signals 39

3.1.1.2 Cumulant slices 40

3.1.1.3 Properties of cumulants 41

3.1.2 Frequency Domain Definitions 42

3.2 Third-Order Derivations 44

3.2.1 Fourier Transform of Cumulant Slices 44
3.2.1.1 Infinite data 44
3.2.1.2 Finite data length 45
3.2.2 The Fourier Transform of $C_{30}[\tau]$ and the Bispectrum 47
3.2.3 Geometric Mean of the Power Spectrum and the Bispectrum 48
3.2.4 Fourier Magnitude Recovery from the Bispectrum 53
3.2.4.1 Existing Methods 53
3.2.4.2 Proposed Methods 57
3.2.4.3 Discussion 62
3.3 Fourth-Order Derivations 63
3.3.1 Fourier Transform of Cumulant Slices 63
3.3.2 DC Component of the Horizontal slice 64
3.4 Conclusion 66

CHAPTER 4 Higher Order Cumulants of Subbanded Speech 67

4.1 Analytical Model for Subbanded Speech 67
4.2 Third-Order Cumulant 69
4.2.1 Unvoiced Speech 69
4.2.2 Transient Speech 70
4.2.3 Steady Voiced Speech 70
4.3 Fourth-Order Cumulant 72
4.3.1 Transient Speech 72
4.3.2 Steady State Voiced Speech 75
4.3.3 Unvoiced Speech 79
4.3.4 Effect of noise on the normalized kurtosis 81
4.4 Summary of the Derivations 82
4.5 Simulation Results Using Speech Signals 84
4.5.1 Voiced Speech 84
4.5.1.1 Skewness and kurtosis 84
4.5.1.2 Diagonal cumulant slice 87
4.5.2 Unvoiced Speech 88
4.6 Conclusion 91

CHAPTER 5 Application of HOC to Speech Enhancement 92

5.1 Motivation and Rationale 92
5.2 Speech Enhancement by Optimal Filtering 94

5.2.1 Optimum Linear Systems 94
5.2.2 Filtering Speech Plus Noise 94
5.2.3 Filtering Subbanded Speech 95
5.2.3.1 General solution for a pth-order filter 95
5.2.3.2 Special cases 96
5.3 Estimating Filter Parameters From Fourth Statistics 97
5.3.1 Speech model and higher correlations 97
5.3.2 Autocorrelation of Speech 98
5.3.3 Probability of Speech Presence 99
5.3.3.1 Rationale 99
5.3.3.2 Probability based on HOS 100
5.3.4 Speech and Noise Energy and SNR Estimation 101
5.4 Inter-frame Smoothing of Parameters 105
5.4.1 SNR Smoothing 105
5.4.2 Autocorrelation Smoothing 106
5.4.3 Low Pass Filtering of Optimum Filter Coefficients 106
5.5 Other Considerations 107
5.5.1 Frequency Masking Psychoacoustics 107
5.5.2 Subbanding Filters 109
5.5.3 Upper and Lower Bounds on Filter gains 109
5.6 Overview of the Algorithm 110
5.7 Experimental Results 112
5.7.1 Data used 112
5.7.2 Performance evaluation 112
5.7.2.1 SNR estimation 113
5.7.2.2 SNR smoothing 116
5.7.2.3 Speech waveforms and spectrogram 117
5.7.3 Discussion 123
5.8 Conclusion 124

CHAPTER 6 Higher Order Cumulants of LPC-filtered Speech 125

6.1 Rationale and Related Work 125
6.2 A Model for the LPC Residual 126
6.2.1 Effect of the LPC order 127
6.2.2 Effect of noise on the LPC residual 127
6.3 Third-Order Cumulant 128
6.3.1 Stationary Voiced Speech 128

6.3.2 Non-Stationary Voiced Speech 131
6.3.3 Unvoiced Speech 131
6.3.3.1 Assuming a non-Gaussian white process 131
6.3.3.2 Assuming a harmonic process 131
6.4 Fourth-Order Cumulant 132
6.4.1 Unvoiced Speech 132
6.4.1.1 Assuming a non-Gaussian white process 132
6.4.1.2 Assuming a harmonic process 132
6.4.2 Voiced Speech 133
6.4.2.1 Common property 133
6.4.2.2 Steady voiced speech 134
6.5 Summary of the Derivations 139
6.5.1 Third-Order Cumulant 139
6.5.2 Fourth-Order Cumulant 139
6.6 Effect of Noise 141
6.6.1 Effect of Noise on γ_3 and γ_4 141
6.6.2 Effect of a non-flat LPC residual 142
6.7 Results Using Speech Signals 143
6.7.1 Voiced Speech 143
6.7.2 Effect of Noise on γ_3 and γ_4 146
6.7.3 Unvoiced Speech 147
6.8 Conclusion 151

CHAPTER 7 Application of HOC to Voice Activity Detection 152

7.1 Motivation and Related Work 152
7.2 Voice Activity Detection using HOS 154
7.2.1 Rationale 154
7.2.2 Soft Detection of Noise Frames 154
7.2.3 Necessary Condition for Voicing 156
7.2.4 HOS-based VAD Algorithm 156
7.3 Experimental Results 159
7.4 Conclusion 166

CHAPTER 8 Implementation Issues of Higher Order Statistics 167

8.1 Efficient Computation of the Third Order Cumulant 168
8.1.1 Motivation 168

8.1.2 The algorithm 168
8.1.3 Comparative Results 172
8.2 Mapping DSP Algorithms onto Configurable Architectures 174
8.2.1 Problem Formulation 176
8.2.2 The Branch-and-bound Search Process 176
8.2.3 A Non-deterministic Allocation Heuristic 177
8.2.3.1 Computing the matching criteria 178
8.2.4 The Scheduling Algorithm 179
8.2.5 Results and Discussion 182
8.2.5.1 Data used 182
8.2.5.2 Scheduling results 183
8.3 Conclusion 186

CHAPTER 9 Conclusion and Future Work 187

9.1 Conclusion 187
9.2 Proposed Future Work 193

APPENDIX A Bias and Variance of the HOS Estimators 194

A.1 Definitions 194
A.2 Bias of the HOS Estimators of a Sinusoid in Noise 195
A.2.1 Case 1: noise only $\{x(n) = g(n)\}$ 195
A.2.2 Case 2: sinusoidal signal only $\{x(n) = s(n)\}$ 196
A.2.3 Case 3: sinusoid and Gaussian noise $\{x(n) = s(n) + g(n)\}$ 199
A.3 Variance of the HOS Estimators of a White Gaussian Process 201

References 205

Publications 211

List of Tables

Table 4-1	$FC^{a}_{4}(w)$ for a single sinusoid with a deterministic phase	75
Table 6-1	Value of $FC_{0}(w)$ at all positive lags	130
Table 6-2	Value of $FC^{b}_{4}(w)$ for all positive lags	135
Table 7-1	Pc's and Pf's for the HOS-based and G.729B VAD	160
Table 8-1	Number of operations for case 1	170
Table 8-2	Number of operations for case 2	172
Table 8-3	Comparative results	172
Table 8-4	Hardware resources available	182

List of Figures

Figure 1-1	Magnitude of the spectrum and bispectrum of a voiced speech segment	5
Figure 2-1	Generic Wiener-based speech enhancement	15
Figure 2-2	Various suppression filters	17
Figure 2-3	Enhancement conditioned on the presence of speech	22
Figure 2-4	Impulse response of a comb filter (L=2)	29
Figure 2-5	Block diagram of Lim's comb filtering	30
Figure 2-6	Variable weighting for the samples in each pitch period	32
Figure 2-7	The LMS harmonic canceller	33
Figure 2-8	Adaptive filtering by spectral modification	35
Figure 3-1	Symmetry regions of the Bispectrum	43
Figure 3-2	Overlapping scheme between consecutive frames	45
Figure 3-3	Fourier transform of the cumulant slice from the bispectrum	48
Figure 3-4	Geometric mean of the power spectrum from the bispectrum	50
Figure 3-5	Algorithm 1 for the Fourier magnitude recovery from the bispectrum	54
Figure 3-6	Algorithm 2 for Fourier magnitude recovery from the bispectrum	55
Figure 3-7	Algorithm 3 for Fourier magnitude recovery from the bispectrum	56
Figure 3-8	Algorithm A for Fourier magnitude recovery from the bispectrum	57
Figure 3-9	Algorithm B for Fourier magnitude recovery from the bispectrum	59
Figure 3-10	Algorithm C for Fourier magnitude recovery from the bispectrum	60
Figure 3-11	Auto-convolution of the power spectrum: the case of a flat-spectrum signal	65
Figure 4-1	Waveform of 10 msec of speech in lower and upper bands	68
Figure 4-2	The bispectrum of two related harmonics with $w_2 = 2\ w_1$	71
Figure 4-3	The value of $\alpha T\coth(\alpha T)$ as a function of (αT)	74
Figure 4-4	$FC_4^a(w)$ for the case of one sinusoid with deterministic phase	76
Figure 4-5	Segments of voiced speech	84
Figure 4-6	Normalized kurtosis in three consecutive lower bands	85
Figure 4-7	Normalized kurtosis in an upper band	85

Figure 4-8	Normalized skewness in a lower band	86
Figure 4-9	Histograms of γ_3 and γ_4 across all bands	86
Figure 4-10	Voiced speech in band 1 along with the computed diagonal slice	87
Figure 4-11	Voiced speech in band 25 along with the computed diagonal slice	88
Figure 4-12	The unvoiced phoneme /h/	89
Figure 4-13	Normalized skewness and kurtosis of unvoiced speech (/h/) in an upper band	89
Figure 4-14	Histograms of γ_3 and γ_4 across the upper bands of unvoiced speech	90
Figure 4-15	Histograms of γ_3 and γ_4 across the upper bands of Gaussian noise	90
Figure 5-1	The ratios $1/(r^2+1)$ and $1/(r^4+1)$ for the possible range of r	99
Figure 5-2	Post filtering estimation of the speech energy and speech autocorrelation	105
Figure 5-3	Auditory filter shapes centered at bands 15 and 40	108
Figure 5-4	Block diagram of the speech enhancement algorithm	111
Figure 5-5	Estimated vs. actual SNR for gaussian noise	113
Figure 5-6	Estimated vs. actual SNR for street noise	114
Figure 5-7	Estimated vs. actual SNR for office noise	115
Figure 5-8	Actual, estimated and smoothed SNR for the case of street noise	116
Figure 5-9	Clean, noisy and processed speech waveforms (Gaussian noise. 10 dB)	117
Figure 5-10	Spectrograms for a section in the above waveform	118
Figure 5-11	Clean, noisy and processed speech waveforms (Street noise 13 dB)	119
Figure 5-12	Spectrograms for a section in the above waveform (street noise)	120
Figure 5-13	Clean, noisy and processed speech waveforms (Fan noise 12 dB)	121
Figure 5-14	Spectrograms for a section in the above waveform (fan noise 12 dB)	122
Figure 6-1	LPC residual: the result of filtering speech by a short-term prediction filter	127
Figure 6-2	Computing the value of $FC_0(w)$ for $w = w_0$	129
Figure 6-3	$\|X(w)*X(w)\|^2$ and $P(w)*P(w)$ for $w = 2w_0$	136
Figure 6-4	The utterance "Help the woman"	144
Figure 6-5	Normalized skewness and kurtosis of the LPC residual	145
Figure 6-6	Histograms of normalized kurtosis of residual (speech vs. Gaussian noise)	145
Figure 6-7	Normalized $C_2[\tau]$, $C_3[\tau]$, and $C_4[\tau]$ for frames 20 and 21	146
Figure 6-8	Normalized skewness and kurtosis at 10 dB SNR Levels	147
Figure 6-9	The unvoiced phonemes /f/ and /h/	148
Figure 6-10	Normalized skewness and kurtosis of the LPC residual of /f/	148
Figure 6-11	Histograms of the normalized skewness and kurtosis of /f/	149

Figure 6-12	Histograms of the normalized skewness and kurtosis of /h/	149
Figure 6-13	Normalized skewness and kurtosis of the LPC residual of Gaussian noise	150
Figure 6-14	Histograms of the normalized skewness and kurtosis of Gaussian noise	150
Figure 7-1	HOS-based VAD state machine	159
Figure 7-2	HOS-based and G.729B VAD in Gaussian noise conditions (20 dB)	162
Figure 7-3	HOS-based and G.729B VAD in street noise conditions (10 dB)	163
Figure 7-4	HOS-based and G.729B VAD in office noise conditions (10 dB)	164
Figure 7-5	HOS-based and G.729B VAD in fan noise conditions (10 dB)	165
Figure 8-1	Product terms for computing $C_3[2,8]$	169
Figure 8-2	Product terms for computing $C_3[6,9]$	171
Figure 8-3	A control / data flow graph	174
Figure 8-4	Flow of control of the allocation algorithm	181
Figure 8-5	Flow graph for the matrix operation $BA^{-1}X$	183
Figure 8-6	Scheduling results using the branch-and-bound model	184
Figure 8-7	Scheduling results using Integer Linear Programming	185

CHAPTER 1 Introduction

1.1 Motivation and Rationale

In the context of mobile telephony, speech signals are often corrupted by surrounding acoustic noise such as engine, traffic and wind as well as by system-introduced noise such as quantization, switching handoff, and channel interference. This in turn has an adverse effect on the perceived quality and intelligibility of speech as well as on the performance of speech processing algorithms throughout the network, such as speech coding and recognition.

If the cellular system encodes the signal prior to its transmission, then further degradation in its performance results, since most speech coders rely on a model for the clean signal which is not suitable for the noisy signal. Similarly, integrated speech recognition systems used for telephone services and trained in high-quality conditions will degrade drastically in noisy environments. This is due to the sensitivity of these systems to differences between testing and training conditions.

The idea of speech enhancement in general is to minimize the effect of noise on the performance of voice communication systems. This entails improving the perceived quality to the human listener as well as providing noise-robust methods for estimating crucial speech parameters such as spectral content, pitch, voicing, and others. The goal therefore of speech enhancement algorithms is to:

- Improve the perceptual aspects of a degraded speech signal
- Improve the performance of speech coders
- Increase the robustness of speech recognition systems.

The *quality* of speech signals is a subjective measure which reflects on the way the signal is perceived by listeners. It can be expressed in terms of how much effort is required in order to understand the message, or how pleasant or comfortable speech sounds to the human ear. *Intelligibility* on the other hand is an objective measure of the amount of information which can be extracted by listeners from the given signal [O'Sh89]. In military contexts, intelligibility is of critical importance, whereas in consumer telephony, it seems that quality takes precedence.

Single-sensor speech enhancement is a particular problem in estimation theory for which an optimum solution could be found if the joint statistics of the signal and noise and a meaningful distortion measure were explicitly available. Since in practice this is not the case, suboptimal solutions based on mathematically tractable distortion measures or on some properties of the auditory system have been proposed. The first group includes such methods as Wiener and Kalman filters, while the second includes comb filtering or spectral shaping methods based on formants and harmonic locations. In all these methods however, the problem of estimating the second statistics of the speech and noise remains an open one, and the effectiveness of the enhancement method is contingent on the proper estimation of these crucial parameters.

Other speech processing applications, such as pitch estimation or voicing detection, are an integral part of a variety of speech communication systems, such as coding and recognition. These problems are challenging in and of themselves, and their difficulty increases in the presence of noise; as such, there is a need for finding robust methods that perform consistently in adverse conditions.

Rationale for Higher Order Statistics

Higher-order statistics (HOS) are a recently developed class of tools that have shown promising potential in such fields as radar, image processing, seismology, and array processing [Nik87][Bril94]. They are of particular value when dealing with a mixture of Gaussian and non-Gaussian random processes and system nonlinearity. Their attractive features include:

- ***Gaussian blindness:*** The higher order cumulants of a Gaussian process are identically zero. Thus, given a signal $x(n)$ corrupted by additive and uncorrelated Gaussian noise $g(n)$: $y(n) = x(n) + g(n)$, the higher order cumulants of $y(n)$ are simply those of $x(n)$:

$$cumulant(y) = cumulant(x) .$$

Analysis in the higher order domain is therefore a way of filtering out all Gaussian noise (both white and coloured). When 3rd-order statistics are used, this statement is true for any symmetrically distributed noise, not only Gaussian.

- ***Phase preservation:*** unlike the power spectrum, the higher-order spectra preserve the complex phase making it possible to recover the original time waveform. For instance, using 3rd-order statistics allows the recovery of the complex phase from the phase of the bispectrum: Consider a sequence *x(n)* and its DTFT *X(w)*; the power spectrum is:

  $P(w) = X(w)X^*(w) = |X(w)|^2 e^{j\alpha(w)}$, where $\alpha(w) = \theta(w) + \theta(-w)$ and $\theta(w)$ is the phase of *X(w)*. For a real signal, $\alpha(w) = 0$. On the other hand, the bispectrum is:

  $$B(w_1, w_2) = |B(w_1, w_2)| e^{j\beta(w_1, w_2)}$$
  $$B(w_1, w_2) = |X(w_1)||X(w_2)||X(-w_1 - w_2)| e^{j(\theta(w_1) + \theta(w_2) - \theta(w_1 + w_2))}$$

  thus allowing the recovery of the phase of the Fourier transform.

- ***Detection of non-linearity and phase coupling:*** The use of 3rd and 4th order statistics allows the detection of system-introduced coupling in frequencies or phases. For instance if a signal consists of two sinusoids: $x(t) = \sin(w_1 t) + \sin(w_2 t)$, then the bispectrum of this signal is identically zero, unless the two frequencies are harmonically related: $w_2 = 2w_1$ (proof in Chapter 4). Similar features are noticed when using the 4th-order cumulant. Therefore, if some of the frequencies at the output of a system are created as a result of non-linearity of the system, then comparison of the HOS of the input and output will unveil such behavior.

The application of HOS to speech analysis has been primarily motivated by their inherent Gaussian blindness and phase preservation features (e.g., [Fal93]). The work in this area has rested on the assumptions that speech has certain HOS properties that are distinct from those of Gaussian noise. While this assumption has been backed by experimental observations (e.g. [Wel85], [Fal93]), there has not been an analytical proof of it, nor a formal framework for using these statistics in a useful manner. For these reasons, it is difficult to assess the limited effectiveness reported in some of these attempts (e.g., [Pal91], [Ful93]).

1.2 Goal of this Thesis

The goal of this thesis is to exploit the HOC properties of speech in the objective of finding new algorithms for quality enhancement and robust estimation of speech parameters in the presence of Gaussian noise. To this end, four major issues are addressed:

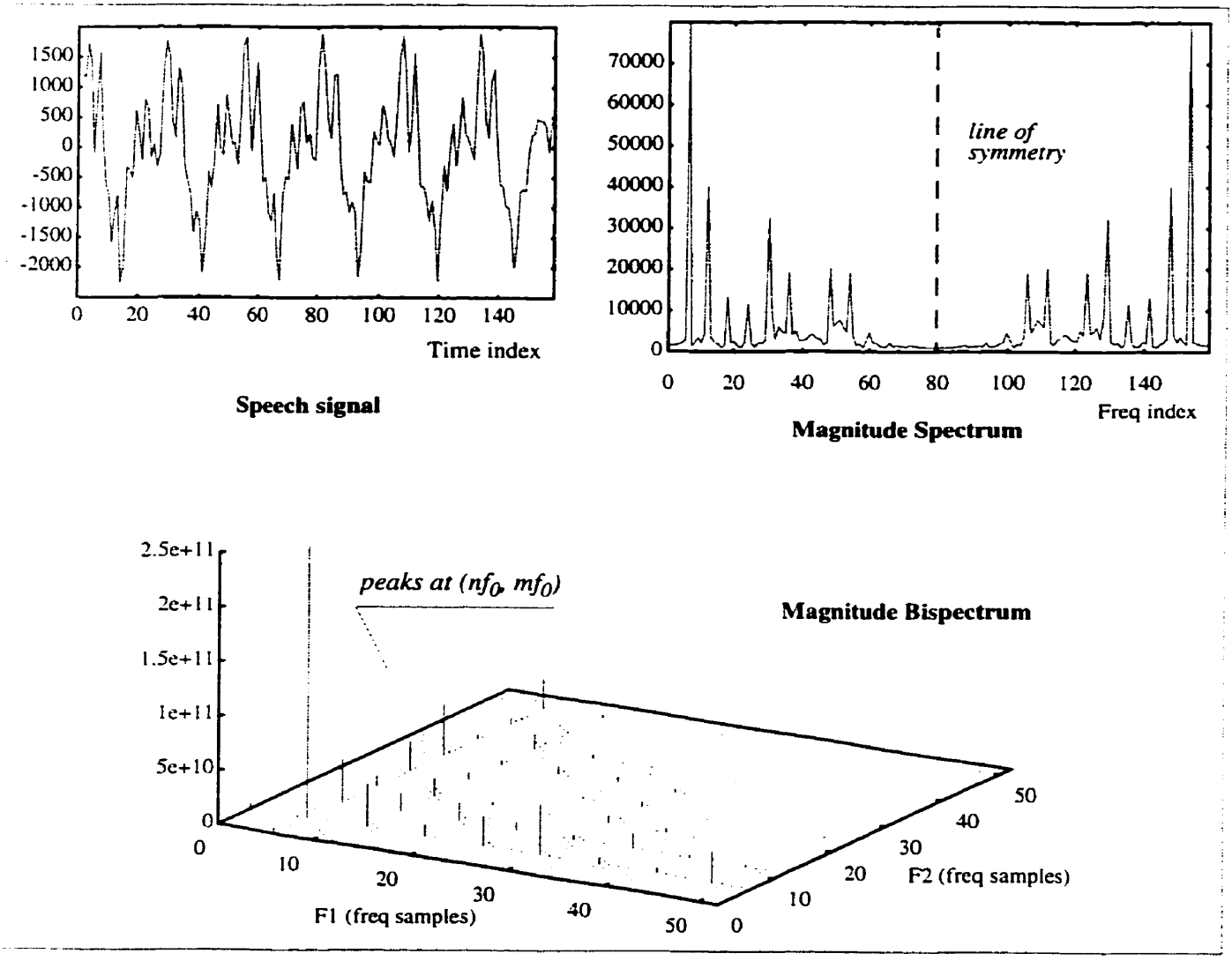
1. ***The burden of proof***: The higher order cumulants of noisy speech are identical to those of the clean speech when the corrupting noise is uncorrelated and Gaussian. For this statement to have a meaningful implication, the cumulants of the speech itself have to be non-zero. Experimental observation and intuitive reasoning suggests that the HOS of speech have to be non-zero. For example, speech samples have been shown to have a Laplacian distribution: $f(x) = ce^{-\alpha|x|}$; as a result, the kurtosis of speech may be shown to be a function of the energy (variance):

$$Kurtosis = E[x^4] - 3(E[x^2])^2 = 3\left(\frac{2}{\alpha^2}\right)^2 = 3(E[x^2])^2$$

and is clearly non-zero. On the other hand, one can take a deterministic approach to speech. If voiced speech is modeled as a sinusoidal signal, then its power spectrum consists of delta functions at harmonic frequencies and its bispectrum may be shown to have similar peaks at multiples of the fundamental frequency. Analysis with actual speech data shows this to be indeed the case (Figure 1-1). In the rest of this thesis, the 3rd and 4th order cumulants of speech are explored and shown to be different from those of Gaussian noise. An analytical model and two speech representations are considered, the LPC residual and a narrow subband paradigm.

2. ***Interpretation of speech HOS***: It is not sufficient to prove that the HOS of speech are non-zero. To make meaningful use of these statistics, one needs a framework where these are expressed in terms of useful speech attributes, such as energy, frequency, pitch, etc. To provide such a framework, it is necessary to assume a model for speech that is both mathematically manageable as well as reasonably representative of actual signals. In this work, the cumulant slices, the kurtosis, skewness and the DC component of some cumulant slices are derived based on a sinusoidal model for speech. Important relations between second and higher-order cumulants are highlighted. As in any scientific approach, the results thus deduced from this model (i.e, the HOC expressions) are verified by experimental data and used to prove the model validity or limitations.

Figure 1-1 Magnitude of the spectrum and bispectrum of a voiced speech segment



3. ***Implementation issues and limitations***: The Gaussian-blindness feature of HOS is true only in the statistical sense. When analysis is done using only one realization and finite data records, issues related to the bias and variance of the HOS estimators need to be accounted for. As an example, the detection of whether a given signal frame consists only of Gaussian noise can only be made in a probabilistic sense with a confidence interval. Another implication is that careful smoothing of the estimators is required prior to using them for inferring speech parameters.

Additional implementation issues related to the computational complexity of HOS and how one may use various efficient computational means to estimate them are also addressed.

4. ***Application of HOC to speech analysis and enhancement:*** Once the HOC properties of speech are established and once the practical limitations of using HOS are identified, the task of designing algorithms for a specific application is a problem in and of itself. For example in the application of voice activity detection, SNR measures have to be combined with HOS metrics to circumvent the problem of false detection of small amplitude segments as voiced speech. The tuning of the algorithm in terms of frame size, and detection thresholds is an important aspect and one that calls for as much art as science. In the application of speech enhancement, the issue of how to formulate the problem so that HOC can be used to deduce various required parameters is an important algorithmic design issue. Furthermore, the details related to the number of subbands, the frame size, the overlap between iterations and the degree of smoothing of the estimated parameters are all aspects that need to be determined by experimentation and intuitive reasoning.

1.3 Structure of this Thesis

Chapter 2 is a review of the various approaches to speech enhancement that are commonly reported in the literature. Reported work on using Higher Order Statistics for speech analysis is also reviewed.

The first part of Chapter 3 is a brief background on HOS, highlighting the time and frequency domain definitions and properties. The second part in that chapter consists of a series of new derivations that relates second and higher order statistics of deterministic signals.

Chapter 4 is an exploratory work into the HOC properties of subbanded speech. It is assumed that speech is divided in narrow bands, such that each band contains one or two harmonics. The expressions for the diagonal slice of the 3^{rd} and 4^{th} order cumulants are derived assuming a sinusoidal model. The peculiarities of these functions in terms of phase and harmonic content are highlighted and the relation between 2^{nd} and 4^{th} order statistics is particularly noted. Actual speech data is used to verify the validity of the expressions and the underlying model.

A new method for speech enhancement based on optimal filtering, subbands, and HOC is proposed in Chapter 5. The key idea is to use the 4^{th}-order statistics to estimate the parameters required for the

enhancement filters, namely the 2nd-order statistics of the speech and noise. It is shown that the kurtosis and the diagonal slice of the fourth cumulant may be used to estimate such parameters as the SNR, speech autocorrelation and the probability of speech presence. Additional aspects of psychoacoustics such as frequency masking are accounted for. The resulting algorithm is tested in typical mobile noise conditions and its performance compared to the TIA standard for noise reduction [IS127].

Chapter 6 takes a similar exploratory approach as Chapter 4 into the HOC properties of speech, with a focus on the LPC residual. It is assumed that the LPC residual has a flat spectrum where ideally all short-term correlation is removed. The expressions for the horizontal slices of the 3rd and 4th order cumulants are derived assuming a sinusoidal model. The peculiarities of these cumulants in terms of phase, periodicity and harmonic contents and their similarity with the autocorrelation function are highlighted. The expressions for the skewness and kurtosis of speech are noted and their use as metrics for detection of voiced speech is discussed. As in Chapter 4, actual speech data is used to assess the validity of the derivations and the underlying sinusoidal model in the LPC residual domain.

The HOC properties of LPC-filtered speech derived in Chapter 6 are exploited in Chapter 7 in the goal of finding a robust algorithm for voice activity detection in the presence of noise. A necessary condition for voicing is derived based on the relation between the skewness and kurtosis of voiced speech. The variance of the HOS estimators is used to yield a likelihood measure for noise frames. The performance of the algorithm is compared to the ITU-T G.729B VAD [Ben97] in various noise conditions. The probability of correct and false classifications is computed for both algorithms relative to hand-labeled speech for various SNR and noise types.

Chapter 8 and Appendix A explore some of the crucial issues in implementing HOS-based methods, namely accuracy and computational complexity. In Appendix A, the bias and variance of the HOS estimators of a mixed process, particularly the skewness and kurtosis of a sinusoid in Gaussian noise, are quantified through a number of derivations. A new unbiased estimator for the kurtosis is proposed and its variance derived for the case of a white Gaussian process. Chapter 8 addresses the computational aspects, namely:

- The complexity in implementing HOS: an algorithm is proposed for efficiently computing the 3rd-order cumulant function with a reduced number of multiplications.

- The problem of executing a DSP algorithm on a parallel architecture. A general scheduling and allocation model is proposed for mapping a set of operations on a configurable multi-unit architecture.

1.4 Contributions of this Thesis

As mentioned before, the goal sought in this thesis is to exploit the HOC properties of speech in the objective of finding new algorithms for quality enhancement and robust speech analysis in the presence of Gaussian noise. To arrive at this goal, fresh insights into the HOC properties of speech signals as well as some general findings about HOS relations with 2^{nd}-order statistics are unveiled. As a result, the contribution of this thesis is on the one hand establishing a framework for using the higher cumulants of speech and on the other in providing two specific algorithms for speech processing to illustrate the effectiveness of these cumulants and their robustness in Gaussian noise conditions. The first algorithm is for enhancing the quality of speech degraded by acoustic additive noise and the second one is a VAD algorithm for classifying speech and noise frames and detecting the voicing characteristics of speech frames.

Unlike the reported work on the use of HOS for speech, the approach taken here is more formal and systematic whereby the HOC properties are first established analytically, then verified using simulations with actual signals, and finally used for estimating specific speech parameters for the two applications considered. The contributions are as follows:

1. General HOS findings

- The expression for the Fourier transform of any horizontal slice of the 3^{rd}-order cumulant (Section 3.2.1) and the horizontal and diagonal slices of the 4^{th}-order cumulant (Section 3.3.1).
- The relation between the Fourier transform of the 3^{rd}-order cumulant slice and the bispectrum (Section 3.2.2).
- The expression for the geometric mean of the power spectrum of a signal from its magnitude bispectrum (Section 3.2.3).
- Three schemes for recovering the magnitude of the Fourier transform of a signal from the magnitude bispectrum (Section 3.2.4).

- Properties about the DC component of the horizontal slice of the 4^{th}-order cumulant (Section 3.3.2).

2. The HOC properties of speech

- The proof that the 3^{rd}-order cumulant of subbanded speech is identically zero (Section 4.2).
- The expression for the diagonal slices of the 4^{th}-order cumulant of subbanded speech according to the sinusoidal model assumed (Section 4.3).
- Relations between the kurtosis of subbanded voiced speech and the speech energy. Analytical and experimental findings are reported (Section 4.3.1 and Section 4.3.2).
- The expression for the diagonal slices of the 4^{th}-order cumulant of subbanded unvoiced speech according to the sinusoidal model and the experimental findings that prove this model invalid (Section 4.3.3 and Section 4.5.2).
- The expressions for the horizontal slices of the 3^{rd} (Section 6.3.1) and 4^{th} order (Section 6.4.2) cumulants of the LPC residual of voiced speech according to the sinusoidal model. The expressions reveal important properties of these cumulants in terms of periodicity, phase and harmonic content and highlight important relations between the energy of speech and its higher order statistics, namely the skewness and kurtosis. The findings are verified by experimental data (Section 6.7).
- The analysis of whether unvoiced speech may be properly modeled as a harmonic process in the LPC domain. The expression for the slices of the cumulants are derived (Section 6.3.3, Section 6.4.1) and the experimental results that prove the model invalid are noted (Section 6.7.3).
- The relation between the DC component of the horizontal slice of the 4^{th}-order cumulant of voiced speech and the energy and bandwidth of the signal, in the case of the LPC residual (Section 6.4.2.1).

3. Application of HOC

- An algorithm for voice activity detection based on the established HOC properties of voiced speech in the LPC residual domain (Chapter 7). The algorithm combines HOS metrics and SNR measures to classify frames as speech or noise and determine whether speech frames are voiced. A condition on voicing is determined based on the relation between the skewness and kurtosis of voiced speech. Results in various noise conditions show the proposed VAD

has overall comparable performance to the ITU-T G.729B: Its probability of false classification is lower in low SNR and Gaussian-like noise, but higher in speech-like noises. The algorithm however is conceptually simpler and is based on more analytical grounds than the heuristic G.729B. Most importantly, the algorithm proposed demonstrates the effectiveness of using HOS and their robustness in low SNR conditions.

- An algorithm for speech quality enhancement based on optimal filters, subband and 4^{th}-order cumulants (Chapter 5). The algorithm uses a subbanding approach in the time domain and MSE-based optimum filters to enhance noisy speech. The 4^{th}-order cumulant is used to estimate such parameters as the SNR, the speech autocorrelation and the probability of speech presence. When compared to the TIA IS-127 noise reduction standard in various noise conditions, the HOS algorithm is shown to be better at preserving the harmonic structure of the speech and results in less speech distortion. It also results in more overall reduction of the noise, particularly in Gaussian-like environments such as street, fan and office noise, but this comes at the cost of slightly more noise artifacts particularly at very low SNR and non-Gaussian noise conditions.

4. HOC Implementation

- An algorithm for efficiently computing the 3^{rd}-order cumulant function with a reduced number of multiplications, by exploiting the redundancy in the product terms (Section 8.1).
- A scheduling and allocation algorithm for mapping a set of operations on a configurable multi-unit architecture (Section 8.2). A new approach based on probabilistic allocation methods is proposed. The scheme follows the general framework as branch-and-bound approaches, whereby a heuristic is used at each decision point to determine the matching of a candidate operation with an available resource. The criterion is a probabilistic decision based on quantifying the opportunity cost of the resource and the scheduling urgency of the operation.

CHAPTER 2 Background and Literature Review

2.1 An Analytical Model For Speech

To provide an analytical framework for using higher order cumulants, it is necessary to assume a model for speech that is both mathematically manageable as well as reasonably representative of actual signals. Speech processing is one area that is dominated by linear models, in spite of the physical and experimental evidence that seems to suggest that non-linearity needs to be accounted for [Tea83][Tea90]. It is argued in [Fac96] that using 3^{rd}-order statistics for speech analysis can reveal information about the nonlinear signal generation mechanisms, but the results in [Fac96] were non-conclusive, even negative about the presence of this non-linearity, or its detection with HOS.

In the area of coding and voice detection, models that describe speech signals as a stream of different Gaussian processes have been proposed [Ram79][Ram80]. The experimental work in [Gab88] using 3^{rd} and 4^{th} order statistics of long-term speech showed however that these models are limited and that the speech process may be more accurately considered as mixtures of spherically invariant Gaussian distributions.

Since the interest in this work is in short-term speech segments, and since the goal is to estimate conventional speech model parameters, it is reasonable to use a linear model that is valid for extracting these parameters. As in any scientific approach, some model has to be assumed at first, then the results deduced from this model -in this case the higher order cumulants- are verified by experimental data and used to prove the model validity and limitations.

Among the sinusoidal models developed to represent speech, the simplest and often referenced one is the so-called *zero-phase harmonic representation* [McA86]. The elegance of this model is in its use of the same expression for both voiced and unvoiced speech and allowing for a soft decision whereby a frame may contain both types of speech. The model is characterized by sine-wave amplitudes, a voicing probability, and a fundamental frequency. Removing (or greatly simplifying) the phase makes it a minimal parameter set for analysis and synthesis [Qua86]. In this representation, a given frame is represented by a sum of harmonically related sine waves. A synthetic phase function is used such that during voiced speech, the sine waves are coherent (in phase) and during unvoiced speech they are incoherent. The speech signal over a short-term window may be expressed as:

$$s(n) = \sum_{m=1}^{M} a_m \cdot \cos\left[(n - n_0)\, w_m + \psi_m + \theta_m\right] \qquad \text{(E 2.1)}$$

where n_0 is the voice onset time, M is the number of sinusoids, a_m the amplitude of the m^{th} sine wave and w_m the excitation frequencies. For a perfectly periodic frame, these are harmonically related, i.e., $w_m = mw_0$, with w_0 the fundamental frequency. The first phase term is due to the onset time n_0, defined as the time when the pitch pulse occurred relative to the beginning of the frame. The second phase component depends on a frequency cutoff w_c and a voicing probability P_v:

$$\psi_m = \begin{cases} 0 & ;for\ w_m \le w_c \cdot P_v \\ U[-\pi, \pi] & ;for\ w_m > w_c \cdot P_v \end{cases} \qquad \text{(E 2.2)}$$

where $U[-\pi, \pi]$ is a uniformly distributed random variable between $-\pi$ and π. Thus, the higher the voicing probability the more sine waves are declared voiced with zero phase. Finally, the third phase component is the system phase θ_m along frequency track m. For simplicity, this component is often assumed to be zero or a linear function of frequency. In the case of steady voiced speech, the sine waves are harmonically related and Eq 2.1 becomes:

$$s(n) = \sum_{m=1}^{M} a_m \cdot \cos\left[(n - n_0)\, mw_0 + \psi_m + \theta_m\right] . \qquad \text{(E 2.3)}$$

Therefore in the framework of the sinusoidal model:

- A *steady (or stationary) voiced* speech segment is modeled as a sum of harmonically related sine waves whose frequencies are multiples of the fundamental and whose phases are determined entirely by the voice onset time n_0.

- A *non-stationary voiced* speech segment is modeled as a sum of sine waves, but only some of those may be harmonically related. For the rest, the phases are assumed deterministic but unknown.
- An *unvoiced* speech segment is modeled as a sum of incoherent sine waves whose phases are assumed random and uniformly distributed. In the more general approach, unvoiced speech is considered as a random -though not necessarily Gaussian- process.

2.2 Speech Quality Enhancement Techniques

The literature on speech enhancement techniques is quite abundant. This section elaborates on the common approaches. These are categorized in the following groups:

1. **Wiener Filtering and Related Methods**: Enhance speech by spectral decomposition and optimal linear filters; these filters are derived by minimizing the MSE or other criteria. Estimation of the SNR and other speech and noise statistics is an important issue in these approaches.
2. **ML and MMSE-based Spectral Estimation:** Formulate the enhancement problem as an ML or an MMSE estimation of the speech spectral envelope or magnitude. These methods involve an estimation of the DFT of speech from the noisy speech spectrum.
3. **Kalman Filtering:** Formulate the enhancement problem in state space form, by modeling the speech as an AR process. Assuming the AR parameters may be estimated, estimate the state vector and the clean speech.
4. **Comb Filtering**: Comb through the spectrum and reinforce the harmonic components of voiced speech. The issue of computing the parameters of the comb filter to account for the imperfect periodicity of speech is a crucial issue.
5. **Wavelet Denoising:** Transform the signal using a set of orthonormal basis functions. Apply a thresholding criteria to reject likely noise components and resynthesize speech. The choice of thresholding rule is a main issue in these methods.
6. **Psychoacoustic Methods:** Apply special filtering that takes into account the peculiarities of perceptually important speech parameters or acoustic criteria of human hearing.

For each approach, the theoretical basis is presented, followed by a summary of one or two typical papers in that category.

2.2.1 Wiener Filtering and Related Methods

Wiener filters are an instance of the general class of optimum filters [Van68], where it is desired to find an estimate $\hat{S}_t$ for a signal S_t given a number of observations, $\{X_{t-a}, \ldots, X_t, \ldots, X_{t+b}\}$. It is assumed that the observations are the sum of the desired signal plus unwanted noise:

$$X_\alpha = S_\alpha + N_\alpha \qquad \alpha \in I \quad I = \{t-a, \ldots, t+b\} . \tag{E 2.4}$$

The estimate $\hat{S}_t$ is obtained by a linear filter acting on the set of observations:

$$\hat{S}_t = \sum_{\beta=t-a}^{t+b} h_{t-\beta} \cdot X_\beta = \sum_{\beta=-b}^{a} h_\beta \cdot X_{t-\beta} . \tag{E 2.5}$$

The filters are *optimum* with respect to minimizing the *mean square error*:

$$E[e_t^2] = E[(S_t - \hat{S}_t)^2] . \tag{E 2.6}$$

The optimum filter is based on the principle of *orthogonality* whereby the error is orthogonal to all observations:

$$E[e_\alpha X_\alpha] = 0 . \tag{E 2.7}$$

It can be shown [Leo89] that in the case where X_t and Z_t are zero-mean jointly wide-sense stationary processes, the filter that minimizes the mean square error must satisfy the equations:

$$R_{S,X}(m) = \sum_{\beta=0}^{p} h_\beta R_X(m-\beta) \qquad m \in \{0, 1, \ldots p\}. \tag{E 2.8}$$

That is the filter coefficients could be found by solving the $p+1$ simultaneous equations (Eq 2.8). Here, $R_S(m)$ and $R_X(m)$ are the autocorrelation functions of the clean signal and the noisy observations and $R_{S,X}(m)$ is the cross correlation between the two. In the case where the signal and noise are independent random processes, that is:

$$R_{S,X}(m) = R_S(m) \quad \text{and} \quad R_X(m) = R_S(m) + R_N(m) , \tag{E 2.9}$$

the equation of the filter becomes:

$$R_S(m) = \sum_{\beta=0}^{p} h_\beta \{R_S(m-\beta) + R_N(m-\beta)\} \qquad m \in \{0, 1, \ldots p\}. \tag{E 2.10}$$

2.2.1.1 Infinite smoothing and frequency-domain filters

If S_t is to be estimated using the entire realization of X_t, that is $I = (-\infty,\infty)$, (this is the case when the entire signal is recorded, and then played back) then:

$$R_{S,X}(m) = \sum_{\beta=-\infty}^{\infty} h_\beta R_X(m-\beta) . \quad \text{(E 2.11)}$$

The Fourier transform yields:

$$P_{S,X}(f) = H(f) P_X(f) . \quad \text{(E 2.12)}$$

If, in addition, the processes are independent and zero-mean, then

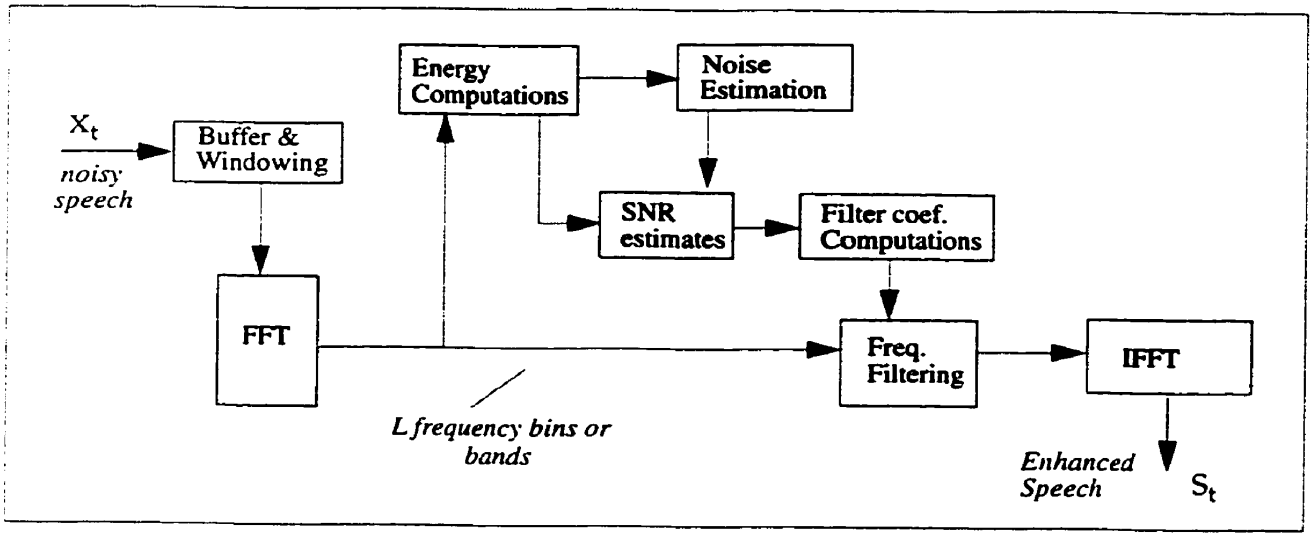
$$P_{S,X}(f) = P_S(f) \text{ and } P_X(f) = P_S(f) + P_N(f) . \quad \text{(E 2.13)}$$

The *optimum* filter is found (using Eq 2.12 and Eq 2.13):

$$H(f) = \frac{P_S(f)}{P_S(f) + P_N(f)} = \frac{SNR(f)}{SNR(f) + 1} . \quad \text{(E 2.14)}$$

Thus assuming the power spectrum of the speech and noise could somehow be estimated, a speech enhancement system may be achieved by spectral decomposition and appropriate scaling of the various frequency coefficients. A generic Wiener-based enhancement system is shown in Figure 2-1.

Figure 2-1 Generic Wiener-based speech enhancement



Power subtraction filters

Consider a filter whose response is the square root of a Wiener filter, thus:

$$|H(f)|_P = \sqrt{\frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2}}. \quad \text{(E 2.15)}$$

The power spectrum of the output of this filter is:

$$P_{out} = P_{in}|H(f)|^2 = (|S(f)|^2 + |N(f)|^2) \bullet \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2} \quad \text{(E 2.16)}$$

or simply

$$P_{out} = |S(f)|^2. \quad \text{(E 2.17)}$$

Thus the filter has suppressed noise in the *power* spectrum sense. This is equivalent to subtracting an estimate of the noise power from the spectrum of the noisy speech.

Magnitude subtraction filters

Consider another filter whose response is given by:

$$|H(f)|_M = 1 - \sqrt{\frac{|N(f)|^2}{|S(f)|^2 + |N(f)|^2}}. \quad \text{(E 2.18)}$$

The *magnitude* spectrum of the output of this filter is:

$$M_{out} = \sqrt{(|S(f)|^2 + |N(f)|^2)} \bullet |H(f)|_M \quad \text{(E 2.19)}$$

or simply

$$M_{out} = \sqrt{(|S(f)|^2 + |N(f)|^2)} - |N(f)| \quad \text{(E 2.20)}$$

and thus this is a magnitude suppression filter. Figure 2-2 below shows a plot of the filter magnitude response for different SNR values.

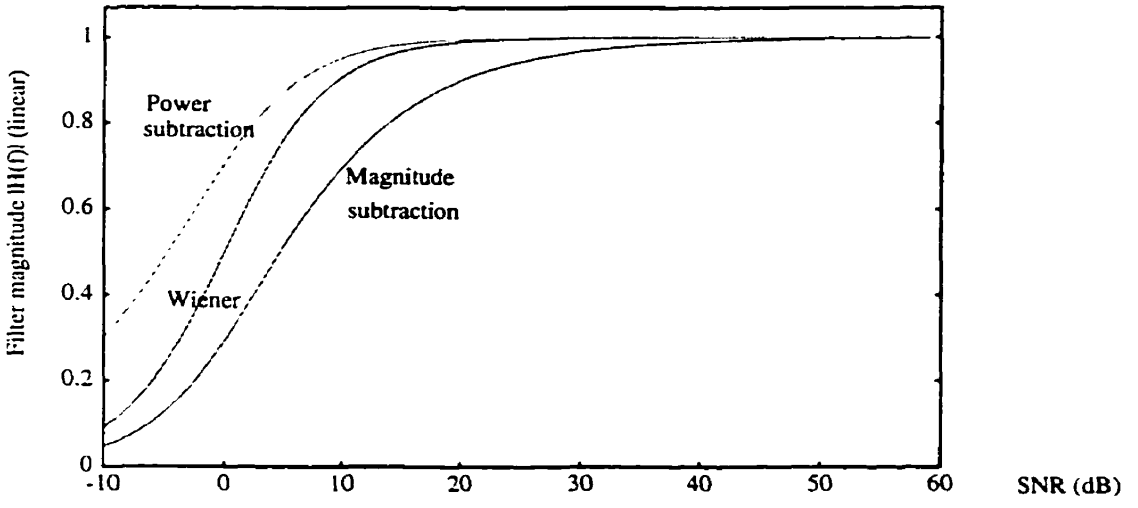
Parametrized Wiener filters

The so-called parametrized Wiener filter has been used in published systems ([Ars95]) and is a modified version of Eq 2.14 with the introduction of two parameters to control the slope of the filter as well as the noise factor:

$$H(f) = \left[\frac{SNR(f)}{SNR(f) + \alpha}\right]^{\beta}; \quad \text{(E 2.21)}$$

α is often termed the oversubtraction factor. It refers to the fact that the noise power is overestimated to achieve a more aggressive filter or to compensate for the fact that the noise spectrum estimation is too conservative. Graphically, this is equivalent to pushing the curve to the right. The β factor controls the slope of the filter. The higher β, the steeper the slope of the filter. A power subtraction filter is obtained when $\beta = 0.5$ and $\alpha = 1$.

Figure 2-2 Various suppression filters



The advantages of the Wiener-based approach is its intuitive basis, its implementation simplicity and its practical effectiveness. Its limitations on the other hand are:

- Noticeable residual noise which consists of narrow-band signals with time-varying frequencies and amplitudes. This is referred to as musical noise. This noise is mostly perceived in weak bands and weak speech frames. In strong speech bands, the noise is masked by the speech energy.
- The dependence on a good noise and SNR estimation, which is a problem in itself. Noise is often estimated during speech pauses, which in turn relies on the assumption that the background noise environment remains locally stationary to the degree that its spectral magnitude expected value just prior to speech activity equals its expected value during speech activity.

- Ignoring the effects of noise on the phase of the signal. While it was shown that using the actual (clean) version of the phase instead of the phase of the noisy speech does not improve the enhancement, very little is understood about the effect of noisy phase on speech. In experiments where only the complex phase was modified, it was found that adding random noise above a given threshold will cause roughness in the reconstructed speech. Zeroing the phase will cause speech to sound monotonous, while randomizing the phase will cause speech to sound rough.

2.2.1.2 White observation and Causal filters

In the special case where the *observation* is white, i.e., $R_X(m) = \sigma_x^2\delta_m$ and the observation range is all the past and present observations: $I = (-\infty, t)$, then the solution is a causal filter given by (from Eq 2.8):

$$R_{S,X}(m) = \sigma_x^2 \sum_{\beta=0}^{\infty} h_\beta \delta_{m-\beta}$$

and the filter coefficients are:

$$h_m = R_{S,X}(m)/\sigma_x^2 \qquad \text{For } m \geq 0. \qquad \text{(E 2.22)}$$

If in addition, the noise and speech processes are statistically independent, then:

$$h_m = R_S(m)/\sigma_x^2 \qquad \text{For } m \geq 0. \qquad \text{(E 2.23)}$$

2.2.1.3 Some Reported systems

- **Magnitude Subtraction**

The suppression schemes in [Bol79] and [Var85] are classical examples of magnitude subtraction systems where noise is estimated by averaging the signal magnitude spectrum during *non-speech* activity. In [Bol79], the window length is chosen to be at least twice as large as the maximum expected pitch period in order to achieve adequate frequency resolution (256 points). Key ingredients of this system include:

— Averaging of the spectrum over a small number of frames is done in order to minimize the residual noise. The fact that speech is non-stationary implies that only a limited time average of no more than three frames is possible.

— A scheme to control the amount of subtracted noise is used. The idea is to reduce the random variation of noise residual between adjacent frames. By comparing the magnitude in those weak bands (i.e., low SNR) over the last three frames, a new value for the magnitude spectrum is chosen as the minimum of these three.

— Additional attenuation is introduced during non-speech activity within a given analysis frame. Whenever speech is absent, an additional amount of 30 dB attenuation is used. To determine the absence of speech, the average SNR is computed and summed over all the frequency bands. Whenever this ratio is less than -12 dB, the frame is classified as having no speech activity and attenuated by 30 dB.

- **Power Subtraction**

In [Bei79], a power subtraction system with the concept of noise overestimation is introduced. Oversubtraction is used to compensate for imperfect noise estimation. This system has three interesting points:

— The variable oversubtraction factor is based on the estimated SNR in each band. The rationale is that strong SNR bands are indicative of a strong speech component and there is no need for aggressive subtraction.

— A spectral floor is used to ensure the presence of a low broadband noise and its effect is to mask any musical noise that may be present.

— A generalized power spectrum subtraction method is introduced:

$$D(w) = G[P^{\gamma}_{s}(w) - \alpha P^{\gamma}_{n}(w)] .$$

- **Various filter-based approaches**

While the classical techniques tackled the key issues of the subtraction concept, they did not address the concept of sub-banding that takes advantage of the masking features of the auditory system. As the newer publications adopted this idea, spectral subtraction got implemented as a linear filtering operation whereby the filters coefficients are computed for each band based on the estimated SNR of this band.

The difference here is that FFT bins are assigned to bands, and a filter coefficient is computed for each band and used to weight all the FFT bins within. The estimate of the signal

energy in the band is computed by averaging the energy across the bins and smoothing in time over few frames. The estimate of the noise energy in each band is computed by averaging the signal energy over these frames that are classified as non-speech.

In [Ars95], a parametrized Wiener filter with noise oversubtraction is used. Some of the classical concepts in [Bei79] are extended to the new form of subtraction. Worth noting is:

— The concept of 'spectral floor' is implemented by setting a lower bound on the filter (i.e., preventing the filter gain from going below a -10 dB limit). In addition to producing a 'broadband noise' effect, this prevents coefficient randomness around small values of the gain which would result in more annoying musicality of the noise.

— The concept of variable oversubtraction is extended to the filter expression. The frequency gains are derived from a parametrized Wiener filter with an oversubtraction based on SNR:

$$H(w) = \sqrt{\frac{P_y(w)}{P_y(w) + \frac{1}{SNR}\alpha P_n(w)}}.$$

In [Vir95], masking across critical bands is taken into account. For each band, a masking threshold is computed depending on the energies of the neighboring bands. Bands where signal energies are below the threshold will be masked and there is no need for aggressive spectral subtraction. The key steps go as follows:

— Perform a normal spectral subtraction without any overestimation, deduce an estimate of the speech energy in each band.

— From the estimated signal energy, compute the masking threshold at each band caused by neighboring speech bands. If the estimated noise is lower than the masked threshold, then there is no need for aggressive overestimation. If however the noise is above the threshold, it will be audible and thus more aggressive subtraction is required.

— Use the threshold as the basis of a spectral subtraction operation where the overestimation factor is adapted for each band and frame.

2.2.2 ML and MMSE Spectral Estimation

While Wiener-based approaches did not make assumptions about the distributions of speech and noise, the following ones are estimation-oriented techniques based on assumed *apriori* distributions of the signals. The solution proposed is an estimator of the spectral components; it is therefore the distribution of this spectral component (DFT coefficient), rather than the time samples, that is of importance.

ML Estimation of the Speech Envelope

The problem of speech enhancement is formulated in [McA80] as a maximum likelihood estimation of the speech spectral envelope. In this model, the noise is assumed to be a *Gaussian* random process and the speech a deterministic waveform of unknown amplitude and phase. Thus, each DFT channel value is given by: $y_k = s_k + w_k$, where the speech is given by: $s_k = A_k e^{j\theta_k}$, A_k representing the speech envelope and θ_k its phase. The estimator is based on maximizing the *posteriori* probability of the channel measurement, $p(y_k|A_k, \theta_k)$ or, after averaging out the phase, the average likelihood function, $\overline{p(y_k|A_k)}$. This leads to the estimator of A_k:

$$\hat{A}_k = \frac{1}{2}\left[|y_k| + \sqrt{|y_k|^2 - \lambda_w(k)}\right], \qquad \text{(E 2.24)}$$

where $|y(k)|^2$ is the measured envelope energy and $\lambda_w(k)$ the estimate of the noise energy. The estimator of the speech spectrum envelope s_k, given the noisy speech spectrum value y_k:

$$\hat{s}_k = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{|y_k|^2 - \lambda_w(k)}{|y_k|^2}}\right] \cdot y_k . \qquad \text{(E 2.25)}$$

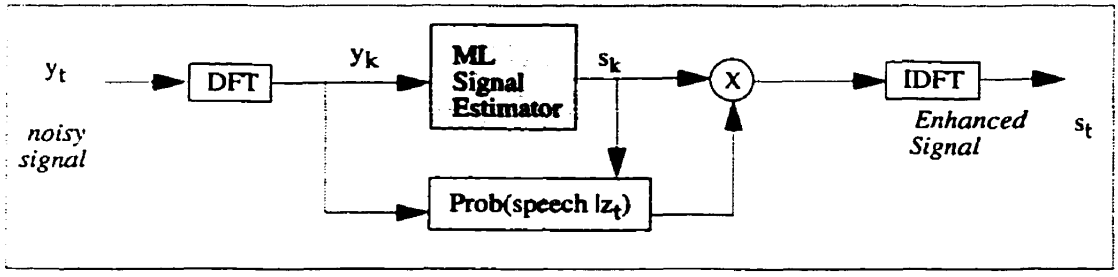
The estimation is further refined to account for the probability of speech presence in channel k. To derive the *posterior* probability of presence of speech, the speech and noise are assumed to be independent Gaussian random processes; thus the *apriori* probability density function of the measured noisy speech envelope is Rician. By assuming the probability of the noise and speech states equally likely, and the *a priori* signal to noise ratio known, the *posteriori* probability of speech presence in band k is:

$$p(H_1) = \frac{exp(-SNR_{prior})\, I_0\left[2\sqrt{SNR_{prior}\left(\frac{|y(k)|^2}{\lambda_w(k)}\right)}\right]}{1 + exp(-SNR_{prior})\, I_0\left[2\sqrt{SNR_{prior}\left(\frac{|y(k)|^2}{\lambda_w(k)}\right)}\right]} . \qquad \text{(E 2.26)}$$

Here, H_1 is the hypothesis that speech is present and the *apriori* signal to noise ratio is $SNR_{prior} = A_k^2/\lambda_w(k)$.

Since the optimal estimator of the clean signal given that this signal is absent from the noisy observations equals zero, the resulting estimator is simply the product of the estimator for the clean signal given that this signal is present in the noisy observations (Figure 2-3).

Figure 2-3 Enhancement conditioned on the presence of speech



MMSE-based Estimation of the Speech Amplitude

The system proposed in [Eph84] is closely related to and builds on the same principle as [McA80], except that the DFT coefficients of the noisy observations are assumed to have a Gaussian distribution. The enhancement problem is formulated as an MMSE estimator of the speech spectral amplitude. Given the same definition of speech, noise and noisy channel measurements as before, the MMSE estimator of the speech amplitude A_k is given by:

$$\hat{A}_k = E\{A_k | Y_0, Y_1, \ldots\} = E\{A_k | Y_k\}$$

where $\{Y_0, Y_1, \ldots\}$ is the set of spectral observations over a few frames, which are assumed to be statistically independent. Given Gaussian assumptions about the spectral values, the estimator can be expressed as a spectral gain $G(p, k)$ that is applied to each short-term spectral value $Y_k(p)$ (here p denotes the frame index):

$$G(p,k) = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+SNR_{post}(p,k)}\right)\left(\frac{SNR_{prior}(p,k)}{1+SNR_{prior}(p,k)}\right)}M\left[(1+SNR_{post}(p,k))\left(\frac{SNR_{prior}(p,k)}{1+SNR_{prior}(p,k)}\right)\right]$$

with

$M[x] = e^{-x/2}\left[(1+x)I_0\left(\frac{x}{2}\right) + xI_1\left(\frac{x}{2}\right)\right]$; I_0 and I_1 are the modified Bessel functions of zero and first order respectively.

Here the two SNR terms are defined as follows:

$$SNR_{post}(p,k) = \frac{|Y(p,k)|^2}{\lambda_w(k)} - 1 \qquad \text{(E 2.27)}$$

is the local estimate of the SNR in the current frame p, at frequency k, using the total energy $|Y(p,k)|^2$ and the estimate of the noise energy $\lambda_w(k)$ at that frequency. The so-called *apriori* SNR represents the information of the unknown spectral magnitude gathered from previous frames and is evaluated as:

$$SNR_{prior}(p,k) = (1-\alpha)P[SNR_{post}(p,k)] + \alpha\frac{[G(p-1,k)]^2|Y(p-1,k)|^2}{\lambda_w(k)}, \qquad \text{(E 2.28)}$$

where $P[x] = x$ when $x > 0$ and 0 otherwise. As in [McA80], the estimator is conditioned on the probability of speech presence at frequency k. Thus the new MMSE gain is:

$$G'(k) = \frac{\Lambda(Y_k, q_k)}{1+\Lambda(Y_k, q_k)}G(k) \qquad \text{(E 2.29)}$$

where $\Lambda(Y_k, q_k)$ is the generalized likelihood ratio defined as:

$$\Lambda(Y_k, q_k) = \mu_k\frac{p(Y_k|H_k^1)}{p(Y_k|H_k^0)} \qquad \text{(E 2.30)}$$

where H_k^1 and H_k^0 denote the hypothesis for speech presence and absence respectively, $\mu_k = (1-q_k)/q_k$, and q_k is the probability of signal absence in the k^{th} spectral band. In [Mal99], a method is proposed to estimate q_k based on a decision-theoretic approach, and in the worst case, $q_k = 0.5$. Using the Gaussian statistical model for the spectral components, it is found that:

$$\Lambda(Y_k, q_k) = \mu_k\frac{e^{\left[\left(\frac{SNR_{prior}(k)}{1+SNR_{prior}(k)}\right)SNR_{post}(k)\right]}}{1+SNR_{prior}(k)}. \qquad \text{(E 2.31)}$$

A detailed study of this approach was reported in [Cap94]. It is shown that the proposed SNR smoothing yields an elimination of the musical noise phenomena without bringing distortion to the speech signal.

2.2.3 Kalman Filters

The motivation for studying a Kalman filter based noise suppression system is that it can handle colored noise and has a reasonable numerical complexity. In addition, it is well suited for speech quality requirements in that it results in low speech distortion and low noise distortion. The use of Kalman filters was proposed in [Pal87] for speech enhancement where experimental results reveal an advantage over Wiener filtering, for the case where the estimated speech parameters are obtained from the clean speech signal (hypothetical case). The method was first proposed for white Gaussian noise, then extended [Koo89] by incorporating a colored noise model.

A key issue in Kalman filtering is that the algorithm relies on a set of parameters that are unknown and have to be estimated from a noisy signal. These include the model parameters for the speech and noise. It is often assumed that the noise is long-time stationary and consequently its parameters are estimated during speech pauses. Speech is viewed as a short-time stationary process, e.g., 10-40 ms. Thus an instantaneous model of the speech has to be obtained from a short segment of noisy measurements.

White Gaussian noise

The noisy speech is modeled ([Pal87], [Koo89]) as the sum of an AR process and a noise process:

$$x(n) = s(n) + v(n) \qquad \text{(E 2.32)}$$

where $x(n)$ denotes the measured signal, $s(n)$ the speech, and $v(n)$ a zero-mean white noise process with variance σ_v^2. Furthermore:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + w(n)\,, \qquad \text{(E 2.33)}$$

where $w(n)$ is a zero-mean white Gaussian process with variance σ_w^2. The above 2 equations can be written in state-space form (bold letters refer to matrices):

$$\begin{aligned} \boldsymbol{s}(n) &= \boldsymbol{F}\boldsymbol{s}(n-1) + \boldsymbol{g}w(n) \\ x(n) &= \boldsymbol{h}^T\boldsymbol{s}(n) + v(n) \end{aligned} \qquad \text{(E 2.34)}$$

with

$$\boldsymbol{s}(n) \equiv \left[s(n-p+1)\;\; s(n-p+2)\;\; \ldots\;\; s(n)\right] \qquad \text{(E 2.35)}$$

$$F = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \dots & a_2 & a_1 \end{bmatrix} \quad \text{and} \quad g = h = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}. \tag{E 2.36}$$

From standard Kalman filtering theory, the state vector estimate is:

$$\hat{s}(n) = F\hat{s}(n-1) + k(n)\,[x(n) - h^T F\hat{s}(n-1)]\,. \tag{E 2.37}$$

The gain and error covariance equations are:

$$k(n) = P(n|n-1)\,h\,[R + h^T P(n|n-1)\,h]^{-1} \tag{E 2.38}$$

$$P(n|n-1) = FP(n-1)\,F^T + gQg^T \tag{E 2.39}$$

$$P(n) = [I - k(n)\,h^T]\,P(n|n-1) \tag{E 2.40}$$

where $k(n)$ is a Kalman gain vector, $P(n|n-1)$ is an *apriori* error covariance matrix, $P(n)$ is an error covariance matrix, $R = \sigma_v^2 I$ is the covariance matrix of the noise sequence and $Q = \sigma_w^2 I$ is the covariance matrix of the driving term $\{w(n)\}$. With the initial condition, $P(0) = [0]_{NxN}$, Eq 2.39 is processed first, followed by Eq 2.38, and then Eq 2.37 and Eq 2.40. The speech sample estimate at time instant n is then obtained by:

$$s(n) = h^T \hat{s}(n)\,. \tag{E 2.41}$$

Colored Gaussian noise

The noisy speech is modeled ([Koo89]) as the sum of 2 AR processes:

$$x(n) = s(n) + v(n) \tag{E 2.42}$$

with $s(n)$ given as in Eq 2.33, and

$$v(n) = \sum_{i=1}^{q} b_i v(n-i) + \eta(n)\,. \tag{E 2.43}$$

The above equation of the noise is rewritten in canonical form:

$$\begin{aligned} v(n) &= F_v v(n-1) + g_v \eta(n) \\ v(n) &= h_v^T v(n) \end{aligned} \tag{E 2.44}$$

with

$$v(n) \equiv \left[v(n-q+1)\ \ v(n-q+2)\ \ \ldots\ v(n)\right]^T \tag{E 2.45}$$

$$F_v = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ b_q & b_{q-1} & b_{q-2} & \ldots & b_2 & b_1 \end{bmatrix} \quad \text{and} \quad g_v = h_v = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \\ 1 \end{bmatrix}. \tag{E 2.46}$$

and $\eta(n)$ is a white Gaussian driving term with zero mean and variance ρ_η^2. Eq 2.34 is combined with Eq 2.44 to yield an aggregate state space model:

$$\begin{bmatrix} s(n) \\ v(n) \end{bmatrix} = \begin{bmatrix} F & 0 \\ 0 & F_v \end{bmatrix} \begin{bmatrix} s(n-1) \\ v(n-1) \end{bmatrix} + \begin{bmatrix} g & 0 \\ 0 & g_v \end{bmatrix} \begin{bmatrix} w(n) \\ \eta(n) \end{bmatrix} \tag{E 2.47}$$

$$x(n) = \begin{bmatrix} h^T & h_v^T \end{bmatrix} \begin{bmatrix} s(n) \\ v(n) \end{bmatrix} \tag{E 2.48}$$

which can be simplified, by defining new vectors and matrices, to:

$$\bar{s}(n) = \bar{F}\bar{s}(n-1) + \bar{G}\bar{w}(n) \tag{E 2.49}$$

$$x(n) = \bar{h}^T \bar{s}(n) \tag{E 2.50}$$

where

$$\bar{s}(n) = \begin{bmatrix} s(n) \\ v(n) \end{bmatrix} \text{ and } \bar{w}(n) = \begin{bmatrix} w(n) \\ \eta(n) \end{bmatrix} \tag{E 2.51}$$

$$\bar{F} = \begin{bmatrix} F & 0 \\ 0 & F_v \end{bmatrix}, \bar{G} = \begin{bmatrix} g & 0 \\ 0 & g_v \end{bmatrix}, \bar{h}^T = \begin{bmatrix} h^T & h_v^T \end{bmatrix}. \tag{E 2.52}$$

This is in the form of a linear system driven by a white Gaussian vector $\bar{w}(n)$. It is assumed that $w(n)$ and $\eta(n)$ are uncorrelated:

$$Q \equiv E\left[\bar{w}(n)\bar{w}^T(n)\right] = \begin{bmatrix} \rho_w^2 & 0 \\ 0 & \sigma_\eta^2 \end{bmatrix}. \tag{E 2.53}$$

The optimal algorithm for this so-called noise-free estimation has the same form as that for the white noise case except that $R = 0$ and different vectors and matrices are used. A new coordinate transforma-

tion is first introduced in order to reduce the dimension of the optimal filter and to guarantee the invertibility of $\boldsymbol{h}^T \boldsymbol{P}(n|n-1)\boldsymbol{h}$. A transformation matrix $\boldsymbol{T}$ is defined as:

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{h}^T \\ \boldsymbol{T}_2 \end{bmatrix} \tag{E 2.54}$$

where $\boldsymbol{T}_2$ can be any $(p+q-m) \times (p+q)$ matrix that yields nonsingular $\boldsymbol{T}$. The dimension of $x(n)$ is denoted by m. Here $\boldsymbol{T}_2 = \left[\boldsymbol{I}_{p+q-1}\ \boldsymbol{0}\right]$ is chosen so that the new state vector becomes identical with part of the original state variables of interest. By using this transformation,

$$\tilde{\boldsymbol{s}}(n) = \boldsymbol{T}\bar{\boldsymbol{s}}(n) \text{ and } \bar{\boldsymbol{s}}(n) = \boldsymbol{T}^{-1}\tilde{\boldsymbol{s}}(n) \tag{E 2.55}$$

and

$$\tilde{\boldsymbol{s}}(n) = \tilde{\boldsymbol{F}}\tilde{\boldsymbol{s}}(n-1) + \tilde{\boldsymbol{G}}\tilde{\boldsymbol{w}}(n) \tag{E 2.56}$$

$$x(n) = \tilde{\boldsymbol{h}}^T\tilde{\boldsymbol{s}}(n) \tag{E 2.57}$$

where

$$\tilde{\boldsymbol{F}} = \boldsymbol{T}\boldsymbol{F}\boldsymbol{T}^{-1} \text{ and } \tilde{\boldsymbol{G}} = \boldsymbol{T}\boldsymbol{G} \tag{E 2.58}$$

$$\tilde{\boldsymbol{h}}^T = \boldsymbol{h}^T\boldsymbol{T}^{-1} = \left[1\ 0\ \ldots\ 0\right]. \tag{E 2.59}$$

The optimal Kalman filter estimate becomes:

$$\hat{\boldsymbol{S}}(n) = \tilde{\boldsymbol{F}}\hat{\boldsymbol{S}}(n-1) + \boldsymbol{k}(n)\left[x(n) - \tilde{\boldsymbol{h}}^T\tilde{\boldsymbol{F}}\hat{\boldsymbol{S}}(n-1)\right]. \tag{E 2.60}$$

The gain and error covariance equations are computed according to:

$$\boldsymbol{k}(n) = \boldsymbol{P}(n|n-1)\tilde{\boldsymbol{h}}\left[\tilde{\boldsymbol{h}}^T\boldsymbol{P}(n|n-1)\tilde{\boldsymbol{h}}\right]^{-1} \tag{E 2.61}$$

$$\boldsymbol{P}(n|n-1) = \tilde{\boldsymbol{F}}\boldsymbol{P}(n-1)\tilde{\boldsymbol{F}}^T + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^T \tag{E 2.62}$$

$$\boldsymbol{P}(n) = \left[\boldsymbol{I} - \boldsymbol{k}(n)\tilde{\boldsymbol{h}}^T\right]\boldsymbol{P}(n|n-1) \tag{E 2.63}$$

Since the state estimate (Eq 2.60) has to be transformed back to the original form, the final estimate becomes:

$$\hat{\boldsymbol{s}}(n) = \left[\boldsymbol{h}^T\ \boldsymbol{0}\right]\boldsymbol{T}^{-1}\hat{\boldsymbol{S}}(n). \tag{E 2.64}$$

Parameter estimation

As mentioned earlier, the problem of parameter estimation is a crucial part of Kalman filtering and improper estimation results in distorted speech. Various estimation methods have been proposed:

In [Sor97], estimation of the $\boldsymbol{b}$ vector (the AR parameters of the noise) and σ_η^2 is performed during speech pauses using a voice activity detector. The autocorrelation function of the noise is computed using a 32 msec block length, then smoothed using an autoregressive scheme. The Levinson-Durbin algorithm is used to infer the parameter set.

In [Koo89], it is assumed that the $\boldsymbol{b}$ vector and σ_η^2 are known or are computed using an extra microphone. The $\boldsymbol{a}$ vector (the AR parameters of speech) and σ_w^2 are calculated iteratively. That is, estimates $\hat{\boldsymbol{a}}_1$ and σ_{w1}^2, are calculated directly from the noisy observations using the Durbin algorithm and then substituted into the appropriate Kalman filter. A new set of coefficients, $\hat{\boldsymbol{a}}_2$ and σ_{w2}^2, are calculated using the filtered output and plugged into the filter. This procedure is iterated until a final set $\hat{\boldsymbol{a}}_n$ and σ_{wn}^2, is obtained.

Experimental results reported in [Sor97] found that this approach leads to good parameter estimation during high SNR segments but performs poorly during low SNR frames and causes a distortion at the filter output.

In [Gan97], Higher Order Cumulants are used to estimate the speech parameters, by considering:

$$cum\,[x(t), x(t-l_1), \ldots, x(t-l_M)] = -\sum_{k=1}^{p} a_k cum\,[x(t), x(t-l_1), \ldots, x(t-l_M)]$$

whenever $M \geq 2$. The cumulants are approximated by substituting the unavailable ensemble averages with sample averages, thus obtaining a set of linear equations that may be used to compute the AR parameters directly from the observed signal. Experimental results found that the method is effective at low SNR conditions (5 dB) when 4^{th}-order cumulants are used. The parameter estimation was superior to using 2^{nd}-order statistics and it was also found that the use of third order statistics is limited in effectiveness.

2.2.4 Comb Filtering

Since the most important audible component of speech is periodic, its harmonic frequencies may be identified for the purpose of either preservation or suppression. One basic method involves comb filtering, in which a dynamic filter is designed to *comb* through the spectrum, modifying energy at equally spaced frequencies to attenuate or enhance them. The frequency response of the filter resembles a *comb*, with large values at a specified F_O and its multiples, and low values between these harmonics.

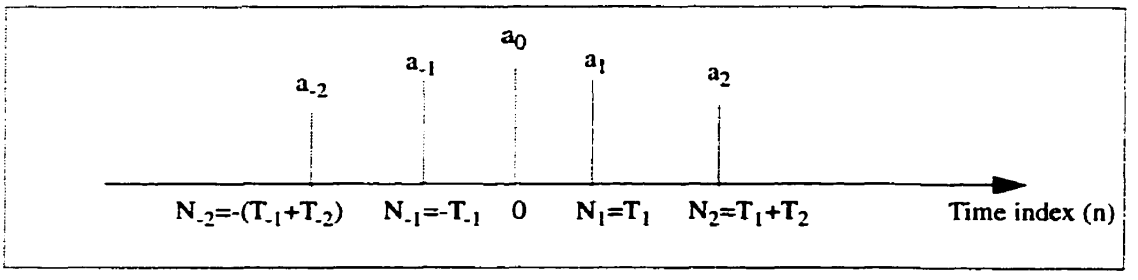
2.2.4.1 Comb filters to reinforce F_0

The operation of the filter can be explained by considering the sample response over one pitch period [Lim78]:

$$h(n) = \sum_{k=-L}^{L} a_k \cdot \delta(n - N_k) \qquad \text{(E 2.65)}$$

where $h(n)$ is the unit sample response, $\delta(n)$ is a unit sample function, a_k are the filter coefficients, the length of the filter is $(2L + 1)$ pitch periods, and the pitch period estimate N_k is computed based on the pitch information of the speech frame. The filter coefficients are unchanged and only N_k is updated once every pitch period based on the pitch information T_k of the speech waveform being processed. A typical impulse response for the case of $L = 2$ is illustrated in Figure 2-4.

Figure 2-4 Impulse response of a comb filter (L=2)



To the extent that the speech waveform is periodic over the $2L + 1$ pitch periods that the filter is applied, the speech samples will add constructively while the noise samples sum to zero. The operation depends on an accurate estimate of the desired signal's period, and its performance is best when this signal has stationary traits. F_O estimation is not always easy and speech signals very often change from one period to the next, in terms of harmonics and spectral envelope.

In the case of a spectral change but constant F_O during the comb's window, the filter spreads the change out over the duration of the window. Thus a rapid change could be smeared in time.

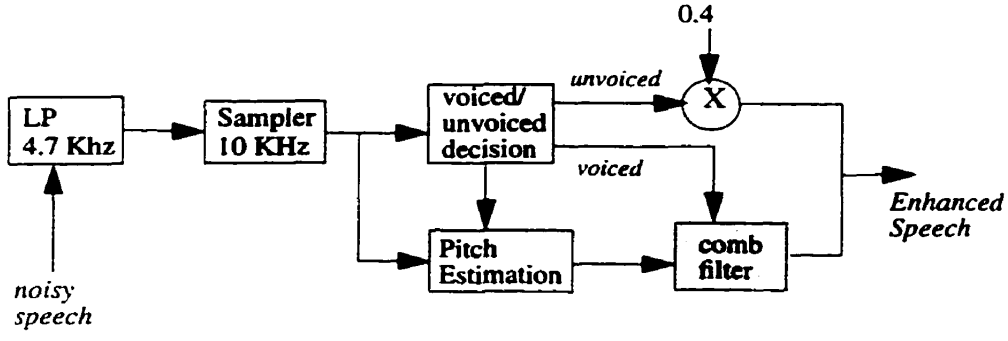
A more difficult problem arises when F_O changes during the course of the window, which may lead to a reduction in the reinforcement of the periods in $y(n)$. Thus comb filters work best only during sections of speech where F_O is not changing rapidly. This problem can be minimize by choosing $L=1$ (i.e. averaging only 3 periods). However, the degree of signal reinforcement is proportional to L. Spectrally, L is inversely proportional to the bandwidth of the comb harmonics. Larger values of L lead to narrow harmonics in the comb filter response, which more effectively suppress energy outside the corresponding harmonics in $x(n)$.

The coefficient set is often chosen from a Hamming-like window where higher weight is given to the middle segment and less weight for the side segment. The coefficient set is normalized so that the sum of the coefficients is unity. For a normalized Hamming window set, each a_k is given by:

$$a_k = \frac{0.54 + 0.46\cos\left(2\pi\frac{k}{2L} + 1\right)}{\sum_{k=-L}^{L} 0.54 + 0.46\cos\left(2\pi\frac{k}{2L} + 1\right)}. \qquad \text{(E 2.66)}$$

Comb filtering can only be applied to the voiced segments of the speech, thus requiring a voiced/unvoiced detector, in addition to a pitch estimator. Figure 2-5 shows the system diagram of the comb filter used in [Lim78].

Figure 2-5 Block diagram of Lim's comb filtering



The results in [Lim78] indicated that even with perfect estimates of the fundamental frequency, the adaptive comb filter does not achieve a significant increase in the intelligibility at any S/N ratio, when the degrading source is additive white noise. A substantial decrease of intelligibility was observed when the length of the filter was between 7 and 13 pitch periods. At a length of 3 periods, intelligibility did not show a noticeable decrease. On the other hand, noticeable increases in S/N ratio were achieved. When the filter length was 3, 7, and 13 pitch periods, the increases were 3.5 dB, 7 dB and 10 dB respectively. Thus, the adaptive comb filter can be useful for the objective of noise reduction without significant decrease in speech intelligibility.

2.2.4.1.1 Variations of the basic method

Some noteworthy variations are proposed to the basic comb filter in order to deal with smearing spectral changes and segment discontinuity.

Adaptive coefficient set

In [Vee89], the coefficient set is chosen to adapt to the data, as opposed to being constant. The argument is that pitch periods may not be similar and thus choosing a constant set does not account for such transitions between periods. The proposed set is derived by minimizing a prediction error, namely the coefficients a_k are chosen such that the samples that are k periods away best predict the current sample. The prediction error may be written as (considering 3 coefficients):

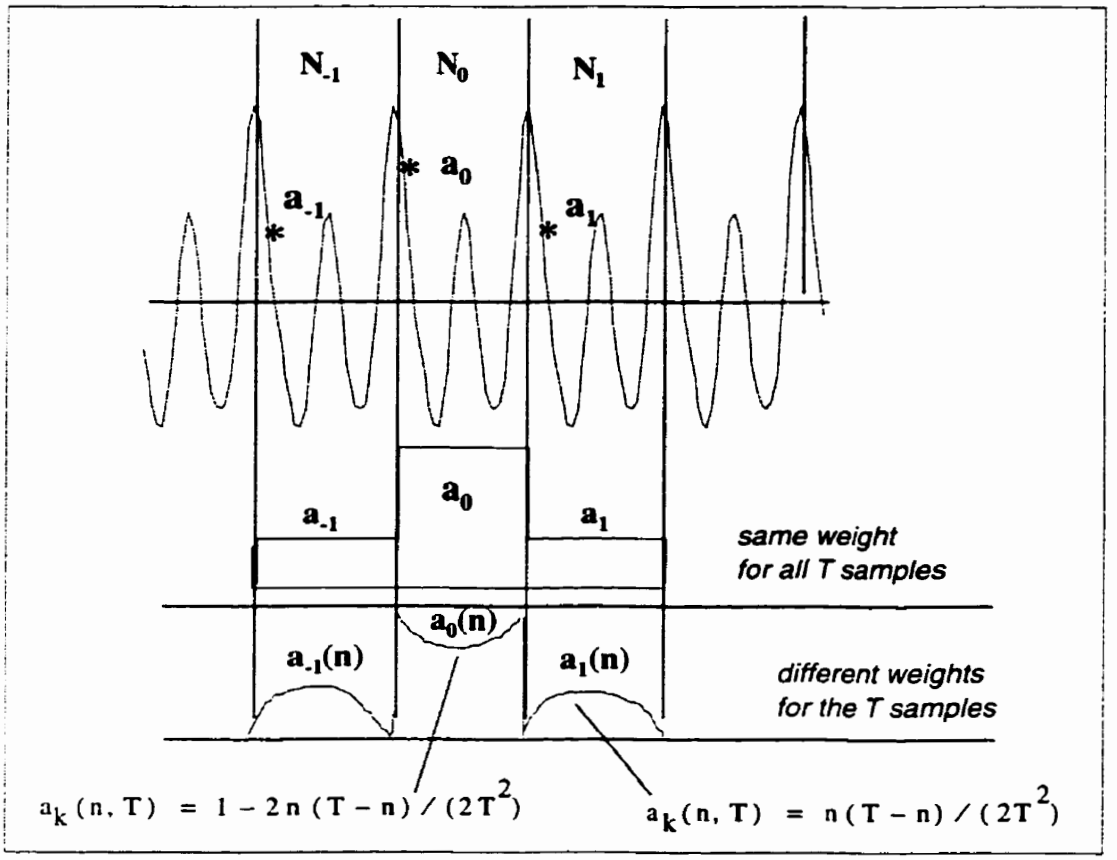
$$Error = \sum_{n} [a_0 x(n) - a_{-1} x(n-T) - a_1 x(n+T)]^2. \qquad \text{(E 2.67)}$$

The summation is taken over the samples of a pitch period. For simplicity a_0 is set to unity. In addition to accounting for the changing characteristics between periods, a variable coefficient set can compensate for any pitch estimation errors, thus making the results less dependent on the precision of the pitch estimate. This is a valid argument since accurate pitch estimation is difficult in noisy conditions.

Different weighting for the samples in a given period

Instead of using the same weight for all the samples in a given period, the method in [Cox81] and [Mal82] uses a class of windows that has a variable weight in each pitch period (Figure 2-6). The rationale is that when comb filtering is done in a pitch synchronous way, segment discontinuity results when the pitch period changes. This class of windows is claimed to minimize this effect.

Figure 2-6 Variable weighting for the samples in each pitch period



Dynamic Time Warping

To better utilize the information contained within voiced speech and account for imperfect periodicity, the time scale of each individual pitch period is warped into the current period being processed. Thus, given an N-length sequence $R[n]$ being processed and an M-length sequence $T[m]$ representing one of the contiguous past or future periods contained within a segment of voiced speech, T[m] is warped into $R[n]$ using an optimal warping function of the form $m = w[n]$ such that a total distance measure D_T is minimized:

$$D_{T_{min}} = \min_{w[n]} \sum_{n=1}^{N} \| (R[n], T[w[n]]) \| .$$

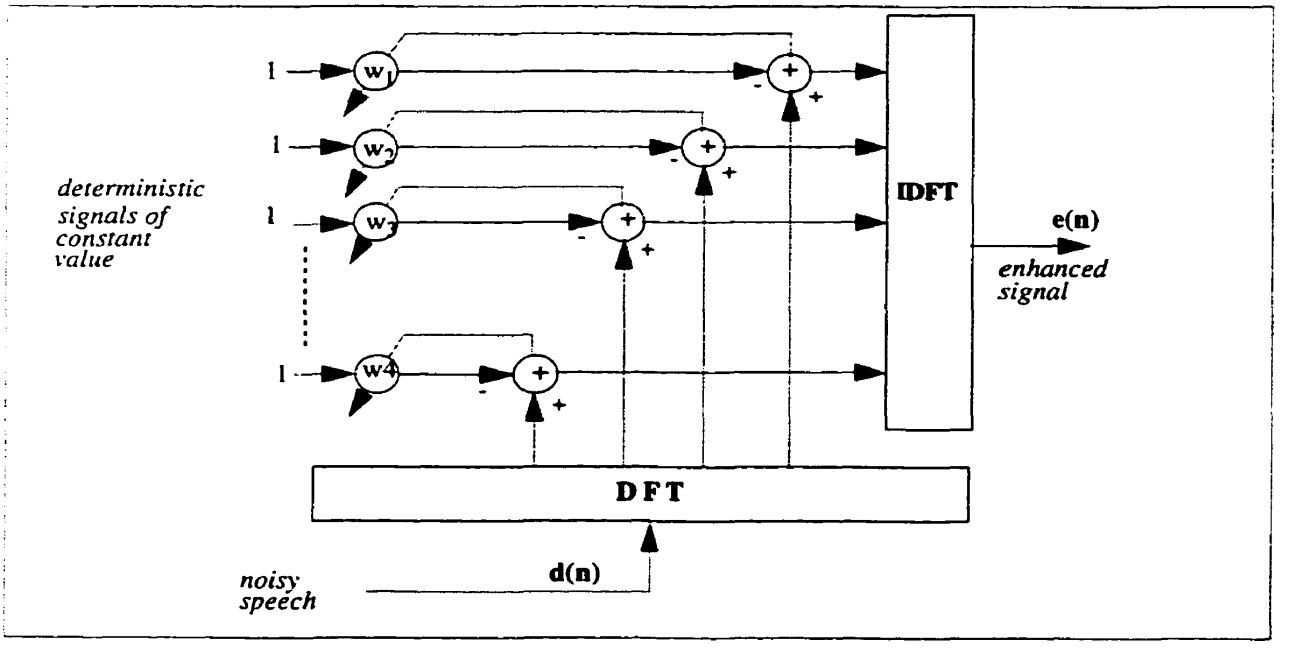
The optimal path is selected as the path which produces the lowest total accumulated error. This requires that all the possible paths through the grid (the (m,n) plane) be computed, thus requiring a

computational load proportional to *NM*. A suboptimal but more manageable approach is proposed in [Gra93] which finds a simple path through the grid. Instead of looking backward through the grid to determine the optimal path on the basis of the total accumulated distance, this suboptimal approach finds a simple path by looking forward through the grid in such a way that the accumulated distance at the point (n,m) between the two consecutive possible points is a local minimum. Thus the next point is chosen on the path on the basis of which point adds the least to the accumulated distance at the grid point (n,m).

2.2.4.2 Comb filters to cancel interfering harmonics

In the harmonic canceller system proposed by [Ami91], a comb filter is used to *eliminate* interfering sinusoids of harmonically related frequencies. This could be used, for example, to reduce the noise from an interfering speaker, provided this speaker's fundamental frequency is known. The filter is implemented as a special case of a frequency domain LMS algorithm. Deterministic signals of constant values are applied to the reference inputs (Figure 2-7) instead of observed values, and the measured signal is applied to the desired signal input $d(t)$.

Figure 2-7 The LMS harmonic canceller



The concept behind this is the fact that if the Fourier transform block length N is chosen to be the fundamental period of the undesired signal, then the harmonic signal transform at each frequency bin will take the same value, regardless of the time reference. The non-harmonic components transforms have a time-varying behavior. To distinguish the two, unit values are applied at the reference inputs. The LMS thus operates on cancelling all harmonic frequency components and leaving out other components to propagate to the error transform with minimum distortion.

2.2.5 Wavelet-based Denoising

In recent years, there has been a renewed interest in the design of structured bases for the linear expansion of signals. In particular, the subject of wavelets and time-scale analysis has received increased interest as a new method of expanding functions onto a set of self-similar orthonormal basis functions. This is largely due to the fact that such techniques offer increased flexibility over more traditional transform methods, combined with the existence of efficient computational structures, in the form of multirate filter banks, which allow rapid computations of these coefficients.

Wavelets provide a tool for non-linear filtering of signals contaminated by noise. In [Mall92], it is shown that effective noise suppression may be achieved by transforming the noisy signal into the wavelet domain, and preserving only the local maxima of the transform. Alternatively, a reconstruction that uses only the large-magnitude coefficients has been shown to approximate well the uncorrupted signal. In other words, noise suppression is achieved by thresholding the wavelet transform of the contaminated signal.

To choose the appropriate threshold, [Don95] has taken a minimax approach to characterizing the signal (rather than the disturbance, which is assumed Gaussian). A threshold is derived that is approximately minimax in the sense that its sample size dependence is of the same order as that of the true minimax. A coefficient C_i is excluded from the reconstruction if $|C_i| \leq \sigma\sqrt{2\log(K)}$ where σ is the standard deviation of the noise, and K is the length of the observation.

The above schemes have relied on the assumption of the normality of the noise and are therefore sensitive to outliers, i.e., to noise distributions whose tails are heavier than the Gaussian distribution. A scheme is proposed in [Sch97] to circumvent this. Essentially, a minimax Description Length

(MMDL) principle is used as the criterion of choice for thresholding wavelet coefficients. Furthermore, the true signal is assumed to have a bounded amplitude, and a thresholding technique from above and below is used to result in bounded estimation errors. The scheme is claimed to be less sensitive to heavy-tailed noises.

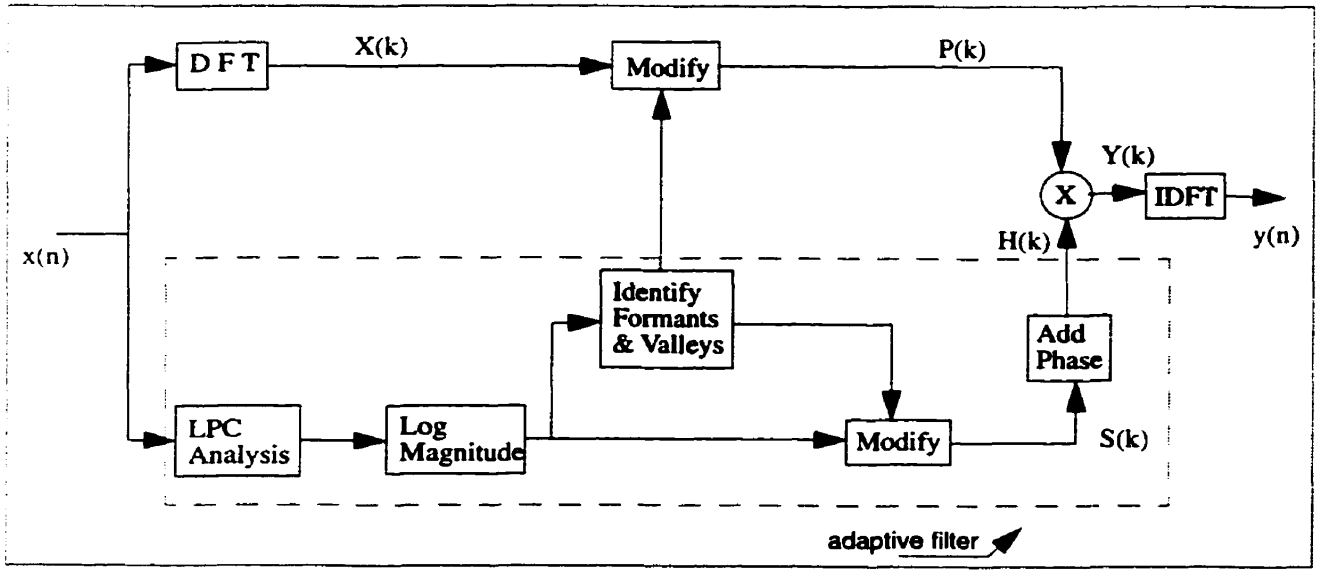
2.2.6 Psychoacoustic-based Methods

Formant-based Spectral Shaping

The frequency domain postfiltering approach proposed in [Wan93] consists of approximating the noisy speech spectrum by LPC analysis and then modifying this spectrum (through frequency filtering) so that the spectral *formants* are *sharpened* and the valleys deepened.

The filter is 'adapted' for every new frame of speech (128 samples) and is represented by a set of DFT coefficients *H(k)* (Figure 2-8). These coefficients are multiplied by a modified version of the speech transform *P(k)* to yield the transform of the enhanced speech. The core operation of the algorithm is to scan the LPC spectrum and detect the 3 or 4 main formants (and valleys) through *peak picking*.

Figure 2-8 Adaptive filtering by spectral modification



Exploiting the Characteristics of Human Hearing

These methods exploit frequency masking or other properties of the auditory system to gain an advantage in locating or attenuating noise bands or in preventing unnecessary attenuation in speech bands. In [Vir95], a cross-band masking threshold is used to control the spectral subtraction process. The idea is to find the best compromise between noise reduction and speech distortion in a perceptual sense. By identifying the bands where natural noise masking occurs, there is no need for an aggressive noise suppression. A similar idea is proposed in [Tso93] where a masking threshold is computed and used to identify the audible noise spectrum. A non-linear suppression rule is then applied where suppression is controlled by a parameter that is a function of the masking threshold identified.

The algorithm proposed in [Che91] uses the property of lateral inhibition of the auditory system. It convolves the so-called function of spatial lateral inhibition with the power spectrum of the noisy speech input to yield an estimate of the clean speech spectrum without apriori knowledge of the noise spectrum.

2.3 The Use of HOS for Speech Analysis

Wells was among the first to use the bispectrum in speech application. His often quoted work [Wel85] used the bispectrum as a voiced/unvoiced detector. The underlying concept is the fact that unvoiced speech has Gaussian characteristics and thus its bispectrum is almost zero (when proper averaging is used), whereas voiced speech has a non-zero bispectrum. The results of that study supported the model that unvoiced speech is produced by Gaussian-like excitation. The work outlines some interesting structural characteristics of the bispectrum for voiced phonemes and suggests that speaker characteristics may be extracted from it, but the results are experimental with no analytical foundation. Similarly, it is reported in [Fal93] that the normalized skewness and kurtosis of short-term speech segments may be used to detect transitional speech events (termed "innovation"). The conclusion thus drawn is based on experimental observations using various utterances and window sizes.

In [Ran95] a method based on Gaussianity tests for the bispectrum and the triple correlation is used to discriminate voiced and unvoiced segments. The method exploits the Gaussian blindness of HOS but not the peculiarities of the HOS of voiced speech to better classify the segments. In [Mor92] a pitch estimation method based on the periodicity of the diagonal slice of the 3^{rd}-order cumulant is described

and leads to more reliable estimation than the autocorrelation of the underlying speech. The method involves computing the autocorrelation of the cumulant slice and as such requires a large number of data points to provide reliable estimation. In addition, the claim of the cumulant slice having similar periodicity as the underlying speech is not clearly demonstrated.

The exploratory work of Fackrell has addressed some of the fundamentals of the HOS of speech signals, particularly with regard to detecting and estimating the non-linearity in the speech production system. In [Fac94], it is argued that the normalized bispectrum (bicoherence) may be used to detect quadrature phase coupling in speech since the bicoherence at a bifrequency (f_1,f_2) includes the contribution of the Fourier value at the sum of these two frequencies. However, the more elaborate work reported in [Fac96] found that the use of the bispectrum is not conclusive, even negative, about the detection of such phenomena in speech.

Seetharaman & Jernigan were likely the first to propose speech enhancement using higher order spectra averaging and speech reconstruction. Their paper [See88] however did not present any result but gave preliminary ideas about magnitude and phase reconstructions.

Fulchiero & Spanias used the least square approach proposed in [Sun90] in order to reconstruct the magnitude of the Fourier transform of a speech signal corrupted by colored Gaussian noise and thus achieve speech enhancement with third-order statistics, based on the idea in [See88]. In their setup [Ful93], colored noise is used and the bispectrums over three consecutive segments are averaged to approximate the statistical averaging of the bispectrum. The authors reported an SNR improvement of about 1 to 2 dB when the input SNR was below 6 dB. The same results were obtained when white noise was used instead of colored. Finally, the enhancer was more effective for blocks of voiced speech, where the enhancement attained 2.7 dB, than in unvoiced speech, that was typically enhanced by 1 to 1.5 dB. This system was tried as a starting point in this thesis, but was found very limited in its effectiveness, given the high computational cost, and other problems associated with bispectrum phase unwrapping [Mat84], [Mar90].

In [Sal94], a speech enhancement based on iterative Wiener filtering is proposed. The AR Spectral estimation is carried out using 3^{rd} and 4^{th} order cumulants. It was found that using the 3^{rd}-order cumulant was better than the 4^{th} and that a good compromise may be achieved among convergence speed, distortion effect and computational complexity.

CHAPTER 3

Higher Order Statistics: Definition and General Derivations

Synopsis

The first part of this chapter provides a brief background on Higher Order Statistics (HOS), highlighting the time and frequency domain definitions, properties, and naming convention and notations used. The second part consists of a series of new derivations and findings that relate second and higher order statistics of real signals.

3.1 Definitions and Notation

3.1.1 Time Domain Definitions

If $x(n)$, $n = 0, \pm1, \pm2, \pm3, \ldots$ is a real stationary discrete-time signal and its moments up to order p exist, then [Nik93]:

$$m_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) \equiv E[x(n)\, x(n+\tau_1) \ldots x(n+\tau_{p-1})] \qquad \text{(E 3.1)}$$

represents the p^{th}-order moment function of the stationary signal, which depends only on the time differences $\tau_1, \tau_2, \ldots, \tau_{p-1}$, $\tau_i = 0, \pm1, \pm2, \ldots$ for all i. Here $E[.]$ denotes statistical expectation and for a deterministic signal, it is replaced by a time summation over all time samples (for energy signals) or time averaging (for power signals):

$$m_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) \equiv \sum_{n=-\infty}^{\infty} \{x(n)\, x(n+\tau_1) \ldots x(n+\tau_{p-1})\} \qquad \text{(E 3.2)}$$

for $p = 3$ and 4, the cumulant function of a non-Gaussian signal $x(n)$ can also be written as:

$$C_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) = m_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) - m^G{}_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) \quad \text{(E 3.3)}$$

where $m_p(\tau_1, \tau_2, \ldots, \tau_{p-1})$ is the p^{th}-order moment function of $x(n)$ and $m^G{}_p(\tau_1, \tau_2, \ldots, \tau_{p-1})$ is the p^{th}-order moment of an equivalent Gaussian process that has the same mean and autocorrelation sequence as $x(k)$. Clearly if $x(n)$ is Gaussian, then all its cumulants higher than two are zero. The following are the relations between the moments and cumulants for $p = 1, 2, 3$, and 4.

- **1st-order cumulant (mean value):**

$$C_1 = m_1 = E\{x(k)\} \quad \text{(E 3.4)}$$

- **2nd-order cumulant (covariance sequence):**

$$C_2(\tau) = m_2(\tau) - (m_1)^2 = C_2(-\tau) \quad \text{(E 3.5)}$$

- **3rd-order cumulant:**

$$C_3(\tau_1, \tau_2) = m_3(\tau_1, \tau_2) - m_1[m_2(\tau_1) + m_2(\tau_2) + m_2(\tau_1 - \tau_2)] + 2(m_1)^3 \quad \text{(E 3.6)}$$

- **4th-order cumulant:**

$$\begin{aligned} C_4(\tau_1, \tau_2, \tau_3) &= m_4(\tau_1, \tau_2, \tau_3) \\ &\quad - m_2(\tau_1) \cdot m_2(\tau_3 - \tau_2) - m_2(\tau_2) \cdot m_2(\tau_3 - \tau_1) - m_2(\tau_3) \cdot m_2(\tau_2 - \tau_1) \\ &\quad - m_1[m_3(\tau_2 - \tau_1, \tau_3 - \tau_2) + m_3(\tau_2, \tau_3) + m_3(\tau_2, \tau_4) + m_3(\tau_1, \tau_2)] \\ &\quad - (m_1)^2[m_2(\tau_1) + m_2(\tau_2) + m_2(\tau_3) + m_2(\tau_3 - \tau_1) + m_2(\tau_3 - \tau_2) + m_2(\tau_2 - \tau_1)] \\ &\quad - 6(m_1)^4 \end{aligned} \quad \text{(E 3.7)}$$

3.1.1.1 Zero-mean signals

If the signal $x(n)$ is zero-mean, i.e. $m_1 = 0$, it follows that:

- The 2nd and 3rd order *cumulants* are identical to the 2nd and 3rd order *moments:*

$$C_2(\tau) = m_2(\tau) \quad \text{and} \quad C_3(\tau_1, \tau_2) = m_3(\tau_1, \tau_2) \quad \text{(E 3.8)}$$

- The 4th-order cumulant is expressed in terms of the 2nd and 4th order moments as:

$$\begin{aligned} C_4(\tau_1, \tau_2, \tau_3) = \; & m_4(\tau_1, \tau_2, \tau_3) - m_2(\tau_1) \cdot m_2(\tau_3 - \tau_2) \\ & - m_2(\tau_2) \cdot m_2(\tau_3 - \tau_1) - m_2(\tau_3) \cdot m_2(\tau_2 - \tau_1) \end{aligned} \quad \text{(E 3.9)}$$

Zero-lag cumulant

By setting all the lags to zero in the above cumulant expressions, the following statistics are obtained:

$$\textbf{Variance: } \gamma_2 = C_2(0) = E\{x^2(k)\} \qquad \textbf{(E 3.10)}$$

$$\textbf{Skewness: } C_3(0,0) = E\{x^3(k)\} \qquad \textbf{(E 3.11)}$$

$$\textbf{Kurtosis: } C_4(0,0,0) = E\{x^4(k)\} - 3\Big(E\{x^2(k)\}\Big)^2. \qquad \textbf{(E 3.12)}$$

When estimating higher-order statistics from finite data records, the variance of the estimators is reduced by normalizing the input data to have a variance of 1, prior to computing the estimators. Equivalently, the 3rd and 4th order statistics may be normalized by the appropriate powers of the data variance, thus the following entities are defined:

$$\textbf{Normalized Skewness: } \gamma_3 \equiv \frac{C_3(0,0)}{[C_2(0)]^{1.5}} = \frac{E[x^3(n)]}{\Big[E\{x^2(n)\}\Big]^{1.5}} \qquad \textbf{(E 3.13)}$$

$$\textbf{Normalized Kurtosis: } \gamma_4 \equiv \frac{C_4(0,0,0)}{[C_2(0)]^2} = \frac{E[x^4(n)]}{\Big[E\{x^2(n)\}\Big]^2} - 3.0. \qquad \textbf{(E 3.14)}$$

3.1.1.2 Cumulant slices

Since the 3rd and 4th order cumulants are multi-dimensional functions, it is customary to use only 2D slices of these, by freezing some of the lags in Eq 3.8 and Eq 3.9.

The horizontal slice of the 3rd-order cumulant is defined as:

$$\boxed{C_{3b}[\tau] \equiv C_3(-b,-\tau) = m_3(b,\tau) = E[x(n)\,x(n-b)\,x(n-\tau)]} \ \textbf{b}^{\textbf{th}}\textbf{ Horizontal Slice.} \qquad \textbf{(E 3.15)}$$

For the 4th- order cumulant, two different slices are used:

The first is obtained by setting all three lags equal ($\tau_1 = \tau_2 = \tau_3 = \tau$):

$$C^a_4[\tau] \equiv C_4(\tau,\tau,\tau) = m_4(\tau,\tau,\tau) - 3m_2(0)\cdot m_2(\tau)$$

$$\boxed{C^a_4[\tau] = E[x(n)\,x^3(n+\tau)] - 3\cdot E[x^2(n)]\cdot E[x(n)\,x(n+\tau)]}\ \textbf{Diagonal Slice.} \qquad \textbf{(E 3.16)}$$

The second slice is obtained by setting $\tau_1 = 0$ and $\tau_2 = \tau_3 = \tau$:

$$C^b_4[\tau] \equiv C_4(0,\tau,\tau) = m_4(0,\tau,\tau) - [m_2(0)]^2 - 2[m_2(\tau)]^2$$

$$C^{h}_{4}[\tau] = E[x^2(n)\,x^2(n+\tau)] - \{E[x^2(n)]\}^2 - 2\{E[x(n)\,x(n+\tau)]\}^2$$ **Horizontal Slice.** (E 3.17)

3.1.1.3 Properties of cumulants

The following are important properties that any *pth*-order cumulants satisfy [Men91]:

1. *Scaled quantities*: The cumulant of scaled quantities equals the product of all the scale factors times the cumulant of the unscaled quantities, i.e., if λ_i, $i = 1, 2, \ldots, p$ are constants and x_i, $i = 1, 2, \ldots, p$ are random variables, then:

$$cum(\lambda_1 x_1, \lambda_2 x_2, \ldots, \lambda_p x_p) = \left\{\prod_{i=1}^{p} \lambda_i\right\} cum(x_1, x_2, \ldots, x_p) . \quad \text{(E 3.18)}$$

2. *Symmetry*: Cumulants are symmetric in their arguments, i.e.,

$$cum(x_1, x_2, \ldots, x_p) = cum(x_{i_1}, x_{i_2}, \ldots, x_{i_p}) \quad \text{(E 3.19)}$$

where $(i_1,\ldots,i_p)$ is a permutation of $(1,\ldots,p)$; interchanging the arguments of the cumulant in any way does not change its value, e.g.: $C_4(\tau_1, \tau_2, \tau_3) = C_4(\tau_3, \tau_1, \tau_2) = C_4(\tau_2, \tau_3, \tau_1)$.

3. *Additiveness*: Cumulants are additive in their arguments, that is the cumulants of sums equal sums of cumulants. For example, even if x_0 and y_0 are not statistically independent, it is true that:

$$cum(x_0 + y_0, z_1, \ldots, z_p) = cum(x_0, z_1, \ldots, z_p) + cum(y_0, z_1, \ldots, z_p) . \quad \text{(E 3.20)}$$

4. *Additive constants*: Cumulants are insensitive to additive constants, that is, for α constant:

$$cum(\alpha + z_1, \ldots, z_p) = cum(z_1, \ldots, z_p) . \quad \text{(E 3.21)}$$

5. *Sums*: The cumulant of a sum of statistically independent quantities equals the sum of the cumulants of the individual quantities, i.e., if the random variables $[x_i]$ are independent of the random variables $[y_i]$ for $i = 1, 2, \ldots, p$ then:

$$cum(x_1 + y_1, x_2 + y_2, \ldots, x_p + y_p) = cum(x_1, x_2, \ldots, x_p) + cum(y_1, y_2, \ldots, y_p) . \quad \text{(E 3.22)}$$

Note that if $x_i, \ldots y_i$ were not independent, then from (Eq 3.20) there would be $2p$ terms on the right-hand side. Statistical independence reduces these terms to just 2.

6. *Independent subsets*: If a subset of the random variables is independent from the rest, then

$$cum(x_1, x_2, \ldots, x_p) = 0 . \quad \text{(E 3.23)}$$

3.1.2 Frequency Domain Definitions

Higher-Order spectra are multi-dimensional Fourier transforms of higher-order cumulants:

$$S_p(w_1, w_2, \ldots, w_{p-1}) = \sum_{\tau_1=-\infty}^{\infty} \cdots \sum_{\tau_{p-1}=-\infty}^{\infty} C_p(\tau_1, \tau_2, \ldots, \tau_{p-1}) \exp\left[-j\sum_{i=1}^{p-1} w_i\tau_i\right] \quad \text{(E 3.24)}$$

- **Power Spectrum: p = 2**

$$S_2(w_1) = \sum_{\tau_1=-\infty}^{\infty} C_2(\tau_1) \exp[-jw_1\tau_1] \quad \text{(E 3.25)}$$

For a deterministic signal $x(n)$, the power spectrum can be expressed in terms of the Fourier transform of the underlying signal as: $S_2(w) = X^*(f)\,X(f)$.

- **Bispectrum: p = 3**

The bispectrum is the 2D-Fourier transform of the third cumulant function:

$$S_3(w_1, w_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} C_3(\tau_1, \tau_2)\, e^{-j\left(w_1\tau_1 + w_2\tau_2\right)}$$

for $|w_1| \le \pi$, $|w_2| \le \pi$, and $|w_1 + w_2| \le \pi$. For a deterministic, zero-DC signal the bispectrum may be expressed in terms of the Fourier transform of the underlying signal since:

$$S_3(w_1, w_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x(n)\,x(n+\tau_1)\,x(n+\tau_2)\, e^{-j\left(w_1\tau_1 + w_2\tau_2\right)}$$

setting $n + \tau_1 = m$ and $n + \tau_2 = k$ and splitting the exponent yields:

$$S_3(w_1, w_2) = \left\{\sum_{m=-\infty}^{\infty} x(m)\, e^{-jw_1 m}\right\}\left\{\sum_{m=-\infty}^{\infty} x(k)\, e^{-jw_2 k}\right\}\left\{\sum_{n=-\infty}^{\infty} x(n)\, e^{j(w_1+w_2)n}\right\}$$

$$= X(w_1)\,X(w_2)\,X^*(w_1 + w_2)$$

and since $x(n)$ is real,

$$S_3(w_1, w_2) = X(w_1)\,X(w_2)\,X(-w_1 - w_2) . \quad \text{(E 3.26)}$$

The symmetry conditions of $S_3(w_1, w_2)$ follow from those of the third cumulant, namely:

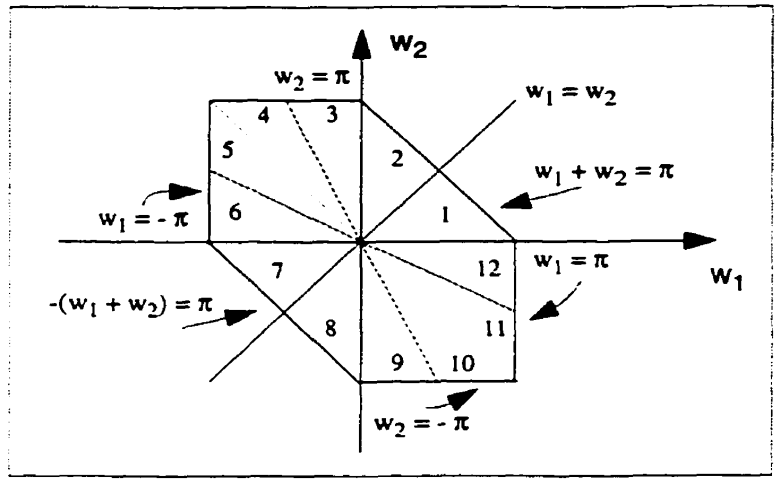
$$S_3(w_1, w_2) = S_3(w_2, w_1)$$

$$= S_3^*(-w_2, -w_1) = S_3^*(-w_1, -w_2)$$

$$= S_3(-w_1 - w_2, w_2) = S_3(w_1, -w_1 - w_2)$$
$$= S_3(-w_1 - w_2, w_1) = S_3(w_2, -w_1 - w_2).$$

Thus, knowledge of the bispectrum in the triangular region $\{ w_2 \geq 0,\ w_2 \geq w_1,\ w_1 + w_2 \leq \pi \}$ is sufficient to describe the rest (Figure 3-1). This region (labeled 1) is often termed the principal region of the bispectrum.

Figure 3-1 Symmetry regions of the Bispectrum



3.2 Third-Order Derivations

3.2.1 Fourier Transform of Cumulant Slices

The horizontal cumulant slice is formed as in Eq 3.15. Thus:

$$C_{3b}[\tau] = \sum_{n=-\infty}^{\infty} x(n)\,x(n-b)\,x(n-\tau) \qquad -\infty \le \tau \le \infty$$

3.2.1.1 Infinite data

• **Theorem 1:** *For a real deterministic signal, the discrete-time Fourier transform of any horizontal slice* $C_{3b}[\tau]$ *may be expressed in terms of the Fourier transform of the underlying signal as:*

$$\boxed{FC_b(w) = \{X(w) \otimes e^{-jbw} X(w)\}\,X(-w)} \qquad \text{(E 3.27)}$$

(It is worth noting that a similar, though not identical, derivation is found in [Men91] for the case of the output of a linear system that is driven by a white process. The similarity is in the sense that both derivations involve a convolution in the Fourier spectrum domain and a complex exponential term).

• **Proof:** As a starting point, the case of $b = 0$ is considered. The Fourier transform of this slice is:

$$FC_0(w) = \sum_{\tau=-\infty}^{\infty}\left(\sum_{n=-\infty}^{\infty} x^2(n)\,x(n-\tau)\right)e^{-j\tau w} = \sum_{n=-\infty}^{\infty} x^2(n)\left\{\sum_{\tau=-\infty}^{\infty} x(n-\tau)\right\}e^{-j\tau w}.$$

The exponent is split by setting n- $\tau = m$

$$FC_0(w) = \left\{\sum_{n=-\infty}^{\infty} x^2(n)\,e^{-jnw}\right\}\left\{\sum_{m=-\infty}^{\infty} x(m)\,e^{jmw}\right\}.$$

The first term is the Fourier transform of $x^2(n)$. The second is the conjugate of the Fourier transform of $x(n)$. Using the convolution property yields:

$$FC_0(w) = \{X(w) \otimes X(w)\}\,X(-w)$$

where the auto-convolution of the spectrum is:

$$X(w) \otimes X(w) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\lambda)\,X(w-\lambda)\,d\lambda.$$

Using the same reasoning and the time shifting property, the DTFT of any slice $C_{3b}[\tau]$ is the one given in Eq 3.27.

3.2.1.2 Finite data length

- **Proposition 1.** *When finite data records are used, the expression for the Discrete Fourier transform of a cumulant slice has a similar form as the DTFT in the infinite data case, provided that:*
  1. *The summation limit in the cumulant expression is constant and not a function of the lag.*
  2. *A sufficient number of lags is computed to cover at least one signal period.*

- **Proof:**

1. ***Computation of cumulant slices from finite data***

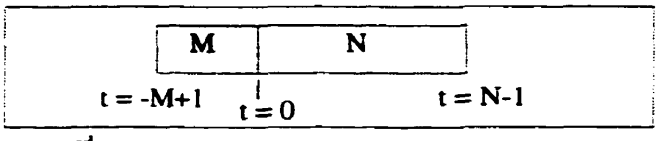
Given an N-length frame of data, the standard estimator [Nik93] for the 3rd-order cumulant is:

$$C_3[\tau_1, \tau_2] = \frac{1}{N}\sum_{n=S_1}^{S_2} x(n)\,x(n+\tau_1)\,x(n+\tau_2)$$

with $S_1 = max(0, -\tau_1, -\tau_2)$ and $S_2 = min(N-1, N-1-\tau_1, N-1-\tau_2)$.

In [Rui95], it is shown that for harmonic signals, this bias estimator generates additional vibrating terms due to the fact that S_1 and S_2 are functions of the lags τ_1 and τ_2. When the underlying signal consists of L damped exponentials, the resulting cumulant has a harmonic content that is different from that of the original signal. The problem is shown to be remedied by setting the limits of the sum to be *constant* and considering only 1D slices of the form: $C[\tau, a\tau + b]$, where a and b are constants. In this thesis, horizontal slices ($a = 0$) are considered and computed using overlapped frames. Every iteration, N new points are read and combined with the last M points of the previous frame (Figure 3-2).

Figure 3-2 Overlapping scheme between consecutive frames



B horizontal slices of the 3rd-order cumulant are computed, with each b slice defined as:

$$C_{3b}[\tau] = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\,x(n-b)\,x(n-\tau) \qquad \text{(E 3.28)}$$

for $b = 0, 1, 2, \ldots B \leq M-1$, and $\tau = 0, 1, 2, \ldots M-1$. The limit of the summation is thus constant for any lag τ.

2. *Transform of horizontal slices*

It is assumed that B horizontal slices of the 3rd-order cumulant are computed using the modified form (Eq 3.28); moreover, both M and N are assumed greater or equal to one period of the signal. Using the discrete frequency notation k, consider the DFT of $C_{30}[\tau]$:

$$FC_0(k) = \frac{1}{M}\sum_{\tau=0}^{M-1} C_{30}[\tau]\, e^{-j2\pi k\tau/N} = \frac{1}{M}\sum_{\tau=0}^{M-1}\left\{\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)\, x(n-\tau)\right\} e^{-j2\pi kn\tau/N}.$$

The order of summation for the indices τ and n is reversed, and the range of summation for the inner function, $x(n - \tau)$, is changed due to its shifted nature:

$$FC_0(k) = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n)\left\{\frac{1}{M}\sum_{\tau=n}^{n+M-1} x(n-\tau)\right\} e^{-j2\pi kn\tau/N}.$$

The exponent is split by setting n- $\tau = m$:

$$FC_0(k) = \left\{\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)\, e^{-j2\pi kn/N}\right\}\left\{\frac{1}{M}\sum_{m=0}^{M-1} x(m)\, e^{j2\pi km/N}\right\}.$$

Using the convolution property yields:

$$FC_0(k) = \{X(k) \otimes X(k)\}\, X(-k)\,. \qquad \textbf{(E 3.29)}$$

The result is thus similar to the case of infinite data length. The first term is the discrete convolution of the Fourier spectrum of $x(n)$ and is defined as:

$$G(k) \equiv X(k) \otimes X(k) = \sum_{\lambda=0}^{N-1} X(\lambda)\, X(k-\lambda)\,.$$

Using the same reasoning and the time shifting property, the DFT of any slice $C_{3b}[\tau]$ becomes:

$$FC_b(k) = \{X(k) \otimes e^{-jbk} X(k)\}\, X(-k)\,. \qquad \textbf{(E 3.30)}$$

Since $x(n)$ is real, the convolution may also be thought of as the cross-correlation between $X(k)$ and its complex conjugate:

$$G(k) = \sum_{\lambda=-N/2}^{N/2} X(\lambda)\, X^*(\lambda-k)\,.$$

This cross-correlation is at maximum for $k= 0$ (the energy of the signal) and has a generally decreasing behavior since the spectrum is more correlated at small frequency lags.

3.2.2 The Fourier Transform of the Horizontal Slice and the Bispectrum

- **Theorem 2:** *Given the bispectrum of a deterministic signal x(n), the DFT of the horizontal slice* $C_{3O}[\tau]$, *denoted as* $FC_0(k)$, *may be recovered by summing the bispectrum points along the diagonal lines* $w = k$:

$$\boxed{FC_0[k] = \frac{1}{K}\sum_{i=0}^{K-1} B(i, k-i)} \qquad \text{(E 3.31)}$$

- **Proof:** On any diagonal line $w = k$, the bispectrum is given by:

$$B(i, k-i) = X(i) \cdot X(k-i) \cdot X(-k) . \qquad \text{(E 3.32)}$$

The sum over all i's in the range $[0,.....,K-1]$ is defined as:

$$Sum[k] \equiv \sum_{i=0}^{K-1} B(i, k-i) . \qquad \text{(E 3.33)}$$

Using Eq 3.32, the sum is:

$$Sum[k] = K \cdot X(-k) \cdot \sum_{i=0}^{K-1} X(i) \cdot X(k-i) . \qquad \text{(E 3.34)}$$

Now, the DFT of the horizontal slice was found to be (Eq 3.29):

$$FC_0[k] = \{X(k) \otimes X(k)\} X(-k) = X(-k) \sum_{i=0}^{K-1} X(i) \cdot X(k-i) . \qquad \text{(E 3.35)}$$

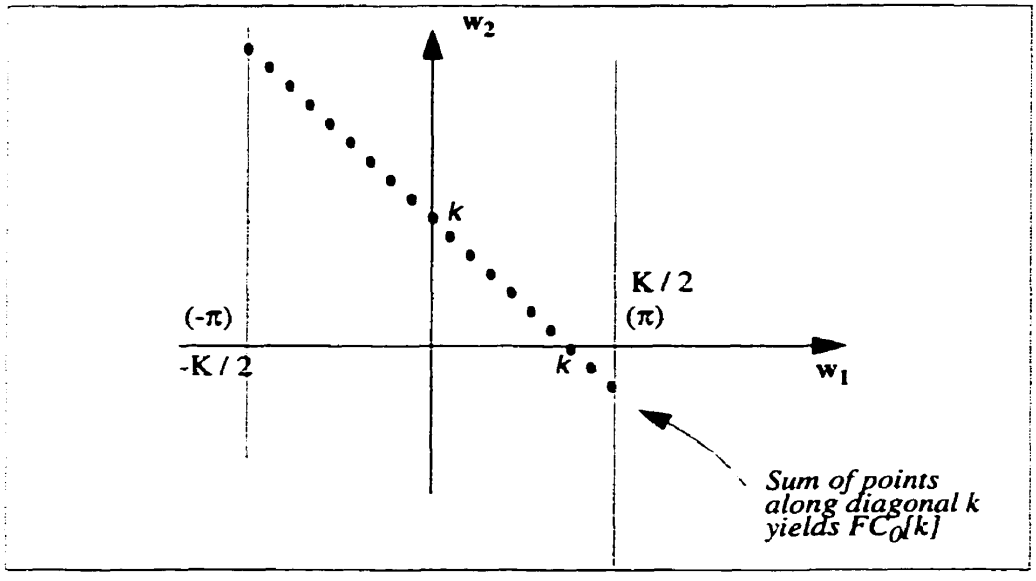
From Eq 3.34 and Eq 3.35:

$$FC_0[k] = \frac{1}{K} Sum[k] = \frac{1}{K}\sum_{i=0}^{K-1} B(i, k-i) .$$

When K is odd, it is more convenient to make the summation over the range $[-K/2, ..., K/2]$ (Figure 3-3), in which case the relation becomes:

$$FC_0[k] = \frac{1}{(K+1)} \sum_{i=-K/2}^{K/2} B(i, k-i) . \qquad \text{(E 3.36)}$$

Figure 3-3 Fourier transform of the cumulant slice from the bispectrum



3.2.3 Geometric Mean of the Power Spectrum and the Bispectrum

• **Theorem 3:** *Given the bispectrum of a deterministic zero-mean signal x(n), the geometric mean of the magnitude spectrum may be recovered entirely from the magnitude bispectrum and the value of the DFT magnitude at an arbitrary frequency point* |X(k)|, *by considering the product of the bispectrum points along the diagonal line w = k, namely:*

$$GM = \left\{ \left(\prod_{i=k+1}^{K} |B(i, k-i)| \right) |X(k)|^{K-k} \right\}^{1/K} \qquad \text{(E 3.37)}$$

• **Proof:** The geometric mean of the Fourier magnitude spectrum is defined as:

$$GM \equiv \left(\prod_{i=1}^{K} |X(i)| \right)^{1/K} . \qquad \text{(E 3.38)}$$

The magnitude bispectrum along any diagonal line $w = k$ *is* given (Figure 3-4) by:

$$|B(i, k-i)| = |X(i)| \cdot |X(k-i)| \cdot |X(k)| . \qquad \text{(E 3.39)}$$

Two products of the bispectrum points along this diagonal line are considered. The first one ($Prod_a[k]$) is confined to the first quadrant (Q1) and over the range of i in $[1, 2, \ldots, k-1]$:

$$Prod_a[k] \equiv \prod_{i=1}^{k-1} |B(i, k-i)|, \tag{E 3.40}$$

or using Eq 3.39

$$Prod_a[k] = |X(k)|^{k-1} \left(\prod_{i=1}^{k-1} |X(i)| \cdot |X(k-i)| \right), \tag{E 3.41}$$

which can further be written as:

$$Prod_a[k] = |X(k)|^{k-1} \left(\prod_{i=1}^{k-1} |X(i)|^2 \right); \tag{E 3.42}$$

as an example $k= 5$:

$$Prod_a[5] = |X(5)|^4 \{|X(1)||X(4)|\} \{|X(2)||X(3)|\} \{|X(3)||X(2)|\} \{|X(4)||X(1)|\}$$

$$= |X(5)|^4 \{|X(1)|^2\} \{|X(2)|^2\} \{|X(3)|^2\} \{|X(4)|^2\} \{|X(5)|^2\}.$$

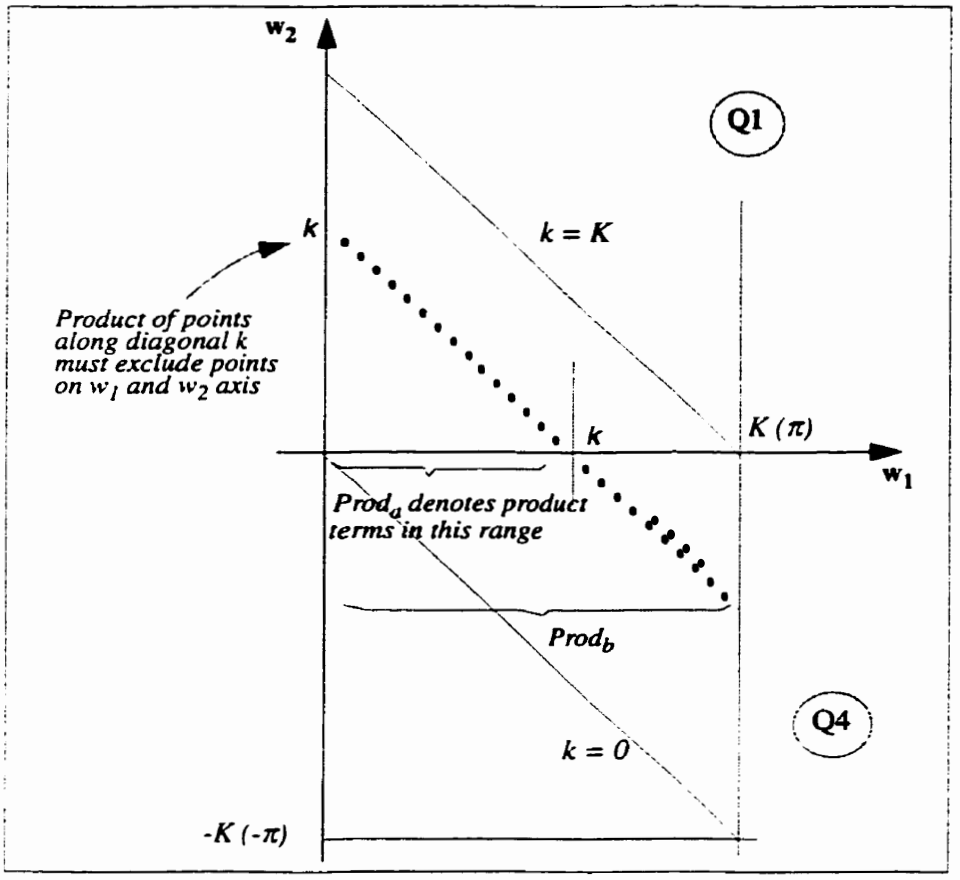
The second product term ($Prod_b[k]$) extends over both the first (Q1) and fourth (Q4) quadrants and over the i's in the range $[1, 2, \ldots, K]$, but excluding the point $(i = k)$ since this will result in a $|X(0)|$ term which will cause a zero expression. Thus:

$$Prod_b[k] \equiv \prod_{\substack{i=1 \\ i \neq k}}^{K} |B(i, k-i)| \tag{E 3.43}$$

or using Eq 3.39

$$Prod_b[k] = |X(k)|^{K-1} \left(\prod_{\substack{i=1 \\ i \neq k}}^{K} |X(i)| \cdot |X(k-i)| \right). \tag{E 3.44}$$

Figure 3-4 Geometric mean of the power spectrum from the bispectrum



The second term in Eq 3.44 (denoted by $Prod_{b2}[k]$) is split into two components as:

$$Prod_{b2}[k] = \left\{\prod_{i=1}^{k-1}|X(i)|\cdot|X(k-i)|\right\}\left\{\prod_{i=k+1}^{K}|X(i)|\cdot|X(k-i)|\right\}. \qquad \text{(E 3.45)}$$

The first product expression (denoted by P_1) can be written as in Eq 3.41:

$$P_1 \equiv \prod_{i=1}^{k-1}|X(i)|\cdot|X(k-i)| = \prod_{i=1}^{k-1}|X(i)|^2. \qquad \text{(E 3.46)}$$

The second product expression in Eq 3.45 is denoted by P_2. Due to the range of the indices, the index in the second term $|X(k-i)|$ is always negative. Using the symmetry of the magnitude spectrum, it may be replaced by the points with positive indices: $|X(i-k)|$. Furthermore, given the range of the

indices of the summation, it is possible to show that this expression may be rewritten as the product of all spectrum points in the range $[1, 2, \ldots, K]$ for any value of k confined in that range. Thus:

$$P_2 \equiv \prod_{i=k+1}^{K} |X(i)| \cdot |X(k-i)|$$

$$= \{|X(K)| \cdot |X(K-k)|\}\ \{|X(K-1)| \cdot |X(K-k-1)|\} \ldots \{|X(k+1)| \cdot |X(-1)|\}$$

$$P_2 = \prod_{i=1}^{K} |X(i)|. \qquad \text{(E 3.47)}$$

This may be verified by considering for example $k = 5$, and $K = 10$:

$$P_2 = \{|X(6)||X(-1)|\}\ \{|X(7)||X(-2)|\}\ \{|X(8)||X(-3)|\}\ \{|X(9)||X(-4)|\}\ \{|X(10)||X(-5)|\}$$

$$= \{|X(6)||X(1)|\}\ \{|X(7)||X(2)|\}\ \{|X(8)||X(3)|\}\ \{|X(9)||X(4)|\}\ \{|X(10)||X(5)|\}$$

$$= |X(1)||X(2)||X(3)||X(4)||X(5)||X(6)||X(7)||X(8)||X(9)||X(10)|$$

$$P_2 = \prod_{i=1}^{10} |X(i)|.$$

Substituting Eq 3.46 and Eq 3.47 into Eq 3.45 and then into Eq 3.44, the expression for $Prod_b[k]$ may be expressed in terms of $Prod_a[k]$ (Eq 3.41) as:

$$Prod_b[k] = |X(k)|^{K-1} \left\{ \prod_{i=1}^{k-1} |X(i)| \cdot |X(k-i)| \right\} \left\{ \prod_{i=k+1}^{K} |X(i)| \cdot |X(k-i)| \right\}$$

$$= |X(k)|^{K-1} \cdot \prod_{i=1}^{k-1} |X(i)|^2 \cdot \prod_{i=1}^{K} |X(i)|$$

$$Prod_b[k] = \frac{|X(k)|^{K-1}}{|X(k)|^{k-1}} \cdot Prod_a[k] \cdot \prod_{i=1}^{K} |X(i)|,$$

from which the expression for the product of the magnitude spectrum is inferred as:

$$\prod_{i=1}^{K} |X(i)| = |X(k)|^{K-k} \cdot \frac{Prod_b[k]}{Prod_a[k]}.$$

The geometric mean of the magnitude spectrum (Eq 3.38) is thus:

$$GM = \left\{ \prod_{i=1}^{K} |X(i)| \right\}^{1/K}$$

$$= \left[|X(k)|^{K-k} \cdot \frac{Prod_b[k]}{Prod_a[k]} \right]^{1/K}$$

$$= \left[|X(k)|^{K-k} \cdot \frac{\prod_{\substack{i=1 \\ i \neq k}}^{K} |B(i, k-i)|}{\prod_{i=1}^{k-1} |B(i, k-i)|} \right]^{1/K}$$

or simply,

$$GM = \left\{ \left(\prod_{i=k+1}^{K} |B(i, k-i)| \right) |X(k)|^{K-k} \right\}^{1/K}.$$

The magnitude of $|X(k)|$ may be evaluated from the bispectrum magnitude as will be explained in Section 3.2.4.

Example 1: A simple example to consider is for $k = 1$. By direct substitution the value of $|X(1)|$ can be evaluated from the bispectrum magnitudes as:

$$|X(1)| = \left[\frac{|B(1, 1)|^3 |B(3, 1)|}{|B(2, 1)||B(2, 2)|} \right]^{\frac{1}{6}}$$

The geometric mean is thus found from the bispectrum points along the diagonal $w = 1$:

$$GM = \left(\left[\prod_{i=2}^{K} |B(i, 1-i)| \right] \left[\frac{|B(1, 1)|^3 |B(3, 1)|}{|B(2, 1)||B(2, 2)|} \right]^{\frac{(K-1)}{6}} \right)^{\frac{1}{K}}.$$

3.2.4 Fourier Magnitude Recovery from the Bispectrum

The bispectrum magnitude is expressed in terms of the signal's Fourier transform magnitude as (using discrete notation):

$$|B_x(i,j)| = |X(i)||X(j)||X(i+j)|. \quad \text{(E 3.48)}$$

A number of approaches may be used to recover the Fourier transform from the bispectrum computations of the observed data. A good survey of these is found in [See88] and [Sun90]. In the following, the classical methods are summarized and three new approaches are proposed. The advantage of each is highlighted and compared to the others.

3.2.4.1 Existing Methods

Algorithm 1: Freezing one variable, and using recursion

The simplest way to recover the magnitude of $|X(w)|$ is by using bispectrum samples of $|B(k,j)|$ on one of the frequency axes. This is done by substituting 0 for one of the frequencies in Eq 3.48; e.g.:

$$|B(0,j)| = |X(0)||X(j)||X(j)| = |X(0)||X(j)|^2 \quad \text{(E 3.49)}$$

$$\text{or } |X(j)| = \sqrt{\frac{|B(0,j)|}{|X(0)|}}$$

where $|X(0)| = \sqrt[3]{|B(0,0)|}$.

The above recursive formula would not be suitable for a zero-DC signal; a more general derivation entails fixing one of the frequencies (i) and using it as an increment to find all other points:

$$|X(j+i)| = \frac{|B(j,i)|}{|X(j)||X(i)|}. \quad \text{(E 3.50)}$$

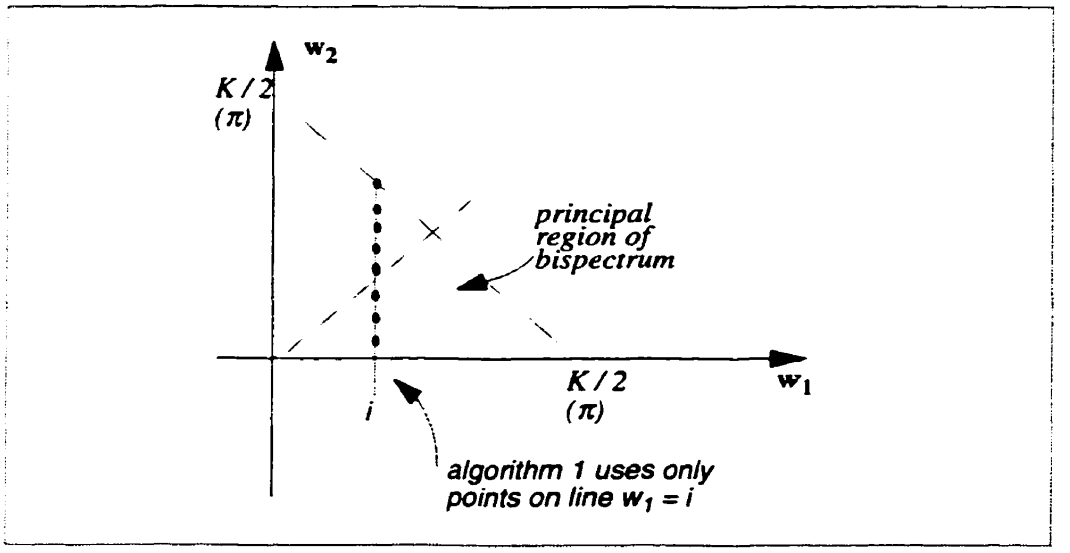
If for example, i is set to 1, $X(1)$ is first computed from $|B(j, i)|$ values:

$$|X(1)| = \left[\frac{|B(1,1)|^3|B(3,1)|}{|B(2,1)||B(2,2)|}\right]^{\frac{1}{6}} \quad \text{(E 3.51)}$$

and then (Eq 3.50) can be rewritten as

$$\boxed{|X(j+1)| = \frac{|B(1,j)|}{|X(1)||X(j)|}} \text{ for } j = 1, \ldots, \frac{N}{2} - 1. \quad \text{(E 3.52)}$$

Figure 3-5 Algorithm 1 for the Fourier magnitude recovery from the bispectrum



This equation essentially uses the points of the bispectrum on the vertical line $w_1 = i$ (Figure 3-5). While simple, it has two shortcomings:

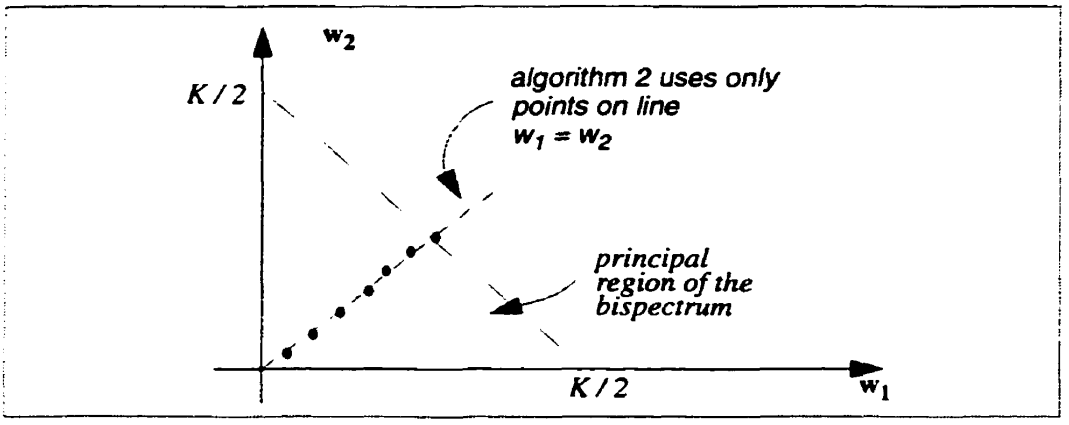
- It breaks down if any of the $|X(j)|$ is zero (or even if $|X(1)|$ is zero). While in practice this is seldom the case (due to roundoff error and leakage effects), this constraint makes it less robust.
- It only uses a small portion of the information available in the principal region of the bispectrum, and as such is susceptible to estimation errors.

Algorithm 2: Using mid-frequency estimates

A closed form estimation of $X(j)$ can be done by recursively using bispectrum estimates at *mid-frequency* points:

$$|X(j)| = \frac{\left|B\left(\frac{j}{2}, \frac{j}{2}\right)\right|}{\left|X\left(\frac{j}{2}\right)\right|^2}. \qquad \text{(E 3.53)}$$

Figure 3-6 Algorithm 2 for Fourier magnitude recovery from the bispectrum



Essentially, this form uses the points on the diagonal line $w_1 = w_2$ (Figure 3-6). Thus:

- It only uses a small portion of the information in the principal domain, and is susceptible to estimation errors.
- Since it relies on mid-point estimates, it is only valid for half the frequency points, unless some form of interpolation is used.
- It breaks down if any of the $|X(j)|$'s are zeros.

Algorithm 3: Least Square approach

The algorithm described in [Sun90] uses a least square approach and all the available information in the principal domain of the bispectrum. First, Eq 3.48 is transformed into a linear equation by using logarithmic representation. Starting with a new notation: $\tilde{B} = \log|B|$, $\tilde{X} = \log|X|$, Eq 3.48 becomes:

$$\tilde{B}(i, j) = \tilde{X}(i) + \tilde{X}(j) + \tilde{X}(i + j) . \qquad \text{(E 3.54)}$$

When considering all the valid values of the two frequencies in the principal region, Eq 3.54 may be written in matrix form: $\tilde{b} = A \cdot \tilde{x}$, where $\tilde{b}$ is the vector of bispectrum estimates computed from the noisy observations, and $\tilde{x}$ is a vector of the estimates of the magnitude Fourier transform:

$$\tilde{b} = \begin{bmatrix} \tilde{B}(1, 1) \\ \tilde{B}(1, 2) \\ \cdots \\ \tilde{B}\left(\frac{K}{4}, \frac{K}{4}\right) \end{bmatrix}, \tilde{x} = \begin{bmatrix} \tilde{X}(1) \\ \tilde{X}(2) \\ \cdots \\ \tilde{X}\left(\frac{K}{2}\right) \end{bmatrix}$$

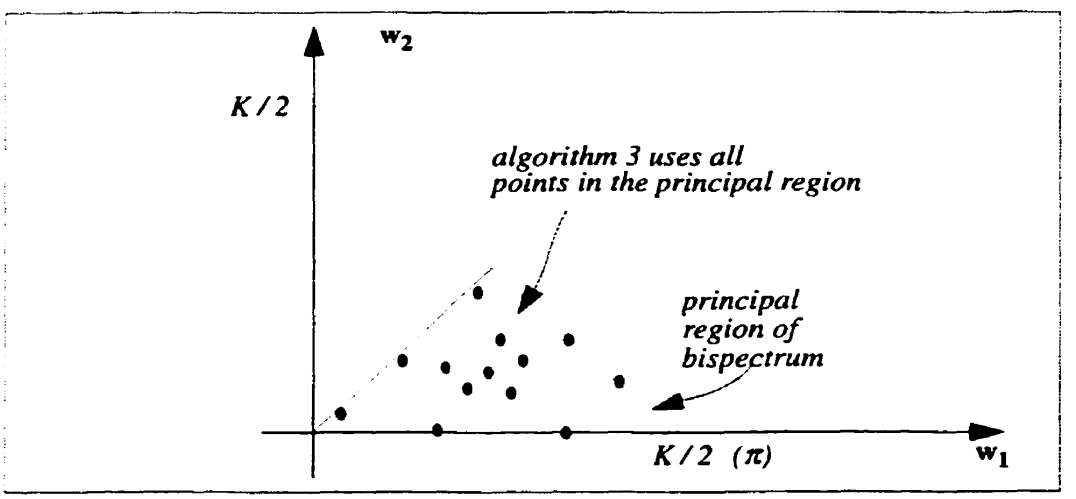
and A is a matrix of the form: $A = \begin{bmatrix} 2 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & 1 \\ 0 & 2 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix}$.

A least square solution [Sun90] can then be expressed in matrix form as:

$$\bar{x} = (A^T A)^{-1} A^T \bar{b}. \qquad \text{(E 3.55)}$$

- The algorithm uses all available information and as such minimizes the variance due to estimation errors.
- The algorithm does assume that none of the $|X(j)|$'s is zero and thus there are no exceptions when taking the logarithm.
- The algorithm uses the point on the frequency axis and thus does not assume the signal may be zero-DC.
- The algorithm entails matrix inversion and it is not clearly shown under what conditions this matrix may be singular.

Figure 3-7 Algorithm 3 for Fourier magnitude recovery from the bispectrum



3.2.4.2 Proposed Methods

Algorithm A: Averaging points on diagonal lines

This algorithm uses the points on diagonal lines in the positive region. On any diagonal line $w = k$, the magnitude expression of Eq 3.48 may be written as (Figure 3-8):

$$|B(i, k-i)| = |X(i)| \cdot |X(k-i)| \cdot |X(k)| . \qquad \text{(E 3.56)}$$

Taking the sum over i over the range [1,..., k-1] on both sides of (Eq 3.56) yields:

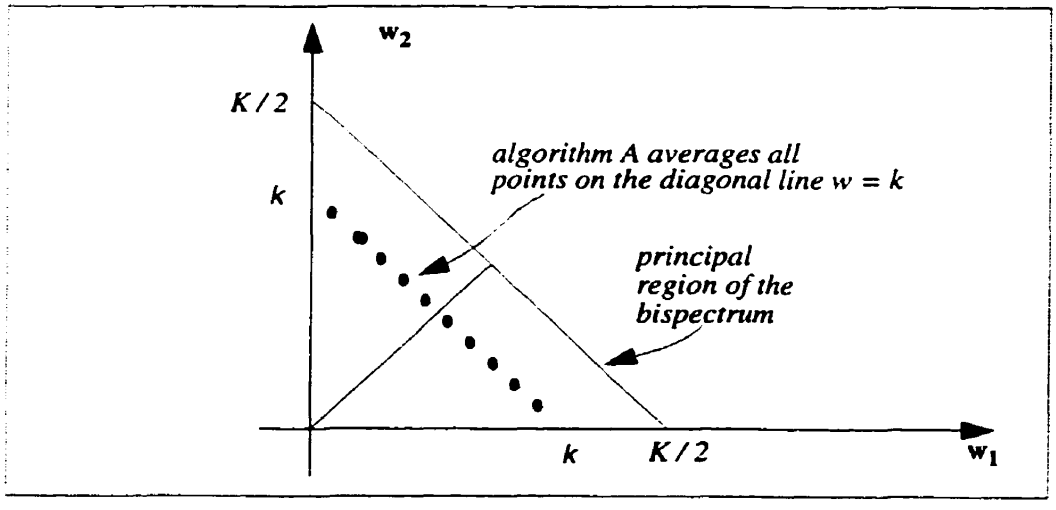
$$\sum_{i=1}^{k-1} |B(i, k-i)| = |X(k)| \left\{ \sum_{i=1}^{k-1} |X(i)| \cdot |X(k-i)| \right\} \qquad \text{(E 3.57)}$$

from which a recursive equation can be derived:

$$|X(k)| = \frac{\sum_{i=1}^{k-1} |B(i, k-i)|}{\sum_{i=1}^{k-1} |X(i)| \cdot |X(k-i)|} \qquad 2 \le k \le K \ . \qquad \text{(E 3.58)}$$

The initial condition, $|X(1)|$ is given by Eq 3.51.

Figure 3-8 Algorithm A for Fourier magnitude recovery from the bispectrum



The characteristics of this algorithm:

- It does not require the knowledge of $|X(0)|$ and is therefore appropriate when the signal has a zero DC i.e. X(0) = 0.

- It will not break down if some of the $|X(k)|$'s are zeros, provided that not all the points on a diagonal are zero, which is practically not possible.

Algorithm B: Geometric mean of the points on diagonal lines

- **Proposition 2.** *By considering the product of the bispectrum points along a diagonal line $w = k$ in the first quadrant, a recursive relation exists between any $|X(k+1)|$ and $|X(k)|$ as:*

$$|X(k+1)| = \left(\frac{|X(k)|^{k-3} \cdot Prod[k+1]}{Prod[k]} \right)^{1/k} \quad k = 3, \ldots, K/2. \tag{E 3.59}$$

- **Proof:** The points on the diagonal lines $w = k$ in the first quadrant are given by:

$$|B(i, k-i)| = |X(i)| \cdot |X(k-i)| \cdot |X(k)|. \tag{E 3.60}$$

The product of the bispectrum along this diagonal over i's in the range [1,..,k-1] is defined as:

$$Prod[k] \equiv \prod_{i=1}^{k-1} |B(i, k-i)| \quad \text{or, using Eq 3.60,}$$

$$Prod[k] = |X(k)|^{k-1} \left(\prod_{i=1}^{k-1} |X(i)| \cdot |X(k-i)| \right) = |X(k)|^{k-1} \left(\prod_{i=1}^{k-1} |X(i)|^2 \right),$$

from which an expression for $|X(k)|$ can be derived as:

$$|X(k)| = \left(\frac{Prod[k]}{\prod_{i=1}^{k-1} |X(i)|^2} \right)^{1/(k-1)} = \left(\frac{\prod_{i=1}^{k-1} |B(i, k-i)|}{\prod_{i=1}^{k-1} |X(i)|^2} \right)^{1/(k-1)} \quad k = 2, \ldots, K/2. \tag{E 3.61}$$

A quicker recursive equation may be derived by considering the product of the bispectrum magnitude on two consecutive diagonals, k and $k+1$. For instance the product along $k+1$:

$$Prod[k+1] = |X(k+1)|^{k} \left(\prod_{i=1}^{k} |X(i)| \cdot |X(k-i)| \right)$$

and dividing $Prod[k+1]$ by $Prod[k]$ yields:

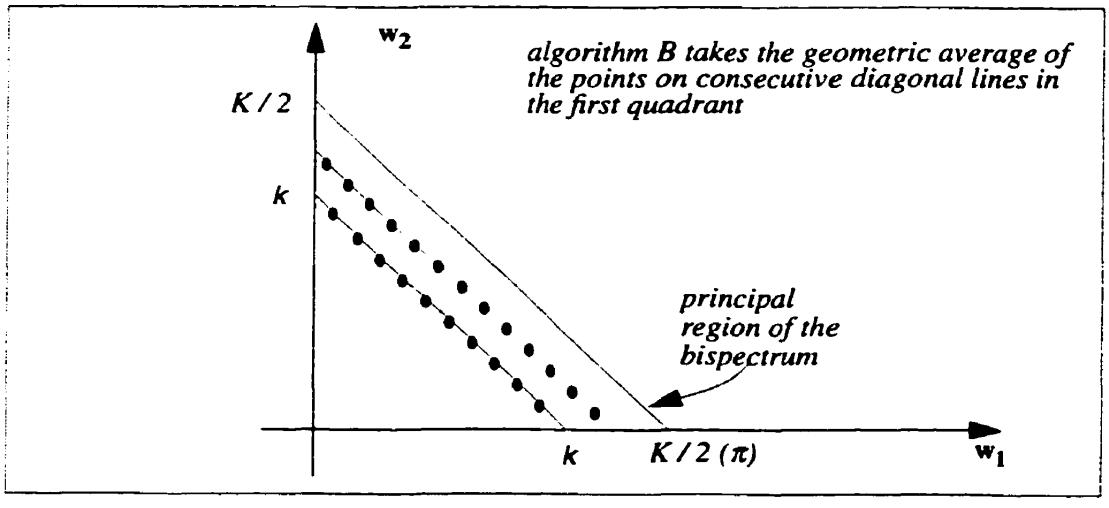
$$\frac{Prod\,[k+1]}{Prod\,[k]} = \frac{|X(k+1)|^k}{|X(k)|^{k-1}} \cdot \frac{\prod_{i=1}^{k}|X(i)|^2}{\prod_{i=1}^{k-1}|X(i)|^2}$$

$$= \frac{|X(k+1)|^k \cdot |X(k)|^2 \cdot \prod_{i=1}^{k-1}|X(i)|^2}{|X(k)|^{k-1} \cdot \prod_{i=1}^{k-1}|X(i)|^2} = \frac{|X(k+1)|^k}{|X(k)|^{k-3}}.$$

Therefore, a relation between $|X(k+1)|$ and $|X(k)|$ may be derived as:

$$\boxed{|X(k+1)| = \left(\frac{|X(k)|^{k-3} \cdot Prod\,[k+1]}{Prod\,[k]}\right)^{1/k}} \quad k = 3, \ldots, K/2.$$

Figure 3-9 Algorithm B for Fourier magnitude recovery from the bispectrum



Algorithm C: relation between any 2 magnitudes in terms of bispectrum products

- **Proposition 3.** *Given the magnitude bispectrum of x(n), then the ratio of any two Fourier magnitude spectrum points, |X(k)| and |X(p)| may be written in terms of the product of bispectrum magnitudes, namely:*

$$\left(\frac{|X(p)|}{|X(k)|}\right)^{k-p-1} = \frac{\prod_{i=1}^{k-p} B[p,i]}{\prod_{i=1}^{k-p} B[i,k-i]} \quad \text{with} \quad p,k \le K/2. \tag{E 3.62}$$

- **Proof:** First, consider the bispectrum points on the vertical line $w_1 = p$ (Figure 3-10).

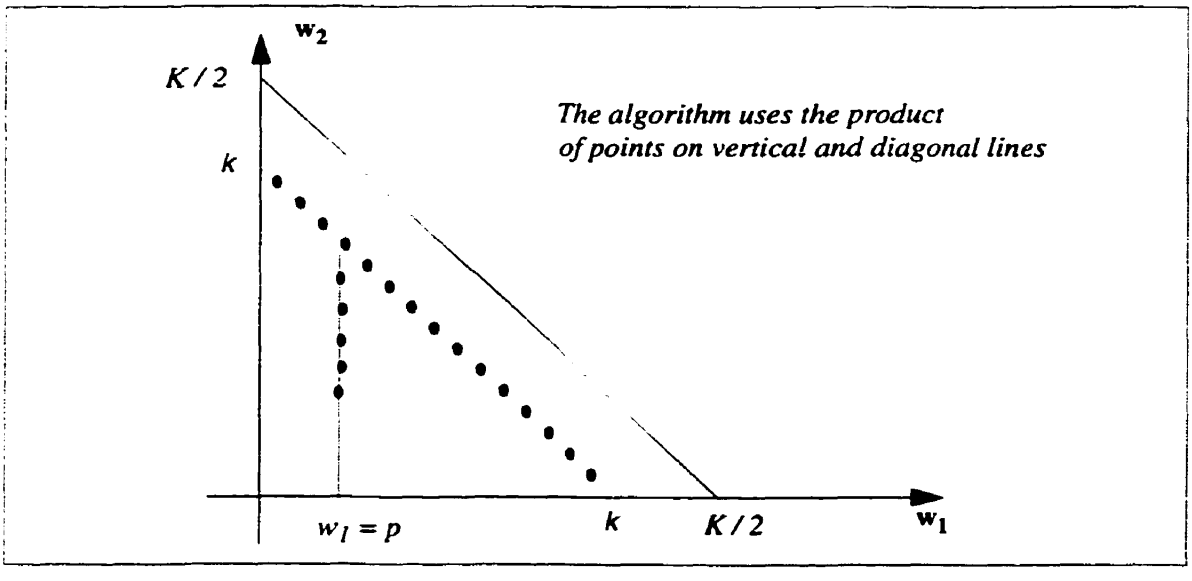
$$A \equiv \prod_{i=1}^{k-p} B[p,i] = |X(p)|^{k-p} \prod_{i=1}^{k-p} |X(i)||X(i+p)| .$$

Changing the limits of the product set to $i = [p+1,\ldots,k-p]$, the above expression becomes:

$$A = |X(p)|^{k-p} \cdot \{|X(1)| \cdot \ldots \cdot |X(p)|\} \cdot \{|X(k)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot \prod_{i=p+1}^{k-p} |X(i)|^2$$

$$= |X(p)|^{k-p-1} \cdot \{|X(1)| \cdot \ldots \cdot |X(p-1)|\} \cdot \{|X(k)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot \prod_{i=p}^{k-p} |X(i)|^2 .$$

Figure 3-10 Algorithm C for Fourier magnitude recovery from the bispectrum



As an example, $p = 3$,

$$A = |X(3)|^{k-3} \cdot \{|X(1)||X(4)| \cdot |X(2)||X(5)| \cdot |X(3)||X(6)| \cdot |X(4)||X(7)| \cdot \ldots \cdot |X(k-3)||X(k)|\}$$

$$= |X(3)|^{k-4} \cdot \left\{|X(1)||X(2)||X(3)| \cdot |X(k-2)||X(k-1)||X(k)| \cdot \prod_{i=4}^{k-3}|X(i)|^2\right\}$$

$$= |X(3)|^{k-5} \cdot \left\{|X(1)||X(2)| \cdot |X(k-2)||X(k-1)||X(k)| \cdot \prod_{i=3}^{k-3}|X(i)|^2\right\}.$$

Now, consider the product terms along the diagonal line $w = k$:

$$B \equiv \prod_{i=1}^{k-p} B[i, k-i]$$

$$B = \left(\prod_{i=p}^{k-p} B[i, k-i]\right) \cdot \left(\prod_{i=1}^{p-1} B[i, k-i]\right) = B_1 \cdot B_2.$$

The second product expression may be written in terms of spectrum magnitudes:

$$B_2 = \{|X(1)| \cdot |X(2)| \cdot \ldots \cdot |X(p-1)|\} \cdot \{|X(k)| \cdot |X(k-1)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot |X(k)|^{p-1}$$

and first product expression (B_1) may be written as: $B_1 = |X(k)|^{k-2p+1} \cdot \prod_{i=p}^{k-p}|X(i)|^2$.

Combining B_1 and B_2 yields:

$$B = |X(k)|^{k-p-1} \cdot \{|X(1)| \cdot \ldots \cdot |X(p-1)|\} \cdot \{|X(k)||X(k-1)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot \prod_{i=p}^{k-p}|X(i)|^2$$

and the ratio A / B is:

$$\frac{A}{B} = \frac{|X(p)|^{k-p-1} \cdot \{|X(1)| \cdot \ldots \cdot |X(p-1)|\} \cdot \{|X(k)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot \prod_{i=p}^{k-p}|X(i)|^2}{|X(k)|^{k-p-1} \cdot \{|X(1)| \cdot \ldots \cdot |X(p-1)|\} \cdot \{|X(k)| \cdot \ldots \cdot |X(k-p+1)|\} \cdot \prod_{i=p}^{k-p}|X(i)|^2}$$

Therefore:

$$\left(\frac{|X(p)|}{|X(k)|}\right)^{k-p-1} = \frac{A}{B} = \frac{\prod_{i=1}^{k-p} B[i, p]}{\prod_{i=1}^{k-p} B[i, k-i]}.$$

3.2.4.3 Discussion

The rationale for proposing three new schemes for Fourier magnitude recovery is that the reported algorithms in the literature have certain shortcomings. These includes their inadequacy for zero-DC signals (algorithm 3), their use of only a small portion of the available information (algorithms 1 and 2) or their computational complexity (algorithm 3). The objective of the three proposed algorithms is to remedy some of these issues, and find an appropriate compromise between using a larger percentage of the available information while keeping a manageable computational cost. Moreover, the proposed algorithms do not make use of the bispectrum points along either frequency axis and as such are suitable for zero-DC signals, which is often the assumed condition when performing HOS analysis.

3.3 Fourth-Order Derivations

3.3.1 Fourier Transform of Cumulant Slices

- **Theorem 4:** *In the case of a deterministic signal, the Fourier transform of the horizontal and diagonal slices may be expressed in terms of the Fourier and power spectra of the underlying signal (X(w) and P(w) respectively). Moreover, the Fourier transform of the horizontal slice is real (zero phase) for any signal.*
- **Proof:**

1. **Diagonal slice:** assuming an energy signal, the diagonal slice is (from Eq 3.16):

$$C^{d}_{4}[\tau] = \sum_{n=-\infty}^{\infty} [x(n)\,x^{3}(n+\tau)] - 3 \cdot m_{2}(0) \cdot \left(\sum_{n=-\infty}^{\infty} [x(n)\,x(n+\tau)] \right).$$

The Fourier transform of the first term is:

$$\sum_{\tau=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} [x(n)\,x^{3}(n+\tau)]\,e^{-jw\tau} = \left(\sum_{n=-\infty}^{\infty} x(n)\,e^{jwn} \right) \left(\sum_{m=-\infty}^{\infty} x^{3}(m)\,e^{-jwm} \right)$$

$$= X(-w)\,\{X(w) \otimes X(w) \otimes X(w)\}$$

where $X(w) \otimes X(w)$ denotes autoconvolution. The transform of the second term is:

$$3 \cdot m_{2}(0) \cdot \left(\sum_{\tau=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} [x(n)\,x(n+\tau)]\,e^{-jw\tau} \right) = 3 \cdot m_{2}(0) \cdot P(w)\ ;$$

thus the Fourier of the diagonal slice is:

$$\boxed{FC^{d}_{4}(w) = X(-w)\,\{X(w) \otimes X(w) \otimes X(w)\} - 3 \cdot m_{2}(0) \cdot P(w)} \qquad \text{(E 3.63)}$$

2. **Horizontal slice:** assuming an energy signal, the horizontal slice is (from Eq 3.17):

$$C^{b}_{4}[\tau] = \left[\sum_{n} x^{2}(n)\,x^{2}(n+\tau) \right] - \left[\sum_{n} x^{2}(n) \right]^{2} - 2 \left[\sum_{n} x(n)\,x(n+\tau) \right]^{2}.$$

The Fourier transform of the first term is:

$$\sum_{\tau=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} [x^{2}(n)\,x^{2}(n+\tau)]\,e^{-jw\tau} = \left(\sum_{n=-\infty}^{\infty} x^{2}(n)\,e^{jwn} \right) \left(\sum_{m=-\infty}^{\infty} x^{2}(m)\,e^{-jwm} \right)$$

$$= ([X(w) \otimes X(w)]^{*})\,(X(w) \otimes X(w))$$

$$= |X(w) \otimes X(w)|^{2}.$$

The Fourier transform of the last terms is:

$$F\left(\left[\sum_n x(n)\,x(n+\tau)\right]^2\right) = F\left(\sum_n x(n)\,x(n+\tau)\right) \otimes F\left(\sum_n x(n)\,x(n+\tau)\right)$$

$$= (P(w) \otimes P(w)) .$$

Therefore the Fourier transform of the horizontal slice is:

$$\boxed{FC^b_4(w) = |X(w) \otimes X(w)|^2 - [m_2(0)]^2\delta(w) - 2\{P(w) \otimes P(w)\}} \quad \text{(E 3.64)}$$

3.3.2 DC Component of the Horizontal slice

- **Theorem 5:** *The DC component of the horizontal 4^{th} order cumulant slice ($C^b_4[\tau]$) may be expressed as the sum of 4^{th} power of the signal spectrum amplitudes.*

- **Proof:** The DC component of $C^b_4[\tau]$ is the value of $FC^b_4(f)$ at $f = 0$. It is first observed that:

$$X(f) \otimes X(f)|_{f=0} = \int_{-\pi}^{\pi} X(\lambda)\,X(-\lambda)\,d\lambda = \int_{-\pi}^{\pi} |X(\lambda)|^2 d\lambda = m_2(0) = SignalEnergy.$$

Therefore, setting $w = 0$ in Eq 3.64 yields,

$$FC^b_4(0) = [m_2(0)]^2 - [m_2(0)]^2 - 2(P(f) \otimes P(f))|_{f=0}$$

$$= -2(P(f) \otimes P(f)|_{f=0}) = -2\int_{-\pi}^{\pi} P(\lambda)\,P(-\lambda)\,d\lambda$$

$$\boxed{FC^b_4(0) = -2\int_{-\pi}^{\pi} |X(\lambda)|^4 d\lambda} \quad \text{(E 3.65)}$$

Special case: Signal with flat spectrum:

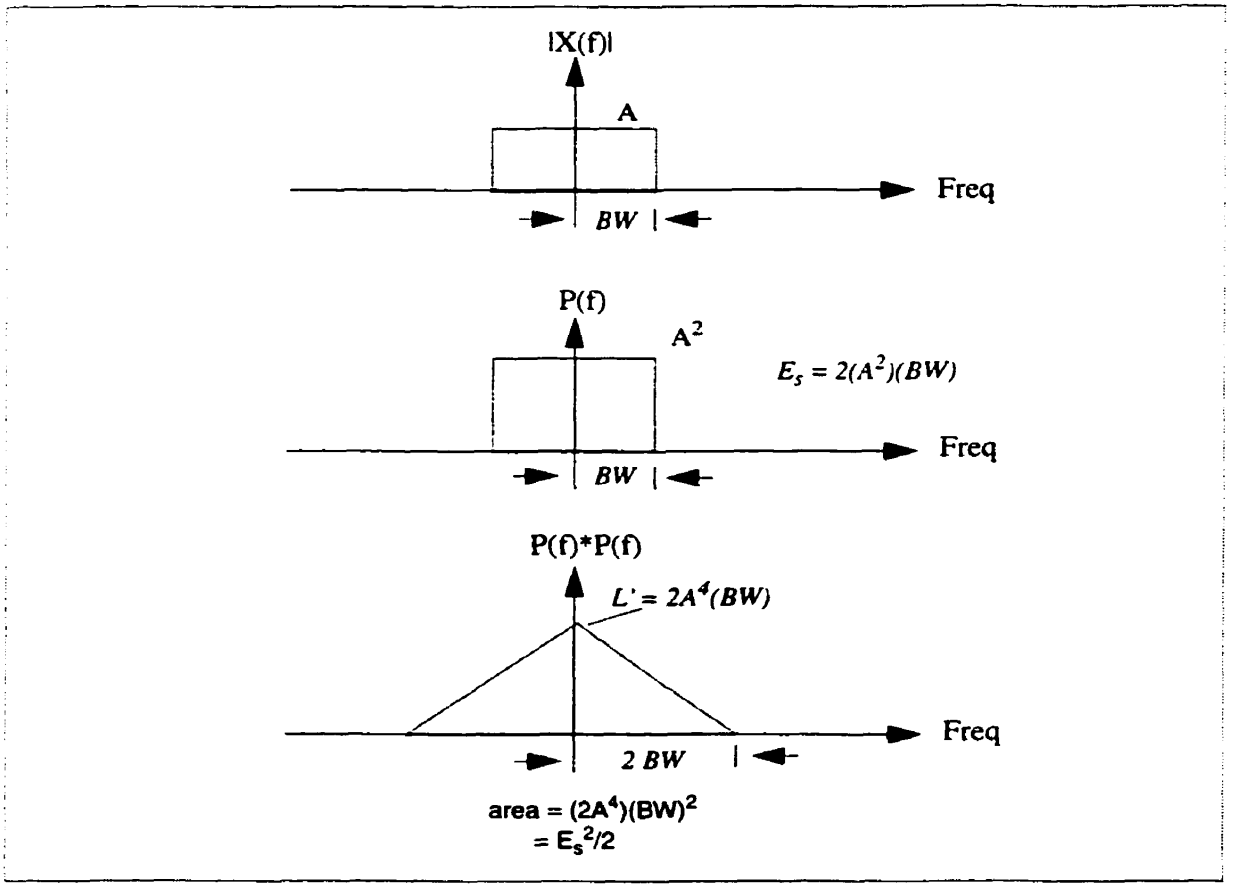
- **Corollary 1:** *When the signal x(n) has a flat spectrum, the DC component of the horizontal 4^{th}-order cumulant slice ($C^b_4[\tau]$) may be written in terms of the signal energy and bandwidth.*

- **Proof:** Assume that $x(n)$ is bandlimited to $[-BW:BW]$. The power spectrum is flat and has magnitude a^2. The energy of the signal is the area under the power spectrum (Figure 3-11), and is given by: $E = 2a^2BW$. From Eq 3.65, the DC component of the horizontal slice is:

$$FC^b_4(0) = -2\int_{-\pi}^{\pi} |P(\lambda)|^2 d\lambda = -2(2a^4BW)$$

$$\text{or} \quad \boxed{FC^b_4(0) = \frac{-E^2}{BW}} \qquad \text{(E 3.66)}$$

Figure 3-11 Auto-convolution of the power spectrum: the case of a flat-spectrum signal



3.4 Conclusion

This chapter provided a background on higher-order statistics and a new set of derivations that relate second and higher-order statistics of real signals. The expressions for the Fourier transform of the horizontal slices of the 3^{rd}-order cumulant as well as the horizontal and diagonal slices of the 4^{th}-order cumulant were derived and shown to be expressed in terms of the Fourier transform of the underlying signal. Similarly, it was shown that the Fourier transform of the horizontal slice of the 3^{rd}-order cumulant may be recovered entirely from the bispectrum points, and that the geometric mean of the power spectrum can be computed directly from the magnitude bispectrum.

Three new schemes were proposed for recovering the magnitude of the Fourier transform of a signal from its magnitude bispectrum. Compared to the ones reported in the literature, they provide a better compromise between using all the available information and finding a computationally manageable solution, in addition to being more suitable for zero-DC signals.

Finally, it was shown that the DC component of the horizontal slice of the 4^{th}-order cumulant can be expressed as the sum of the 4^{th} power of the signal spectrum amplitudes, and that in the case of a flat-spectrum signal, this DC component can be written in terms of the signal energy and bandwidth.

Some of the HOS expressions derived in this chapter are used throughout this thesis. As for the rest, their value is in providing insight about the relation between second and higher-order statistics and a quantitative way in the frequency domain to interpret the HOS of a signal in terms of its Fourier spectrum.

CHAPTER 4 Higher Order Cumulants of Subbanded Speech

Synopsis

This chapter is an exploratory work into the HOC properties of subbanded speech. It is assumed that speech is divided in narrow bands, such that each band contains one or two harmonics. The expressions for the diagonal slices of the 3rd and 4th order cumulants are derived assuming a modified sinusoidal model. Special properties of these cumulants in terms of phase and harmonic contents are highlighted and the relation between the 2nd and 4th order statistics is particularly noted. Actual speech data is used to verify the derivations and assess the validity of the underlying model.

4.1 Analytical Model for Subbanded Speech

According to the sinusoidal model of [McA86], a short speech segment is modeled as a sum of sinusoids that are coherent (in-phase) during voiced speech and incoherent during unvoiced speech:

$$s(n) = \sum_{m=1}^{M} a_m \cos\left[(n - n_0)\, w_m + \Psi_m + \theta_m \right], \qquad \text{(E 4.1)}$$

where n_0 is the voice onset time, M is the number of sinusoids, w_m is the frequency and a_m is the amplitude of the m^{th} sine wave. The first phase term in Eq 4.1 is due to the onset time n_0, defined as the time when the pitch pulse occurred relative to the beginning of the frame. The second phase component depends on a frequency cutoff w_c and a voicing probability, denoted by P_v, so that the higher the voicing probability the more sine waves are declared voiced with zero phase. The third phase component is the system phase θ_m at frequency track m, often assumed zero or a linear function of frequency.

If speech were divided in narrow bands such that at most two harmonics fall in each band, then in light of the model, voiced speech is expressed as the sum of two sinusoids with deterministic phase and unvoiced speech as the sum of two sinusoids with random (and uncorrelated) phases. Speech waveforms in both upper and lower bands (Figure 4-1) suggest that this model is plausible, at least for steady state voiced speech. Thus:

Steady Voiced Speech: $$s(n) = a_1 \cos(nPw_1 + \phi_1) + a_2 \cos(nPw_2 + \phi_2) \quad \text{(E 4.2)}$$

with $\phi_1 = cw_1$, $\phi_2 = cw_2$, both deterministic.

Unvoiced speech: $$s(n) = a_1 \cos(nPw_1 + \phi_1) + a_2 \cos(nPw_2 + \phi_2) \quad \text{(E 4.3)}$$

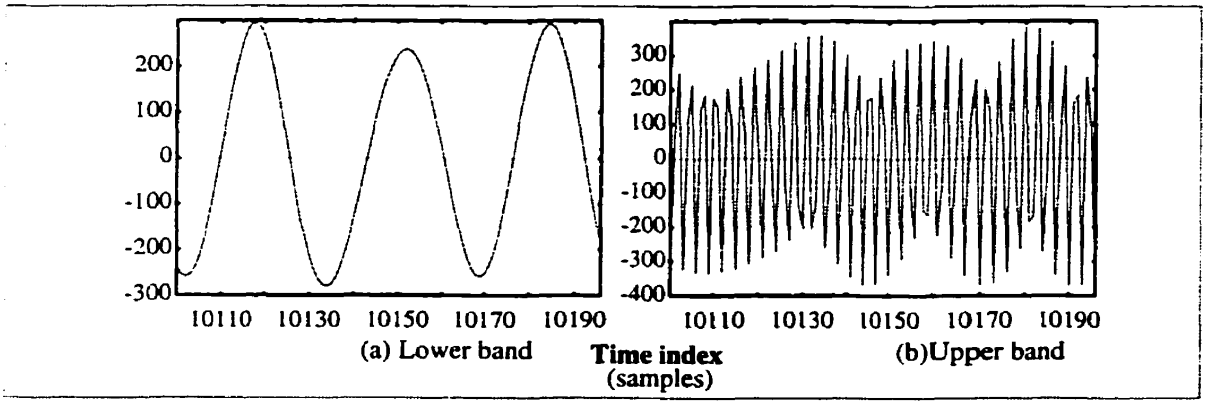
with $\phi_1, \phi_2 \in [-\pi, \pi]$, uniformly distributed,

where P is the sampling period. To account for transitional segments, transient speech is modeled here as an exponentially decaying (or growing) sinusoid:

Transient Voiced speech: $$s(n) = a \cdot e^{-\alpha nP} \cdot \cos(nPw_0 + \phi) \quad \text{(E 4.4)}$$

with $\phi = cw_0$, constant.

Figure 4-1 Waveform of 10 msec of speech in lower and upper bands



It is worth noting that the term 'voiced speech' used here refers to a simple two-harmonic signal in a narrow band. This is unlike the conventional usage of this term which usually refers to a speech signal with a fundamental frequency and many harmonics. Similarly, the model assumed for transient speech is a simplistic model in that it only uses a single sinusoid and assumes an exponential decay. These restrictions are made to simplify the mathematics while providing a third simple signal model which is different from the steady two-sinusoid models for voiced and unvoiced speech.

4.2 Third-Order Cumulant

4.2.1 Unvoiced Speech

- **Theorem 1:** *According to the sinusoidal model assumed, the third-order cumulant of unvoiced speech is identically zero.*
- **Proof:**

1. Case of a single sinusoid:

Using continuous time notation for simplicity, the signal $x(t)$ is a random process given by: $x(t) = a\cos(w_0 t + \theta)$, with θ uniformly distributed in the interval $[-\pi, \pi]$. The third cumulant function is: $C_3[\tau_1, \tau_2] = E[x(t)\,x(t+\tau_1)\,x(t+\tau_2)]$ and can be evaluated as:

$$C_3[\tau_1, \tau_2] = \frac{a^3}{2\pi}\int_{-\pi}^{\pi} \cos(w_0 t + x)\cos(w_0[t+\tau_1] + x)\cos(w_0[t+\tau_2] + x)\,dx .$$

By making repeated use of the trigonometric identity:

$\cos(a)\cos(b) = \frac{1}{2}\cos(a+b) + \frac{1}{2}\cos(a-b)$, the above equation becomes

$$C_3[\tau_1, \tau_2] = \frac{a^3}{4\pi}\int_{-\pi}^{\pi} \cos(w_0 t + x)\,[\cos(2w_0 t + w_0\tau_1 + w_0\tau_2 + 2x) + \cos(w_0\tau_1 - w_0\tau_2)]\,dx$$

$$= \frac{a^3}{8\pi}\left[\int_{-\pi}^{\pi}[\cos(3w_0 t + w_0\tau_1 + w_0\tau_2 + 3x)]\,dx + \int_{-\pi}^{\pi}[\cos(w_0 t + w_0\tau_1 + w_0\tau_2 + x)]\,dx\right.$$
$$\left. + \int_{-\pi}^{\pi}[\cos(w_0 t + w_0\tau_1 - w_0\tau_2 + x)]\,dx + \int_{-\pi}^{\pi}[\cos(w_0 t - w_0\tau_1 + w_0\tau_2 + x)]\,dx\right]$$

and making use of the fact that $\int_{-\pi}^{\pi}[\cos(C + x)]\,dx = 0$, the above simply cancels out and:

$$\boxed{C_3[\tau_1, \tau_2] = 0} \qquad \text{(E 4.5)}$$

2. Case of two sinusoids

When unvoiced speech is modeled as two uncorrelated sinusoids, then the cumulant of the sum is the sum of the two cumulants [Men91]. The cumulant of the sum is therefore zero.

4.2.2 Transient Speech

- **Theorem 2:** *According to the sinusoidal model assumed, the horizontal slice of the 3rd-order cumulant of transient speech is identically zero for any practical values of the model parameters.*
- **Proof:** Using the continuous time notation, and assuming a zero phase for simplicity the signal is modeled as: $x(t) = a \cdot e^{-\alpha t} \cdot \cos(w_0 t)$.

The horizontal cumulant slice is given by: $C_{30}[\tau] = \frac{1}{T}\int_0^T x^2(t) \cdot x(t-\tau)\, dt$ and becomes:

$$C_{30}[\tau] = \frac{1}{T}\int_0^T e^{-2\alpha t}[\cos(w_0 t)]^2 e^{-\alpha(t-\tau)} \cos[w_0(t-\tau)]\, dt$$

$$= \frac{1}{T} e^{\alpha\tau}\int_0^T e^{-3\alpha t}\left[\left(\frac{1}{2} + \frac{1}{2}\cos 2w_0 t\right)\cos[w_0(t-\tau)]\right] dt$$

$$= \frac{1}{T} e^{\alpha\tau}\int_0^T e^{-3\alpha t}\left[\frac{1}{2}\cos[w_0(t-\tau)] + \frac{1}{4}\cos[w_0(t+\tau)] + \frac{1}{4}\cos[w_0(3t-\tau)]\right] dt .$$

It is to note here that: $\int_0^T e^{ct}\cos(w_0 t)\, dt = \left(\frac{e^{ct}[c\cos(w_0 t) + w_0 \sin(w_0 t)]}{c^2 + w_0^2}\right)\Bigg|_0^T = \frac{c[e^{cT} - 1]}{c^2 + w_0^2}$

whenever T is a multiple of the signal period. If, in addition, the signal decays entirely at the end of the interval T, then the integral reduces to $\frac{-c}{c^2 + w_0^2}$ and is upper bounded in magnitude by $\frac{1}{c}$ for any value of the frequency w_0. In the model assumed here, the interval length is typically 20 to 50 msec. For the signal to decay within this interval, the decay constant is of the order: $c \sim 50$. At this typical value, the upper bound on the integral is very small ($= 0.02$). The integral thus has negligible values for any practical ranges of the model parameters and is therefore assumed zero. Thus the three sub-integrals from the above equation are zero and so is the third cumulant slice: $C_{30}[\tau] = 0$.

4.2.3 Steady Voiced Speech

- **Theorem 3:** *According to the sinusoidal model assumed, the 3rd-order cumulant of steady state voiced speech is non-zero if and only if one frequency is twice the other:* $w_2 = 2w_1$.
- **Proof:** The case is better understood when considering the bispectrum of the signal, which may be expressed in terms of the Fourier transform of the underlying signal:

$$B(f_1, f_2) = X(f_1)\, X(f_2)\, X^*(f_1 + f_2) .$$

The signal consists of the sum of two sinusoids. Consider the case where these two are related and using the continuous time notation for simplicity,

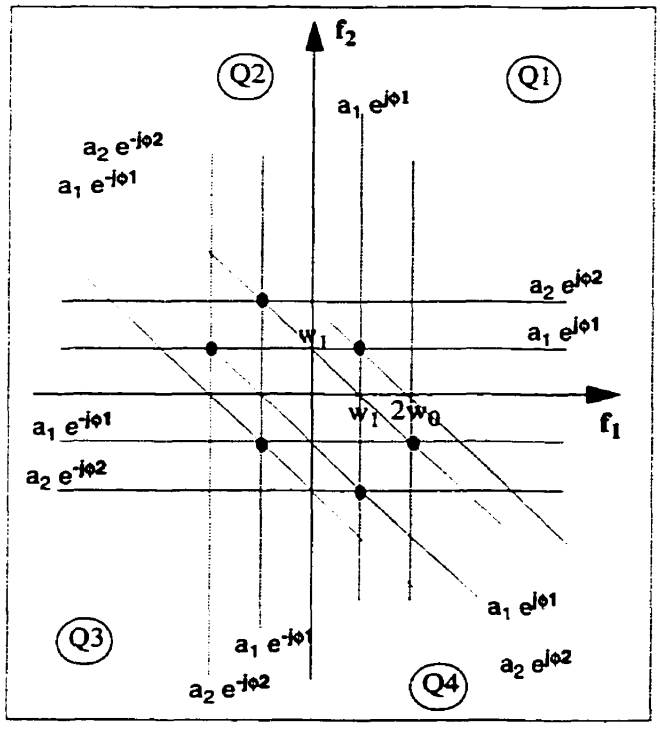
$$x(t) = a_1 \cos(tw_1 + \phi_1) + a_2 \cos(tw_2 + \phi_2) .$$

The spectrum $X(f)$ then consists of two pairs of delta functions:

$$X(f) = \frac{1}{2} \sum_{k=1}^{2} a_k [e^{j\phi_k} \delta(f - kw_1) + e^{-j\phi_k} \delta(f + kw_1)] .$$

A plot of the bispectrum (Figure 4-2) in the bifrequency domain shows that each of the factors: $X(f_1)$, $X(f_2)$, $X(-f_1-f_2)$ consists of three pairs of parallel *delta* lines, corresponding to the three harmonics, with $X(f_1)$ represented by the vertical lines, $X(f_2)$ the horizontals, and $X(-f_1-f_2)$ the diagonals. The bispectrum is non-zero wherever vertical, horizontal and diagonal lines intersect in each of the four quadrants. It is easy to see that if the two frequencies are not related, then there will be no intersection and the bispectrum will be zero.

Figure 4-2 The bispectrum of two related harmonics with $w_2 = 2\ w_1$



The expression of the 3rd-order cumulant (in the non-zero case) may be derived from the inverse 2D Fourier transform of the bispectrum. Because of the symmetry between the quadrants, each *delta* pair from the 1st and 3rd or from the 2nd and 4th quadrants would yield a *cosine* term. Thus:

$$C_3[\tau_1, \tau_2] = \frac{1}{4}\{a_1^2 a_2 \cos(w_0\tau_1 + w_0\tau_2) + a_1^2 a_2 \cos(w_0\tau_1 - 2w_0\tau_2) + a_1^2 a_2 \cos(2w_0\tau_1 - w_0\tau_2)\},$$

and the skewness is: $C_3[0, 0] = \frac{3}{4}a_1^2 a_2$.

4.3 Fourth-Order Cumulant

4.3.1 Transient Speech

- **Theorem 4:** *In the case where transient speech is modeled as an exponentially decaying sinusoid, the diagonal slice of the fourth-order cumulant may be expressed in terms of the signal energy, the damping factor and the signal frequency:*

$$\boxed{C_4^a[\tau] = 3\left[\frac{1}{2}\alpha T \coth(\alpha T) - 1\right] e^{-\alpha\tau}\cos(w_0\tau)\,[E_S]^2} \qquad \text{(E 4.6)}$$

- **Proof:** The signal is given by (using the continuous time notation): $x(t) = ae^{-\alpha t}\cos(w_0 t)$. The phase term is discarded here for simplicity and its presence does not affect the final results. For a deterministic signal, $C^a_4[\tau]$ is given by (from Section 3.1.1.2):

$$C_4^a[\tau] = \left[\frac{1}{T}\int_0^T x^3(t)\,x(t+\tau)\,dt\right] - 3\left[\frac{1}{T}\int_0^T x^2(t)\,dt\right]\left[\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt\right]. \qquad \text{(E 4.7)}$$

The second moment is evaluated first as:

$$\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt = \frac{1}{T}\int_0^T a^2 e^{-\alpha t}e^{-\alpha(t+\tau)}\cos(w_0 t)\cos(w_0[t+\tau])\,dt$$

$$= \frac{a^2}{T}\int_0^T e^{-\alpha\tau}e^{-2\alpha t}\left[\frac{1}{2}\cos(w_0\tau) + \frac{1}{2}\cos(w_0[2t+\tau])\right]dt$$

$$= \frac{a^2 e^{-\alpha\tau}}{T}\left[\frac{1}{2}\cos(w_0\tau)\int_0^T e^{-2\alpha t}dt + \frac{1}{2}\int_0^T e^{-2\alpha t}\cos(w_0[2t+\tau])\,dt\right].$$

As in Theorem 2, it is assumed that $\int_0^T e^{ct}\cos(w_0 t)\,dt \approx 0$ for practical values of the model parameters and whenever T is a multiple of the signal period. Furthermore, $\int_0^T e^{ct}dt = \frac{e^{cT}-1}{c}$; so after rearranging terms, the second moment becomes:

$$\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt = \frac{a^2 e^{-\alpha\tau}\cos(w_0\tau)}{4\alpha T}\left(1-e^{-2\alpha T}\right). \qquad \text{(E 4.8)}$$

The average energy of the signal in segment [0, T] is found by setting the lag to zero:

$$E_S \equiv \frac{1}{T}\int_0^T x^2(t)\,dt = \frac{a^2}{4\alpha T}\left(1-e^{-2\alpha T}\right). \qquad \text{(E 4.9)}$$

The 4^{th}-order moment function is evaluated as:

$$\frac{1}{T}\int_0^T x^3(t)\,x(t+\tau)\,dt = \frac{1}{T}\int_0^T a^4 e^{-3\alpha t}e^{-\alpha(t+\tau)}\left[\cos w_0 t\right]^3\left[\cos w_0(t+\tau)\right]dt$$

$$= \frac{a^4 e^{-\alpha\tau}}{T}\int_0^T e^{-4\alpha t}\left[\left(\frac{3}{4}\cos w_0 t + \frac{1}{4}\cos 3w_0 t\right)\cos w_0(t+\tau)\right]dt$$

$$= \frac{a^4 e^{-\alpha\tau}}{T}\int_0^T e^{-4\alpha t}\left[\frac{3}{4}\cos w_0 t\cos w_0(t+\tau) + \frac{1}{4}\cos 3w_0 t\cos w_0(t+\tau)\right]dt$$

$$= \frac{a^4 e^{-\alpha\tau}}{T}\int_0^T e^{-4\alpha t}\left[\frac{3}{8}\cos w_0\tau + \frac{3}{8}\cos w_0(2t+\tau) + \frac{1}{8}\cos w_0(2t-\tau) + \frac{1}{8}\cos w_0(4t+\tau)\right]dt$$

$$= \frac{3a^4 e^{-\alpha\tau}}{8T}\int_0^T e^{-4\alpha t}\cos w_0\tau\,dt = \frac{3a^4 e^{-\alpha\tau}}{8T}\cos w_0\tau\frac{1-e^{-4\alpha T}}{4\alpha}$$

$$= \frac{3a^4 e^{-\alpha\tau}\cos w_0\tau}{32\alpha T}\left(1-e^{-4\alpha T}\right);$$

it is to note that the factor: $F = \frac{a^4}{16\alpha T}\left(1-e^{-4\alpha T}\right)$ may be written in terms of the signal energy, since:

$$\frac{F}{\left(E\left[x^2(n)\right]\right)^2} = \frac{\frac{a^4}{16\alpha T}\left(1-e^{-4\alpha T}\right)}{\left(\frac{a^2}{4\alpha T}\left(1-e^{-2\alpha T}\right)\right)^2} = \alpha T\frac{\left(1-e^{-4\alpha T}\right)}{\left(1-e^{-2\alpha T}\right)^2} = \alpha T\frac{\left(e^{2\alpha T}-e^{-2\alpha T}\right)}{\left(e^{\alpha T}-e^{-\alpha T}\right)^2}$$

$$= \alpha T \frac{2\sinh(2\alpha T)}{[2\sinh(\alpha T)]^2} = \frac{\alpha T}{4}\frac{4\sinh(\alpha T)\cosh(\alpha T)}{[\sinh(\alpha T)]^2} = \alpha T \coth(\alpha T) .$$

Therefore:

$$\frac{a^4}{16\alpha T}(1 - e^{-4\alpha T}) = [\alpha T \coth(\alpha T)][E_S]^2 \qquad \text{(E 4.10)}$$

and using Eq 4.10, the 4th-order moment function may be written in terms of the signal energy as:

$$\frac{1}{T}\int_0^T x^3(t)\,x(t+\tau)\,dt = \frac{3}{2}[\alpha T \coth(\alpha T)]\,e^{-\alpha\tau}\cos w_0\tau\,[E_S]^2 . \qquad \text{(E 4.11)}$$

The 4th-order cumulant function is then derived by substituting Eq 4.8, Eq 4.9 and Eq 4.11 into Eq 4.7:

$$C_4^a[\tau] = \frac{3}{2}[\alpha T \coth(\alpha T)]\,[e^{-\alpha\tau}\cos w_0\tau]\,[E_S]^2 - 3\left(\frac{a^2 e^{-\alpha\tau}}{4\alpha T}(1 - e^{-2\alpha T})\right)^2 \cos(w_0\tau) .$$

The second term in the above equation may be written in terms of signal energy (Eq 4.9) as:

$$\left(\frac{a^2 e^{-\alpha\tau}}{4\alpha T}(1 - e^{-2\alpha T})\right)^2 \cos(w_0\tau) = e^{-\alpha\tau}\cos(w_0\tau)\,[E_S]^2 .$$

Thus the 4th-order cumulant slice can be expressed in terms of the signal energy, the damping factor and the signal frequency as given in Eq 4.6. Furthermore:

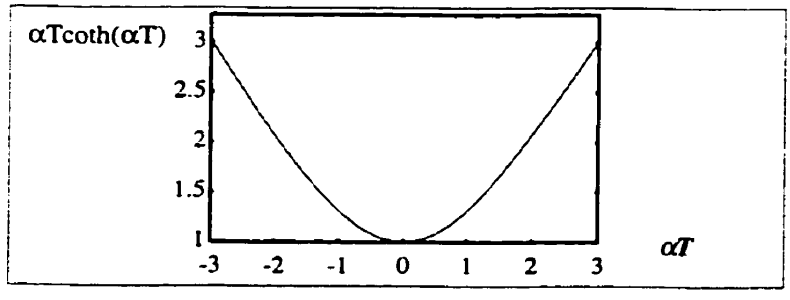
The kurtosis is: $$C_4^a[0] = \left(\frac{3}{2}\alpha T \coth(\alpha T) - 3\right)[E_S]^2 \qquad \text{(E 4.12)}$$

The normalized kurtosis is: $$C_4^a[0] / [E_S]^2 = \frac{3}{2}\alpha T \coth(\alpha T) - 3 \qquad \text{(E 4.13)}$$

The value of the term $\alpha T \coth(\alpha T)$ as a function of αT is shown in Figure 4-3. It is to note here the limiting values of this term:

$$\alpha T \coth(\alpha T) = \begin{cases} 1 & ; \text{for small } \alpha T \\ \alpha T & ; \text{for large } \alpha T \end{cases}$$

Figure 4-3 The value of $\alpha T \coth(\alpha T)$ as a function of (αT)



4.3.2 Steady State Voiced Speech

- **Theorem 5:** *In the case of a single sinusoid with deterministic phase, the diagonal slice of the 4^{th} cumulant may be expressed in terms of signal energy and frequency:*

$$\boxed{C_4^a[\tau] = -1.5\cos(w_0\tau)\,[E_S]^2} \qquad \text{(E 4.14)}$$

- **Proof:** The signal is given by: $x(n) = a \cdot \cos(nPw_0 + \phi)$, with ϕ deterministic and unknown. For a zero-mean deterministic signal, the diagonal slice of the 4^{th}-order cumulant is defined (from Section 3.1.1.2) as:

$$C_4^a(\tau) = \left[\frac{1}{N}\sum_n x(n)\,x^3(n+\tau)\right] - 3\left[\frac{1}{N}\sum_n x^2(n)\right]\cdot\left[\frac{1}{N}\sum_n x(n)\,x(n+\tau)\right]$$

and its Fourier transform (Section 3.3.1):

$$FC_4^{\,a}(w) = X(w)\,[X(w) \otimes X(w) \otimes X(w)] - 3\,[m_2(0)]\,P(w)\,.$$

The value of $FC_4^{\,a}(w)$ for the case of a single sinusoid is illustrated in Figure 4-4. The values at the different lags are shown in Table 4-1 below.

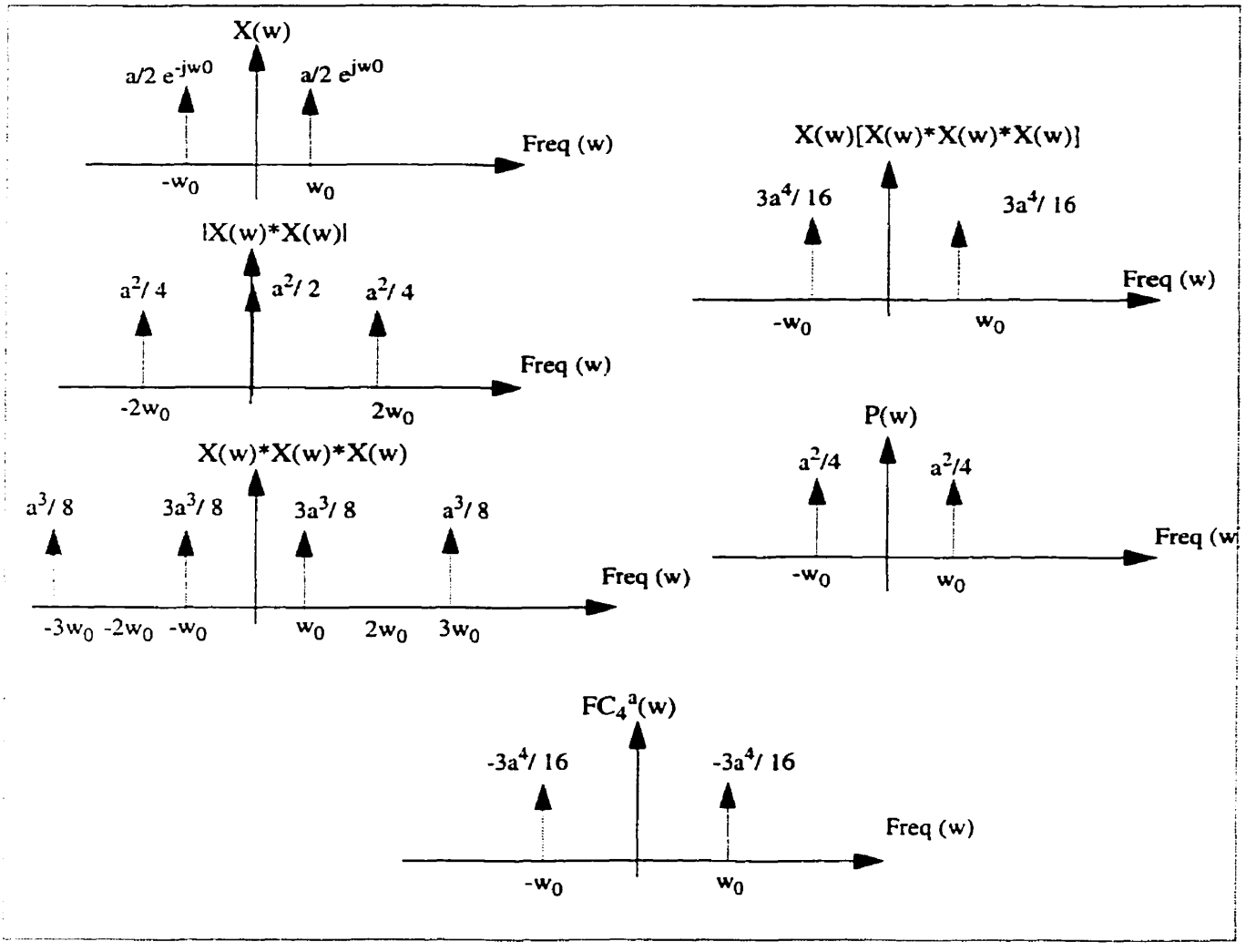
Table 4-1 $FC^a_4(w)$ for a single sinusoid with a deterministic phase

Lag (w)	X(w)*X(w)	X(w)*X(w)*X(w)	X(w)[X(w)*X(w)*X(w)]	P(w)	$FC^a_4(w)$
$-3w_0$	0	$a^3/8$	0	0	0
$-2w_0$	$a^2/4$	0	0	0	0
$-w_0$	0	$3a^3/8$	$3a^4/16$	$a^2/4$	$-3a^4/16$
0	$a^2/2$	0	0	0	0
w_0	0	$3a^3/8$	$3a^4/16$	$a^2/4$	$-3a^4/16$
$2w_0$	$a^2/4$	0	0	0	0
$2w_0$	0	$a^3/8$	0	0	0

The cumulant slice is found by inverse transforming $FC_4^{\,a}(w)$:

$$C_4^a[\tau] = \frac{-3a^4}{8}\,[\cos(w_0\tau)]\,. \qquad \text{(E 4.15)}$$

Figure 4-4 $FC_4^a(w)$ for the case of one sinusoid with deterministic phase



Since the signal energy is $E_S = (a^2/2)$, then the 4th-order cumulant may be written in terms of signal energy and frequency (Eq 4.14). Furthermore:

The kurtosis is: $\boxed{C_4^a[0] = \frac{-3a^4}{8} = -1.5\,[E_S]^2}$ **(E 4.16)**

The normalized kurtosis is: $\boxed{C_4^a[0] / [E_S]^2 = -1.5}$ **(E 4.17)**

• **Theorem 6:** *In the case where voiced speech is modeled as the sum of two sinusoids with deterministic but unknown phases, the diagonal slice of the 4^{th}-order cumulant may be written in terms of the speech amplitudes and frequencies:*

$$\boxed{C_4^a[\tau] = \frac{-3}{8}\left[a_1^4\cos(w_1\tau) + a_2^4\cos(w_2\tau)\right]} \quad \text{(E 4.18)}$$

and has zero phase regardless of the phase of the underlying segment. Moreover, the kurtosis is upper and lower bounded by scale factors of the signal energy:

$$\boxed{-1.5\,[E_S]^2 \le Kurtosis \le -0.75\,[E_S]^2} \quad \text{(E 4.19)}$$

• **Proof:** The signal is modeled as (using the continuous time notation):

$$x(t) = a_1\cos(w_1 t + \phi_1) + a_2\cos(w_2 t + \phi_2)\ .$$

To simplify notation, the following variables are defined:

$$X_1 \equiv w_1 t + \phi_1\ ,\ X_1' \equiv w_1(t+\tau) + \phi_1\ ,\ X_2 \equiv w_2 t + \phi_2\ ,\ X_2' \equiv w_2(t+\tau) + \phi_2\ ,$$

The expression for the cumulant slice is given by (from Section 3.1.1.2):

$$C_4^a[\tau] = \left[\frac{1}{T}\int_0^T x^3(t)\,x(t+\tau)\,dt\right] - 3\left[\frac{1}{T}\int_0^T x^2(t)\,dt\right]\left[\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt\right]. \quad \text{(E 4.20)}$$

The second moment is first evaluated as:

$$\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt = \frac{1}{T}\left[\begin{array}{l}\displaystyle\int_0^T a_1^2\cos(X_1)\cos(X_1')\,dt + \int_0^T a_2^2\cos(X_2)\cos(X_2')\,dt + \\ \displaystyle\int_0^T a_1a_2\cos(X_1)\cos(X_2')\,dt + \int_0^T a_1a_2\cos(X_1')\cos(X_2)\,dt\end{array}\right].$$

Using the trigonometric identity for $\cos(a)\cos(b)$ and noting that any integral of the form $\int_0^T \cos(Awt + \phi)\,dt$ is identically zero for integer values of A whenever T is a multiple of the signal period, the second moment becomes:

$$\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt = \frac{a_1^2}{2}\cos(w_1\tau) + \frac{a_2^2}{2}\cos(w_2\tau) \quad \text{(E 4.21)}$$

and the signal energy is:

$$\frac{1}{T}\int_0^T x^2(t)\,dt = \left(\frac{a_1^2}{2} + \frac{a_2^2}{2}\right). \qquad \text{(E 4.22)}$$

To evaluate the 4th-order moment, the following are noted:

$$x^3(t) = a_1^3[\cos X_1]^3 + a_2^3[\cos X_2]^3 + 3a_1^2a_2[\cos X_1]^2[\cos X_2] + 3a_1a_2^2[\cos X_1][\cos X_2]^2.$$

Thus,

$$\int_0^T x^3(t)\,x(t+\tau)\,dt = \int_0^T a_1^4[\cos X_1]^3[\cos X_1']\,dt + \int_0^T a_1^3a_2[\cos X_1]^3[\cos X_2']\,dt$$
$$+\int_0^T a_2^4[\cos X_2]^3[\cos X_2']\,dt + \int_0^T a_2^3a_1[\cos X_2]^3[\cos X_1']\,dt$$
$$+\int_0^T 3a_1^3a_2[\cos X_1]^2[\cos X_2][\cos X_1']\,dt + \int_0^T 3a_2^2a_1^2[\cos X_1]^2[\cos X_2][\cos X_2']\,dt$$
$$+\int_0^T 3a_2^2a_1^2[\cos X_2]^2[\cos X_1][\cos X_1']\,dt + \int_0^T 3a_2^3a_1[\cos X_2]^2[\cos X_1][\cos X_2']\,dt.$$

Assuming that T is a multiple of signal period, and using the appropriate trigonometric identities, the following are noted:

$$\int_0^T [\cos X_2]^3[\cos X_1']\,dt = \int_0^T [\cos X_1]^3[\cos X_2']\,dt = 0$$

$$\int_0^T [\cos X_1]^2[\cos X_2][\cos X_1']\,dt = \int_0^T [\cos X_2]^2[\cos X_1][\cos X_2']\,dt = 0$$

$$\frac{1}{T}\int_0^T [\cos X_1]^3[\cos X_1']\,dt = \frac{3}{8}\cos(w_1\tau)$$

$$\frac{1}{T}\int_0^T [\cos X_2]^3[\cos X_2']\,dt = \frac{3}{8}\cos(w_2\tau)$$

$$\frac{1}{T}\int_0^T [\cos X_1]^2[\cos X_2][\cos X_2']\,dt = \frac{1}{4}\cos(w_2\tau)$$

$$\frac{1}{T}\int_0^T [\cos X_2]^2 [\cos X_1] [\cos X_1'] dt = \frac{1}{4}\cos(w_1\tau) .$$

The 4th-order moment function becomes:

$$\frac{1}{T}\int_0^T x^3(t)\, x(t+\tau)\, dt = \frac{3a_1^4}{8}\cos(w_1\tau) + \frac{3a_2^4}{8}\cos(w_2\tau) + \frac{3a_1^2a_2^2}{4}\cos(w_1\tau) + \frac{3a_1^2a_2^2}{4}\cos(w_2\tau) \quad \text{(E 4.23)}$$

substituting Eq 4.21, Eq 4.22 and Eq 4.23 into Eq 4.20 yields the expression for the cumulant slice given in Eq 4.18. Furthermore:

The kurtosis is: $$\boxed{C_4^a[0] = \frac{-3(a_1^4 + a_2^4)}{8}} \quad \text{(E 4.24)}$$

The normalized kurtosis is: $$\boxed{C_4^a[0] / [E_S]^2 = \frac{-3(a_1^4 + a_2^4)}{4(a_1^2 + a_2^2)}} \quad \text{(E 4.25)}$$

To determine the bounds on the kurtosis, the two extreme cases on the amplitudes are considered:

- **Case 1:** $a_1 \approx a_2$. The signal energy is: $E_S \approx a_1^2$. The kurtosis becomes: $C_4^a[0] \approx -3a_1^4/4$, or in terms of signal energy: $C_4^a[0] = -0.75[E_S]^2$.
- **Case 2:** $a_1 \gg a_2$. The signal energy is: $E_S \approx a_1^2/2$. The kurtosis becomes: $C_4^a[0] \approx -3a_1^4/8$, or in terms of signal energy: $C_4^a[0] \approx -1.5[E_S]^2$.

Therefore, for any values of the two amplitudes, the kurtosis is bounded by the signal energy as given in Eq 4.19.

4.3.3 Unvoiced Speech

- **Theorem 7:** *In the case where unvoiced speech is modeled as the sum of two sinusoids with random phases* $x(n) = a_1\cos(nPw_1 + \phi_1) + a_2\cos(nPw_2 + \phi_2)$ *with* $\phi_1 \in [-\pi, \pi]$ *and* $\phi_2 \in [-\pi, \pi]$, *the diagonal slice of the 4th-order cumulant may be written in terms of the signal amplitudes and frequencies:*

$$\boxed{C_4^a[\tau] = \frac{-3}{8}[a_1^4\cos(w_1\tau) + a_2^4\cos(w_2\tau)]} \quad \text{(E 4.26)}$$

• **Proof:**

1. Single sinusoid with random phase

The signal is modeled as (using the continuous time notation): $x(t) = a\cos(wt + \theta)$. The diagonal slice is given (from Section 3.1.1.2) by:

$$C_4^a[\tau] = E[x^3(t)\,x(t+\tau)] - 3 \cdot E[x^2(t)] \cdot E[x(t)\,x(t+\tau)] \qquad \text{(E 4.27)}$$

The 2nd-order moment function is first evaluated as:

$$E[x(t)\,x(t+\tau)] = a^2 E[\cos(w_0 t + \theta)\cos(w_0(t+\tau) + \theta)]$$

$$= \frac{a^2}{2}E[\cos(2w_0 t + w_0\tau + 2\theta)] + \frac{a^2}{2}E[\cos(w_0\tau)]$$

$$= \frac{a^2}{2}\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos[2w_0 t + w_0\tau + 2y]\,dy + \frac{a^2}{2}\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos[w_0\tau]\,dy$$

$$E[x(t)\,x(t+\tau)] = \frac{a^2}{2}\cos[w_0\tau]\,. \qquad \text{(E 4.28)}$$

The average energy of the process is

$$E_s \equiv E[(x(t))^2] = a^2/2\,. \qquad \text{(E 4.29)}$$

The 4th-order moment function is evaluated as:

$$E[x^3(t)\,x(t+\tau)] = a^4 E[\{\cos(w_0 t + \theta)\}^3\{\cos(w_0(t+\tau) + \theta)\}]$$

$$= a^4 E\left[\left(\frac{3}{4}\cos(w_0 t + \theta) + \frac{1}{4}\cos 3(w_0 t + \theta)\right)\cos(w_0(t+\tau) + \theta)\right]$$

$$= \frac{a^4}{4}\left[\begin{array}{c} \frac{3}{2}E[\cos(2w_0 t + w_0\tau + 2\theta)] + \frac{3}{2}E[\cos(w_0\tau)] + \\ \frac{1}{2}E[\cos(4w_0 t + 3w_0\tau + 4\theta)] + \frac{1}{2}E[\cos(2w_0 t + 3w_0\tau + 3\theta)] \end{array}\right]$$

$$E[x^3(t)\,x(t+\tau)] = \frac{3a^4}{8}\cos(w_0\tau) \qquad \text{(E 4.30)}$$

and substituting Eq 4.28, Eq 4.29 and Eq 4.30 in Eq 4.27, the diagonal slice becomes:

$$C_4^a[\tau] = \frac{-3a^4}{8}\cos(w_0\tau)\,. \qquad \text{(E 4.31)}$$

2. Case of two sinusoids

Since the two sinusoids are statistically independent, then the cumulant of the sum is the sum of the cumulants. Using the results for the case of one sinusoid (Eq 4.31), the resulting cumulant is the one given in Eq 4.26. Furthermore:

The kurtosis is: $$\boxed{C_4^a[0] = \frac{-3a^4}{8}(a_1^4 + a_2^4)} \quad \text{(E 4.32)}$$

The normalized kurtosis is: $$\boxed{C_4^a[0] / [E_S]^2 = \frac{-3(a_1^4 + a_2^4)}{4(a_1^2 + a_2^2)}} \quad \text{(E 4.33)}$$

It is to note that the expression for the 4th-order cumulant is the same as in the case of steady state voiced speech, that is whether the phases are deterministic or random does not change the cumulant. The derivations are also in agreement with the more general expressions for harmonic signals found in [Swa91]. As before, the kurtosis has lower and upper bounds that are a function of the speech energy:

$$-1.5[E_S]^2 \le Kurtosis \le -0.75[E_S]^2. \quad \text{(E 4.34)}$$

4.3.4 Effect of noise on the normalized kurtosis

In the case of steady voiced and unvoiced speech, it is shown that an upper and lower bound on the kurtosis can be specified in terms of the speech energy. When normalized by the square of the second moment ($[E_S]^2$), the normalized kurtosis is independent of signal energy and is bounded by:

$$-1.5 \le \gamma_4 \le -0.75. \quad \text{(E 4.35)}$$

When the signal consists of both speech and noise, then $x(n) = s(n) + g(n)$. If *s(n)* and *g(n)* are statistically independent, then the energy of *x(n)* is the sum of speech and noise energies: $E_x = E_s + E_g$. Second-order statistics are thus directly affected and in an additive way by the presence of noise. Higher-order statistics on the other hand are immune to Gaussian noise, which has zero HOS. Since cumulants are cumulative [Men91], it follows that the bounds on the kurtosis still hold in terms of speech energy. However, when normalizing the kurtosis by the total signal energy E_x, the effect of the noise term in the denominator does not cancel out with the speech energy term E_s in the numerator in Eq 4.25 and Eq 4.34. It is easy to see that the bounds on the normalized kurtosis (Eq 4.35) can now be extended to include an SNR term as follows:

$$-1.5\frac{[E_S]^2}{[E_S + E_g]^2} \le \gamma_4 \le -0.75\frac{[E_S]^2}{[E_S + E_g]^2}$$

$$-1.5\left[\frac{SNR}{SNR+1}\right]^2 \le \gamma_4 \le -0.75\left[\frac{SNR}{SNR+1}\right]^2. \qquad \text{(E 4.36)}$$

Therefore depending on the SNR, the normalized kurtosis could be larger than -0.75, but it is always negative and cannot be smaller than -1.5.

4.4 Summary of the Derivations

The above derivations were based on modeling a short-term segment of speech according to a sinusoidal model, where it was assumed that a narrow subbanding is used such that at most two harmonics fall in each band. Steady voiced speech is modeled as a sum of two sinusoids of deterministic but unknown phase. Unvoiced speech is modeled as the sum of two incoherent sinusoids of random and uniformly distributed phases. Transient speech is modeled as an exponentially decaying sinusoid. From the analytical derivations, the following are noted:

- The 3rd-order cumulant of subbanded speech is identically zero, for unvoiced and transient segments. For steady state voiced segments, this cumulant is non-zero only in one condition, that is when the two frequencies are harmonically related. This is only likely to happen in the first two bands of speech, but the rare occurrence of this condition suggests that 3rd-order statistics are in general not useful in the subband domain.
- The 4th-order cumulant function of subbanded speech is non-zero, and may be expressed in terms of the parameters of the underlying speech.
- In the case of transient speech, the 4th-order cumulant slice may be expressed in terms of the signal energy, frequency and damping factor:

  The diagonal slice: $C_4^a[\tau] = 3\left[\frac{1}{2}\alpha T \coth(\alpha T) - 1\right] e^{-\alpha\tau}\cos(w_0\tau)\,[E_S]^2$.

  The kurtosis: $C_4^a[0] = \left(\frac{3}{2}\alpha T \coth(\alpha T) - 3\right)[E_S]^2$.

  The normalized kurtosis: $C_4^a[0]/[E_S]^2 = \frac{3}{2}\alpha T \coth(\alpha T) - 3$.

  Therefore, it is conceivable to estimate the signal energy and damping factor from the cumulant values at the first few lags. Moreover, due to the value of the factor: $\alpha T \coth(\alpha T)$, the normalized kurtosis may assume any positive or negative values, including zero. For this

reason, it is not easy to distinguish transient speech from Gaussian noise based on the normalized kurtosis.

- In the case of steady state voiced speech, the cumulant slice may be expressed in terms of the 4^{th} power of the signal harmonic amplitudes and frequencies. This cumulant has the same harmonic nature as the underlying speech and has zero phase regardless of the phase of the underlying signal. The kurtosis is written in terms of the 4^{th} powers of the signal amplitudes and by considering the two limiting cases on these amplitudes, an upper and lower bound for the kurtosis in terms of signal energy may be deduced:

Diagonal slice: $$C_4^a[\tau] = \frac{-3a_1^4}{8}\cos(w_1\tau) - \frac{3a_2^4}{8}\cos(w_2\tau) .$$

Kurtosis: $$C_4^a[0] = \frac{-3(a_1^4 + a_2^4)}{8} .$$

The normalized kurtosis: $$C_4^a[0] / [E_S]^2 = \frac{-3(a_1^4 + a_2^4)}{4(a_1^2 + a_2^2)} .$$

$$-1.5[E_S]^2 \leq Kurtosis \leq -0.75[E_S]^2 .$$

Consequently, the kurtosis of voiced speech is always negative and the normalized kurtosis is contained in the interval [-1.5, -0.75]. When noise is present, the normalized kurtosis may assume any negative value in the interval [-1.5, 0].

- The 4^{th}-order cumulant slice of unvoiced speech has the same expression and behaviour as steady state voiced speech. It is an interesting fact to note that regardless of whether the phases are deterministic or random, the expression for the cumulant still holds:

Diagonal slice: $$C_4^a[\tau] = \frac{-3}{8}[a_1^4\cos(w_1\tau) + a_2^4\cos(w_2\tau)] .$$

The kurtosis: $$C_4^a[0] = \frac{-3}{8}(a_1^4 + a_2^4) .$$

4.5 Simulation Results Using Speech Signals

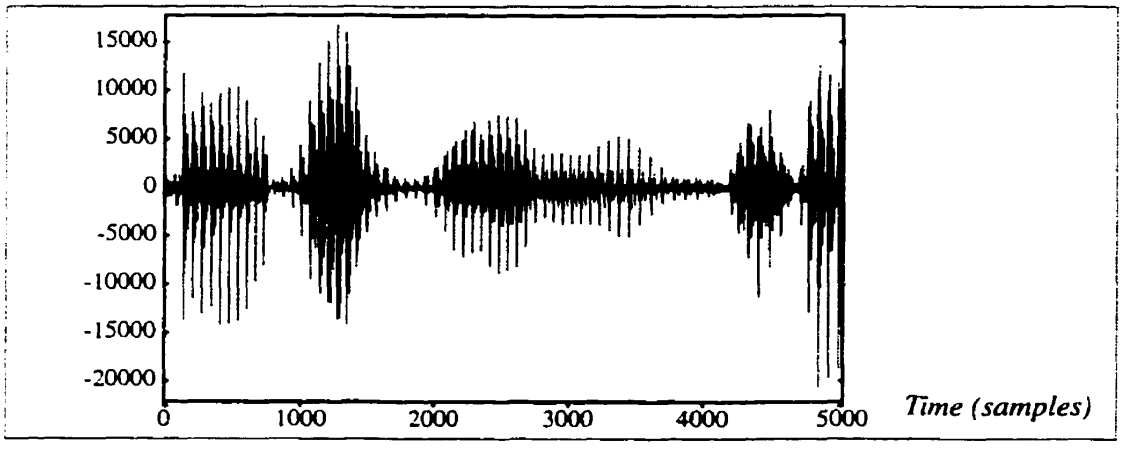
Clean speech recorded and sampled at 8 kHz is used to verify the above derivations. The speech is sub-banded using 50 cosine modulated filters. In each band, the normalized skewness and normalized kurtosis are computed as well as the diagonal slice of the 4^{th}-order cumulant. Analysis is done in frames of 120 points with 33% overlap. Computations are carried out for segments of voiced speech, unvoiced speech and Gaussian noise.

4.5.1 Voiced Speech

4.5.1.1 Skewness and kurtosis

A section of voiced speech is shown in Figure 4-5. The normalized kurtosis in three consecutive bands in the lower spectrum is computed for a number of consecutive frames. The results are shown in Figure 4-6. It is clear that the values are within the range expected for voiced speech (i.e, between -0.75 and -1.5) for the majority of the time. Clearly some small or even positive values may occur due to the presence of transitional segments or pauses.

Figure 4-5 Segments of voiced speech



The normalized kurtosis in the upper bands (Figure 4-7) assumes both positive and negative values. The fact that the negative ones are in the vicinity of -1.0 reinforces the validity of the sinusoidal model. The fact that the kurtosis also takes on large positive value in some segments indicates the transient nature of the speech in the upper spectrum. The skewness of voiced speech (shown in Figure 4-8) is near zero, as expected.

Figure 4-6 Normalized kurtosis in three consecutive lower bands

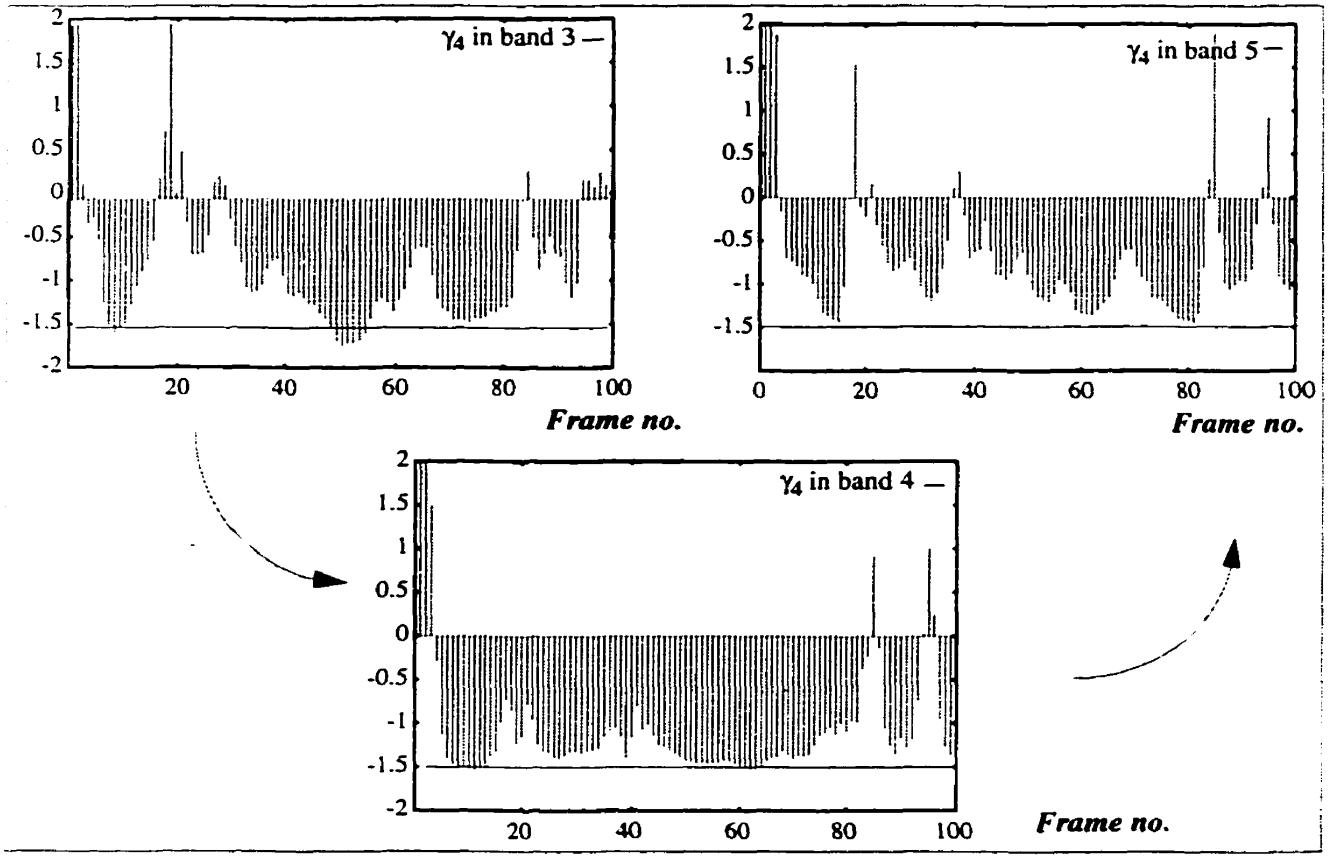


Figure 4-7 Normalized kurtosis in an upper band

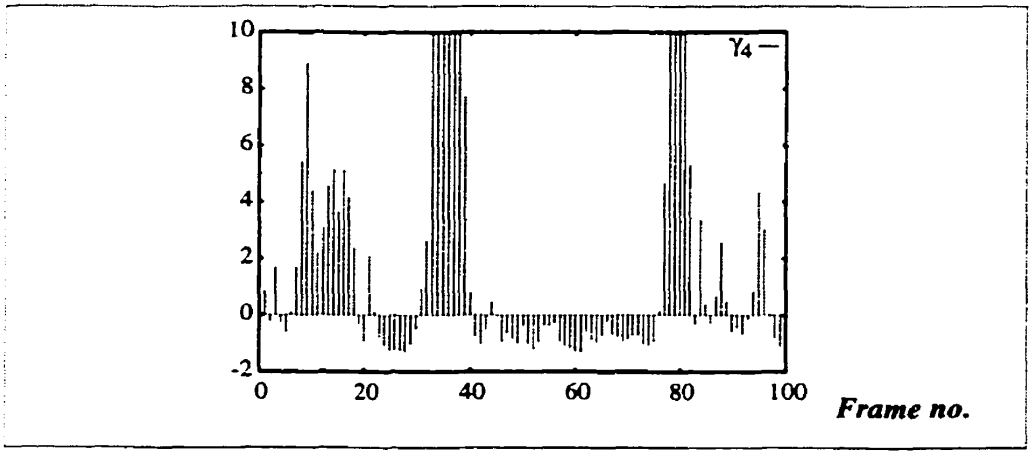
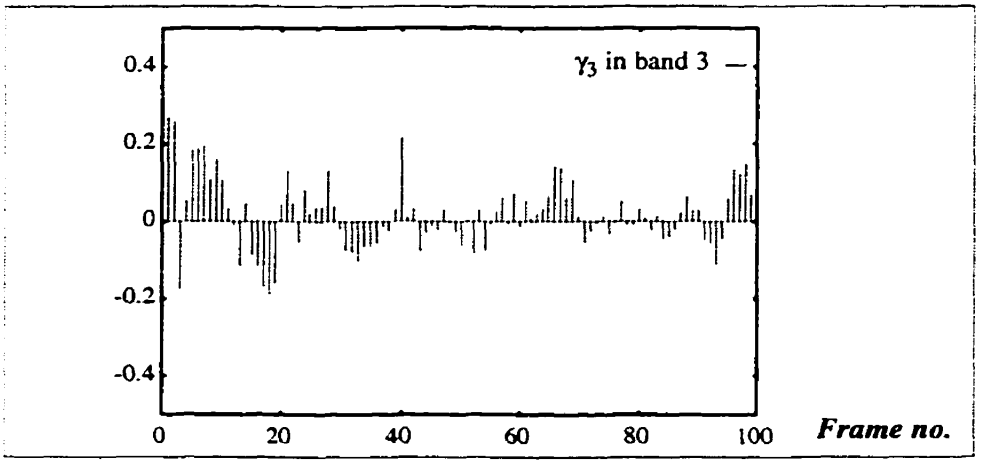
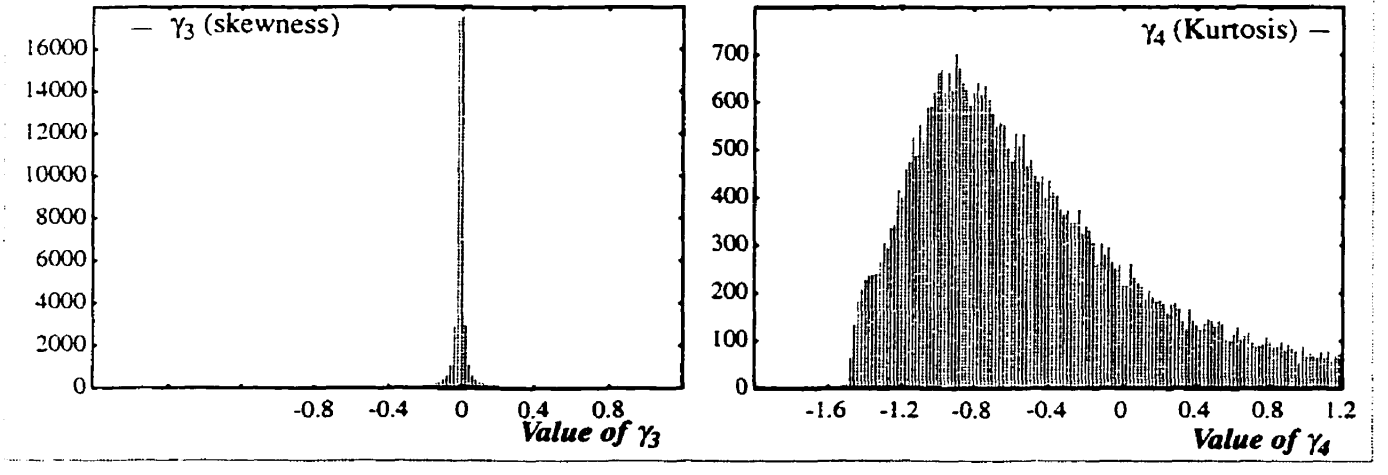


Figure 4-8 Normalized skewness in a lower band



The histograms for the normalized skewness and kurtosis for a speech file of about 1000 frames were collected. These are shown in Figure 4-9 below. The skewness is clearly zero, as expected; the normalized kurtosis has a distribution around -0.8, which is within the range that is expected given that a typical 7 seconds of speech would contain a mixture of voiced, unvoiced and non-speech frames.

Figure 4-9 Histograms of γ_3 and γ_4 across all bands

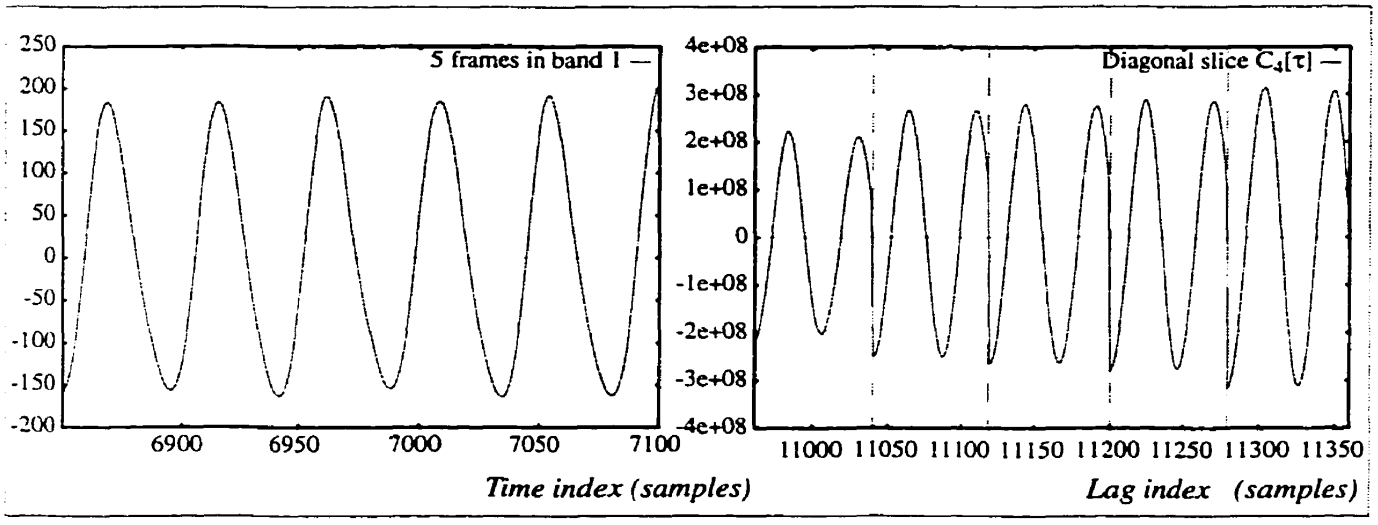


4.5.1.2 Diagonal cumulant slice

The diagonal slice of the 4^{th}-order cumulant ($C_4^a[\tau]$) is computed for τ in the range [0, 80] for a number of consecutive frames in a lower band and an upper band of voiced speech.

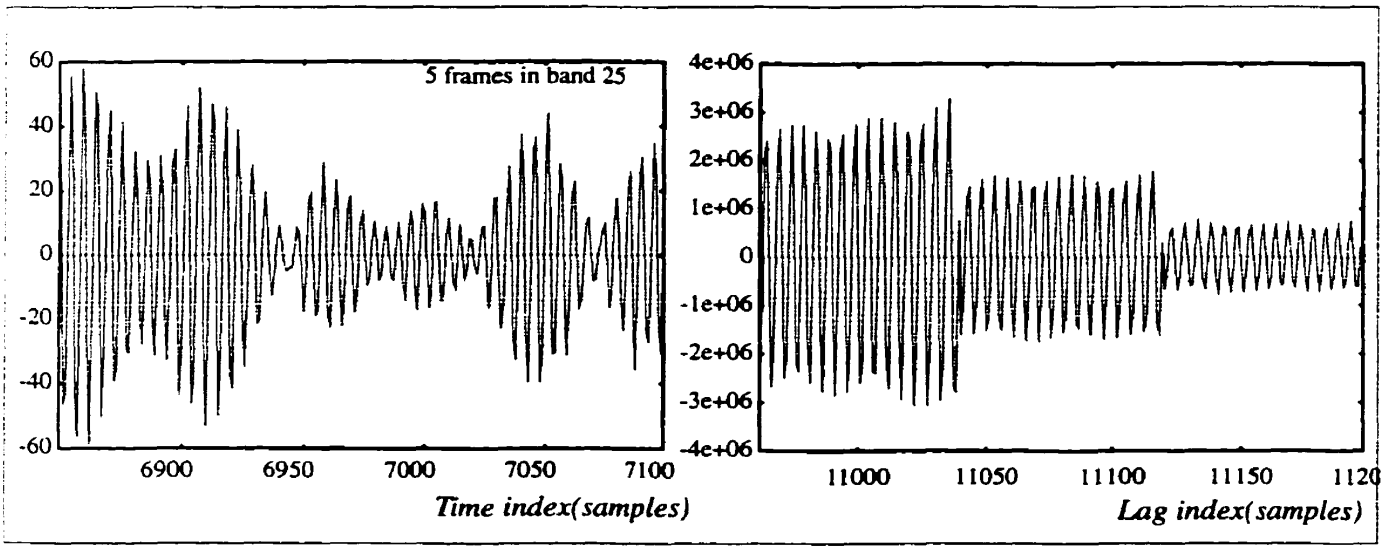
A sequence of voiced speech in band 1 (120 Hz center) is shown in Figure 4-10. It is evident here that the signal may be represented as a single sinusoid. The diagonal slice is computed for lags 0 to 80. The start of each slice is indicated by dotted lines, and it is clear the slice has 0^o (180^o) phase as expected with the value at zero lag (the kurtosis) having the largest magnitude.

Figure 4-10 Voiced speech in band 1 along with the computed diagonal slice



A sequence in band 25 is shown in Figure 4-11 along with the diagonal cumulant slice of three segments. According to the derivations, the cumulant slice is the sum of two sinewaves and has zero phase. It is clear from the plot of the slice that it can be reasonably modeled as predicted by the analytical derivations.

Figure 4-11 Voiced speech in band 25 along with the computed diagonal slice



4.5.2 Unvoiced Speech

Two sustained fricatives, namely /h/ and /f/ were recorded for a few seconds and used for better analysis of the HOS of unvoiced speech. The speech waveform for /h/ is shown in Figure 4-12 and its normalized skewness and kurtosis in an upper band of the spectrum are shown in Figure 4-13. It is not clear what is the behavior of the two statistics when looking at the values for consecutive frames, but it is certain that the value for the kurtosis does not converge to -1.5 as predicted by the derivations based on the sinusoidal model. To get a better understanding, the values of the two entities across all the bands of the upper spectrum were computed and two histograms generated. These are shown in Figure 4-14. The same analysis was done using Gaussian noise instead of sustained unvoiced speech. The histograms are shown in Figure 4-15. By comparing the two sets of histograms, it is concluded that sustained unvoiced speech in the subband domain is likely Gaussian and cannot be modeled as a harmonic process as suggested by the sinusoidal model.

However, since unvoiced segments are usually short and occur at speech boundaries, their HOS is generally non-zero. This fact is in agreement with the reported findings in [Fal93].

Figure 4-12 The unvoiced phoneme /h/

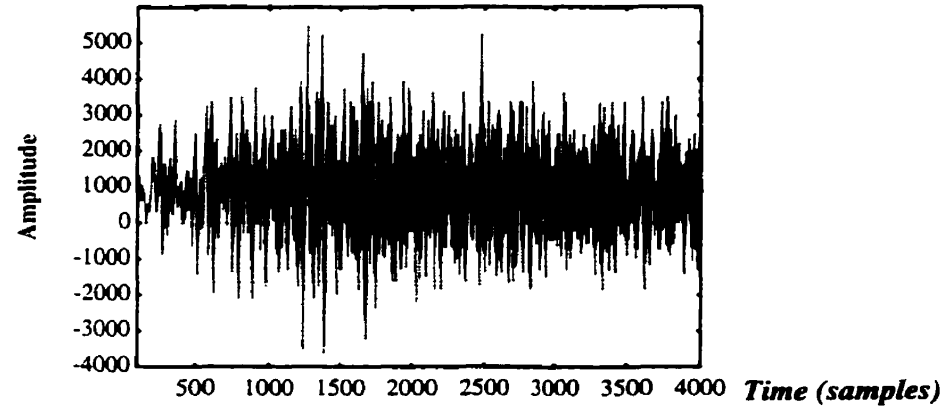


Figure 4-13 Normalized skewness and kurtosis of unvoiced speech (/h/) in an upper band

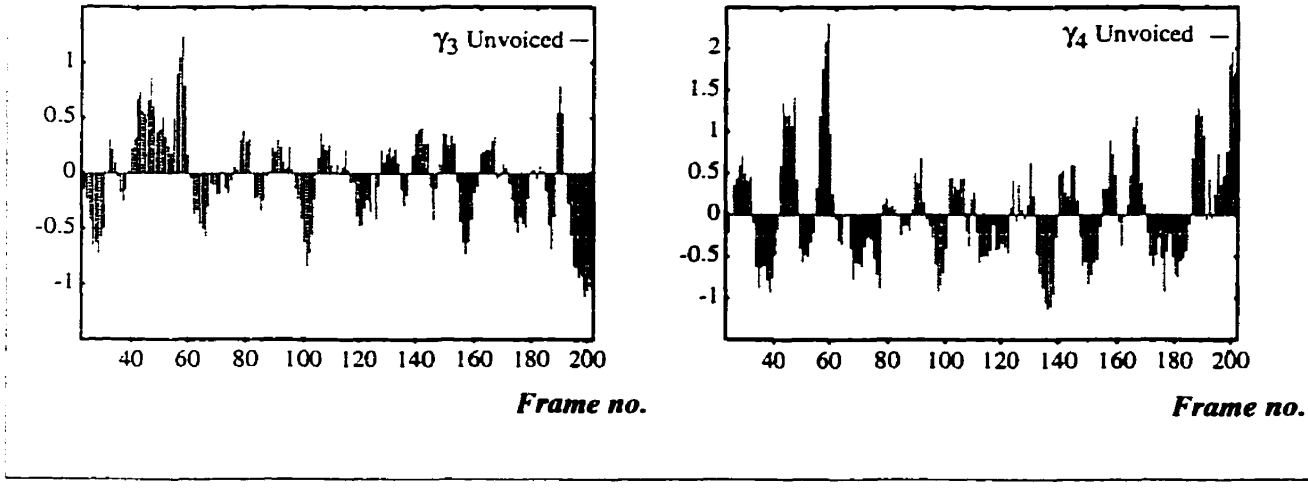


Figure 4-14 Histograms of γ_3 and γ_4 across the upper bands of unvoiced speech

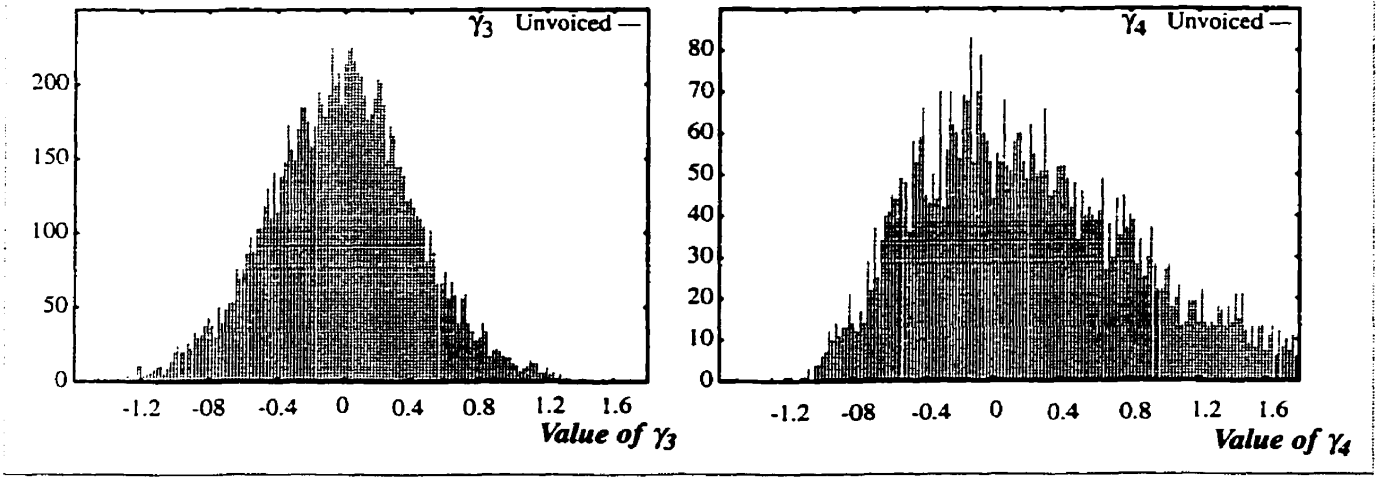
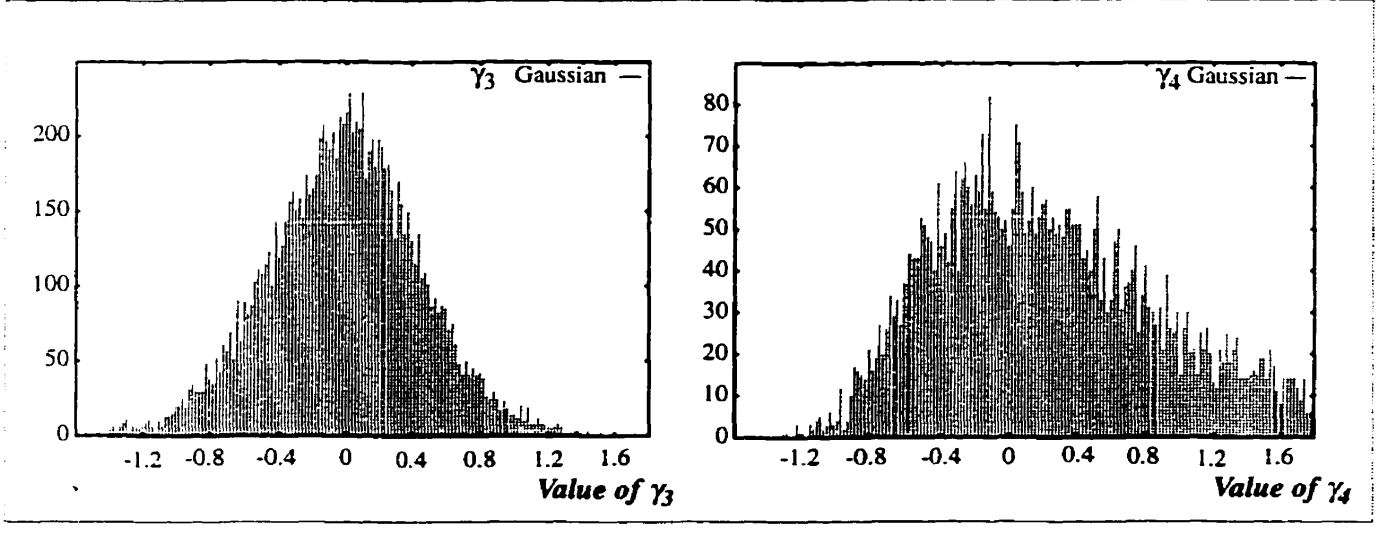


Figure 4-15 Histograms of γ_3 and γ_4 across the upper bands of Gaussian noise



4.6 Conclusion

The McAulay sinusoidal model [McA86] is fairly accurate for modeling steady voiced speech segments but the HOS analysis of unvoiced speech using speech data showed that this signal has a Gaussian nature and cannot be considered a harmonic process as the model suggests. When a narrow subbanding scheme is used, steady voiced speech may be modeled as the sum of two sinusoids in each band. The model however is not appropriate for transitional segments, which may be more appropriately represented by an exponentially decaying sinusoid.

The 3^{rd}-order cumulant of subbanded speech is identically zero, except in rare situations, which suggests that this cumulant is generally not useful in the subband domain.

The 4^{th}-order cumulant of voiced speech is non-zero and may be expressed in terms of speech parameters. In the transient case, it is expressed in terms of signal energy and frequency. In the steady state case, it is expressed in terms of signal harmonic frequencies and amplitudes. Therefore, it is conceivable to estimate these parameters using the value of the 4^{th}-order cumulant at a few lags.

The normalized kurtosis of transient speech may assume any positive or negative values and as such is not a sufficient metric for distinguishing speech from noise. The normalized kurtosis of steady voiced is contained in the interval [-1.5, -0.75]. When noise is present the normalization may cause smaller amplitude, but the metric is always negative and cannot be smaller than -1.5.

The properties thus derived analytically and verified by simulation prove to be quite interesting: The 4^{th}-order statistic of speech allows detecting the presence of speech harmonics and provides an upper and lower bound on the speech energy. These, in turn, provide bounds on the noise present in that band. These findings will be exploited in the next chapter in the context of speech enhancement.

CHAPTER 5 Application of Higher Order Cumulants to Speech Enhancement

Synopsis

A new method for speech enhancement based on optimal filtering, subbands, and Higher-Order Cumulants is proposed in this chapter. The key idea is to use the HOC to estimate the parameters required for the enhancement filters, namely the 2^{nd}-order statistics of the speech and noise. It is shown that the kurtosis and the diagonal slice of the 4^{th}-order cumulant may be used to estimate such parameters as the SNR, speech autocorrelation and the probability of speech presence when speech is divided into narrow bands as explained in Chapter 4. The resulting algorithm is tested in typical mobile noise conditions and proves effective under such types as street, office and fan noises. Performance comparisons with the TIA noise reduction algorithm [IS127] are reported for the noise types used.

5.1 Motivation and Rationale

Speech enhancement by spectral decomposing and filtering [Cap94] [Eph84][Yan93][Sal94] remains a common and effective approach for enhancing speech degraded by acoustic additive noise when only the noisy speech is available. This general class is based on optimal filters and encompasses such methods as Wiener filtering, spectral subtraction, and maximum likelihood (ML) estimations. A common set of requirements in this class includes:

- An appropriate suppression rule that is based on some criteria of optimality.
- An estimation of the speech and noise power spectral densities, or equivalently their respective autocorrelation.

- A quantification of the probability of speech presence in a given band, to further attenuate non-speech bands.
- A method for reducing residual noise by appropriately smoothing the estimated quantities and/or exploiting the psychoacoustic properties of human hearing.

The choice of suppression rules is governed by many factors, such as computational efficiency, optimality criteria, and the exploiting of the human hearing properties. In the reported literature, the range includes heuristic rules (e.g. [Hoe97]) as well as formally derived ones. The ML estimation approaches in [McA80] and [Eph84] attempt to better exploit the statistical properties of the DFT of noisy speech. These approaches assume a statistical model of the Fourier coefficients of speech and derive optimal estimators of the magnitude spectrum based on that model.

Another important contribution is the smoothing approach proposed in [Eph84] whereby the variation in SNR between successive frames is reduced significantly by averaging the locally computed SNR (SNR_{post}) with the SNR estimated in the previous frame after the filtering operation (SNR_{est}). The method results in a significant reduction in the noise artifacts, particularly in low-SNR frames, as was shown in [Cap94].

While much of the published work has focused on appropriate suppression rules, little has been done in the other aspects, not the least being the estimation of the 2^{nd}-order statistics of the speech and noise, such as the SNR, which remains a crucial aspect for effective enhancement, or the quantification of the uncertainty of speech presence which is shown to improve the suppression of noise residuals [Sca96] under a number of suppression rules.

The idea of using HOC for estimating these parameters hinges on being able to separate speech and noise based on these statistics and express the HOC of speech in terms of the desired parameters. It was shown in Chapter 4 that when a subbanding scheme was used and some analytical model for speech assumed, a number of speech parameters may be derived from the 4^{th}-order cumulant. For example, it was shown that this cumulant may be expressed in terms of speech amplitudes for the case of steady voiced speech and in terms of the speech energy in the case of transient speech. It was also shown that the kurtosis of subbanded speech is different from that of Gaussian noise and may be used as discriminator.

These basic ideas are combined in the context of a speech enhancement algorithm that is based on using optimal filters in each subband to estimate the clean speech. The parameters of these filters are derived from the 4^{th}-order cumulant and smoothed appropriately to yield effective enhancement.

5.2 Speech Enhancement by Optimal Filtering

5.2.1 Optimum Linear Systems

In the general linear estimation problem, a discrete-time, zero-mean process X_α is observed over a certain time interval $I = \{n-a, \ldots, n+b\}$ and $(a+b+1)$ observations $\{X_{n-a}, \ldots, X_n, \ldots, X_{n+b}\}$ are used to compute an estimate $\hat{S}_n$ of some other process S_n. This estimate is required to be linear, thus:

$$\hat{S}_n = \sum_{\beta=n-a}^{n+b} h_{n-\beta} \cdot X_\beta = \sum_{\beta=-b}^{a} h_\beta \cdot X_{n-\beta}. \qquad \text{(E 5.1)}$$

The figure of merit of the estimator is the mean-square error $E[e_n^2] = E[(S_n - \hat{S}_n)^2]$. The problem is thus to seek the optimum filter h_β which minimizes the MSE.

The derivations of the optimal filter that meets this objective is based on the orthogonality condition: By making the error orthogonal to all observations X_α, it can be shown [Leo89] that the filter coefficients may be computed by solving the $(a+b+1)$ linear equations:

$$R_{SX}[\tau] = \sum_{\beta=0}^{p} h_\beta R_X[\tau-\beta] \qquad \tau \in \{0, 1, \ldots p\} \qquad \text{(E 5.2)}$$

where $(p+1)$ is the filter order, $R_X[\tau]$ is the autocorrelation of the observation $X(n)$ and $R_{SX}[\tau]$ is the cross correlation of the observed and desired processes $X(n)$ and $S(n)$ respectively.

5.2.2 Filtering Speech Plus Noise

The problem of extracting a signal from noise entails estimating a desired process $S(n)$ from the $(p+1)$ most recent noisy observations:

$$X_\alpha = S_\alpha + N_\alpha \qquad \alpha \in I = \{n-p, \ldots, n\}. \qquad \text{(E 5.3)}$$

If S_α and N_α are independent random processes, then:

$$R_{SX}[\tau] = R_S[\tau] \text{ and } R_X[\tau] = R_S[\tau] + R_N[\tau],$$

and Eq 5.2 becomes:

$$R_S[\tau] = \sum_{\beta=0}^{p} h_\beta \{R_S[\tau-\beta] - R_N[\tau-\beta]\} \qquad \tau \in \{0, 1, \ldots p\} \ . \tag{E 5.4}$$

This set of $(p+1)$ linear equations in $(p+1)$ unknowns h_β may be solved by matrix inversion.

5.2.3 Filtering Subbanded Speech

5.2.3.1 General solution for a p^{th}-order filter

It is assumed that speech is divided into narrow subbands as explained in Chapter 4. Let the autocorrelation of speech be denoted by $R_S[\tau]$. Since the bands are narrow, it is assumed that noise has a flat spectral characteristic in each band, so its autocorrelation is:

$$R_N[\tau] = E_N \delta[\tau] \ . \tag{E 5.5}$$

Assuming that speech and noise are statistically independent, then the system of equations given by Eq 5.4 may be written in the following matrix form:

$$\begin{bmatrix} R_S[0]+E_N & R_S[1] & R_S[2] & \ldots & R_S[p] \\ R_S[1] & R_S[0]+E_N & R_S[1] & \ldots & R_S[p-1] \\ R_S[2] & R_S[1] & R_S[0]+E_N & \ldots & R_S[p-2] \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ R_S[p] & R_S[p-1] & R_S[p-2] & \ldots & R_S[0]+E_N \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \ldots \\ h_p \end{bmatrix} = \begin{bmatrix} R_S[0] \\ R_S[1] \\ R_S[2] \\ \ldots \\ R_S[p] \end{bmatrix} . \tag{E 5.6}$$

Both sides are divided by the speech energy E_S $(= R_S[0])$. In addition, let $\Gamma = E_N/E_S$ denote the noise-to-signal ratio and $Ra[\tau] = R_S[\tau]/R_S[0]$, the normalized speech autocorrelation at lag τ. Eq 5.6 becomes:

$$\begin{bmatrix} 1+\Gamma & Ra[1] & Ra[2] & \ldots & Ra[p] \\ Ra[1] & 1+\Gamma & Ra[1] & \ldots & Ra[p-1] \\ Ra[2] & Ra[1] & 1+\Gamma & \ldots & Ra[p-2] \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ Ra[p] & Ra[p-1] & Ra[p-2] & \ldots & 1+\Gamma \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \ldots \\ h_p \end{bmatrix} = \begin{bmatrix} 1 \\ Ra[1] \\ Ra[2] \\ \ldots \\ Ra[p] \end{bmatrix} . \tag{E 5.7}$$

The solution to the above matrix equation: $[A]\,[H] = [R]$ is simply $[H] = [A]^{-1}[R]$.

Since matrix A is symmetric, it is orthogonally diagonalizable:

$$[A] = [Q]\,[D]\,[Q]^t$$

where:

- $[\boldsymbol{D}] = diag(\lambda_1, \lambda_2, \ldots, \lambda_{p+1})$ and $(\lambda_1, \lambda_2, \ldots, \lambda_{p+1})$ are the eigenvalues of $\boldsymbol{A}$. The inverse of $\boldsymbol{D}$ is simply: $[\boldsymbol{D}]^{-1} = diag(1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_{p+1})$,
- $[\boldsymbol{Q}] = (\underline{u_1}, \underline{u_2}, \ldots, \underline{u_{p+1}})$ is the matrix whose columns are the orthonormal eigenvectors of $\boldsymbol{A}$. Since $\boldsymbol{Q}$ is orthogonal then $[\boldsymbol{Q}]^{-1} = [\boldsymbol{Q}]^t$.

The solution to the above matrix system (Eq 5.7) becomes:

$$\boxed{[\boldsymbol{H}] = [\boldsymbol{Q}]^t [\boldsymbol{D}]^{-1} [\boldsymbol{Q}] [\boldsymbol{B}]} \qquad \text{(E 5.8)}$$

Thus, to construct the optimum filters, one needs to estimate the noise-to-signal ratio (or its inverse, the SNR), and the value of the normalized autocorrelation at p lags. In the following section, these entities are shown to be the by-product of the 4th-order cumulant.

5.2.3.2 Special cases

2-tap filter

For the case where $p = 1$, Eq 5.7 becomes:

$$\begin{bmatrix} \Gamma + 1 & Ra(1) \\ Ra(1) & \Gamma + 1 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \end{bmatrix} = \begin{bmatrix} 1 \\ Ra(1) \end{bmatrix}. \qquad \text{(E 5.9)}$$

The solution is found by direct matrix inverse:

$$\begin{bmatrix} h_0 \\ h_1 \end{bmatrix} = \frac{1}{(1+\Gamma)^2 - Ra^2(1)} \begin{bmatrix} \Gamma + 1 & -Ra(1) \\ -Ra(1) & \Gamma + 1 \end{bmatrix} \begin{bmatrix} 1 \\ Ra(1) \end{bmatrix}$$

$$= \frac{1}{(1+\Gamma)^2 - Ra^2(1)} \begin{bmatrix} \Gamma + 1 - Ra^2(1) \\ \Gamma \cdot Ra(1) \end{bmatrix}. \qquad \text{(E 5.10)}$$

Single-tap filter

Eq 5.4 simplifies to a single equation with one unknown: $E_S = h_0 \{E_S + E_N\}$. The filter gain is thus:

$$h_0 = \frac{E_S}{E_S + E_N} = \frac{SNR}{SNR + 1}. \qquad \text{(E 5.11)}$$

5.3 Estimating Filter Parameters From Fourth Statistics

5.3.1 Speech model and higher correlations

The narrow subbanding scheme discussed in Chapter 4 is used here. The speech representation and the corresponding expressions for the autocorrelation and the diagonal slice of the 4^{th}-order cumulant are given by (from Section 4.3):

1. Transient Speech:

- Speech model: $s(n) = a \cdot e^{-\alpha n P} \cdot \cos(nPw_0 + \phi)$ with $\phi = cw_0$, and c constant.
- Correlations:

Autocorrelation: $R_S[\tau] = E_S \cdot e^{-\alpha\tau}\cos(w_0\tau)$. (E 5.12)

Speech energy: $E_S = (a^2/4\alpha T)(1 - e^{-2\alpha T})$. (E 5.13)

Diagonal slice 4^{th} cumulant: $C^a_{4S}[\tau] = 3\left[\frac{1}{2}\alpha T \coth(\alpha T) - 1\right]e^{-\alpha\tau}\cos(w_0\tau)\,[E_S]^2$. (E 5.14)

Kurtosis: $C^a_{4S}[0] = \left(\frac{3}{2}\alpha T \coth(\alpha T) - 3\right)[E_s]^2$. (E 5.15)

2. Steady Voiced Speech:

- Speech model: $s(n) = a_1\cos(nPw_1 + \phi_1) + a_2\cos(nPw_2 + \phi_2)$

  with $\phi_1 = cw_1$, $\phi_2 = cw_2$, both deterministic and unknown.
- Correlations:

The autocorrelation: $R_S[\tau] = \frac{a_1^2}{2}\cos(w_1\tau) + \frac{a_2^2}{2}\cos(w_2\tau)$. (E 5.16)

Speech energy: $E_S = [a_1^2 + a_2^2]/2$. (E 5.17)

Diagonal slice of 4^{th} cumulant: $C^a_{4S}[\tau] = \frac{-3}{8}[a_1^4\cos(w_1\tau) + a_2^4\cos(w_2\tau)]$. (E 5.18)

Kurtosis: $C^a_{4S}[0] = -3[a_1^4 + a_2^4]/8$. (E 5.19)

5.3.2 Autocorrelation of Speech

• **Proposition 1.** *The normalized autocorrelation of speech at small values of the lag may be derived from the diagonal slice of the 4th-order cumulant normalized by the kurtosis:*

$$\boxed{Ra[\tau] = \frac{C^a_{4S}[\tau]}{C^a_{4S}[0]}} \qquad \text{(E 5.20)}$$

• **Proof:** For the case of transient speech, the autocorrelation and the 4th order cumulant are given by Eq 5.12 and Eq 5.14 respectively. The normalized autocorrelation is:

$$Ra[\tau] \equiv \frac{R[\tau]}{R[0]} = e^{-\alpha\tau}\cos(w_0\tau)$$

and the 4th-order cumulant normalized by the kurtosis is:

$$\frac{C^a_{4S}[\tau]}{C^a_{4S}[0]} = e^{-\alpha\tau}\cos(w_0\tau) .$$

Clearly the two expressions are equal. For the case of steady voiced speech, the autocorrelation is given by Eq 5.16. First, it is to note here that since the bands are narrow, the two frequencies are very close and one can be expressed as a small increment (Δw) of the other. Thus:

$$R_S[\tau] = \frac{a_1^2}{2}\cos[w_1\tau] + \frac{a_2^2}{2}\cos[(w_1+\Delta w)\tau]$$

$$= \frac{a_1^2}{2}\cos[w_1\tau] + \frac{a_2^2}{2}\{\cos[w_1\tau]\cos[\Delta w\tau] - \sin[w_1\tau]\sin[\Delta w\tau]\} \ ;$$

since $\Delta w\tau$ is small for small values of the lag τ, then: $\sin[\Delta w\tau] = \Delta w\tau$ and $\cos[\Delta w\tau] = 1$. The autocorrelation becomes: $R_S[\tau] = \left(\frac{a_1^2}{2} + \frac{a_2^2}{2}\right)\cos[w_1\tau] - \left(\frac{a_2^2}{2}\right)(\Delta w\tau)\sin[w_1\tau]$.

The normalized autocorrelation is found by dividing $R_S[\tau]$ by the speech energy:

$$Ra[\tau] \equiv \frac{R[\tau]}{R[0]} = \cos[w_1\tau] - \left(\frac{1}{1 + a_1^2/a_2^2}\right)(\Delta w\tau)\sin[w_1\tau] \ . \qquad \text{(E 5.21)}$$

A similar approach is used for the 4th-order cumulant slice, which can be written as:

$$C^a_{4S}[\tau] = \frac{-3}{8}(a_1^4 + a_2^4)\cos[w_1\tau] + \frac{3a_2^4}{8}(\Delta w\tau)\sin[w_1\tau] .$$

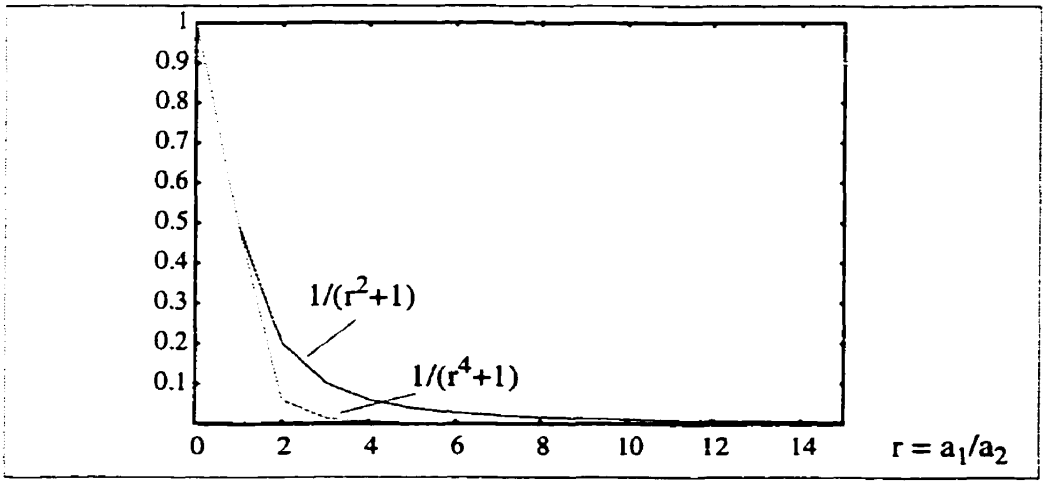
Normalization by the kurtosis yields:

$$\frac{C_{4S}^{a}[\tau]}{C_{4S}^{a}[0]} = \cos[w_1\tau] + \left(\frac{1}{1+a_1^4/a_2^4}\right)(\Delta w\tau)\sin[w_1\tau]\,. \qquad \text{(E 5.22)}$$

Comparing Eq 5.21 and Eq 5.22 shows that the first term is the same. The second term differs only in the power of the ratio of the amplitudes. A plot of the ratio of the amplitudes in the two equations (Figure 5-1) shows they differ by a maximum of 0.1, and only for a small range of the amplitudes. It is to note here that the second term is in general very small due to the $(\Delta w\tau)$ factor. For a typical case of 80 Hz bands and a sampling frequency of 8 kHz, this factor in expressed in terms of integer values of lags as: $\Delta w\tau = 2\pi \cdot 100 \cdot \frac{m}{8000} = 0.063 \cdot m$. Given therefore the multiplication with $(\Delta w\tau)$, it is clear that the second term in both equations is identical for any practical purposes, therefore for steady voiced speech:

$$Ra[\tau] \approx \frac{C_{4S}^{a}[\tau]}{C_{4S}^{a}[0]}.$$

Figure 5-1 The ratios $1/(r^2+1)$ and $1/(r^4+1)$ for the possible range of r



5.3.3 Probability of Speech Presence

5.3.3.1 Rationale

The uncertainty of speech presence in a given band is often used in conjunction with the suppression rule adopted. The rationale being that by further attenuating bands unlikely to contain any speech, the noise residuals become less audible. This method was proposed in [McA80], [Eph84] and [Mal99] and

is shown to be effective with various suppression rules in [Sca96]. The derivations in [McA80] and [Eph84] assumed a Gaussian model for the Fourier coefficients of noisy speech. Though this assumption is intuitively sound, various statistical analysis have led to different conclusions as mentioned in [Eph84] and so the probability based on this model is to be taken only in approximation.

The derivations proposed here only assume that speech in any subband has a different distribution from that of Gaussian noise. By not making explicit assumption about the distribution of subbanded speech, the results are not dependent on the validity of any assumed distribution.

5.3.3.2 Probability based on HOS

Since the kurtosis of Gaussian noise is different from that of subbanded speech, it is sensible to use this statistic as discriminator of noise-only bands. Given the variance of the HOS estimators, the decision can only be made with a confidence level that would account for the variance of the kurtosis estimator when computing it from a finite length record. If one has a knowledge of the mean and variance of the kurtosis estimator in the case of noise only and speech plus noise, then a likelihood ratio can be used to decide between the two hypotheses. Given however that the variance of the kurtosis estimator for the case of speech plus noise is not possible to quantify, an approximation is used here, whereby the probability of speech presence is the complement of the probability of noise only being present:

$$Prob\,[SpeechPresence] \approx 1 - Prob\,[NoiseOnly]\,.$$

- **Proposition 2.** *The probability of a noise-only band is determined by the estimate of the kurtosis, scaled by its variance expressed in terms of the noise energy. Denoting this scaled value by 'b', then a reasonable quantification of this probability is:*

$$\boxed{Prob\,[NoiseOnly] = erfc\,(|b|)} \qquad \text{(E 5.23)}$$

- **Proof:** Given a Gaussian process $g(n)$, the estimators of the 2nd, 3rd and 4th order moments are:

$$M_{kg} = \frac{1}{N}\sum_{n=0}^{N-1} [g(n)]^k \text{ estimator for } \quad E[\{g(n)\}^k] \qquad \text{(E 5.24)}$$

for $k = 2, 3$ and 4. In Appendix A, it is shown that these estimators are unbiased, and for the case of a white Gaussian process, their mean and variance may be expressed in terms of the data variance, υ_g. In addition, the estimator for the kurtosis may be computed from the moments. To ensure an unbiased estimate, the modified estimator proposed in Eq A.4 is used:

$$\bar{KU} = \left(1 + \frac{2}{N}\right) M_{4g} - 3\,(M_{2g})^2\,. \qquad \text{(E 5.25)}$$

As shown in Appendix A, this estimator is unbiased, with zero mean and known variance given in Eq A.27. The distribution of this estimator is not obvious, since it consists of the difference of two variables, one Gaussian and one chi-square. However, an approximation is used here and the estimator is assumed normally distributed. A unit-variance version of this variable is defined as:

$$\bar{KU}_a = \frac{\bar{KU}}{\sqrt{\frac{3\upsilon_g^4}{N}\left(104 + \frac{452}{N} + \frac{596}{N^2}\right)}} . \qquad \text{(E 5.26)}$$

Therefore, given a realization of the estimated kurtosis and the corresponding scaled value, denoted by 'b', one can find the probability that the frame is Gaussian as:

$$Prob\,[NoiseOnly] = Prob\,[|\bar{KU}_a| \geq b] . \qquad \text{(E 5.27)}$$

Graphically, this is equivalent to computing the area under the tail of the Gaussian curves of $\bar{KU}_a$. Clearly when $b = 0$, the area is unity. The area under the tail of the curve can be evaluated using the $erfc(x)$ function. For example, when b > 0, $Prob\,[Noise] = \frac{2}{\sqrt{\pi}}\int_b^{\infty} e^{-x^2/2} dx$. Therefore,

$$Prob\,[NoiseOnly] = erfc\,(|b|) .$$

In the above discussion, it is assumed that the true variance of the noise (υ_g) is known apriori. In reality, this is not the case and one has only a (hopefully good) estimate of the noise energy ($\hat{E}_N$) that is done using the approach explained in the next section. This estimator is not equal to the true variance, but relatively speaking it is fairly good compared to the kurtosis estimator, which is only estimated from a short data frame.

5.3.4 Speech and Noise Energy and SNR Estimation

In the traditional approach to speech enhancement by spectral decomposition, the local SNR is computed by estimating the noise spectrum during periods of non-speech frames and combining that with the noisy speech power spectrum to infer the SNR. In practical situations however, this is seldom sufficient as noise changes during speech frames as well. The resulting poor SNR estimation limits the effectiveness of the suppression filters, regardless of what rule is used, often resulting in annoying noise artifacts.

A number of approaches have attempted to remedy the problem of SNR estimation without speech detection. In [Hir95], an iterative estimation is used whereby the spectral amplitude at each frequency

is compared to the current noise estimate at that frequency; the noise is estimated whenever the amplitude is below a given threshold. The approach works well most of the time but being based on relative energy levels cannot distinguish between rising noise energy and the presence of speech. In [Mar93], a spectral analysis is performed and the noise estimated by measuring the spectral floor in individual bands of the power spectrum. This scheme requires a relatively long segment of speech and a good frequency resolution to overcome the errors introduced by the DFT windowing effects; consequently, only the long term noise spectrum can be estimated.

The idea in using HOS for SNR estimation hinges on being able to separate signal and noise energies based on these statistics. Given the fact that the 4^{th} cumulant of noisy speech is simply that of the clean speech, and given that the HOS of Gaussian noise are zero, then the energies are estimated as follows:

- In bands where only noise is present, the noise energy is estimated using the total energy in that band. These bands are detected using the procedure in Section 5.3.3 above.
- In bands where steady voiced speech is present, an upper bound on the speech energy is derived from the kurtosis, from which a lower bound on the noise is deduced. These bands are detected using the value of the normalized kurtosis.
- In bands where transient speech is present, the energy of speech is derived from the first few lags of the 4^{th}-order cumulant. The noise energy is deduced from the total energy and the estimated speech energy. These bands are detected by exclusion.

Once the energy of the speech and noise are estimated in a given band k, the SNR is computed as:

$$SNR(k) = \frac{\hat{E}_S(k)}{\hat{E}_N(k)} \qquad \text{or} \qquad SNR(k) = Pos\left[\frac{E_X(k)}{\hat{E}_N(k)} - 1\right] \tag{E 5.28}$$

where $Pos[x] = x$ when $x > 0$ and 0 otherwise. $\hat{E}_N$ is the estimate of the noise energy, $\hat{E}_S$ the estimate of the speech energy and E_X the total band energy. These three ideas are detailed below.

Noise Energy from noise-only bands

The probability that a band contains only noise can be quantified using the value of the kurtosis and the variance of the kurtosis estimator as explained in Proposition 2. In these bands, the estimate of the noise energy is updated using the total signal energy. In addition to the probability of noise, the normalized skewness and kurtosis are used to discriminate noise-only bands, thus:

If ($Prob\,[noise] > T_{noise}$ and $|\gamma_3| < T_{\gamma_3}$ and $|\gamma_4| < T_{\gamma_4}$)

$$\hat{E}_N(j) = (1-\beta)\,\hat{E}_N(j-1) + \beta E_X,$$

where j denotes the iteration index and β is an integration constant that is a function of the probability of noise of this band.

Speech and noise energies from steady state speech bands

Proposition 3. *In the case where steady voiced speech is present in a band, then a lower bound on the noise energy in this band may be computed from the kurtosis and the total band energy:*

$$\lfloor \hat{E}_N \rfloor = E_X - \sqrt{\left[\frac{C_{4X}[0]}{-0.75}\right]}. \qquad \text{(E 5.29)}$$

Moreover, the detection of this band can be done using the normalized kurtosis.

Proof: When a band contains steady voiced speech, the kurtosis is upper and lower bounded by a scale factor of the signal energy (from Section 4.3.2):

$$-1.5\,[E_S]^2 \le Kurtosis \le -0.75\,[E_S]^2. \qquad \text{(E 5.30)}$$

The normalized kurtosis is then contained in the range [-1.5, -0.75]. When noise is present, the normalization may yield a smaller amplitude for the normalized kurtosis as explained in Section 4.3.4. However the normalized kurtosis is always negative and cannot be smaller than -1.5.

Since cumulants are cumulative and since the kurtosis of Gaussian noise is zero, then when the signal consists of both speech and noise, the total energy is the sum of the two energies, whereas the kurtosis of noisy speech is simply that of clean speech:

$$E_X = E_S + E_N. \qquad \text{(E 5.31)}$$

$$C_{4X} = C_{4S}. \qquad \text{(E 5.32)}$$

From Eq 5.30, one can compute an upper bound on the speech energy from the kurtosis: $E_S \le \sqrt{\frac{C_{4X}[0]}{-0.75}}$ and from it a lower bound on the noise is deduced using Eq 5.31. Thus the noise energy is updated as follows:

If ($\gamma_4 < 0$ and $\gamma_4 \ge -1.5$ and $Prob\,[noise] < T_{noise}$)

$$\lfloor \hat{E}_N \rfloor = E_X - \sqrt{\frac{C_{4X}[0]}{-0.75}}$$

$$\hat{E}_N(j) = (1-\beta)\,\hat{E}_N(j-1) + \beta \lfloor \hat{E}_N \rfloor$$

where j denotes the iteration index and β an integration constant with: $\beta < 0.1$.

Speech and noise energies from transient speech bands

•**Proposition 4.** *In the case where transient speech is present in a given band, the speech energy may be estimated by considering the first three lags of the diagonal slice of the 4^{th}-order cumulant.*

•**Proof:** The diagonal slice is (Eq 5.14):

$$C_{4S}^{a}[\tau] = 3\left[\frac{1}{2}\alpha T \coth(\alpha T) - 1\right] e^{-\alpha\tau} \cos(w_0\tau) [E_S]^2 .$$

Let R_1 and R_2 denote the following ratios:

$$R_1 = \frac{C_{4S}^{a}(1)}{C_{4S}^{a}(0)} = e^{-\alpha}\cos(w_0) \qquad \text{and} \qquad R_2 = \frac{C_{4S}^{a}(2)}{C_{4S}^{a}(0)} = e^{-2\alpha}\cos(2w_0) .$$

The value of α is found from these since:

$$R_2 - 2R_1^2 = e^{-2\alpha}\{\cos(2w_0) - 2[\cos(w_0)]^2\} = -e^{-2\alpha} \quad \text{and} \quad \alpha = \frac{\ln(2R_1^2 - R_2)}{-2} .$$

Once α is determined, the speech energy in this frame is estimated from the kurtosis of noisy speech (using Eq 5.15):

$$\tilde{E}_S = \sqrt{Pos\left[\frac{C_{4X}^{a}[0]}{\frac{3}{2}\alpha T \coth(\alpha T) - 3}\right]} . \qquad \text{(E 5.33)}$$

The noise energy is estimated from the total energy and the speech energy and the two are smoothed:

$$\hat{E}_S(j) = (1-\alpha)\hat{E}_S(j-1) + \alpha\tilde{E}_S \qquad \text{(E 5.34)}$$

$$\hat{E}_N(j) = (1-\beta)\hat{E}_N(j-1) + \beta\tilde{E}_N \qquad \text{(E 5.35)}$$

where j denotes the iteration index and α and β are integration constants; simulation shows that values of $\alpha = 0.5$ and $\beta < 0.05$ gives best results. To determine the best values, a comparison between the actual and the estimated SNR is examined (Section 5.7.2.1) in controlled conditions and the values of the two constants varied until the estimated SNR is closest to the actual SNR.

5.4 Inter-frame Smoothing of Parameters

5.4.1 SNR Smoothing

The SNR smoothing scheme proposed in [Eph84] is used whereby the variation in SNR between successive frames is reduced significantly by averaging the locally computed SNR (SNR_{post}) with the SNR estimated in the previous frame after the filtering operation (SNR_{est}):

$$SNR_{prior}(k) = (1-\gamma)\,SNR_{post}(k) + \gamma SNR_{est}(k)\,, \qquad \text{(E 5.36)}$$

where $SNR_{post}(k)$ refers to the local SNR from this iteration, using the estimated speech and noise energies or alternatively using the total energy and the estimated noise energy:

$$SNR_{post}(k) = \frac{\hat{E}_S(k,j)}{\hat{E}_N(k,j)} \quad \text{or} \quad SNR_{post}(k) = Pos\left[\frac{E_X(k,j)}{\hat{E}_N(k,j)} - 1\right], \qquad \text{(E 5.37)}$$

where the indices k and j are used to denote the band index (k) and the iteration index (j).

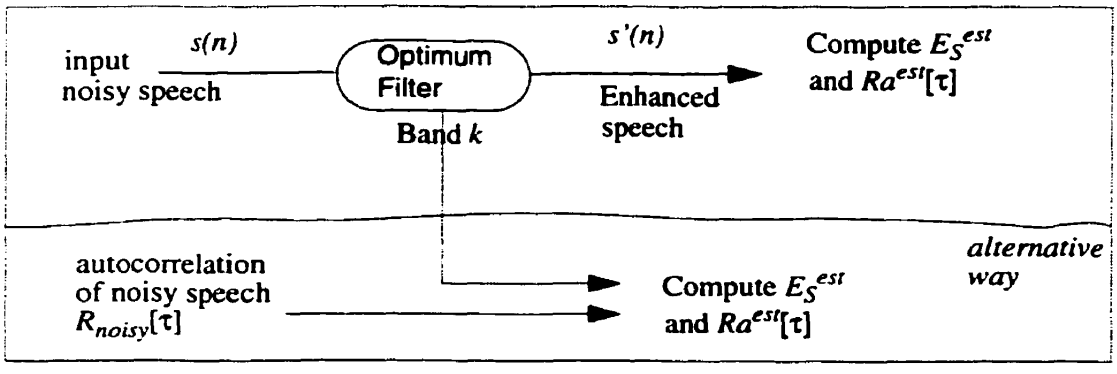
$SNR_{est}(k)$ is computed in the *previous* iteration *after* the filtering operation (Figure 5-2), i.e., using the energy of the filtered speech instead of the energy of the noisy speech.

The estimated *speech* energy may be computed as the energy at the filter output:

$$E_S^{est} = \frac{1}{N}\sum_{n=0}^{N-1}[s'(n)]^2 \quad \text{and thus:} \quad SNR_{est}(k) = \frac{E_S^{est}(k)}{\hat{E}_N(k,j)}. \qquad \text{(E 5.38)}$$

Alternatively, E_S^{est} may be computed directly from the autocorrelation of noisy speech ($R_{noisy}[\tau]$) and the value of the filter coefficients, as shown in the next section.

Figure 5-2 Post filtering estimation of the speech energy and speech autocorrelation



5.4.2 Autocorrelation Smoothing

Using the same principle as in the case of the SNR, the normalized autocorrelation function $Ra[\tau]$ is smoothed using the value computed locally (from the 4th-order cumulant) and the value computed in the previous iteration *after* optimal filtering:

$$Ra^{prior}[\tau, k] = (1-\gamma)\, Ra^{post}[\tau, k] + \gamma Ra^{est}[\tau, k] , \qquad \text{(E 5.39)}$$

where k refers to the band index and τ is the time lag. From Eq 5.20:

$$Ra^{post}[\tau] = \frac{C_{4X}^{a}[\tau]}{C_{4X}^{a}[0]} , \qquad \text{(E 5.40)}$$

where $C_{4X}^{a}[\tau]$ refers to the diagonal slice of the 4th order cumulant of the noisy speech.

In the previous iteration, $Ra^{est}[\tau]$ is computed using the output of the optimal filters (Figure 5-2):

$$Ra^{est}[\tau] = \frac{R_S[\tau]}{R_S[0]} , \text{ with } \quad R_S[\tau] = \frac{1}{N-\tau}\sum^{N-\tau} s'(n)\, s'(n+\tau) . \qquad \text{(E 5.41)}$$

Alternatively, $Ra^{est}[\tau]$ may be computed directly from the autocorrelation of noisy speech, denoted by $R_{noisy}[\tau]$ in Figure 5-2 and the coefficients of the optimum filter h_β. Since *s'(n)* is the output of the linear filter *h(n)*, the autocorrelation of *s'(n)* may be written in terms of the autocorrelation of *s(n)* and the filter coefficients:

$$R_S[\tau] = \sum_{j=0}^{H-1}\sum_{i=0}^{H-1} h_j \cdot h_i \cdot R_{noisy}[\tau + j - i] \qquad \text{(E 5.42)}$$

where H is the filter length. As a direct consequence, the estimated speech energy E_S^{est} (Eq 5.38) may be computed directly from Eq 5.42 as: $E_S^{est} = R_S[0]$.

5.4.3 Low Pass Filtering of Optimum Filter Coefficients

In order to avoid rapid time variation of the optimal filter coefficients, each coefficient is smoothed using a 2-tap low-pass filter:

$$\overline{h_\beta} = 0.7 \cdot h_\beta(j) + 0.3 \cdot h_\beta(j-1) \qquad \beta = 0, \ldots, p \qquad \text{(E 5.43)}$$

where $h_\beta(j)$ is the coefficient computed in this iteration (j) and $h_\beta(j-1)$ is the one computed in the previous iteration (j-1). Simulation showed that a 2-tap lowpass filter is sufficient to smooth any significant time variations.

5.5 Other Considerations

5.5.1 Frequency Masking Psychoacoustics

In computing $SNR_{post}(k)$ and $SNR_{est}(k)$, the masking effect of the auditory system is taken into account by convolving both speech and noise energies with the critical band filter at each frequency.

Simultaneous masking is the phenomena whereby a tone or noise masks a neighboring one that is within a critical distance from it in frequency. The peripheral auditory system is modeled as a bank of overlapping bandpass filters whose bandwidth is related to the critical bandwidth within which masking occurs. The bandwidth of these filters, referred to as the equivalent rectangular bandwidth (ERB), is an increasing function of frequency and has been determined in a number of experiments [Moo81]. The function may be approximated in polynomial form by:

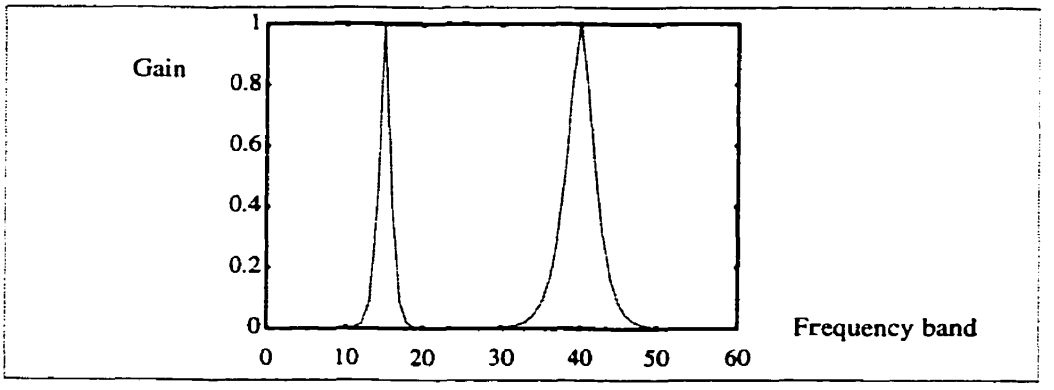
$$ERB(F)|_{Hz} = 6.23F^2 + 93.39F + 28.52 \qquad \text{(E 5.44)}$$

where F is the center frequency in kHz. In [Moo83], an exponentially decreasing function is proposed to model the shape of the auditory filters:

$$W(g) = (1 + pg)\,e^{-pg} \qquad \text{(E 5.45)}$$

where g is the deviation in frequency from the filter center frequency divided by the center frequency and p is a parameter that is expressed in terms of the ERB: $p = \frac{4f}{ERB}$ where f is the center frequency, expressed in Hz. Figure 5-3 illustrates two such filters (here discrete frequency spacing is used (bands) with each band representing 80 Hz).

Figure 5-3 Auditory filter shapes centered at bands 15 and 40



The effect of the auditory filters is accounted as follows: Once the total energy is computed and the noise energy estimated, then at each frequency a set of new entities is computed, referred to as the *perceptual* total and noise energy measures. These are computed by convolving the critical filter at this frequency (f) with the total energy and noise energies respectively (the iteration index j is dropped):

$$E_x^{\,p}(f) = W(f) \otimes E_x(f) \quad \text{and} \quad E_n^{\,p}(f) = W(g) \otimes \hat{E}_n(f) \,. \tag{E 5.46}$$

Using the discrete notation k adopted here for frequency, this becomes:

$$E_x^p(k) = \sum_{m=0}^{K-1} W\left(\frac{|k-m|}{k+0.5}\right) E_x(m) \quad \text{and} \quad E_n^p(k) = \sum_{m=0}^{K-1} W\left(\frac{|k-m|}{k+0.5}\right) \hat{E}_n(m) \,. \tag{E 5.47}$$

The local SNR in band k is then computed as:

$$SNR_{post}(k) = Pos\left[\frac{E_x^p(k)}{E_n^p(k)} - 1\right]. \tag{E 5.48}$$

A similar procedure is used at the end of the current iteration to compute $SNR_{est}(k)$.

5.5.2 Subbanding Filters

A filter bank consisting of M cosine-modulated filter pairs is used for analysis and synthesis. The expression for the analysis and synthesis filters is the one proposed in [Koi92]:

$$H(l) = 2P(l)\cos\left(\frac{(2k+1)\,(2l-L+1+M)\,\pi}{4M}\right) \quad \text{(E 5.49)}$$

$$F(l) = 2P(l)\cos\left(\frac{(2k+1)\,(2l-L+1-M)\,\pi}{4M}\right) \quad \text{(E 5.50)}$$

where M is the number of filters (bands), L is the length of the filter and $P(l)$ is the prototype filter. For the case where $L = 2M$, the expression of $P(l)$ is simple and given by:

$$P(l) = \frac{1}{M\sqrt{2}}\sin\left(\frac{(l+0.5)\,\pi}{2M}\right). \quad \text{(E 5.51)}$$

5.5.3 Upper and Lower Bounds on Filter gains

Recall the expression for the filter coefficients (Section 5.2.3.2):

$$\begin{bmatrix} h_0 \\ h_1 \end{bmatrix} = \frac{1}{(1+\Gamma)^2 - Ra^2(1)}\begin{bmatrix} \Gamma + 1 - Ra^2(1) \\ \Gamma \cdot Ra(1) \end{bmatrix}.$$

In the above expression, two cases are worth noting:

- Very High SNR: In this case, $\Gamma \rightarrow 0$ and the coefficients simplify to: $h_0 = 1$, $h_1 = 0$.
- Low SNR: In this case, $\Gamma \rightarrow \infty$ and the coefficients simplify to: $h_0 = h_1 = 0$.

In practice, a lower bound is used for the filter coefficients to prevent too much audible variations in the noise levels. In order to identify the two conditions, the value of the SNR as well as the probability of noise-only frames are used as follows:

If ($SNR_{prior} \geq T_{SNR1}$ and $Prob[noise] < T_{Noise}$)

$h_0 = 1$, $h_1 = MinGain$

If (($SNR_{prior} < T_{SNR2}$ and $Prob[noise] > T_{noise2}$) Or ($Prob[noise] > T_{noise3}$))

$h_0 = h_1 = MinGain$

5.6 Overview of the Algorithm

A block diagram of the algorithm is shown in Figure 5-4. Speech recorded and sampled at 8 kHz is used. Analysis is carried out on a frame-by-frame basis with 70 new points read every iteration. In each band k, a frame overlap of 30% is used, thus the 70 new points are combined with the last 30 of the previous iteration to form a frame of $N = 100$ points. A filter bank consisting of M=50 cosine-modulated filters is used to divide the signal into 50 bands of 80 Hz width each. The analysis filters are given by Eq 5.49.

In each band k, the following computations are performed every iteration:

- The DC component is removed and the 2nd and 4th moment functions and the 3rd moment are computed; the diagonal slice of the 4th-order cumulant is inferred from these for three lags:

$$M_2[k][\tau] = \frac{1}{N'}\sum_{n=0}^{N'-1} x_k(n)\,x_k(n+\tau)$$

$$M_3[k][0] = \frac{1}{N}\sum_{n=0}^{N'-1} [x_k(n)]^3$$

$$M_4[k][\tau] = \frac{1}{M}\sum_{n=0}^{N'-1} [x_k(n)]^3 x_k(n+\tau)$$

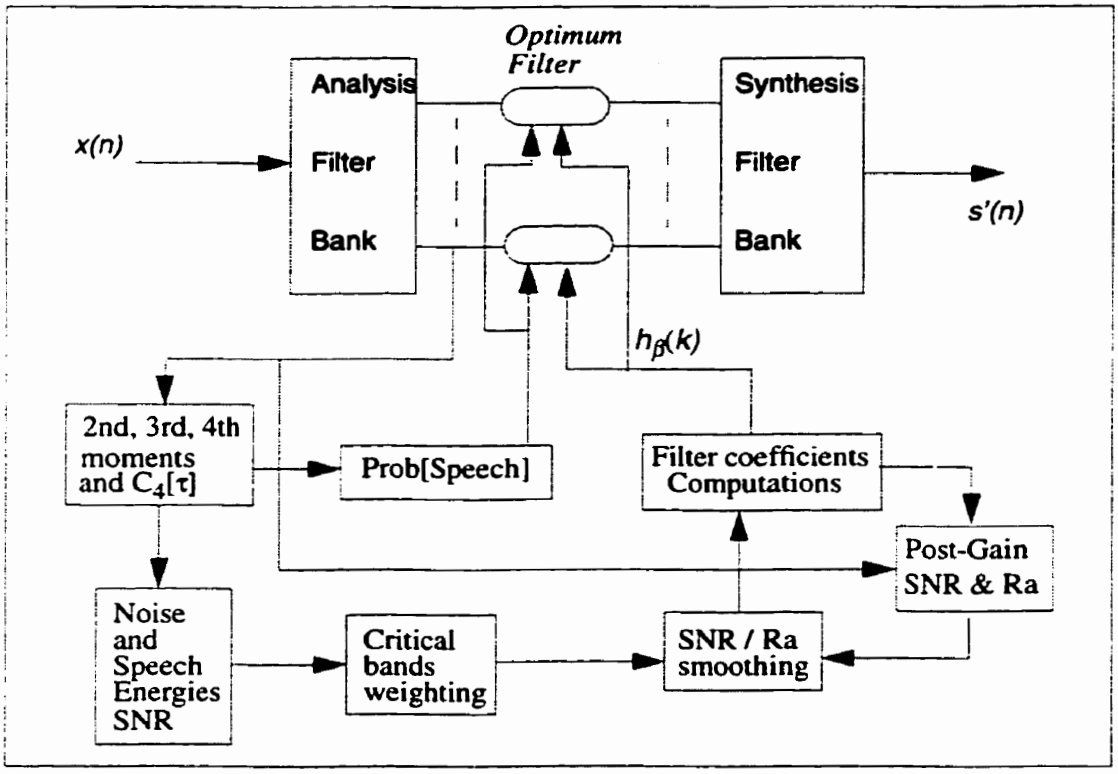
  where $N' = N - max(\tau)$. Note that $M_2[k][\tau]$ is also the autocorrelation of the noisy speech: $R_{noisy}[k][\tau] \equiv M_2[k][\tau]$, that is later used for computing $Ra^{est}[\tau]$, as explained in Section 5.4.2. To reduce the bias of the estimator (as explained in Section A.2.2), each of the moment functions is smoothed using a first-order autoregressive averaging to yield $\overline{M_2}[k][\tau]$ and $\overline{M_4}[k][\tau]$. The diagonal slice is then computed as:

$$C_4^a[k][\tau] = \left(1+\frac{2}{M}\right)\overline{M_4}[k][\tau] - 3\overline{M_2}[k][0]\cdot\overline{M_2}[k][\tau].$$

- The probability of noise is computed using the kurtosis and the estimated noise energy $\hat{E}_N$ as explained in Section 5.3.3. This probability is then used when upper and lower bounding the filter gain as explained in Section 5.5.3.
- The noise and speech energies are estimated using the procedure in Section 5.3.4.
- Frequency masking is accounted for in the SNR computations as described in Section 5.5.1.
- The SNR is smoothed (Section 5.4.1) to yield $SNR_{prior}(k)$ from which Γ is deduced.

- The autocorrelation is computed and smoothed (Section 5.3.2 and Section 5.4.2).
- A 2-tap optimal filter is used (Eq 5.10). The filter coefficients are computed using the estimated parameters as explained in Section 5.5.3. The coefficients are smoothed as explained in Section 5.4.3.
- The 70 samples are filtered using the newly computed optimum filter.
- A synthesis bank (Eq 5.50) reconstructs the signal and yields the enhanced speech.

Figure 5-4 Block diagram of the speech enhancement algorithm



5.7 Experimental Results

5.7.1 Data used

Speech

Clean recorded Harvard sentences spoken by male and female speakers are used. For each set, two sentences spoken by the same speaker are used. The following sets are used:

- Female 1: *A king ruled the state in the early days. The ship was torn apart on the sharp reef.*
- Female 2: *Mend the coat before you go out. The wrist was badly strained and hung limb.*
- Male 1: *He picked up the dice for a second roll. These coins will be needed to pay his debts.*
- Male 2: *The nag pulled the freight car along. Twist the valve and release hot steam.*

Noise

Noise is recorded and sampled at 8 kHz and mixed linearly with clean speech at SNR levels between 10 and 13 dB. The following noise types are used:

- Gaussian noise: Synthetically generated Gaussian noise of a roughly flat spectrum.
- Street noise: recorded noise from a street corner.
- Office noise: includes machines noise and conversations.
- Fan noise: recorded in a lab, and contains a dominant periodic component.

5.7.2 Performance evaluation

Evaluating the performance of speech enhancement algorithms is a difficult task, given the lack of standardized metrics for objective evaluation. In most cases, subjective listening is used and the following criteria are used in deciding on the overall effectiveness of the algorithm:

1. The degree of noise reduction.
2. The degree of distortion -if any- to the enhanced speech.
3. The degree of distortion -if any- to the remaining noise.

To provide a frame of reference, the results are compared to the TIA noise reduction algorithm [IS127]. This algorithm is based on spectral subtraction, using heuristic rules for gain computations and using SNR and stationarity measures for noise estimation.

5.7.2.1 SNR estimation

Since the SNR in each band is a crucial factor in the computation of the optimum filter coefficients, the estimation is examined under a number of noise types and compared to the true SNR. Figure 5-5 shows the estimated (SNR_{post}) vs. the actual SNR for a number of consecutive frames and for two bands in the case of Gaussian noise.

Figure 5-5 Estimated vs. actual SNR for gaussian noise

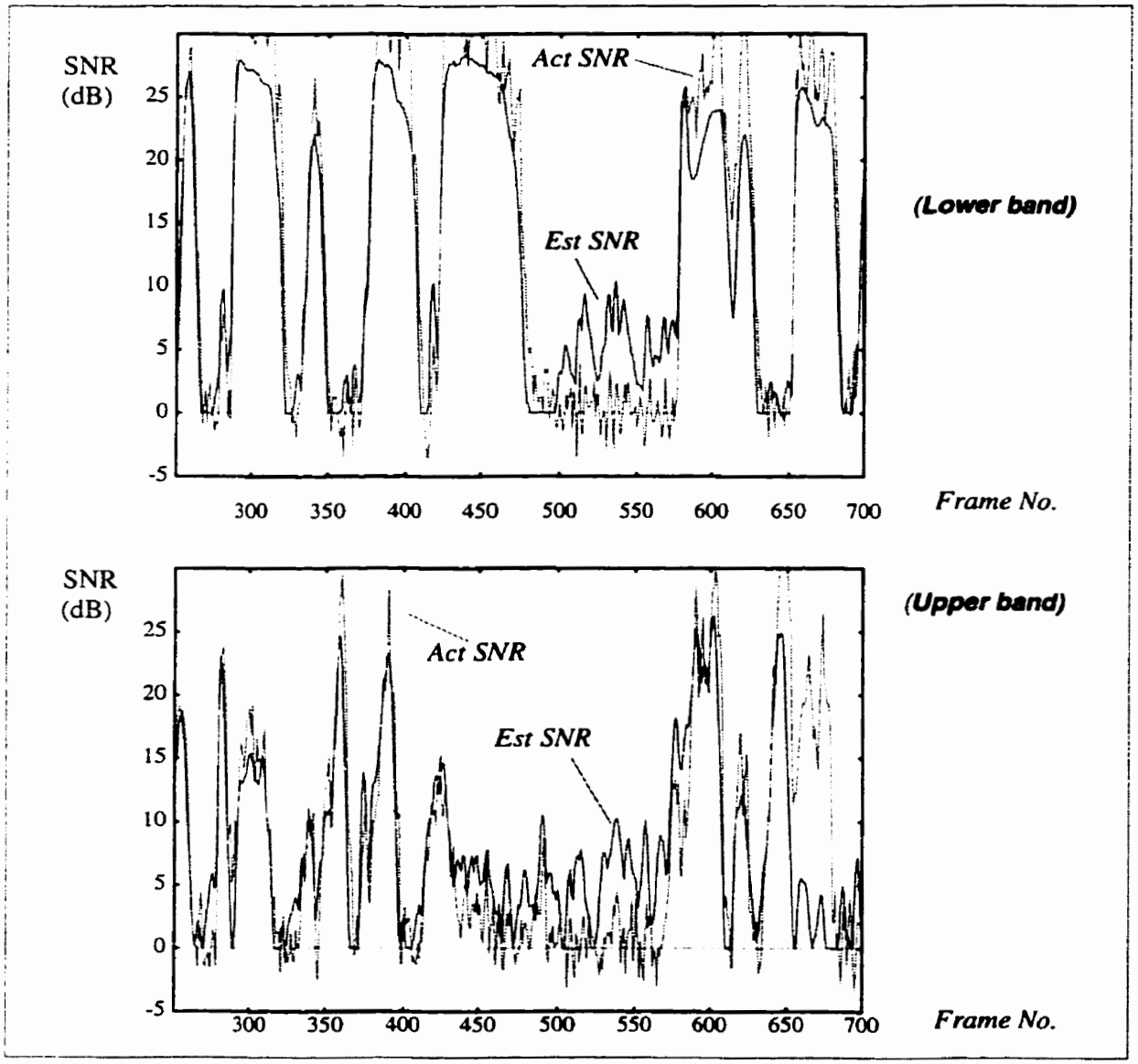


Figure 5-6 shows similar results for the case of street noise and Figure 5-7 for office noise.

Figure 5-6 Estimated vs. actual SNR for street noise

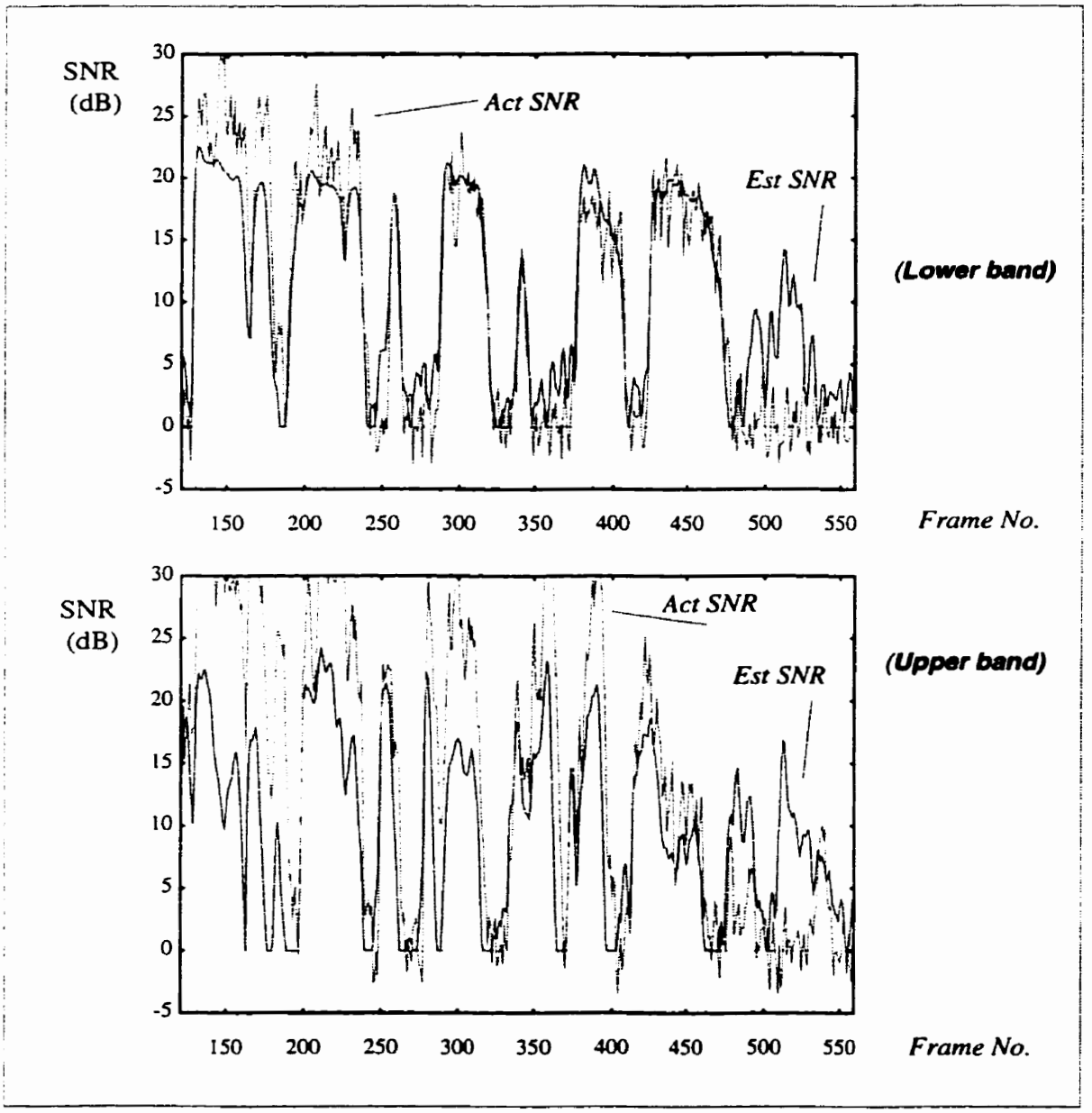
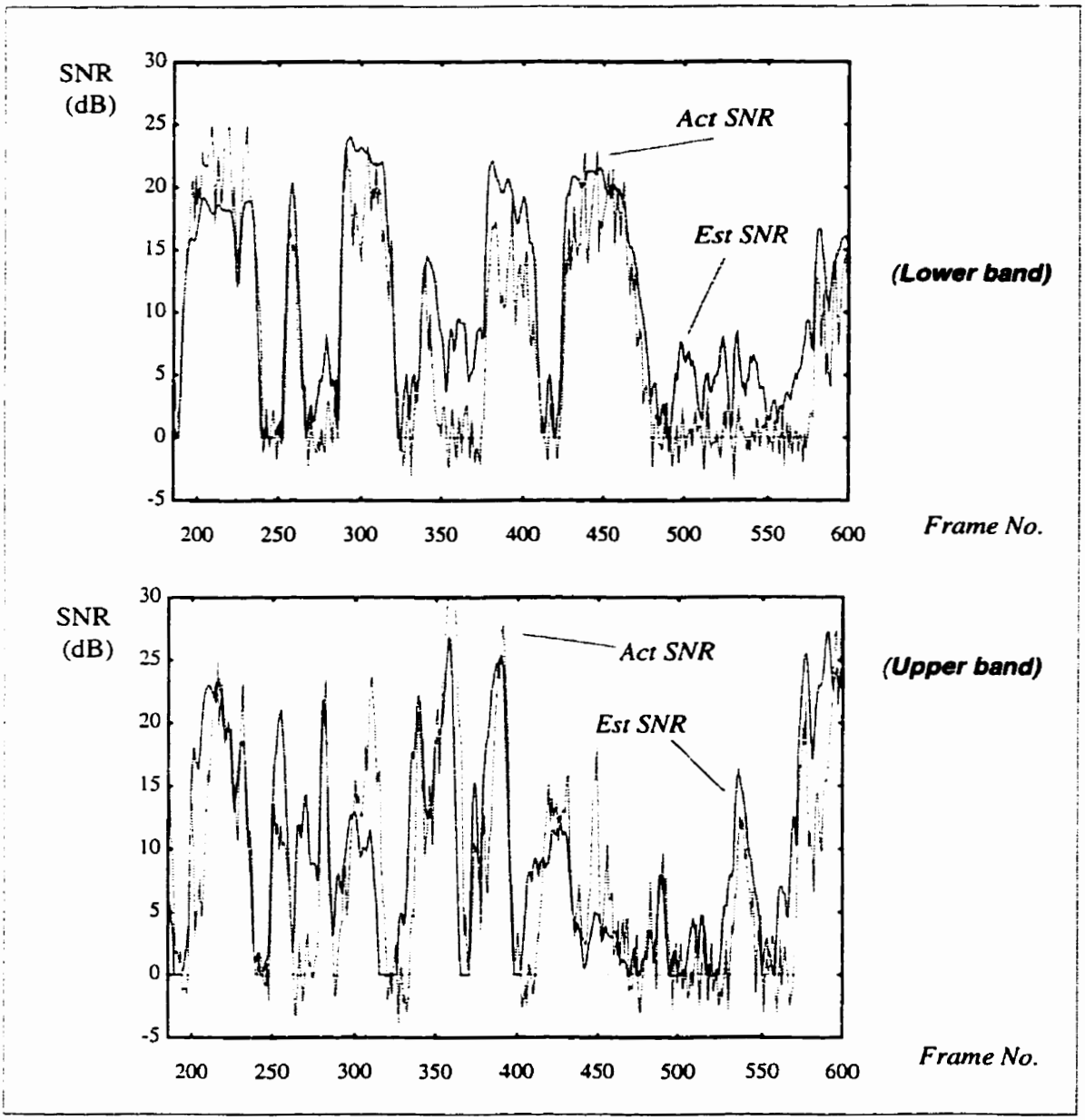


Figure 5-7 Estimated vs. actual SNR for office noise

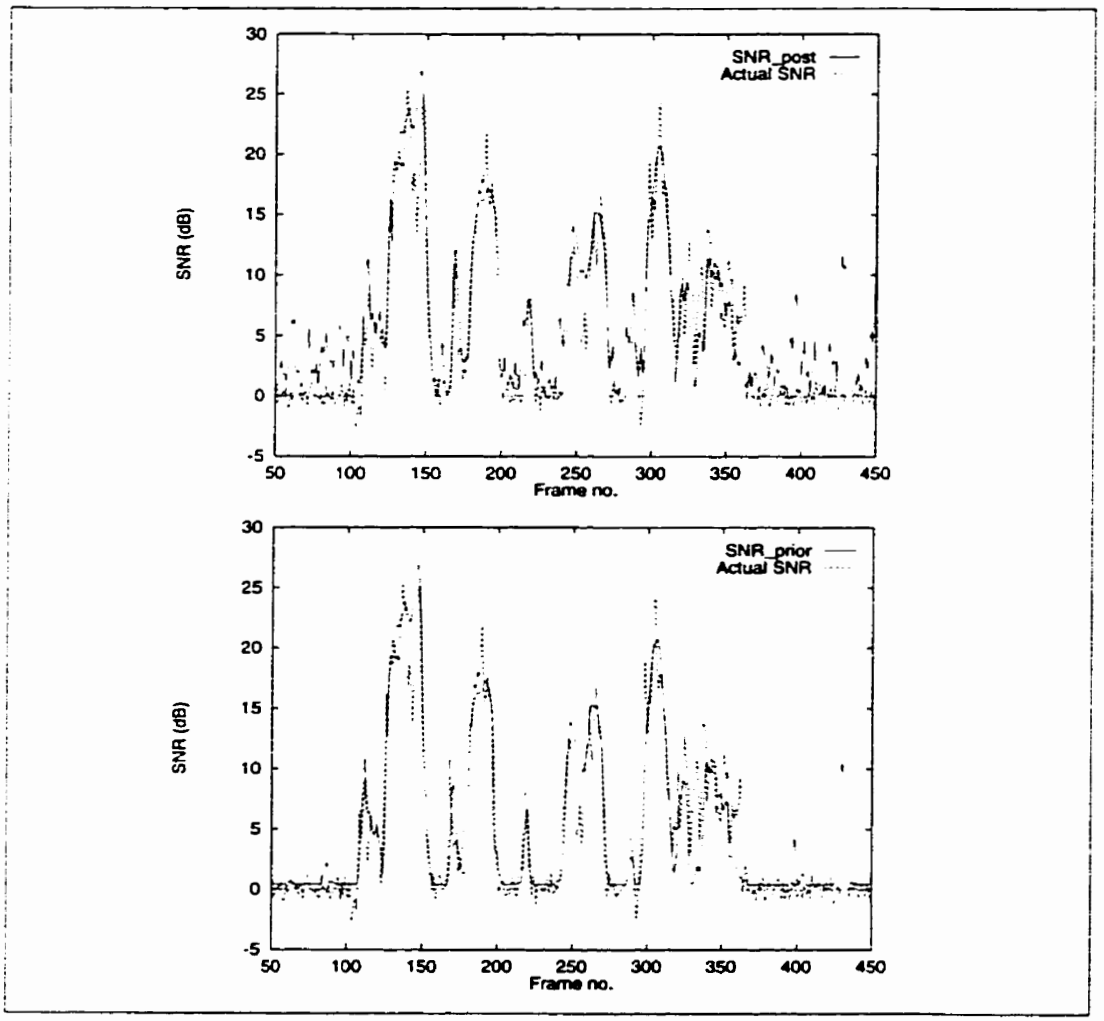


It is observed in general that the SNR estimation is more accurate in the lower bands of the spectrum, likely due to the fact that the sinusoidal model assumed is mostly valid and complete in the lower spectrum. Another point worth noting is that the estimated SNR is mostly inaccurate in the low SNR regions (i.e., < 5 dB).

5.7.2.2 SNR smoothing

The effect of the SNR smoothing is examined to ensure that the estimated SNR (shown above) is further improved by the smoothing used. The case of street noise is shown in Figure 5-8 below. It is clear that the smoothed SNR (SNR_{prior}) is closer to the actual SNR than the SNR_{post}, particularly in the lower SNR regions. These regions are important since musical noise is more perceptible there than in the strong SNR bands, where it is masked by speech. The fact that the smoothed SNR is closer to the actual SNR and is more stable implies that the effect of noise musicality is greatly reduced.

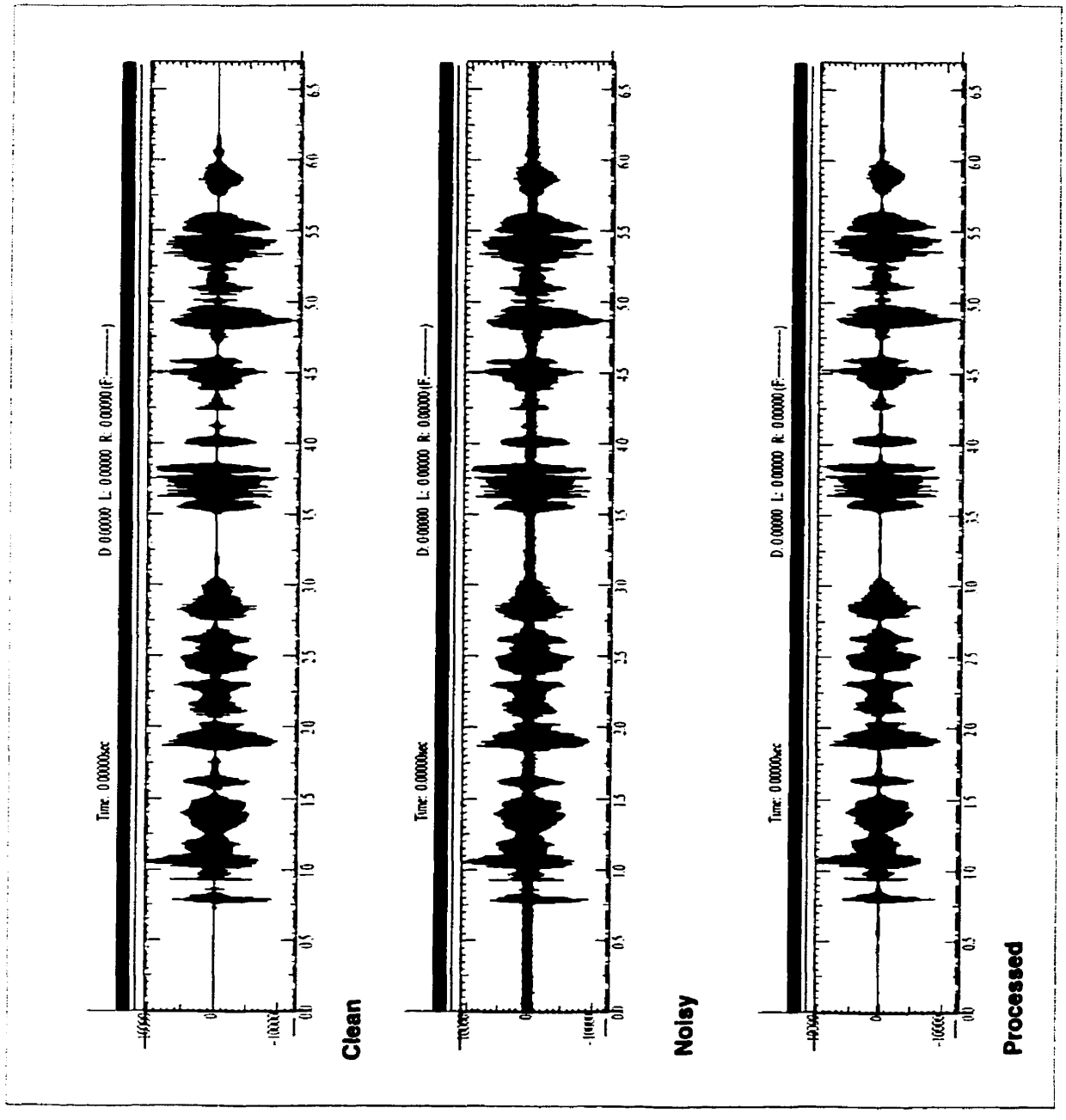
Figure 5-8 Actual, estimated and smoothed SNR for the case of street noise



5.7.2.3 Speech waveforms and spectrogram

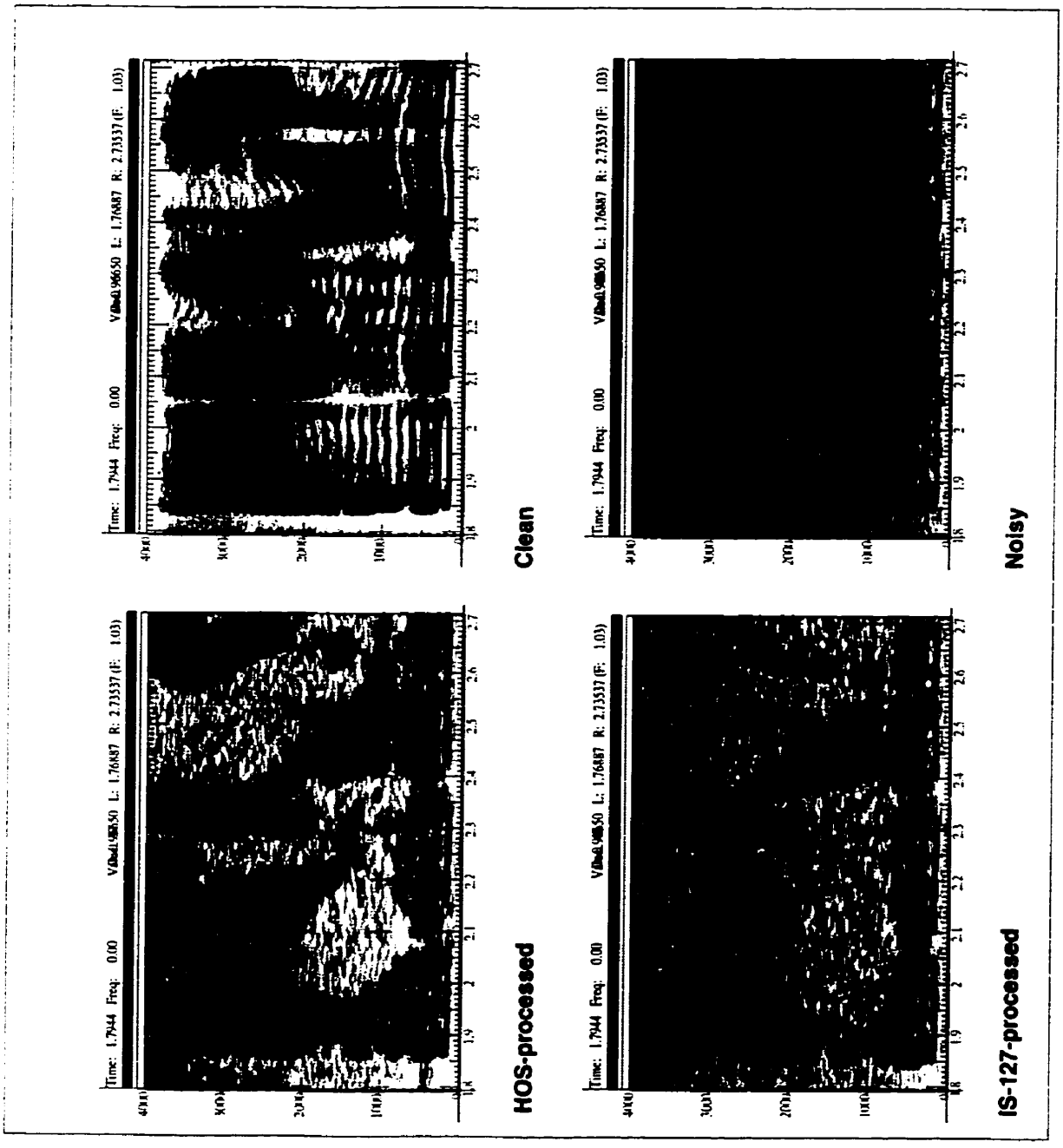
Figure 5-9 shows the clean, noisy and processed waveforms for the case of Gaussian noise at 10 dB.

Figure 5-9 Clean, noisy and processed speech waveforms (Gaussian noise. 10 dB)



The spectrograms for a short segment of this waveform are shown in Figure 5-10. The HOS-processed speech spectrogram is shown along with the one generated by the IS-127 algorithm.

Figure 5-10 Spectrograms for a section in the above waveform



Similar plots for the waveforms (Figure 5-11) and spectrograms (Figure 5-12) are shown for the case of street noise at 13 dB.

Figure 5-11 Clean, noisy and processed speech waveforms (Street noise 13 dB)

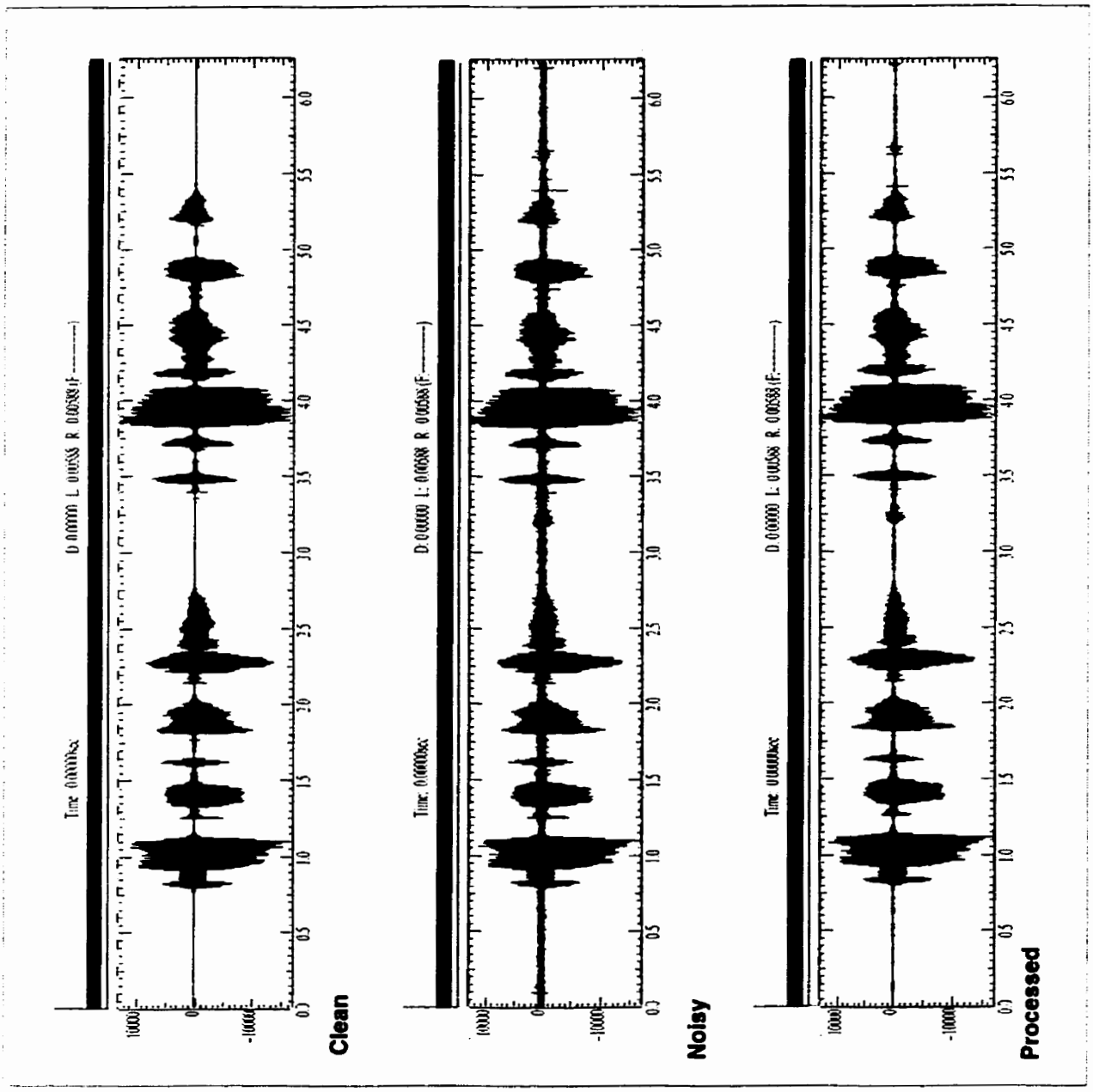
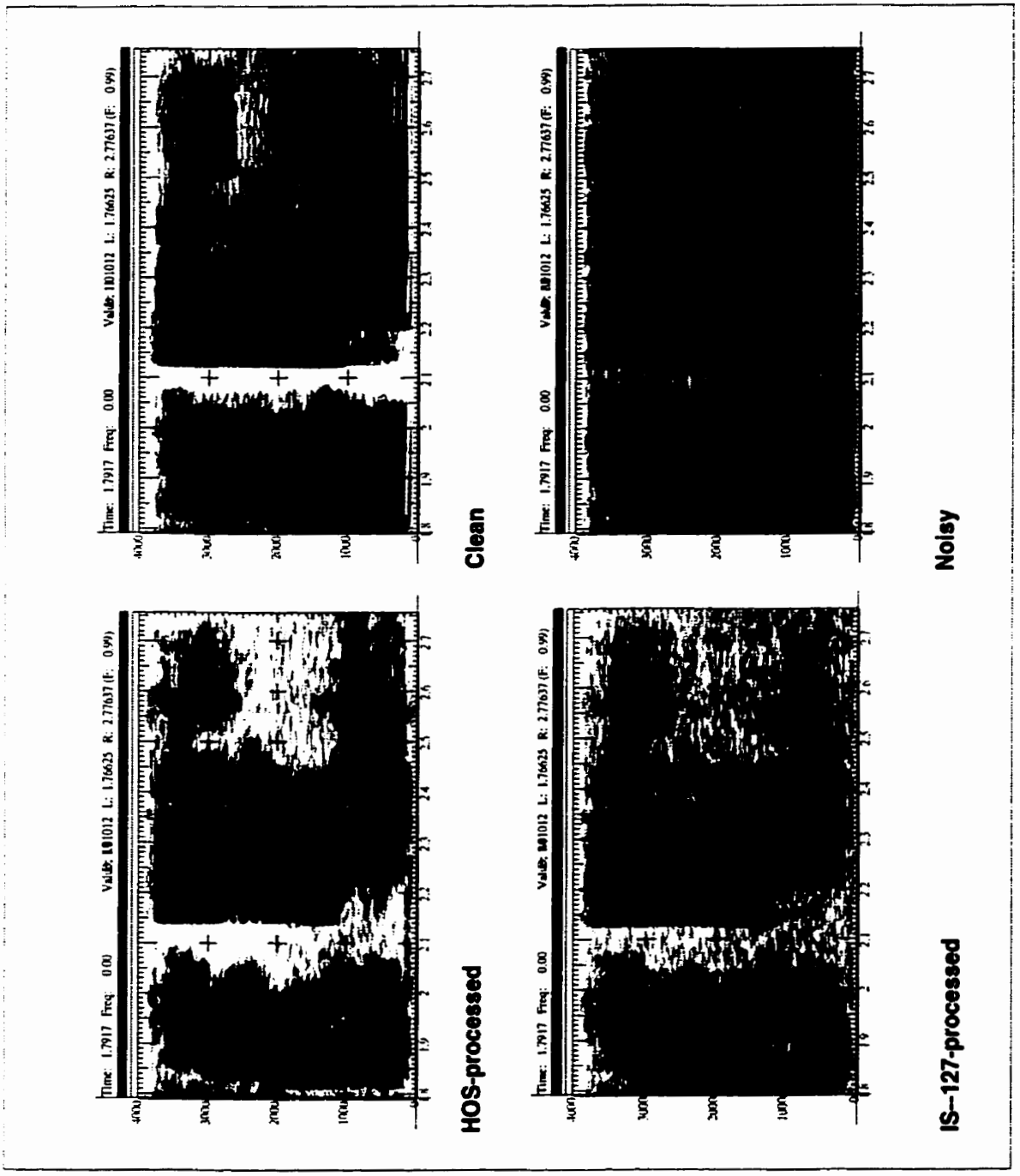


Figure 5-12 Spectrograms for a section in the above waveform (street noise)



Finally the case of fan noise at 12 dB is shown in Figure 5-13 and Figure 5-14.

Figure 5-13 Clean, noisy and processed speech waveforms (Fan noise 12 dB)

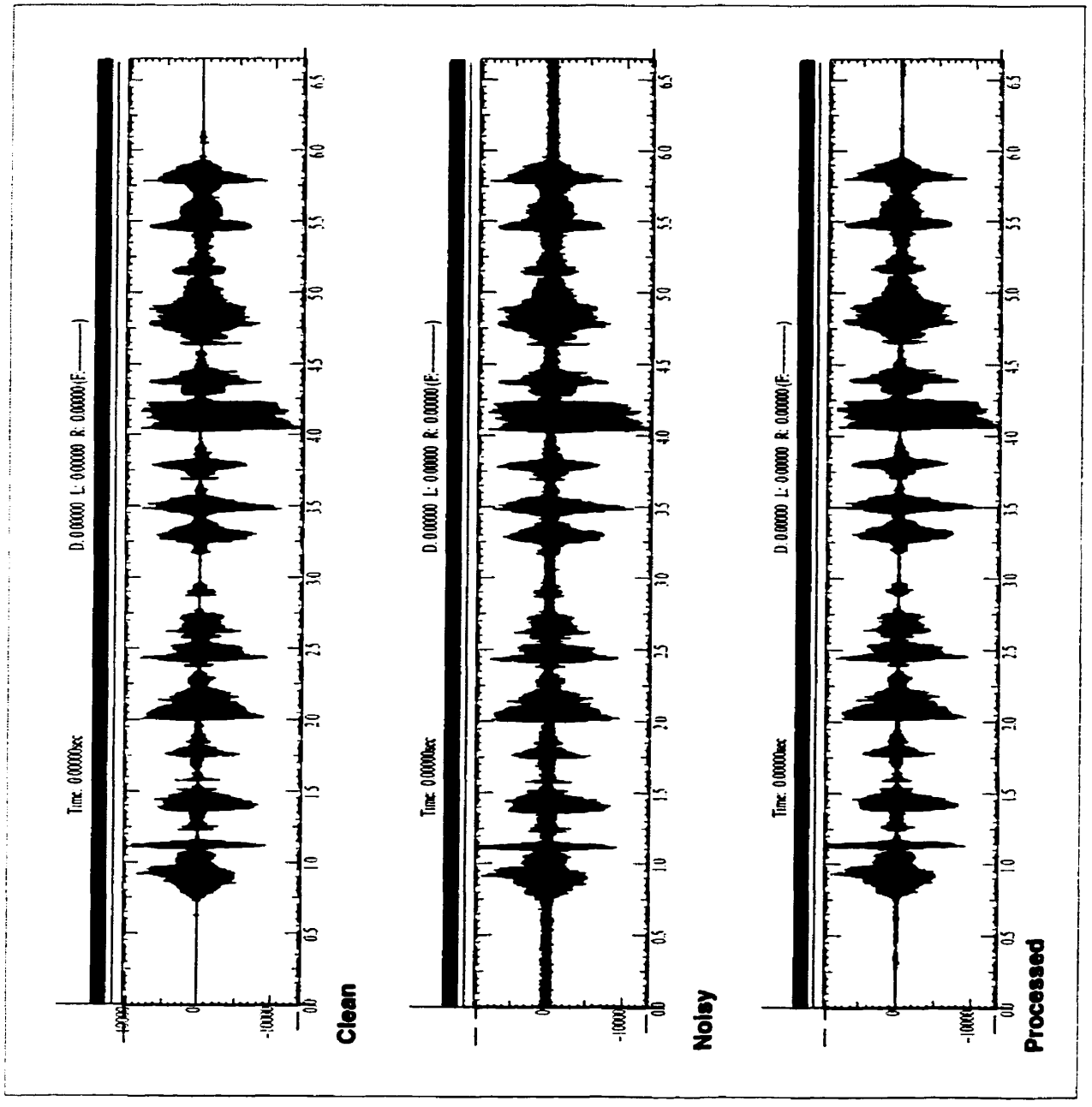
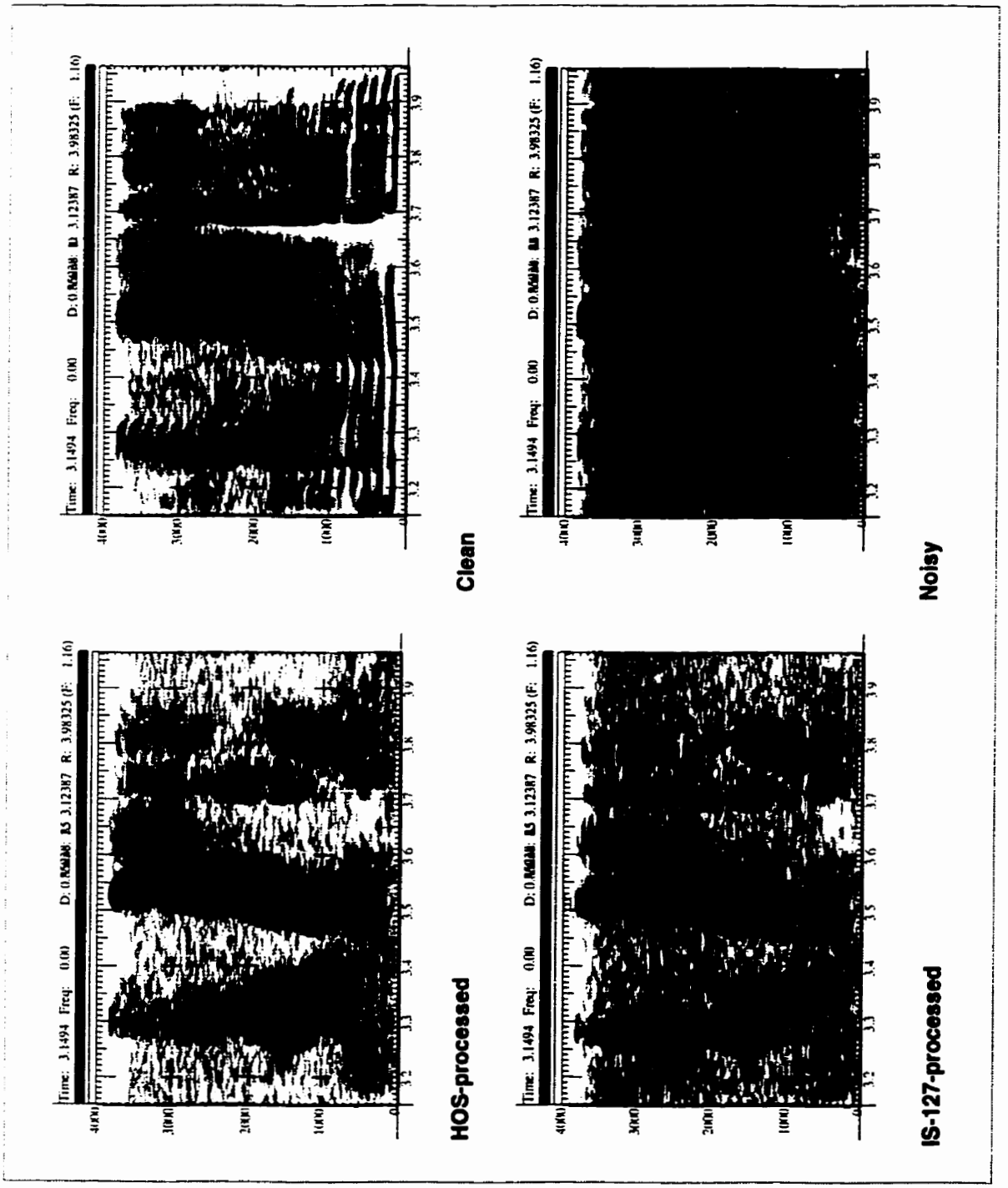


Figure 5-14 Spectrograms for a section in the above waveform (fan noise 12 dB)



5.7.3 Discussion

At the range of the noise levels considered in these experiments, [10, 14 dB], it is found that the degree of noise reduction is overall higher for the HOS-based algorithm than for the TIA's IS-127. This is particularly true for the Gaussian-like noises such as street and fan noise. This fact can be seen in the spectrograms (e.g., Figure 5-12) where the non-speech regions are 'whiter' in the HOS processed case, demonstrating more noise removal. The effectiveness however diminishes in the case of non-Gaussian noise, such as office noise where dominant conversations and impulsive machine noise are not detected as noise given their non-zero HOS. In general, it is found that the noise effectiveness comes at the cost of a slightly more audible noise artifacts in the case of the HOS algorithm. This is due to two phenomena: the first being that the noise estimation is not as accurate in the upper bands as it is in the lower ones, thus resulting in more noise left over in these regions of the spectrum, which in turn results in changing the structure (and thus the perceptual character) of the noise in the resulting processed speech. The second factor is the high variance of the HOS estimator that results in SNR fluctuation during periods of non-speech. Even though the smoothing schemes are effective in reducing this fluctuation, there are nevertheless variations that cause more noise musicality than the TIA algorithm, though it is not very pronounced. Finally, the HOS algorithm is better at preserving the harmonic structure of the speech and results in noticeably less speech distortion in all noise types. This phenomena can be seen by examining the spectrograms (e.g., Figure 5-14) where the harmonics are more visible in the HOS processed case and that in spite of the fact that more noise is removed than in the IS-127 case. This fact is an important one and clearly demonstrates that 4^{th}-order statistics are effective in isolating speech bands containing speech harmonics and preventing overattenuation of these bands or their use as noise bands for noise estimation.

In low SNR conditions (<10 dB), the advantages of the HOS approach start to diminish: Errors related to noise estimation become more pronounced and result in more audible artifacts. In spite of the fact that safety guards are put in place to prevent speech distortions, the estimation errors become too large and generate more noticeable speech loss and higher musicality. It is to note however that the SNR levels chosen here (around 12 dB) cover the majority of the cases in a typical mobile telephony context, and it is rare to be in situations where the SNR is less than 10 dB. It is therefore concluded that the HOS algorithm has merit in that it is effective in the majority of practical situations.

5.8 Conclusion

This chapter presented an algorithm for enhancing speech corrupted by Gaussian noise, using optimal filtering in the time domain, subbands and higher order cumulants. The idea is to use the 4^{th} order statistics to estimate the required parameters for the enhancement filters, such as the SNR, autocorrelation of speech and the probability of speech presence.

The rationale for using a subband approach is twofold:

1. Speech can be easily modeled analytically and the expressions for the HOC that were developed in Chapter 4 are useful in this context, since they are expressed in terms of vital speech parameters, such as energy and autocorrelation.
2. Noise may be assumed flat since the bands are quite narrow. This in turn simplifies the formulation of the optimal filter problem in a way that the problem can then be expressed in terms of matrix algebra involving symmetric and easily invertible matrices. The fact that the noise is uncorrelated also allows quantifying the bias and variance of the HOS estimators (Appendix A).

The resulting algorithm is shown to be effective on typical noises encountered in mobile telephony such as street, office and fan noise. This finding is not surprising, as these noise types contain a significant Gaussian component, being generated by a large number of independent sources that make up this noise (e.g., street noise is the aggregate of pedestrian, traffic, wind and other such sources). The algorithm does not however eliminate the non-Gaussian components in these noises, such as dominant conversations, as these processes are impulsive and do not have zero HOS.

The computational complexity of the algorithm is mostly due to the analysis / synthesis stages which require a relatively large number of filters (40 -> 60). Efficient implementations of these filters will significantly reduce the operations required.

The performance is compared to the TIA IS-127 standard for noise reduction. The results show that the HOS algorithm is better at preserving the harmonic structure of the speech and results in less speech distortion. It also results in overall more reduction of the noise, but that comes at the cost of more noise artifacts, particularly at very low SNR where the variance of the HOS estimators starts to cause large errors in the estimation of the noise and the identification of harmonic bands.

CHAPTER 6 Higher Order Cumulants of LPC-filtered Speech

Synopsis

This chapter takes a similar exploratory approach as Chapter 4 into the HOC properties of speech, with a focus on the LPC residual. It is assumed that an LPC analysis is performed and that the LPC residual has a flat-envelop spectrum where ideally all the short-term correlation is removed. The expression for the horizontal slices of the 3^{rd} and 4^{th} order cumulants are derived assuming the McAulay sinusoidal model. The peculiarities of these cumulants in terms of phase, periodicity and harmonic contents are highlighted. The expressions for the skewness and kurtosis of speech are noted and their use as metrics for detection of voiced speech is discussed. As in Chapter 4, actual speech data is used to assess the validity of the derivations and the underlying sinusoidal model in the LPC residual domain, as far as the HOC are concerned.

6.1 Rationale and Related Work

It is reported in [Fal93] that the normalized skewness and kurtosis of short-term speech segments may be used to detect transitional speech events (termed "innovations"). The conclusion thus drawn is based on the observation that these two statistics take on non-zero values at the boundaries of speech segments where non-stationarity occurs, though no analytical ground is given to support or refute the results. In [Wel85], the bispectrum of voiced segments is observed to have characteristics that distinguish it from unvoiced and Gaussian segments. It is thus concluded that the bispectrum may be used as a voicing indicator, without however providing the analytical backing for this claim.

The work presented in this chapter extends in both scope and formalism relative to the work reported in the literature. The approach adopted is to first establish a reasonable model for speech and use it to analytically infer the HOC properties of speech.The approach of using the LPC residual is motivated by the fact that the (flat) signal in that domain is easier to model analytically thus making the HOC derivations more tractable. The approach is also motivated by experimental results that revealed that the use of the residual provides more distinct characteristics than the original speech signal.

6.2 A Model for the LPC Residual

The model for speech in [McA86] is assumed here, whereby a short speech segment is modelled by a sum of sinusoids that are coherent (in-phase) during voiced speech and incoherent during unvoiced speech:

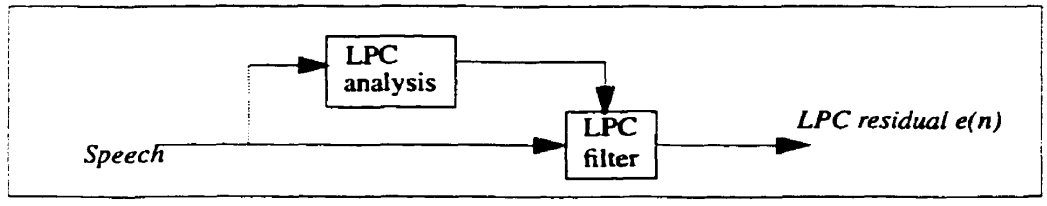
$$s(n) = \sum_{m=1}^{M} a_m \cos\left[(n-n_0)\,w_m + \psi_m + \theta_m\right], \qquad \text{(E 6.1)}$$

where n_0 is the voice onset time, M *is* the number of sinusoids, w_m is the frequency and a_m the amplitude of the m^{th} sine wave. The first phase term in Eq 6.1 is due to the onset time n_0, defined as the time when the pitch pulse occurred relative to the beginning of the frame. The second phase component depends on a frequency cutoff w_c and a voicing probability, denoted by P_v, so that the higher the voicing probability the more sine waves are declared voiced with zero phase. The third phase component is the system phase θ_m along frequency track m, often assumed zero or a linear function of frequency.

The LPC residual signal is the result of filtering the speech signal by the LPC prediction filter (Figure 6-1). Assuming a proper LPC analysis is performed, the residual signal has a flat-envelop spectrum, since all short-term correlation is removed. In the light of the sinusoidal model,

- The residual of *voiced* speech may be modeled as a deterministic signal, consisting of M sinusoids with *equal* amplitudes. The frequencies of these sinusoids may or may not be harmonically related, depending on whether speech is steady or non-stationary.
- The residual of *unvoiced* speech may be modeled as a harmonic process, consisting of M sinusoids with random (and uniformly distributed) phases. In the more general case, *unvoiced* speech may be modeled as a *non-Gaussian white* process
- Gaussian noise at the input becomes white Gaussian in the residual. This feature is particularly interesting in the computation of the variance of the HOS estimators.

Figure 6-1 LPC residual: the result of filtering speech by a short-term prediction filter



6.2.1 Effect of the LPC order

It is worth noting here that the statement about modelling speech in the residual as a sum of equal-amplitude harmonics is true, provided that the number of sinusoids contained in the original signal is greater than half the order of the LPC analysis. Otherwise, the residual signal would be zero, as the prediction filter would exactly match all the poles therein (for simplicity, the windowing effects on the accuracy of the LPC analysis are not accounted for). Simulations showed that for an LPC order of 8 to 12, this phenomena does not occur often, though it may occasionally happen that some voiced segments result in near-zero residual. For this reason, a 10^{th}-order LPC analysis is used in the experimental part of this work. This choice is also motivated by the use of this order of analysis in most speech coders.

6.2.2 Effect of noise on the LPC residual

The assumption about the flat spectral feature of the speech and noise in the residual holds when the original signal consists of either one. If both speech and noise are present, and an autocorrelation-based method is used for LPC analysis, then the residual signal will have a flat spectrum but only in an aggregate sense. The spectral characteristics of the speech component will be highly affected by the SNR and the spectral content of the noise. For the flat nature of speech to hold, a robust method for LPC analysis is required to yield a filter that will only match the speech spectrum. In [Pal91], an approach based on 3^{rd}-order cumulants is used and results in a robust LPC filter that is shown effective in noisy conditions. In the rest of this work, it is assumed that such methods are used and that the flat characteristics of the speech residual hold in all conditions. The effect of a non-flat residual on the analytical derivations will be discussed in the appropriate section.

6.3 Third-Order Cumulant

6.3.1 Stationary Voiced Speech

Voiced speech is modeled as a sum of coherent sine waves whose frequencies are harmonically related to the fundamental. Furthermore, a linear system phase is assumed; as a result, the phases of the sine waves in (Eq 6.1) are entirely determined from the onset time n_0 and a constant due to the system phase.

- **Theorem 1:** *According to the sinusoidal model, the horizontal slice $C_3[\tau]$ of the 3rd-order cumulant of the LPC residual of a steady voiced segment that is bandlimited to $f_s/4$ has M harmonics and the same periodicity as the residual itself. The amplitude of each harmonic may be written in terms of the signal energy (variance) and the number of harmonics M. Moreover, $C_3[\tau]$ has zero phase and reaches maxima at multiples of the pitch lag, namely:*

$$C_3[\tau] = 2c\left(\frac{E_s}{M}\right)^{3/2} \sum_{m=1}^{M} [2M - 1 - m] \cos(m w_0 \tau) \qquad \text{(E 6.2)}$$

where $c = 2^{3/2}/8$ and E_s is the signal energy: $E_s \equiv m_2(0) = M(a^2/2)$.

- **Proof:** Consider the horizontal slice $C_3[\tau] = \frac{1}{N}\sum_n x^2(n)\, x(n-\tau)$ of the 3rd-order cumulant of a deterministic signal. The Fourier transform of this slice can be shown (Section 3.2.1) to be:

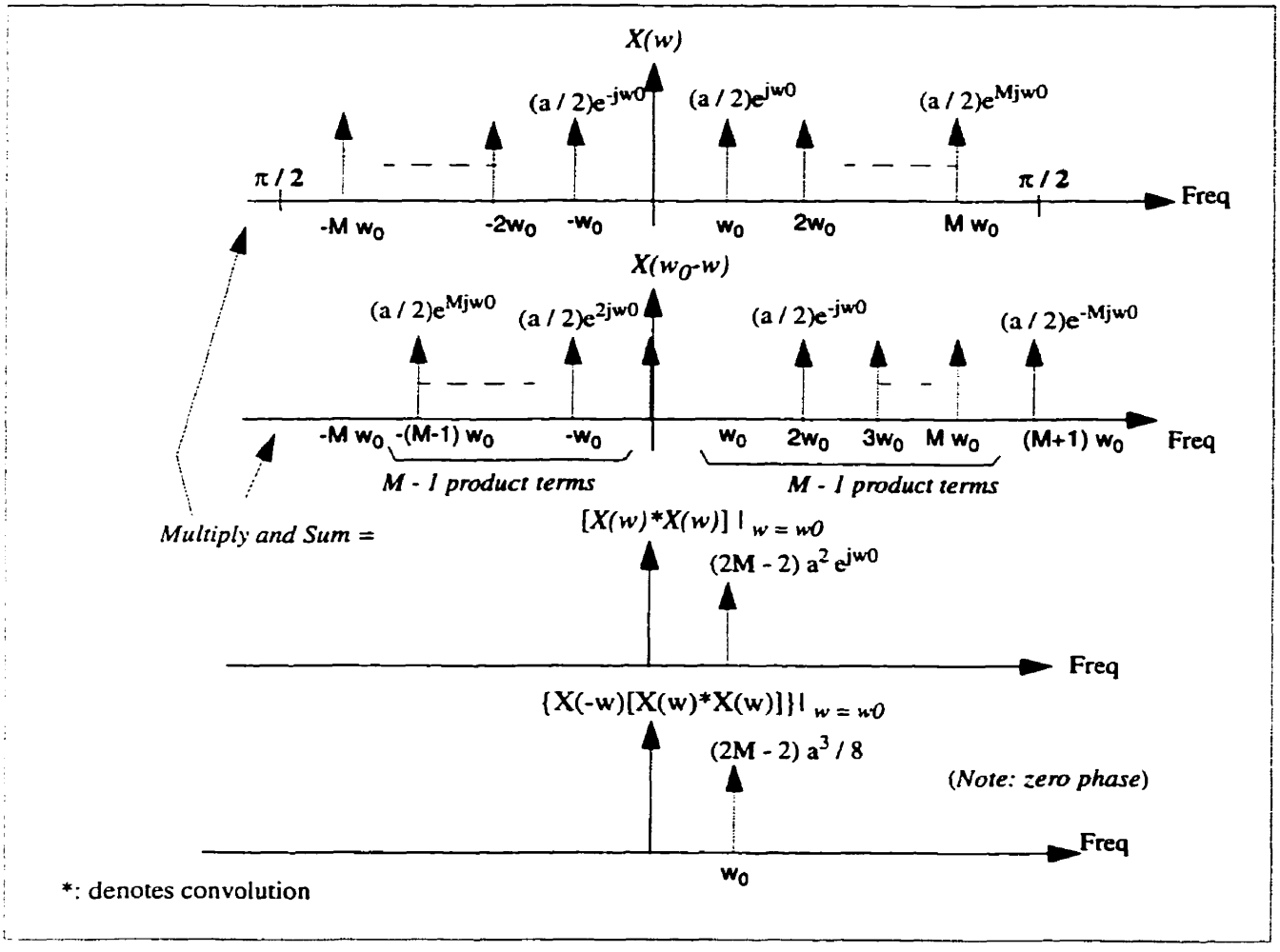
$$FC_0(w) = \{X(w) \otimes X(w)\} X(-w) . \qquad \text{(E 6.3)}$$

Since the signal $x(n)$ consists of M harmonics, its spectrum consists of M impulses on each of the positive and negative frequencies; moreover, the flat spectrum of the LPC residual implies equal magnitude impulses. Therefore, $X(w) = (a/2)\, e^{jkw}$ for $w = \pm(w_0, 2w_0, \ldots, Mw_0)$, and k is a constant that depends on the onset time and the system delay. The convolution of the spectrum is non-zero only at the multiples of the fundamental frequency w_0. In addition, it is assumed that the signal is bandlimited to $\pi/2$ or $f_s/4$ and as a result, there are only $2M$ positive and $2M$ negative lags that lead to non-zero values of the autoconvolution $X(w) \otimes X(w)$. The bandlimited assumption is required to limit the convolution terms to $2M$, otherwise more non-zero terms will result on each side, due the mirror image of the spectrum. Figure 6-2 illustrates the value of the convolution for lag $w = w_0$.

Due to the multiplication by $X(-w)$ in Eq 6.3, only the first M lags on both the negative and positive frequency sides are non-zero. In addition, the phase of $X(w) \otimes X(w)$ cancels out with the phase of $X(-w)$ at any lag and the resulting spectrum $FC_0(w)$ has zero phase. Table 6-1 shows the values of

$FC_0(w)$ for all positive values of the lag w. Moreover, the value of $FC_0(w)$ may be expressed in terms of signal energy since $a^3 = \left(\frac{2E_s}{M}\right)^{3/2}$.

Figure 6-2 Computing the value of $FC_0(w)$ for $w = w_0$



Due to spectral symmetry, the results are the same for negative lags; consequently the inverse transform of $FC_0(w)$ leads to the sum of M cosine terms given by Eq 6.2, from which it is evident that $C_3[\tau]$ has maximum at $\tau = 0, P, 2P...$ (where P is the pitch period i.e. $2\pi/w_0$).

Table 6-1 Value of $FC_0(w)$ at all positive lags

Lag (w)	X(w)*X(w)	X(-w)	$FC_0(w)$
0	$2Ma^2/4$	0	0
w_0	$[(M-1)+(M-1)]\,a^2/4$ *Phase:* e^{jkw0}	$a/2$ *Phase:* e^{-jkw0}	$(2M-2)\,a^3/8$
$2w_0$	$[(M-2)+(M-1)]\,a^2/4$ *Phase:* e^{j2kw0}	$a/2$ *Phase:* e^{-j2kw0}	$(2M-3)\,a^3/8$
$3w_0$	$[(M-3)+(M-1)]\,a^2/4$ *Phase:* e^{j3kw0}	$a/2$ *Phase:* e^{-j3kw0}	$(2M-4)\,a^3/8$
.....	.....		.....
$(M-1)w_0$	$[1+(M-1)]\,a^2/4$ *Phase:* $e^{j(M-1)kw0}$	$a/2$ *Phase:* $e^{-j(M-1)kw0}$	$M\,a^3/8$
$M\,w_0$	$[0+(M-1)]\,a^2/4$ *Phase:* e^{jMkw0}	$a/2$ *Phase:* e^{-jMkw0}	$(M-1)\,a^3/8$

Moreover, the amplitude of the normalized cumulant slice, $C'_3[\tau]$, is only a function of M, i.e., the effect of signal energy has been eliminated. This result is interesting since it draws on a similar effect when dealing with a random process: the normalized 3rd-order statistics is equivalent to having an input data process with a variance of one. Finally, the zero-phase characteristic of the third cumulant is in agreement with the general property derived in [Nik87] that the bispectrum (and thus the third-order cumulant) is insensitive to time shifts.

- **Corollary 2:** *The skewness of the LPC residual of steady voiced speech may be written as a function of the energy of the residual and the number of harmonics M. The normalized skewness is only function of M and is greater than zero for any practical values of M (which is function of the pitch).*

- **Proof:** The skewness is found by setting $\tau = 0$ in Eq 6.2:

$$C_3[0] = 2c\left(\frac{E_s}{M}\right)^{3/2}\sum_{k=1}^{M}[2M-1-k]$$

$$= 2c\left(\frac{E_s}{M}\right)^{3/2}\{M(M-1)+[1+2+3+\ldots+(M-1)]\}$$

$$= 2c\left(\frac{E_s}{M}\right)^{3/2}\{M(M-1)+\frac{(M-1)M}{2}\}$$

$$= 3c\left(\frac{E_S}{M}\right)^{3/2}\{M(M-1)\} = 3c(E_S)^{3/2}\left[\frac{(M-1)}{\sqrt{M}}\right]$$

The skewness is: $$\boxed{C_3[0] = 3c(E_s)^{3/2}\left[\frac{(M-1)}{\sqrt{M}}\right]} \quad \text{(E 6.4)}$$

The normalized skewness is: $$\boxed{\gamma_3 \equiv \frac{C_3[0]}{E_S^{3/2}} = 3c\frac{(M-1)}{\sqrt{M}}} \quad \text{(E 6.5)}$$

6.3.2 Non-Stationary Voiced Speech

In the case of a non-steady voiced segment, not all harmonics are related and as result, the value of $C_3[\tau]$ may be zero for some of the harmonics. Clearly, if the frame is completely non-stationary, then no three frequencies are related and the 3rd-order cumulant is zero for all lags. In practice however, we expect that even for non-stationary voiced segments, a subset of the harmonics are related to the fundamental and thus $C_3[\tau]$ is rarely zero for voiced speech.

6.3.3 Unvoiced Speech

6.3.3.1 Assuming a non-Gaussian white process

Since the LPC signal is assumed to have a flat envelop, the bispectrum is 2D-flat across all bifrequencies:

$$B(w_1, w_2) = E[X(w_1)X(w_2)X(-w_1-w_2)] = \gamma.$$

The third-order cumulant, being the inverse Fourier transform of the bispectrum, thus consists of a 2D delta function of amplitude γ:

$$C_3[\tau_1, \tau_2] = \gamma\delta(\tau_1, \tau_2). \quad \text{(E 6.6)}$$

6.3.3.2 Assuming a harmonic process

The LPC residual is modeled as a sum of sinusoids with random phases. In Section 3.2.3, it was shown that the 3rd-order cumulant of a single such sinusoid is zero. Since all sinusoids are statistically independent, then the cumulant of the sum is the sum of the cumulants [Men91] and the 3rd cumulant of the LPC residual of unvoiced speech is identically zero:

$$C_3[\tau_1, \tau_2] = 0. \quad \text{(E 6.7)}$$

6.4 Fourth-Order Cumulant

6.4.1 Unvoiced Speech

6.4.1.1 Assuming a non-Gaussian white process

Both the power spectrum and the trispectrum of the LPC residual of unvoiced speech are flat-envelop: The trispectrum is: $T_e(w_1, w_2, w_3) = \gamma$ and the power spectrum: $P_e(w) = \alpha$. As a result, the fourth moment of the residual is a delta function:

$$m_4(\tau_1, \tau_2, \tau_3) = \gamma\delta(\tau_1, \tau_2, \tau_3) \quad \text{(E 6.8)}$$

and the second moment (i.e., the autocorrelation) is another delta function:

$$m_2(\tau) = \alpha\delta(\tau) . \quad \text{(E 6.9)}$$

Using Eq 3.16, the diagonal slice of the 4th-order cumulant of the residual may then be written as:

Diagonal Slice: $\boxed{C^a_4[\tau] = [\gamma - 3\alpha^2]\,\delta(\tau)}$ (E 6.10)

Kurtosis: $\boxed{C^a_4[0] = \gamma - 3\alpha^2}$ (E 6.11)

6.4.1.2 Assuming a harmonic process

The LPC residual of unvoiced speech is modeled as a sum of non-harmonically related sinusoids with uniformly distributed random phases. In Section 4.3.3, it was shown that for the case of a single sinusoid with random phase, the diagonal slice of the 4th order cumulant is given by:

$$C^a_4[\tau] = -1.5\cos(w_0\tau)\,[E_S]^2 .$$

For the case of M independent sinusoids, the cumulant of the sum is the sum of cumulants. As a result, the above equation can be extended for unvoiced speech by simply summing the cumulants of the individual sinewaves. The flat LPC spectrum implies all amplitudes are equal. Thus:

The diagonal slice: $\boxed{C^a_4[\tau] = -1.5\,[E_S]^2 \sum_{m=1}^{M} \cos[w_m\tau]}$ (E 6.12)

The kurtosis: $\boxed{C_4[0] = -1.5M\,[E_S]^2}$ (E 6.13)

6.4.2 Voiced Speech

The LPC residual of voiced speech is a deterministic signal, that is modeled as either:

1. A sum of coherent sinusoids with frequencies that are multiple of the fundamental.
2. A sum of non-harmonically related sinusoids with unknown (but deterministic) phases and frequencies.
3. A general signal, not necessarily harmonic, with a flat spectrum.

6.4.2.1 Common property

- **Theorem 2:** *If voiced speech is modeled as a deterministic harmonic signal, then the DC component of the horizontal slice of the 4^{th}-order cumulant ($C^b_4[\tau]$) of the LPC residual of voiced speech (both steady and non-stationary) may be written in terms of the signal energy and the number of harmonics.*
- **Proof:** In Section 3.3.2, it is shown that the DC value of the horizontal slice of the 4^{th}-order cumulant slice of a deterministic signal may be written in terms of the integral of the 4^{th} powers of the magnitude spectrum:

$$FC^b_4(0) = -2\int_{-\pi}^{\pi} P(\lambda)\,P(-\lambda)\,d\lambda = -2\int_{-\pi}^{\pi} |X(\lambda)|^4 d\lambda. \qquad \text{(E 6.14)}$$

Here, the spectrum $X(f)$ consists of delta functions at frequencies f_m and amplitudes $a_m/2$, therefore:

$$FC^b_4(0) = -\frac{1}{4}\sum_{m=1}^{M} a_m^4$$

where a_m is the amplitude of the M^{th} sinusoid. In the LPC residual, all these are equal, thus:

$$\boxed{DC\{C^b_4[\tau]\} = FC^b_4(0) = -M(a^4/4) = \frac{-[E_s]^2}{M}} \qquad \text{(E 6.15)}$$

- **Theorem 3:** *If voiced speech is modeled as a general deterministic signal, then the DC component of the horizontal slice of the 4^{th}-order cumulant ($C^b_4[\tau]$) of the LPC residual of voiced speech (both steady and non-stationary) may be written in terms of the signal energy and bandwidth.*
- **Proof:** This follows directly from Corollary 1 in Section 3.3.2, given that the LPC residual has a flat-envelop spectrum of magnitude a, and is bandlimited to [-BW:BW]. The power spectrum is flat-envelop and has magnitude a^2. The DC component of the horizontal slice is:

$$FC^{b}_{4}(0) = -2\int_{-\pi}^{\pi} |P(\lambda)|^2 d\lambda = -2(2a^4 BW) = -\frac{E^2}{BW}. \quad \text{(E 6.16)}$$

6.4.2.2 Steady voiced speech

- **Theorem 4:** *According to the sinusoidal model, the 4th-order cumulant slice $C^{b}_{4}[\tau]$ of the LPC residual signal of steady voiced speech that is bandlimited to $f_s/4$ is made of 2M-1 harmonics and has the same periodicity as the underlying signal. The value at each harmonic may be written in terms of the energy of the signal and the number of harmonics. Moreover, $C^{b}_{4}[\tau]$ has zero phase and maxima at multiples of the pitch lag.*
- **Proof:** For a deterministic signal, $C^{b}_{4}[\tau]$ is:

$$C^{b}_{4}[\tau] = \left[\frac{1}{N}\sum_{n} x^2(n)\,x^2(n+\tau)\right] - \left[\frac{1}{N}\sum_{n} x^2(n)\right]^2 - 2\left[\frac{1}{N}\sum_{n} x(n)\,x(n+\tau)\right]^2 \quad \text{(E 6.17)}$$

and its Fourier transform is (from Section 3.3.1):

$$FC^{b}_{4}(w) = |X(w) \otimes X(w)|^2 - [m_2(0)]^2\delta(w) - 2\{P(w) \otimes P(w)\}. \quad \text{(E 6.18)}$$

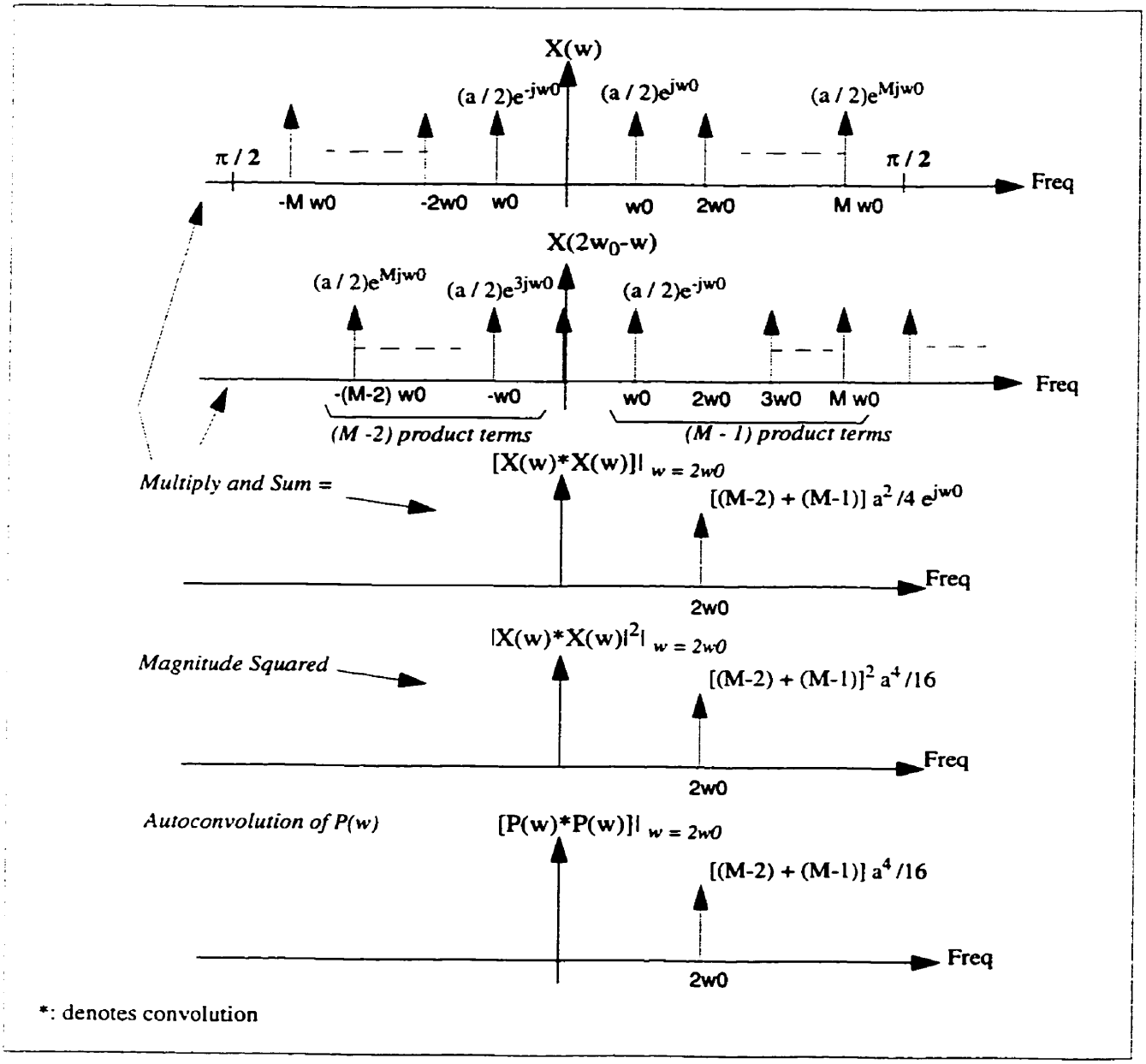
Since the signal $x(n)$ consists of M harmonics, its spectrum is made of M delta functions on each of the positive and negative frequencies; moreover, the flat spectrum of the LPC residual implies equal magnitude impulses. Therefore, $X(w) = (a/2)e^{jkw}$ for $w = \pm(w_0, 2w_0, \ldots, Mw_0)$, and k is a constant that depends on the onset time and the system delay. As in Theorem 1, it is assumed that the signal is bandlimited to $\pi/2$ or $f_s/4$ and as a result, there are only $2M$ positive and $2M$ negative lags that lead to non-zero values of the autoconvolution $X(w) \otimes X(w)$ as well as the autoconvolution $P(w) \otimes P(w)$. Therefore, $FC_4(w)$ has $2M$ non-zero values on each side of the spectrum, and only at multiples of w_0 as seen in Figure 6-3.

Clearly, the phase of $FC^{b}_{4}(w)$ is zero for all frequencies w since each term in Eq 6.18 has zero phase. Table 6-2 shows the various values of $FC^{b}_{4}(w)$ for all positive lags w. Due to spectral symmetry, the values are the same for negative lags. The signal energy is: $m_2(0) = E_S = M(a^2/2)$ and $[m_2(0)]^2 = E_S^2 = M^2(a^4/4)$, it follows that the magnitudes at the various harmonics (column 3) may be expressed in terms of E_S^2. As seen from Table 6-2, there are $2M-1$ non-zero values. However, depending on the value of M, some other harmonics may be zero as well.

Table 6-2 Value of $FC^b_4(w)$ for all positive lags

Lag (w)	$\|X(w)*X(w)\|^2$	$[m_2(0)]^2\delta(w)$	$P(w)*P(w)$	$FC^b_4(w)$
0	$4\,M^2\,a^4/16$	$M^2\,a^4/4$	$2\,M\,a^4/16$	$-\,M\,a^4/4$
w_0	$[(M-1)+(M-1)]^2\,a^4/16$	0	$[(M-1)+(M-1)]\,a^4/16$	$[(2M-2)(2M-4)]\,a^4/16$
$2w_0$	$[(M-2)+(M-1)]^2\,a^4/16$	0	$[(M-2)+(M-1)]\,a^4/16$	$[(2M-3)(2M-5)]\,a^4/16$
$3w_0$	$[(M-3)+(M-1)]^2\,a^4/16$	0	$[(M-3)+(M-1)]\,a^4/16$	$[(2M-4)(2M-6)]\,a^4/16$
.....	.....	0	.....	.....
.....	.....	0	.....	.....
$(M-1)w_0$	$[1+(M-1)]^2\,a^4/16$	0	$[1+(M-1)]\,a^4/16$	$[M(M-2)]\,a^4/16$
$M\,w_0$	$[0+(M-1)]^2\,a^4/16$	0	$[0+(M-1)]\,a^4/16$	$[(M-1)(M-3)]\,a^4/16$
$(M+1)w_0$	$M^2\,a^4/16$	0	$M\,a^4/16$	$M(M-2)\,a^4/16$
$(M+2)w_0$	$(M-1)^2\,a^4/16$	0	$(M-1)\,a^4/16$	$(M-1)(M-3)\,a^4/16$
$(M+3)w_0$	$(M-2)^2\,a^4/16$	0	$(M-2)\,a^4/16$	$(M-2)(M-4)\,a^4/16$
.....	.....	0	.....	.....
$(2M-1)w_0$	$4\,a^4/16$	0	$2\,a^4/16$	0
$2Mw_0$	$a^4/16$	0	$a^4/16$	$-\,a^4/16$

Figure 6-3 $|X(w)*X(w)|^2$ and $P(w)*P(w)$ for $w = 2w_0$



*: denotes convolution

- **Corollary 3:** *The kurtosis of the LPC residual of steady voiced speech may be expressed in terms of speech energy and the number of harmonics. The normalized kurtosis is a function of the number of harmonics only and is greater than zero for any practical value of the pitch, namely:*

The kurtosis: $$\boxed{C_4(0) = E_s^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]} \quad \text{(E 6.19)}$$

The normalized kurtosis: $$\boxed{\gamma_4 \equiv \frac{C_4[0]}{E_s^2} = \frac{4}{3}M - 4 + \frac{7}{6M}} \quad \text{(E 6.20)}$$

- **Proof:** The value of the 4th moment may be determined in the frequency domain, by summing the coefficients of the Fourier transform of $\frac{1}{N}\sum_n x^2(n)\,x^2(n+\tau)$ since:

$$\frac{1}{N}\sum_n x^2(n)\,x^2(n+\tau) \leftrightarrow |X(w) \otimes X(w)|^2 \text{ Transform pair.}$$

Therefore, the value at $\tau = 0$ is: $\frac{1}{N}\sum_n x^4(n) = \int_{-\pi}^{\pi} |X(w) \otimes X(w)|^2 dw$.

The value of the Fourier coefficients $|X(w) \otimes X(w)|^2$ is given in the first column of Table 6-2. Due to spectral symmetry, the value of the sum over all frequency lags is simply twice the value over the positive lags plus the value at lag 0. Furthermore, the sum over the positive lags may be divided in 2 groups: *Group 1* includes lags w_0 through $(M-1)w_0$ and *Group 2* includes lags Mw_0 through $2Mw_0$. Therefore:

$$\int_{-\pi}^{\pi} |X(w) \otimes X(w)|^2 dw = |X(w) \otimes X(w)|^2\big|_{w=0} + 2\int_{Group1} |X(w) \otimes X(w)|^2 dw + 2\int_{Group2} |X(w) \otimes X(w)|^2 dw$$

The integral over the lags of Group 2 is (from Table 6-2):

$$\int_{Group2} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{16}\{1 + 2^2 + 3^2 + \ldots + M^2\} = \frac{a^4}{16}\left(\frac{M(M+1)(2M+1)}{6}\right),$$

and the integral over the lags of Group 1 is:

$$\int_{Group1} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{16}\{\underbrace{(M-1)^2 + M^2 + (M+1)^2 + (M+2)^2 + \ldots + (M+(M-2))^2}_{M \text{ terms}}\}$$

$$= \frac{a^4}{16}\sum_{i=1}^{M}[i + (M-2)]^2 = \frac{a^4}{16}\sum_{i=1}^{M}[i^2 + 2i(M-2) + (M-2)^2]$$

$$= \frac{a^4}{16}\left[\frac{M(M+1)(2M+1)}{6} + M(M-2)(M+1) + M(M-2)^2\right]$$

$$\int_{Group1} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{16}\left[\frac{M(M+1)(2M+1)}{6} + M(M-2)(2M-1)\right].$$

Combining Groups 1 and 2 yields:

$$\int_{Group1+Group2} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{16}\left[\frac{M(M+1)(2M+1)}{3} + M(M-2)(2M-1)\right]$$

$$\int_{Group1+Group2} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{16}\left[M^3\left(2+\frac{2}{3}\right) + M\left(2+\frac{1}{3}\right) - 4M^2\right]. \quad \text{(E 6.21)}$$

Multiplying Eq 6.21 by the scale factor of 2, and adding the value at lag zero yields the final result for the 4th moment:

$$\boxed{\frac{1}{N}\sum_n x^4(n) = \int_{-\pi}^{\pi} |X(w) \otimes X(w)|^2 dw = \frac{a^4}{8}\left[\frac{8}{3}M^3 - 2M^2 + \frac{7}{3}M\right]} \quad \text{(E 6.22)}$$

Noting that the value of the second moment (signal energy) is: $E_s \equiv \frac{1}{N}\sum_n x^2(n) = Ma^2/2$, Eq 6.22 may be written in terms of E_s^2 as:

$$\frac{1}{N}\sum_n x^4(n) = \frac{a^4M^2}{8}\left[\frac{8}{3}M - 2 + \frac{7}{3M}\right] = \frac{E_s^2}{2}\left[\frac{8}{3}M - 2 + \frac{7}{3M}\right].$$

The kurtosis is determined by first setting $\tau = 0$ in Eq 6.17:

$$C_4[0] = \frac{1}{N}\sum_n x^4(n) - 3\left[\frac{1}{N}\sum_n x^2(n)\right]^2. \quad \text{(E 6.23)}$$

and using Eq 6.22 for the value of the 4th moment, the Kurtosis becomes:

$$C_4[0] = E_s^2\left[\frac{4}{3}M - 1 + \frac{7}{6M}\right] - 3E_s^2$$

$$Kurtosis \equiv C_4[0] = E_s^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]$$

and the normalized Kurtosis is simply:

$$\gamma_4 \equiv \frac{C_4[0]}{E_s^2} = \frac{4}{3}M - 4 + \frac{7}{6M}.$$

6.5 Summary of the Derivations

6.5.1 Third-Order Cumulant

- **Steady voiced speech**

The horizontal slice $C_3[\tau]$ of the third-order cumulant of the LPC residual of a steady voiced segment has M harmonics and the same periodicity as the residual itself. The amplitude of each harmonic may be written in terms of the signal energy (variance) and the number of harmonics M. Moreover, $C_3[\tau]$ has zero phase and reaches maximum at multiples of the pitch lag, namely:

$$C_3[\tau] = 2c\left(\frac{E_S}{M}\right)^{3/2} \sum_{m=1}^{M} [2M - 1 - m] \cos(m w_0 \tau) ,$$

and the normalized skewness is: $\gamma_3 \equiv \dfrac{C_3[0]}{E_S^{3/2}} = 3c\dfrac{(M-1)}{\sqrt{M}}$.

- **Non-stationary voiced:**

If no three harmonic frequencies are harmonically related, then the 3rd-order cumulant is zero. In reality it is expected to be non-zero, partially due to non-linearity and partially to some randomly related frequencies.

- **Unvoiced Speech:**

If unvoiced speech is modeled as a harmonic process, then its 3rd-order cumulant is zero. If it is modeled as a non-Gaussian white process, then its cumulant is a 2D-delta function:

$$C_3[\tau_1, \tau_2] = \gamma\delta(\tau_1, \tau_2) .$$

6.5.2 Fourth-Order Cumulant

- **Voiced Speech (steady and nonstationary)**

If voiced speech is modeled as a *harmonic signal*, then the DC component of the horizontal slice of the 4th-order cumulant ($C^b_4[\tau]$) of the LPC residual can be written in terms of the energy and number of harmonics:

$$DC\{C^b_4[\tau]\} = -M(a^4/4) = \frac{-[E_S]^2}{M}.$$

If voiced speech is modeled as a general deterministic signal, then the DC component of the horizontal slice of the 4th cumulant ($C^b_4[\tau]$) of the LPC residual may be written in terms of the signal energy and bandwidth:

$$DC\{C^b_4[\tau]\} = -2(2a^4B) = -\frac{E_S^2}{B}.$$

- **Steady voiced**
  - —The 4th cumulant slice $C^b_4[\tau]$ of the LPC residual is made of $(2M-1)$ harmonics and has the same periodicity as the underlying signal. The value at each harmonic may be written in terms of the energy of the signal and the number of harmonics. Moreover, $C^b_4[\tau]$ has zero phase and maximums at multiples of the pitch lags.
  - —The normalized kurtosis may also be expressed in terms of the number of harmonics M, and is also greater than zero for any practical value of M, namely:

$$\gamma_4 \equiv \frac{C^b_4(0)}{[m_2(0)]^2} = \frac{4}{3}M - 4 + \frac{7}{6M}.$$

- **Unvoiced Speech (assuming a non-Gaussian white process)**

  Diagonal Slice: $C^a_4[\tau] = [\gamma - 3\alpha^2]\,\delta(\tau)$.

  The kurtosis: $C_4[0] = \gamma - 3\alpha^2$.

- **Unvoiced Speech (assuming a harmonic process):**

  The diagonal slice: $C^a_4[\tau] = -1.5[E_S]^2 \sum_{m=1}^{M} \cos[w_m\tau]$.

  The kurtosis: $C_4[0] = -1.5M[E_S]^2$.

6.6 Effect of Noise

6.6.1 Effect of Noise on γ_3 and γ_4

When the signal consists of both speech and noise, then $x(n) = s(n) + g(n)$. If $s(n)$ and $g(n)$ are statistically independent, then the energy of $x(n)$ is the sum of speech and noise energies: $E_x = E_s + E_g$. Second-order statistics are thus directly affected and in an additive way by the presence of noise. Higher-order statistics on the other hand are immune to Gaussian noise, which has zero HOS. Since cumulants are cumulative [Men91], it follows that the 3rd and 4th order cumulants of $x(n)$ are simply those of $s(n)$. As a result, the above derivations for $C_3[0]$ and $C_4[0]$ still hold in the presence of Gaussian noise. However, when normalizing these two quantities by the signal energy E_x, the effect of the noise term in the denominator does not cancel out with the speech energy term E_s in the numerator in Eq 6.4 and Eq 6.19. It is easy to see that the expressions for the normalized skewness (Eq 6.5) and kurtosis (Eq 6.20) of noisy speech can now be extended to include an SNR term as follows:

$$\gamma_3 \equiv \frac{C_3[0]}{E_x^{3/2}} = 3c\left(\frac{E_s}{E_s + E_N}\right)^{3/2}\left[\frac{M-1}{\sqrt{M}}\right] \quad \text{or simply,}$$

$$\boxed{\gamma_3 = 3c\left(\frac{SNR}{SNR+1}\right)^{3/2}\left[\frac{M-1}{\sqrt{M}}\right]} \qquad \text{(E 6.24)}$$

$$\gamma_4 \equiv \frac{C_4[0]}{E_x^2} = \frac{C_4[0]}{[E[s^2(n)] + E[g^2(n)]]^2} = \left(\frac{4}{3}M - 4 + \frac{7}{6M}\right)\left(\frac{E[s^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2$$

$$\boxed{\gamma_4 = \left(\frac{SNR}{SNR+1}\right)^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]} \qquad \text{(E 6.25)}$$

Therefore, the effectiveness of these two metrics to detect voicing decreases with the *SNR*. Due to the power of 2, the normalized kurtosis is more adversely affected by the presence of noise than the skewness.

If the noise is non-Gaussian but has a symmetrical distribution, then only its 3rd-order statistics are zero and the above reasoning holds for the normalized skewness but not the kurtosis. Consider the hypothetical case of Laplacian noise. Using the cumulative property of the HOS and the fact that the kurtosis of a Laplacian process may be written in terms of its energy[1] as: $C_{4_g}[0] = 3(E[g^2(n)])^2$, then the normalized kurtosis of noisy speech becomes:

1.By evaluating the 2nd and 4th moments of the Laplacian pdf.

$$\gamma_4 \equiv \frac{C_{4S}[0] + C_{4g}[0]}{E_x^2} = \frac{C_{4S}[0] + 3E[g^2(n)]}{[E[s^2(n)] + E[g^2(n)]]^2}$$

$$= \left(\frac{4}{3}M - 4 + \frac{7}{6M}\right)\left(\frac{E[s^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2 + \left(\frac{3E[g^2(n)]}{E[s^2(n)] + E[g^2(n)]}\right)^2$$

$$\gamma_4 = \left(\frac{SNR}{SNR+1}\right)^2\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right] + 3 \cdot \left(\frac{1}{SNR+1}\right)^2.$$

The effect is similar to the case of Gaussian noise, in that the effectiveness of the kurtosis degrades with the SNR. However, the problem of distinguishing speech from noise based on this metric needs to be reformulated since the normalized kurtosis of the noise is no longer zero, but 3 in this case.

6.6.2 Effect of a non-flat LPC residual

It was mentioned in Section 6.2.2 that noise may adversely affect the LPC analysis, which will result in a residual that is flat in an aggregate sense, though the speech component itself is not flat. Even in the case where noise is not present, one would not expect the LPC residual to be perfectly flat since the LPC analysis itself is seldom perfect. It is therefore necessary to assess the implication of a non-flat residual on the properties of the HOS of speech derived so far.

In the above derivations of the 3rd and 4th order cumulants, the flat envelop characteristic of the signal caused the values of the autoconvolution of the spectrum to be expressed in terms of the signal energy. This in turn implied that the amplitude at each harmonic of the cumulants may be expressed in terms of the speech energy and number of harmonics, which led to closed-form equations for the skewness and kurtosis. If the flat spectrum assumption is no longer true, then clearly the autoconvolution can no longer be expressed in terms of the signal energy but will be a combination of the amplitudes at the various signal harmonics. It is easy to see however that the overall behavior of these cumulants will not be significantly changed, specifically:

- The zero-phase characteristic of the cumulant slices still holds and as a result the skewness and kurtosis still represent the maximum values of these functions.
- The harmonic nature of the cumulant slices remains unchanged, though the harmonic magnitudes may be different.

- The skewness and kurtosis are still positive but may not be expressed in terms of the speech energy. Instead, their value is a non-closed-form function of the speech amplitudes at the various harmonics. As a result, the normalization by the energy (in the expressions of γ_3 and γ_4) will no longer cancel out the effect of energy in the numerator and the normalized metrics are no longer independent of signal amplitude as is the case in Eq 6.5 and Eq 6.20.

6.7 Results Using Speech Signals

The derivations presented above are verified using recorded clean speech. A 10^{th}-order LPC analysis is performed on speech sampled at 8 kHz. The normalized 3^{rd} and 4^{th} order cumulants functions of the LPC residual are computed using frames of 20 msec (160 samples) with a 25% overlap. The residual is low-passed using a 60-tap FIR filter with a cutoff at 1.8 kHz. In order to avoid the problem of additional harmonics reported in [Rui95], the cumulant slices are computed as:

$$C_3[\tau] = \frac{1}{N-K}\sum_{n=K}^{N-1} x^2(n)\,x(n-\tau)$$

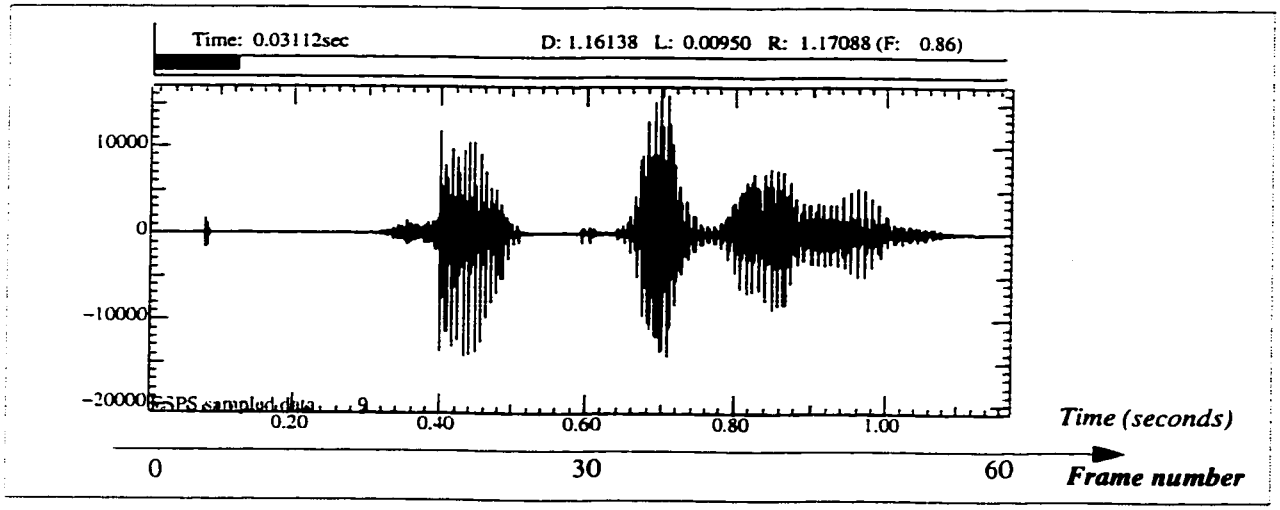
$$C_4[\tau] = \left[\frac{1}{N-K}\sum_{n=0}^{N-K} x^2(n)\,x^2(n+\tau)\right] - \left[\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)\right]^2 - 2\left[\frac{1}{N-K}\sum_{n=0}^{N-K} x(n)\,x(n+\tau)\right]^2$$

where K is the maximum lag and N is the number of points in the frame. The idea is to keep the limit of the summation constant for all lags τ.

6.7.1 Voiced Speech

The waveform for the utterance "*help the woman*" spoken by a male speaker is shown in Figure 6-4. Mild Gaussian noise was added (at 30 dB SNR) to avoid zero level signals. The normalized skewness and kurtosis of the LPC residual are shown in Figure 6-5 for a number of frames.

Figure 6-4 The utterance "Help the woman"



From these it is observed that the skewness and kurtosis of voiced segments are both greater than zero as expected. It is also worth noting that the two normalized metrics take on large positive (kurtosis) or negative (skewness) values for transient and small amplitude segments (for example at frame 30 in Figure 6-5). This is mainly due to the small energy of these frames that is used for normalization, thus resulting in large normalized HOS for any transient segment, i.e., when only some of the samples are non-zero. Consequently, the normalized metrics by themselves are not sufficient for detecting voiced frames as they take on erroneously large values for small transient segments.

To examine the distribution of the kurtosis, histograms of the frame-by-frame values of the normalized kurtosis were generated for 500 frames (10 seconds). Another histogram was generated for the normalized kurtosis when Gaussian noise is used prior to LPC filtering. These histograms are shown in Figure 6-6 and clearly show the difference in the 4^{th}-order statistics between speech and Gaussian noise. It is to note here that the speech utterance contains silence periods and thus the kurtosis would be zero some of the time, as evident from the histograms.

Figure 6-5 Normalized skewness and kurtosis of the LPC residual

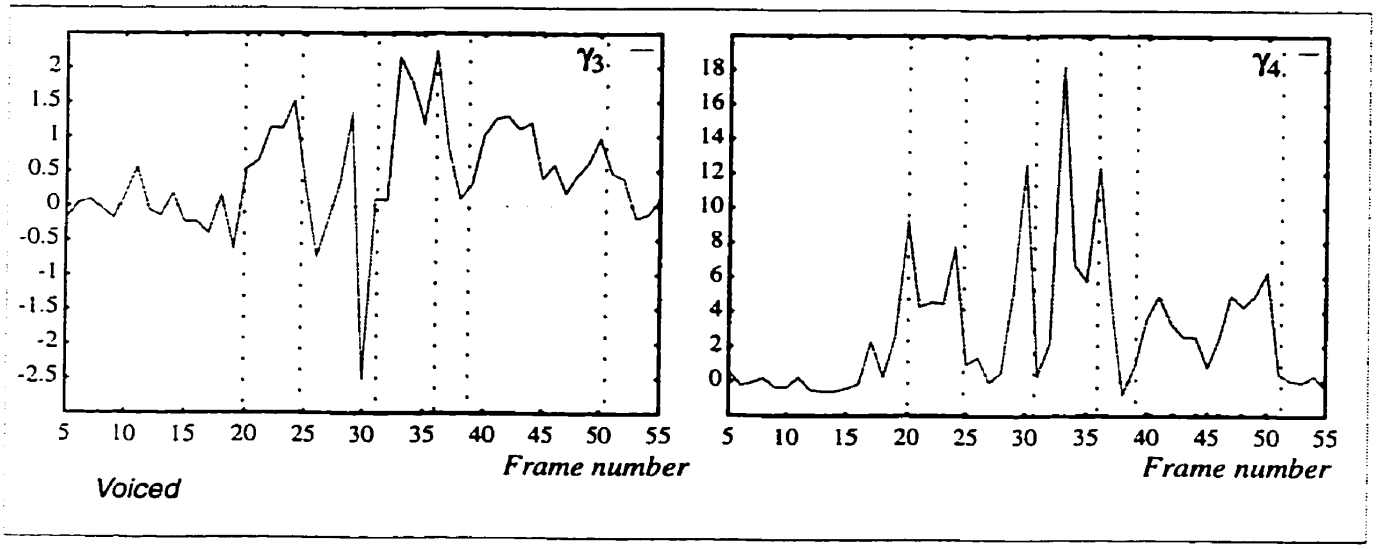
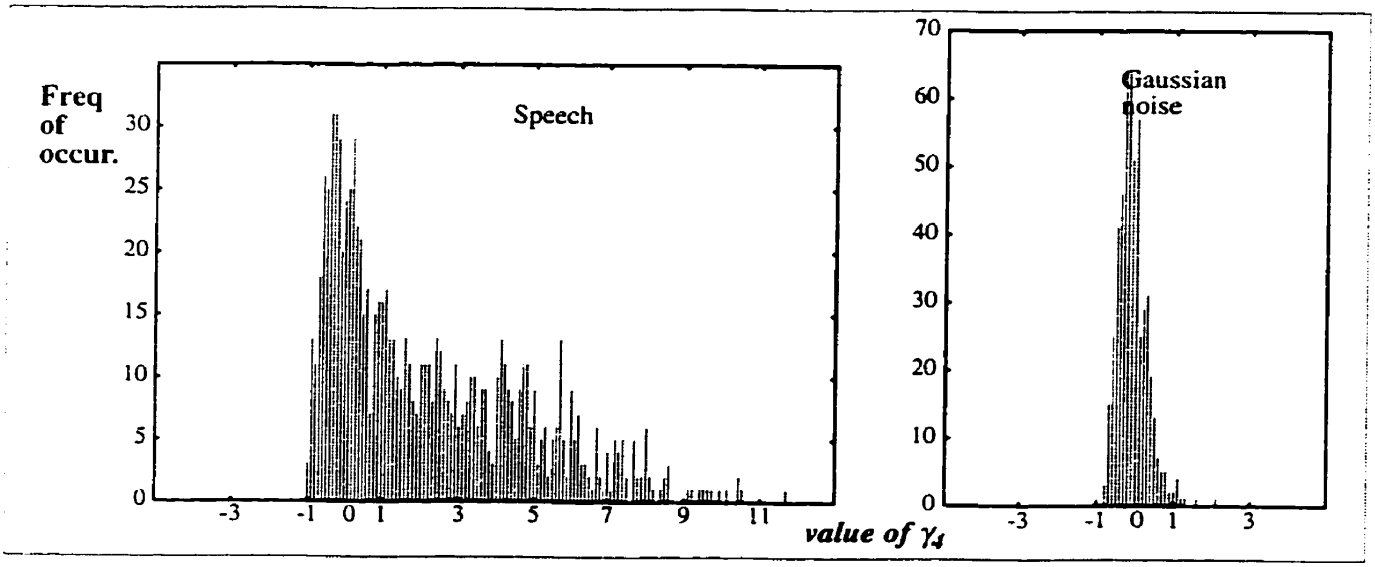
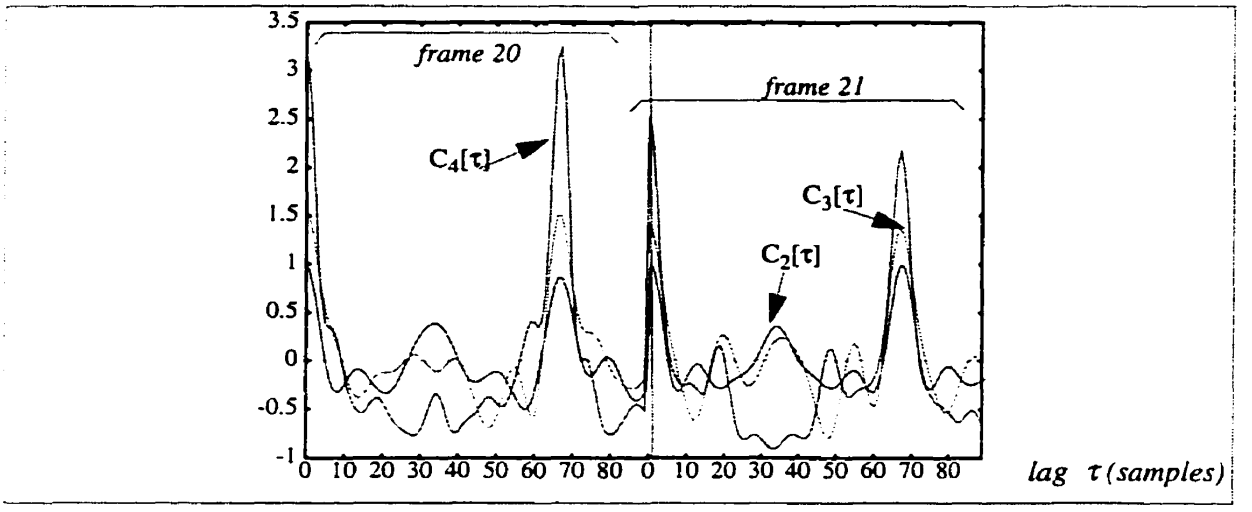


Figure 6-6 Histograms of normalized kurtosis of residual (speech vs. Gaussian noise)



The 2^{nd}, 3^{rd} and 4^{th} order cumulant slices are evaluated for a range of possible pitch lags. The slices of 3^{rd} and 4^{th} order cumulants were normalized by the signal variance as was mentioned earlier. Figure 6-7 compares the normalized $C_2[\tau]$ (the autocorrelation) with the normalized $C_3[\tau]$ and $C_4[\tau]$ slices for the case of two consecutive voiced frames (20 and 21 in Figure 6-4). As may be seen from Figure 6-7, all three functions have a maximum at the pitch lag in addition to having zero phase (max at lag 0).

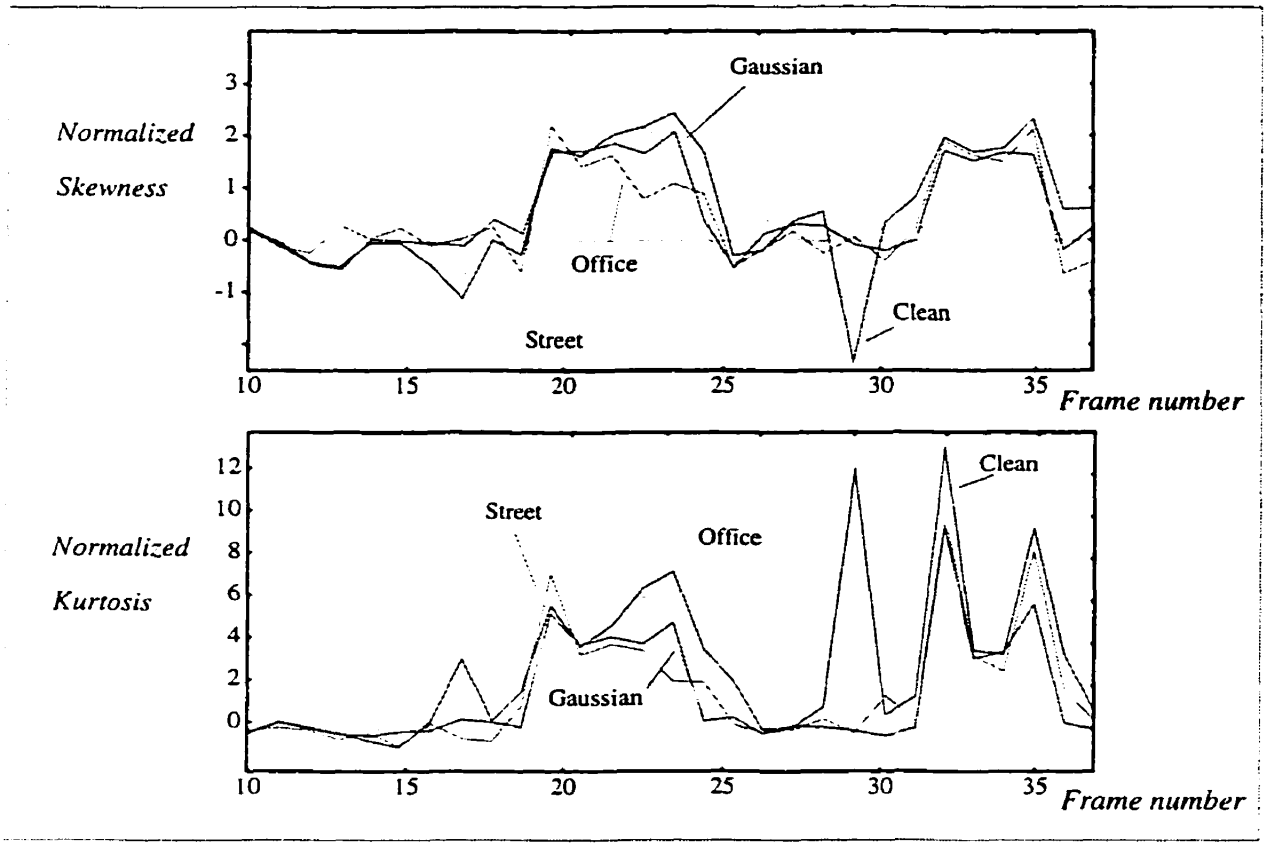
Figure 6-7 Normalized $C_2[\tau]$, $C_3[\tau]$, and $C_4[\tau]$ for frames 20 and 21



6.7.2 Effect of Noise on γ_3 and γ_4

The skewness and kurtosis are computed at an SNR of 10 dB under street, Gaussian and office noise conditions. The results are shown in Figure 6-8 and compared to the case of clean speech. At this SNR level, one would not expect significant degradation of the normalized metrics (Eq 6.24 and Eq 6.25). Indeed, for the case of Gaussian noise, these appear quite robust. Some degradation is observed for the other noise types, but the overall behavior is not significantly altered under any of the types shown.

Figure 6-8 Normalized skewness and kurtosis at 10 dB SNR Levels



6.7.3 Unvoiced Speech

To better analyze the HOS of unvoiced speech, two sustained fricatives, namely /f/ and /h/ were recorded for a few seconds and their LPC residual used for computing the skewness and kurtosis. The waveforms for the two phonemes are shown in Figure 6-9. The normalized skewness and kurtosis for /f/ is shown in Figure 6-10. To better interpret the results, histograms for the two entities are computed and shown in Figure 6-11 for /f/ and in Figure 6-12 for /h/. When the HOS quantities of unvoiced speech are compared to those of Gaussian noise, shown in Figure 6-13 and Figure 6-14, it seems that the LPC residual of unvoiced speech is likely Gaussian since its HOS are zero. However since in reality unvoiced speech occurs in small segments and often at transitional boundaries, it is expected that its HOS are non-zero. This phenomena is confirmed by simulation (for example frames 18 and 19 in Fig-

ure 6-4) and is in agreement with the experimental findings in [Fal93] about the non-zero normalized HOS of transitional speech segments.

Figure 6-9 The unvoiced phonemes /f/ and /h/

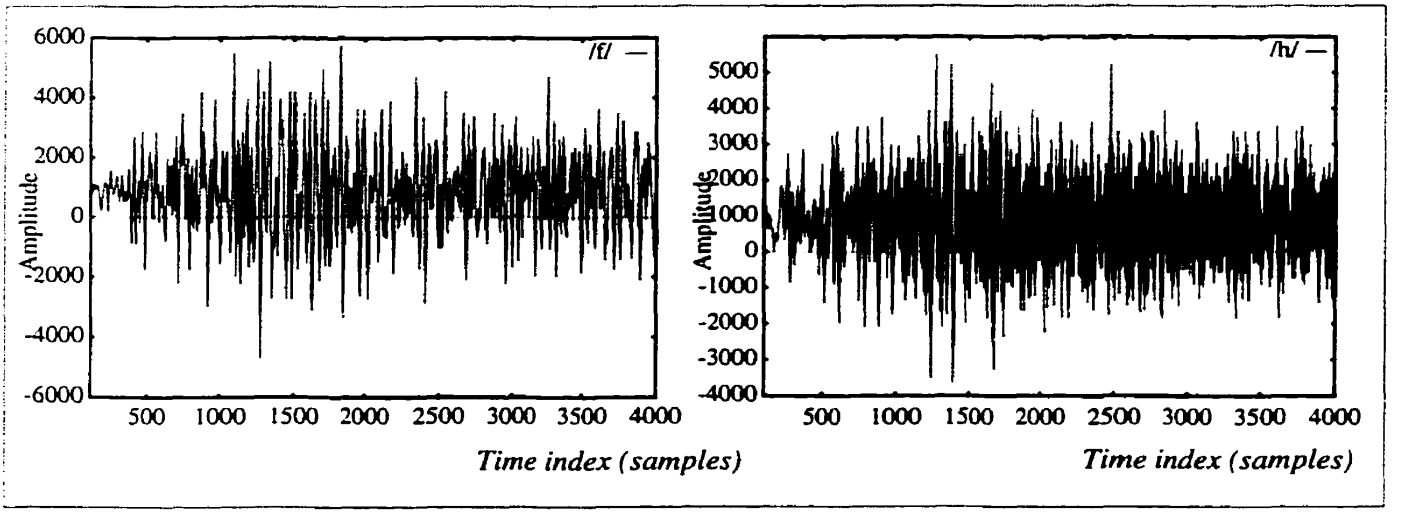


Figure 6-10 Normalized skewness and kurtosis of the LPC residual of /f/

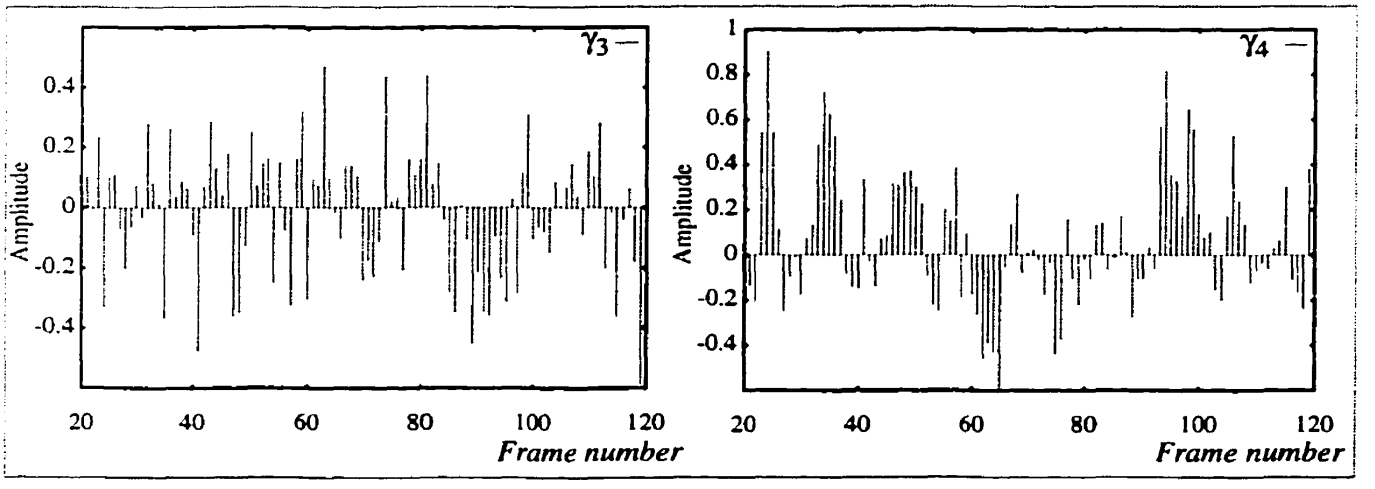


Figure 6-11 Histograms of the normalized skewness and kurtosis of /f/

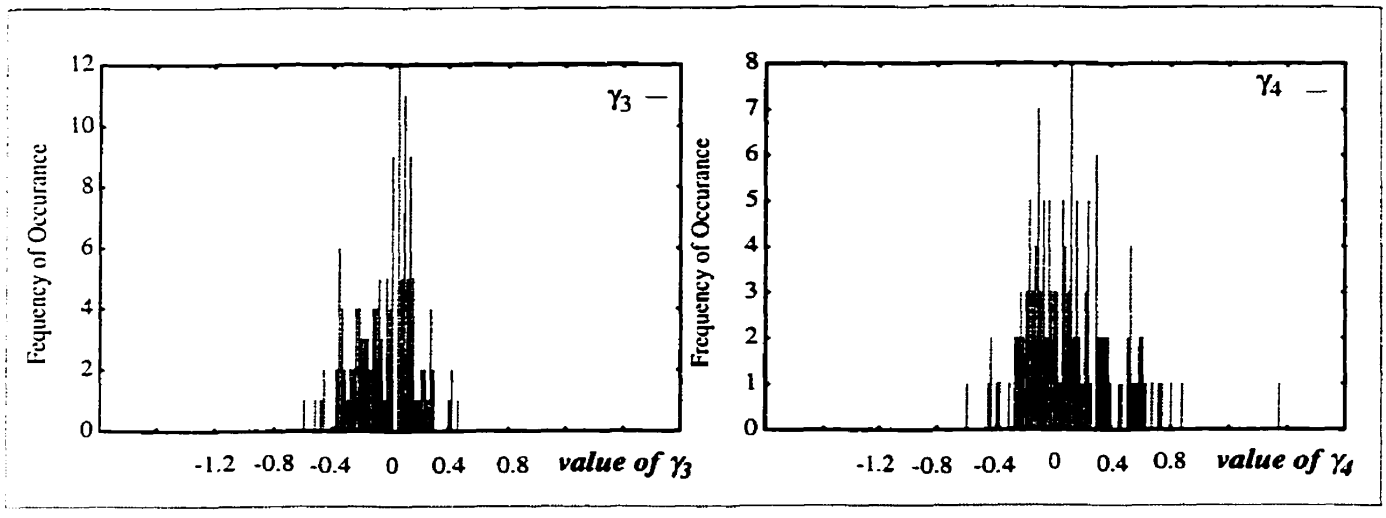


Figure 6-12 Histograms of the normalized skewness and kurtosis of /h/

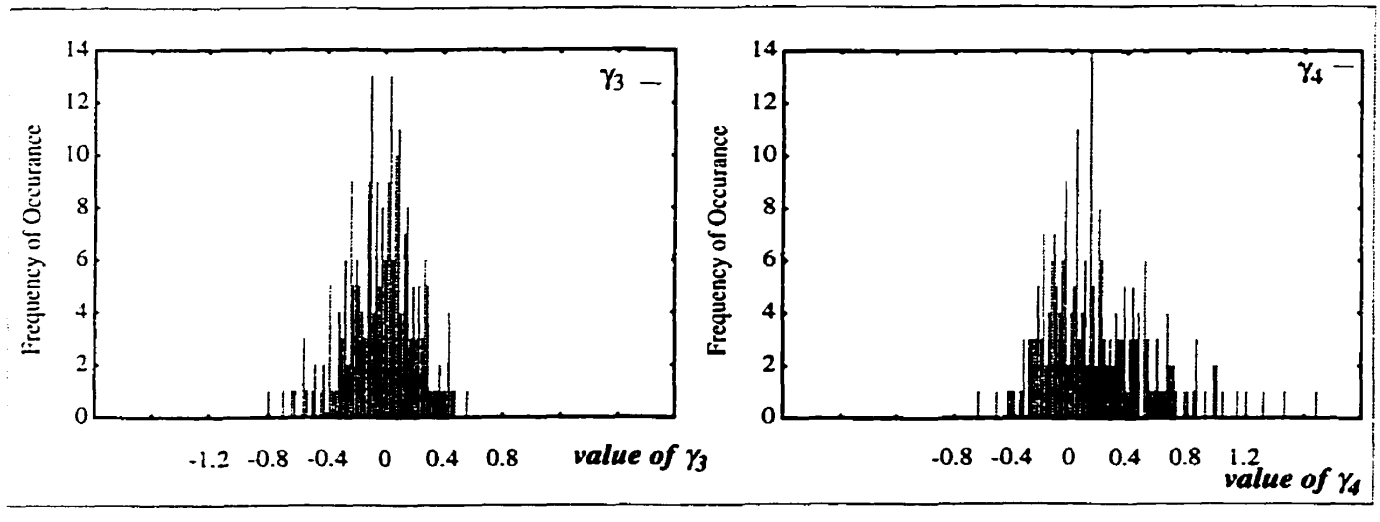


Figure 6-13 Normalized skewness and kurtosis of the LPC residual of Gaussian noise

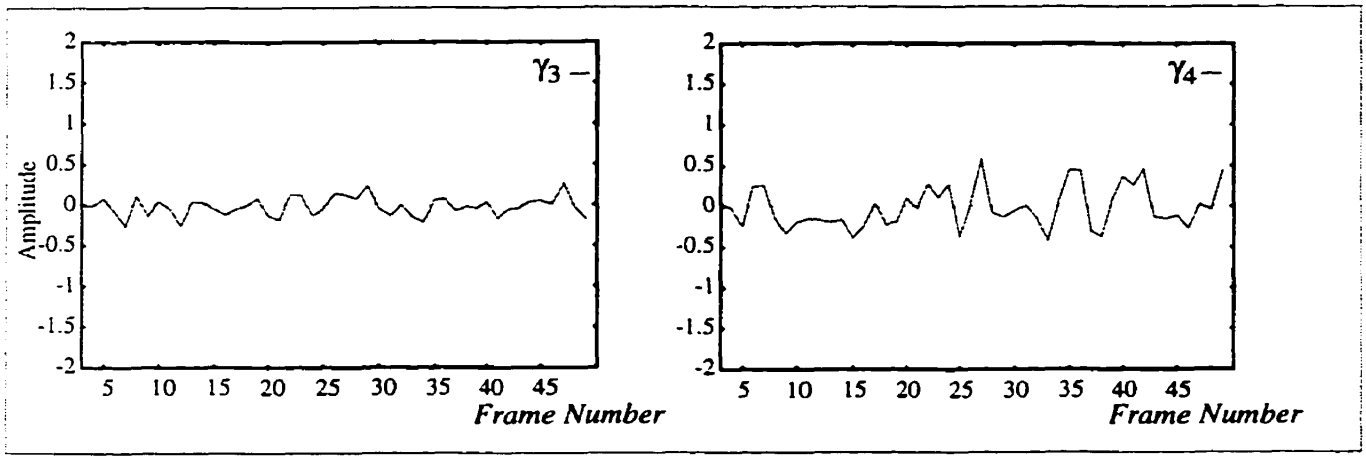
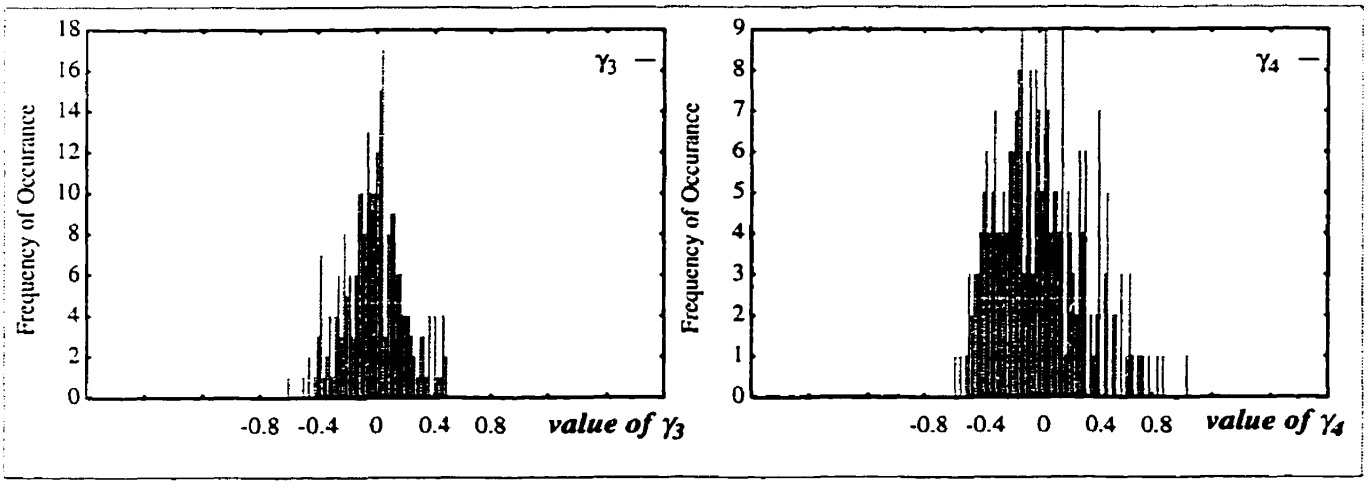


Figure 6-14 Histograms of the normalized skewness and kurtosis of Gaussian noise



6.8 Conclusion

The LPC residual of steady voiced speech is far from being Gaussian as shown analytically by the HOC derivations and verified by simulation. The zero-phase characteristic of the HOC slices of voiced speech allows one to use these slices in a similar manner as the autocorrelation function for pitch estimation. This zero-phase feature however degrades as speech becomes non-stationary, with the 3^{rd}-order cumulant being more sensitive to this condition than the 4^{th}-order one, and quickly going to a near-zero value for non-stationary segments.

The horizontal slice of the 4^{th}-order cumulant holds its zero-or-180^{o} phase characteristics for non-stationary speech, since it consists of three terms, each having inherently zero phase as evident from its Fourier transform. For this reason, this slice keeps a significant magnitude during non-stationary speech, but its periodicity no longer reflects that of the original speech due to different spacing of the harmonics in the auto-convolution of the spectrum. Thus, its use for pitch estimation becomes unreliable. Moreover, the fact that the phase can shift to 180^{o} implies a negative value of the kurtosis during non-stationarity, though simulation showed that this rarely happens.

The derivations are based on the sinusoidal modeling of the original speech and the LPC residual, and the fact that the LPC residual of steady voiced speech consists of harmonically related sinusoids of nearly-equal amplitudes. This assumption is contingent on the number of sinusoids in the original speech signal being higher than the order of the LPC analysis, otherwise the LPC residual takes near-zero values. This phenomena is sometimes encountered for certain sustained vowels (which can be modeled by a low order sinusoidal model), thus causing erroneous values of the HOS.

The skewness and kurtosis of steady voiced speech are clearly greater than zero and may be used as discriminators for voicing. The normalized entities are independent of signal amplitudes, but their effectiveness degrades in noisy conditions and in non-stationary speech conditions.

The LPC residual of unvoiced speech (such as fricatives) has zero HOS and as such it was concluded that unvoiced speech is Gaussian-like rather than a harmonic-like as suggested by McAulay's sinusoidal model.

CHAPTER 7 Application of Higher Order Cumulants to Voice Activity Detection

Synopsis

The HOC properties of LPC-filtered speech derived in the previous chapter are exploited here in the goal of finding a robust algorithm for voice activity detection in the presence of noise. A necessary condition for voicing is derived based on the relation between the skewness and kurtosis of voiced speech. The variance of the HOS estimators is used to yield a likelihood measure for noise frames. The performance of the algorithm is compared to the ITU-T G.729B VAD [Ben97] in various noise conditions. The probability of correct and false classifications is computed for both algorithms relative to hand-labeled speech for various SNR and noise types.

7.1 Motivation and Related Work

Voice activity detection (VAD) is an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation. In the GSM-based wireless system, for instance, a VAD module [Fre89] is used for discontinuous transmission to save battery power. Similarly, a VAD device is used in any variable bit rate codec [PN-3292] to control the average bit rate and the overall coding quality of the speech. In wireless systems based on Code Division Multiple Access, this scheme is important for enhancing the system capacity by minimizing the interference.

In the early VAD algorithms, short-term energy, zero-crossing rate and the LPC coefficients were among the common features used for speech detection [Rab77]. Cepstral features [Hai93], formant

shape [Hoy94], and a least-square periodicity measure [Tuc92] are some of the more recent metrics used in VAD designs. In the proposed G.729B VAD [Ben97], a set of metrics including line spectral frequencies (LSF), low-band energy, zero-crossing rate and full-band energy is used along with heuristically determined regions and boundaries to make a VAD decision for each 10 msec frame.

While previous work in the area of speech analysis, such as voicing detection or pitch estimation, attempted to exploit some of the observed features of the HOS of speech signals, little was done to provide a formal framework for using these cumulants: In [Wel85], a voiced/unvoiced detector using the bispectrum was developed and based on the observation that unvoiced phonemes are produced by a Gaussian-like excitation and thus result in a small bispectrum whereas the same is not true for voiced phonemes. In [Ran95] a method based on Gaussianity tests for the bispectrum and the triple correlation was used to discriminate voiced and unvoiced segments. The method exploits the Gaussian blindness of HOS but not the peculiarities of the HOS of voiced speech to better classify the segments. In [Fal93], the normalized skewness and kurtosis of short-term speech segments was used to detect transitional speech events (termed innovation), based on the observation that these two statistics take on non-zero values at the boundaries of speech segments, but no analytical ground was given to support the results. In [Mor92] a pitch estimation method based on the periodicity of the diagonal slice of the 3^{rd}-order cumulant was described and yielded more reliable pitch estimates than the autocorrelation, but the claim of the 3^{rd}-order cumulant slice having similar periodicity as the underlying speech was not clearly demonstrated.

The key idea in using higher-order statistics for a robust voicing detection hinges on being able to separate signal and noise based on these statistics. The findings in the previous chapter clearly demonstrate that these statistics are indeed different from those of Gaussian noise when the LPC residual of the speech signal is considered.

7.2 Voice Activity Detection using HOS

7.2.1 Rationale

It was shown in Chapter 6 that the skewness and kurtosis of the LPC residual of voiced speech can be expressed in terms of the number of harmonics M and signal energy and are greater than zero for any practical value of M (which is a function of pitch). The normalized statistics may be expressed in terms of M only (Eq 6.5, Eq 6.20). This is clearly distinct from the case of Gaussian noise, where both of these entities are zero. It seems sensible then to make use of these two statistics as one way of detecting voicing. The advantage of using the normalized ones is that they are independent of the signal energy and therefore absolute thresholds may be used. However, when using normalized statistics, one has to account for the effect of noise (Eq 6.24, Eq 6.25). Alternatively, one may consider the variance of the estimators of the skewness and kurtosis and normalize the computed entities to yield unit-variance estimators. Another point worth exploiting is the relation between the skewness and kurtosis for voiced speech and the use of this relation as a necessary condition for classifying a frame as voiced. These two ideas are further detailed below and are the basis of the proposed VAD algorithm.

7.2.2 Soft Detection of Noise Frames

The skewness and kurtosis of Gaussian noise are zero only in a statistical average sense. Since in practice finite length frames are used, the decision that a given frame is noise can only be made in a probabilistic sense with a confidence level that takes into account the variance and distribution of the estimators of the skewness and kurtosis. Given a Gaussian process $g(n)$, the estimators of the 2nd and 4th moments are:

$$M_{kg} = \frac{1}{N}\sum_{n=0}^{N-1} [g(n)]^k \text{ estimator for } E[\{x(n)\}^k], \qquad \text{(E 7.1)}$$

for $k = 2$ and 4. In Appendix A, it is shown that these estimators are unbiased, and for the case of white Gaussian noise, their mean and variance may be expressed in terms of the process variance, υ_g:

$$E[M_{3g}] = 0 \text{ and } E[M_{4g}] = 3\upsilon_g^2$$

$$Var[M_{3g}] = \frac{15\upsilon_g^3}{N} \text{ and } Var[M_{4g}] = \frac{96\upsilon_g^4}{N}. \qquad \text{(E 7.2)}$$

As a result, the estimator of the skewness $\hat{SK} = M_{3_g}$ is unbiased, with zero mean and known variance. Since this estimator is the sum of a large number of independent identically distributed (iid) random variables, then by the Central limit theorem [Leo89], the following random variable:

$$\hat{SK}_a = \frac{M_{3_g}}{\sqrt{15\upsilon_g^{\,3}/N}} \qquad \text{(E 7.3)}$$

is Gaussian with zero mean and unit variance. Therefore, given the estimate of the skewness of a given frame and the corresponding scaled value denoted by 'a', one can find the probability that the frame is Gaussian noise as:

$$Prob\,[Noise] = Prob\,[|\hat{SK}_a| \geq a]\ . \qquad \text{(E 7.4)}$$

Graphically, this is equivalent to computing the area under the tail of the Gaussian curve of $\hat{SK}_a$. Clearly when $a = 0$ the area is unity. The area under the tail of the curve can be evaluated using the *erfc(x)* function. For example, when $a > 0$, $Prob\,[Noise] = \frac{2}{\sqrt{\pi}}\int_a^{\infty} e^{-x^2/2}dx$.

Thus, $Prob\,[Noise] = erfc\,(|a|)$.

Similarly, the estimator for the kurtosis is first computed from the 2nd and 4th moments. To ensure an unbiased estimate, the modified estimator proposed in Eq A.4 is used:

$$\hat{KU}_U = \left(1 + \frac{2}{N}\right)M_{4_g} - 3\,(M_{2_g})^2 . \qquad \text{(E 7.5)}$$

This estimator is unbiased, with zero mean and known variance given in Eq A.27. The distribution of this estimator is not straightforward, since it consists of the difference of two variables, one Gaussian and one Chi-square. However, an approximation is used here and the estimator is assumed normally distributed[1]. A unit-variance version of this zero-mean variable is defined as:

$$\hat{KU}_{Ua} = \frac{\hat{KU}_U}{\sqrt{\frac{3\upsilon_g^{\,4}}{N}\left(104 + \frac{452}{N} + \frac{596}{N^2}\right)}} \qquad \text{(E 7.6)}$$

Therefore, given the value of the estimate of the kurtosis of a given frame and the corresponding scaled value, denoted by 'b', the probability that the frame is noise as: $Prob\,[Noise] = erfc\,(|b|)$.

1.This assumption is verified to be reasonable by simulation (e.g., Figure 6-6).

The discussion so far pointed out that given the estimate of the skewness and kurtosis, one can determine the probability of the frame being Gaussian noise using the normalized values of these estimates and the *erfc* functions. Moreover, it is assumed that the true variance of the noise (υ_g) is known *apriori*. In reality, this is not the case but one has only a (hopefully good) estimate of the noise energy, which is estimated during frames declared non-speech. This estimate is not equal to the true variance, but is relatively good compared to the estimates of the skewness and kurtosis, which are only deduced from a short data frame.

7.2.3 Necessary Condition for Voicing

The skewness and kurtosis of voiced speech are expressed in terms of energy and number of harmonics and may be used for detecting voiced frames. In order to eliminate the effect of energy, one may consider the normalized metrics (γ_3 and γ_4), but in the presence of noise, these metrics become less effective for detecting voiced frames (Eq 6.24, Eq 6.25). Alternatively, the ratio of the appropriate power of the skewness to that of the kurtosis may be considered as one way of eliminating the effect of signal energy in Eq 6.4 and Eq 6.19, while avoiding the effect of noise. Consider the ratio:

$$SKR = \frac{skewness^2}{kurtosis^{1.5}} = \frac{9(M-1)^2}{8M\left[\frac{4}{3}M - 4 + \frac{7}{6M}\right]^{1.5}} \qquad \text{(E 7.7)}$$

which is independent of signal energy and is only a function of M. Examining the SKR reveals that for the practical range of M ($M \geq 4$), the ratio is confined to the range $[0 \ldots 1]$, and thus this is a necessary condition for classifying a frame as a voiced frame. When only Gaussian noise is present, the ratio is undetermined since both operands are zero. In reality, this zero condition is never the case due to the variance of the estimators. However, the ratio may take on any values, including the range for voiced speech (i.e., $[0 \ldots 1]$); for this reason, this is a necessary but not sufficient condition for detecting voiced frames.

7.2.4 HOS-based VAD Algorithm

Since sustained unvoiced speech has been shown to have Gaussian-like characteristics, it cannot be distinguished from Gaussian noise using HOS. However, as discussed in Chapter 6, this is seldom the case in reality where unvoiced speech occurs at speech transitional boundaries which have non-zero HOS. Therefore the VAD detection proposed here may be based on HOS and can be formulated as a finite 2-state machine. The algorithm proposed combines the use of the skewness, kurtosis, their nor-

malized versions γ_3 and γ_4, the SNR, and the SKR, for detecting speech frames and making a voicing decision.

Data format

Speech sampled at 8 kHz is used. A 10-order LPC analysis is performed once every 20 msec, thus generating a 20 msec residual. Voice activity detection is done every 10 msec using the residual and a 20% overlap (i.e., 80 new points are combined with 20 from the past iteration).

HOS Computations

Every 10 msec iteration, the estimators for the 2^{nd}, 3^{rd} and 4^{th} moments are computed using the N = 100 points and Eq A.1. An autoregressive scheme is used to smooth the estimates of the moments. From these, the unbiased estimate of the kurtosis (Eq A.4) is deduced. The estimate of the skewness is simply the 3^{rd} moment (Eq A.2). The two metrics are then normalized by the signal energy to yield:

$$\gamma_3 = \frac{\hat{SK}}{M_{2x}^{1.5}} \text{ and } \gamma_4 = \frac{K\hat{U}_U}{M_{2x}^2}. \qquad \text{(E 7.8)}$$

Noise and SNR Estimation

The noise power is estimated using frames declared non-speech (i.e., when in the noise state). Moreover, it is assumed that the first three frames are non-speech and are used to initialize the noise power estimate. Whenever a frame is declared non-speech, its energy is used to update the noise energy. An averaging scheme is used to smooth the estimate with an integration constant that is a function of the noise likelihood of that frame.

$$\tilde{\upsilon}_g(k) = (1-\beta)\,\tilde{\upsilon}_g(k-1) + \beta M_{2X} \qquad \text{(E 7.9)}$$

where k is the iteration index, M_{2X} is the frame energy, $\tilde{\upsilon}_g$ is the estimate of the noise energy, and $\beta = 0.1 \cdot Prob\,[Noise]$. At every iteration, the current estimate of the noise energy is used to compute the SNR of that frame:

$$SNR = Pos\left[\frac{M_{2X}}{\tilde{\upsilon}_g} - 1\right] \qquad \text{(E 7.10)}$$

where $Pos\,[x] = x$ for $x > 0$ and 0 otherwise. Since the residual is low-pass filtered at 2 kHz, the above SNR is for the lower spectrum only. Using a similar reasoning a 'total SNR' metric is computed using the non-filtered residual and the energy of the full band.

Probability of noise-only frame

Once the skewness and kurtosis are computed, the variances of these estimates are computed using the noise energy $\bar{\upsilon}_g$, according to Eq 7.3 and Eq 7.6, to yield the zero-mean, unit variance estimates $\hat{SK}_a$ and $\hat{KU}_{Ua}$ respectively. From these two scaled values, the probability of the frame being noise is:

$$Prob\,[Noise] \;=\; [\,erfc\,(a) + erfc\,(b)\,]\,/2 \qquad \text{(E 7.11)}$$

where a and b are the computed values of $\hat{SK}_a$ and $\hat{KU}_{Ua}$ respectively.

SKR ratio

The ratio is computed directly from the non-normalized estimates of the skewness and kurtosis:

$$SKR \;=\; \frac{[\hat{SK}]^{2}}{[\hat{KU}_{U}]^{1.5}}. \qquad \text{(E 7.12)}$$

Speech/Noise State machine

The VAD algorithm is implemented as a 2-state machine (Figure 7-1). The following operations are carried out in each state:

- **Speech State**
  - —The SKR ratio (Eq 7.12) is used to determine if the current speech frame is voiced.
  - —The noise likelihood (Eq 7.11) along with the values of γ_3 and γ_4 (Eq 7.8) are used to determine whether the frame is Gaussian. After a hangover period (2 to 3 frames), transition to the *Noise* state occurs:

$$\text{If } \{\, Prob\,[Noise] > T_{Gaus} \text{ and } \gamma_3 < T_{\gamma_3} \text{ and } \gamma_4 < T_{\gamma_4} \,\}$$

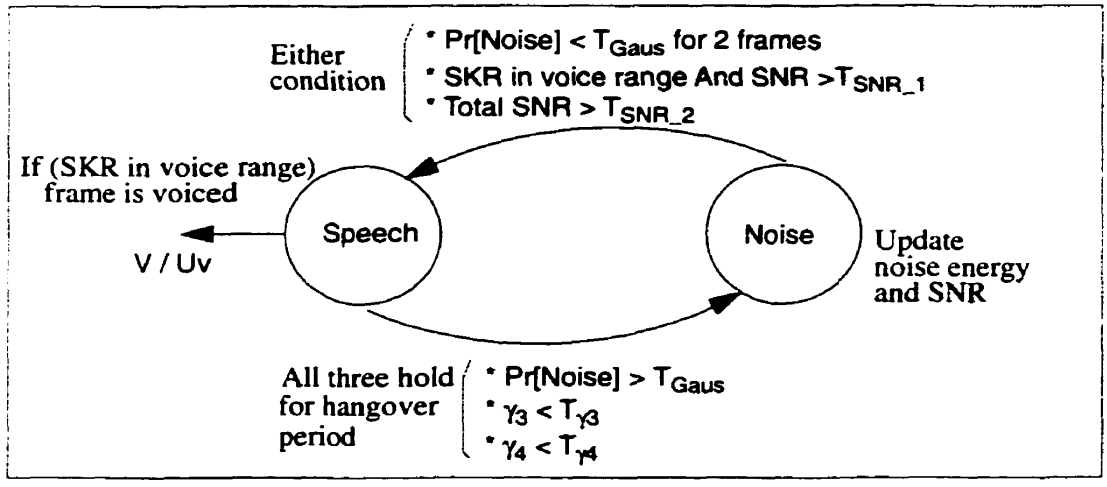
- **Noise State**
  - —The noise energy is updated according to the Prob[Noise] (Eq 7.9).
  - —The SKR ratio, the Gaussian likelihood and the SNR (Eq 7.10) values are used to determine whether the frame is speech. The occurrence of any of the following three conditions triggers a transition:

$Prob\,[Noise] < T_{Gaus}$ for two consecutive frames

SKR in voicing range and $SNR > T_{SNR1}$ (an indication of voiced frame)

$TotalSNR > T_{SNR2}$ (strong speech frame)

Figure 7-1 HOS-based VAD state machine



7.3 Experimental Results

To evaluate the effectiveness of the HOS-based VAD, the probability of correct and false detection was calculated for a number of noisy speech scenarios. To obtain these two metrics, a reference decision was made for clean speech material of 25 seconds containing utterances spoken by male and female speakers and the following Harvard sentences:

—Note closely the size of the gas tank. Wipe the grease off his dirty face.

—Pluck the bright rose with leaves. Two plus seven is less than ten.

—He picked up the dice for a second role. These coins will be needed to pay his debts.

—Both brothers wore the same size. In some form or another, we need fun.

Each segment of 10 msec was manually labeled and the following metrics are used:

- Pc_{Speech}: Probability of correctly detecting speech frames. Computed as the ratio of correct speech detections to the total number of hand-labeled speech frames.
- Pc_{Noise}: Probability of correctly detecting noise frames. Computed as the ratio of correct noise detections to the total number of hand-labeled noise frames.
- *Pf*: Probability of false detection. Computed as the ratio of incorrectly classified speech or noise frames to the total number of frames.

Noisy speech is produced by digitally mixing noise files with the clean speech file at various SNR levels. For each noise type and SNR, the *Pc*'s and *Pf*'s of the proposed VAD are compared to those computed for the G.729B VAD [Ben97]. The results from all these scenarios are summarized in Table 7-1. The noise types used here are all recorded noises, with the exception of the Gaussian noise which is synthetically generated. Three SNR levels are used, namely 20 dB, 10 dB and 5 dB. The SNR value is computed based on the ratio of the total energy of speech to that of the noise over the entire utterance. In addition to computing the above metrics, marker files are generated from running each algorithm on the given speech in order to visually inspect the performance.

Table 7-1 Pc's and Pf's for the HOS-based and G.729B VAD

Noise Environment		HOS-based VAD			G729B VAD		
Type	SNR	Pc Noise (%)	Pc Speech(%)	Pf (%)	Pc Noise (%)	Pc Speech(%)	Pf (%)
Gaussian	20 dB	86.43	93.82	9.23	86.81	91.31	10.54
	10 dB	93.64	72.71	16.63	90.47	70.73	21.09
	5 dB	95.86	58.18	26.23	90.37	59.8	27.54
Street	20 dB	64.29	99.32	15.16	85.37	91.64	11.0
	10 dB	71.89	93.62	15.36	79.69	77.39	21.65
	5 dB	81.14	78	19.22	80.07	68.02	27.00
Office	20 dB	20.79	99.79	32.88	56.4	96.74	19.94
	10 dB	27.14	97.55	31.57	66.50	82.96	23.85
	5 dB	26.37	93.35	34.36	59.58	76.92	30.25
Fan	20 dB	79.01	98.1	9.79	85.65	94.43	9.19
	10 dB	88.35	83.02	14.77	90.18	72.50	20.18
	5 dB	92.68	62.66	24.92	90.37	61.57	26.51

At high SNR, both algorithms perform roughly the same in Gaussian and fan noise conditions, with a probability of false detection around 10%. While this figure may seem high, it is partly due to the subjective factor in hand-labeling the ambiguous speech segments, such as the small amplitude ones. However, incorrect decisions about these are not very drastic in the context of variable rate coding, for example. Figure 7-2 illustrates the decision for both VAD's in mild Gaussian noise (20 dB). The clean waveform is shown along the HOS and G.729B VAD markers. In office noise where dominant conversations occur, the HOS-based detector falsely classifies noise segments as speech. This is due to the

fact that the noise in this case is speech-like and has a non-zero HOS. Clearly, the G.729B VAD gives better performance in terms of false classification at all SNR levels in this case. Figure 7-4 illustrates the case for this noise at 10 dB SNR and it is clear the HOS-based VAD is biased towards speech decisions. For the case of street noise at 20 dB SNR, a somehow similar though less pronounced behaviour is observed where the HOS performance is biased towards speech. While it is arguable that street noise is close to being Gaussian, being the sum of many independent sources, it also contains periods of bursty non-stationary noise caused by the sound of pedestrians or accelerating vehicles. The reason for the HOS-based decisions being biased towards speech in non-Gaussian and low noise conditions may be explained by considering the state machine transitions (Figure 7-1): A non-Gaussian noise will result in a low noise likelihood measure which will delay or inhibit transition from the speech to the noise states, resulting in more speech classifications. The consequences of this behaviour however are not as severe as if the problem occurred on the other transition, since then it will result in falsely classifying speech as noise, which has more detrimental consequences.

In low SNR conditions, the HOS-based VAD performs overall better in Gaussian, street and fan noise. The case of street noise at 10 dB SNR is shown in Figure 7-3. The difference in classification is particularly noted in the last non-speech segment where the G.729B has a rather erratic behaviour and results in wrong oscillations between the states. The case of fan noise is illustrated in Figure 7-5. In this scenario, the HOS-based system performs overall better at 10 dB and 5 dB SNRs (Table 7-1). From the figure, it also results in smoother transitions between the two states.

A final point worth noting is that the degradation in performance between high and low SNR's for street and fan noises is not as severe in the HOS-based VAD as it is in the G.729B one. This can be seen in Table 7-1 by comparing performance between the 20 dB and 10 dB cases. This suggests that while the metrics used in the ITU-T VAD are effective in isolating speech, they are not robust to noise and quickly degrade in low SNR conditions.

Figure 7-2 HOS-based and G.729B VAD in Gaussian noise conditions (20 dB)

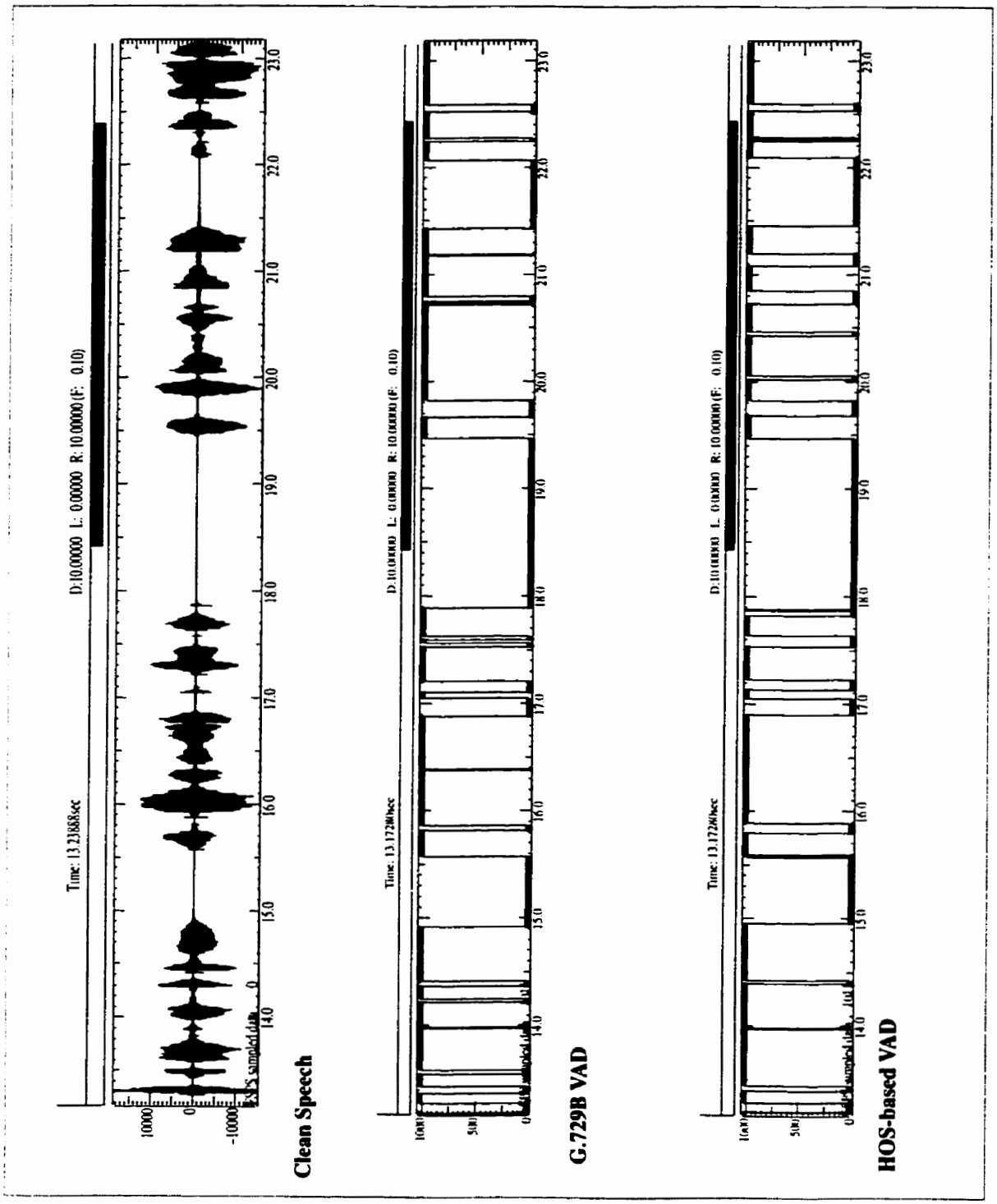


Figure 7-3 HOS-based and G.729B VAD in street noise conditions (10 dB)

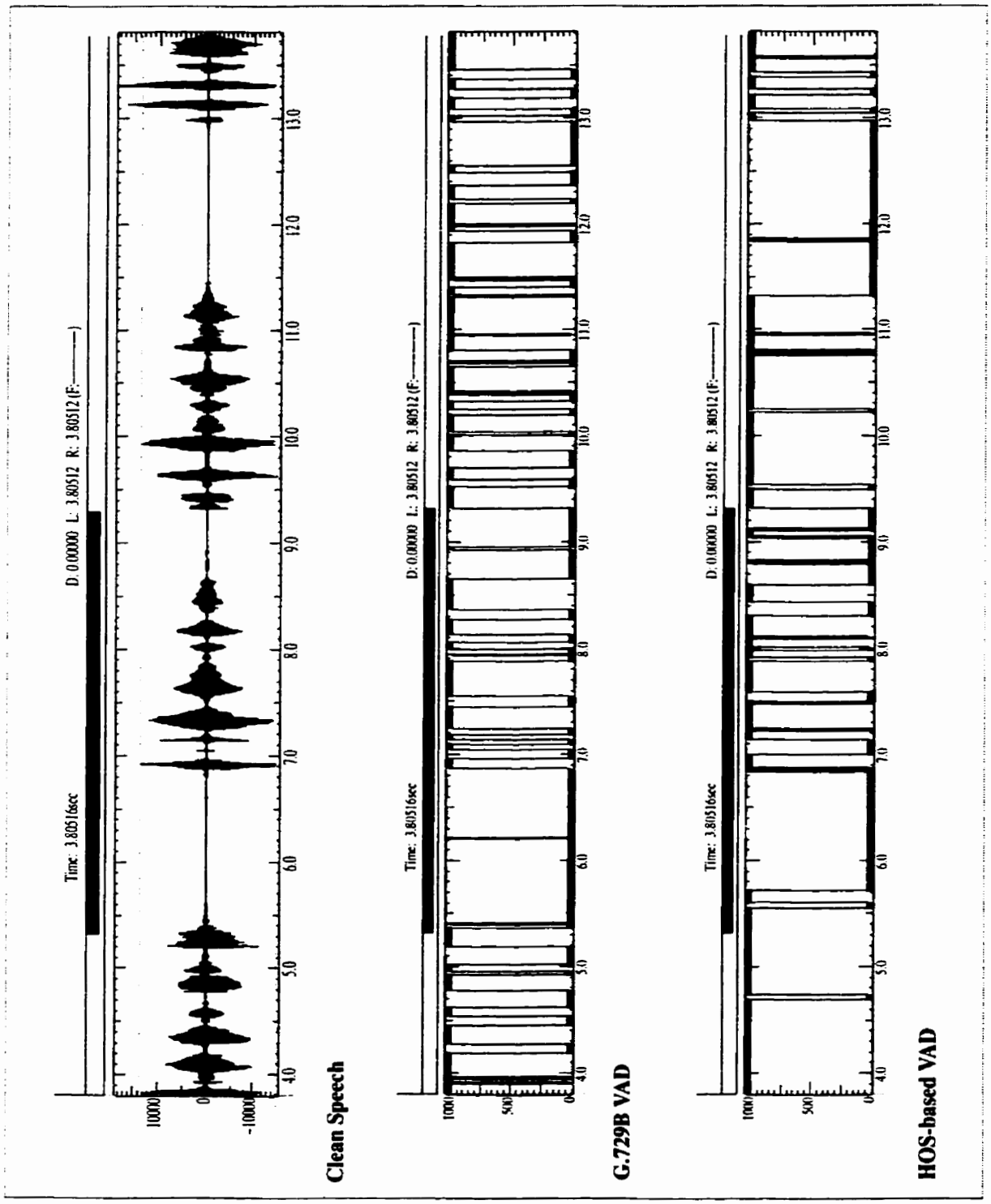


Figure 7-4 HOS-based and G.729B VAD in office noise conditions (10 dB)

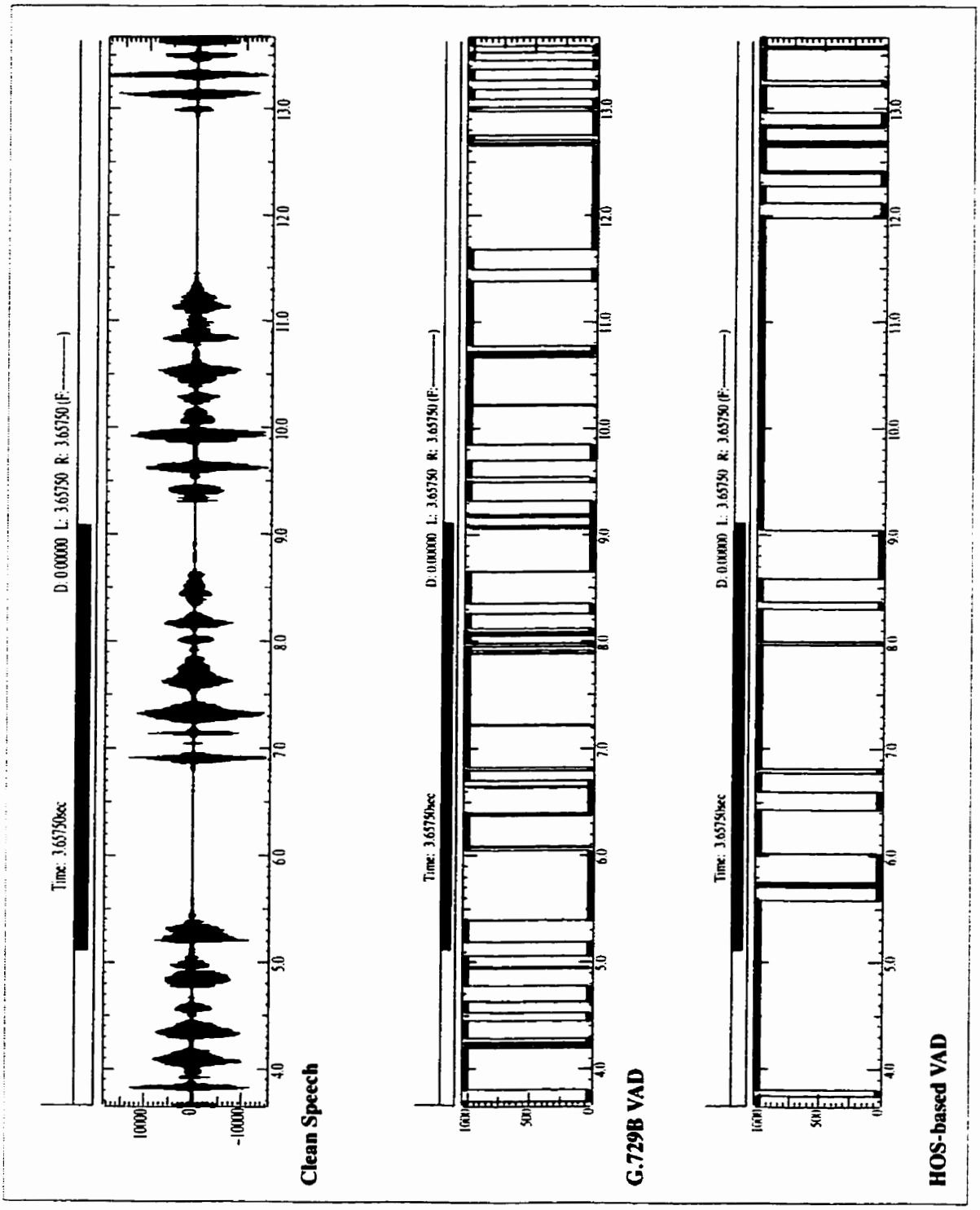
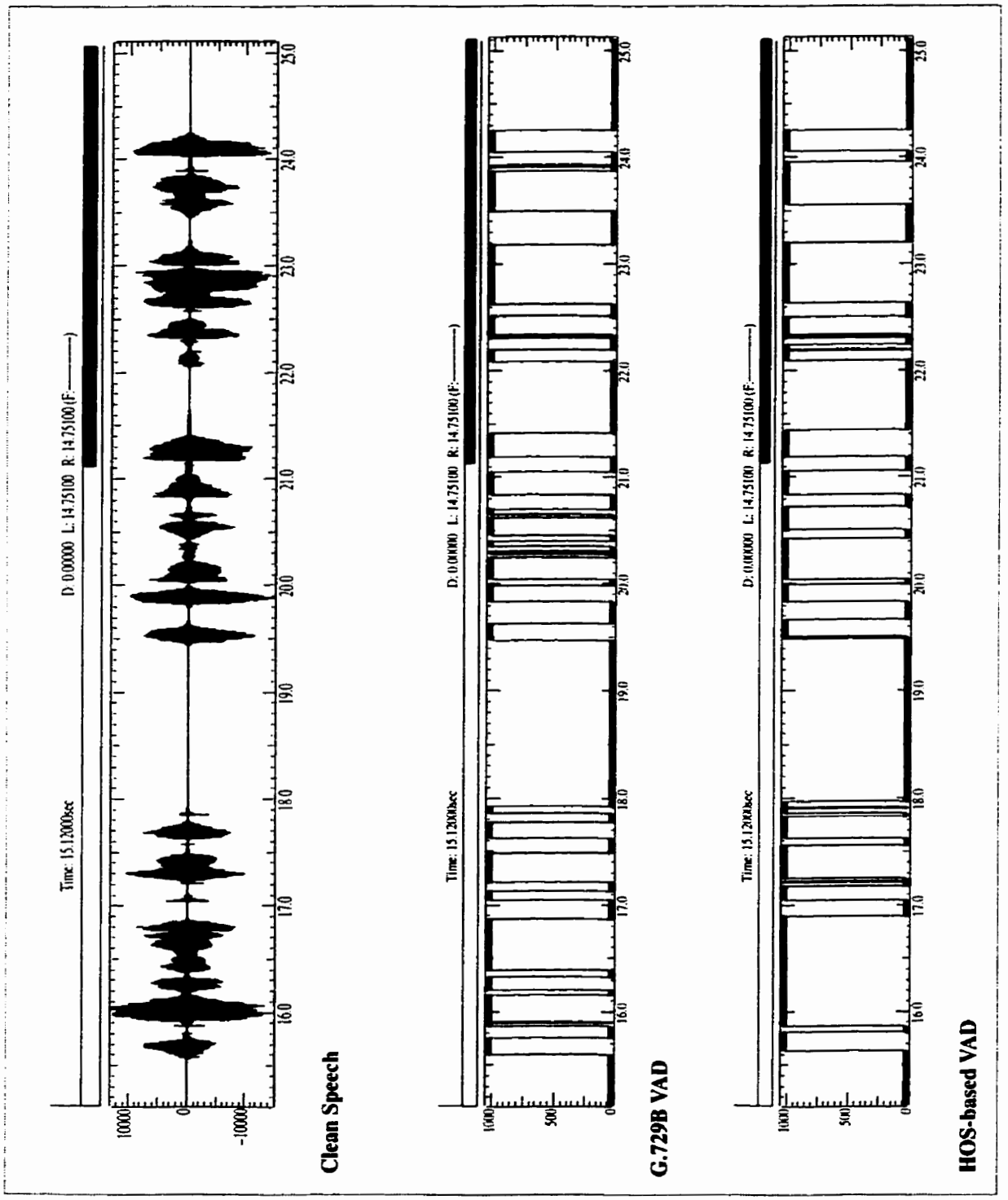


Figure 7-5 HOS-based and G.729B VAD in fan noise conditions (10 dB)



7.4 Conclusion

A new algorithm for voice activity detection was described in this chapter. This proposed VAD scheme makes use of the HOC properties of speech derived in Chapter 6, particularly the fact that the skewness and kurtosis of voiced speech are non-zero and may be expressed in terms of signal energy and number of harmonics. The relation between the two metrics is used as a necessary condition for voicing. Moreover, the variance of the HOS estimators is used to quantify the noise likelihood of a given frame. The algorithm combines these concepts with low-band and full-band SNR measures to classify frames into one of the two states and determine whether speech frames are voiced.

Compared to the G.729B VAD, the proposed algorithm is based on a more analytical framework, is conceptually simpler and uses a smaller parameter set, therefore making it easier to tune. Performance results show that the proposed algorithm has overall comparable performance to the G.729B: Its probability of false classification is lower in low SNR and Gaussian-like noise, but higher in speech-like noises. The fact that a simple HOS VAD based on a minimum parameter set can match the performance of the current standard suggests that HOS metrics have promising potential in yielding VAD algorithms that would surpass the current state of the art.

CHAPTER 8 Implementation Issues of Higher Order Statistics

Synopsis

Some of the crucial issues in implementing HOS-based methods include accuracy and computational complexity. The first is due to the bias and variance when estimating HOS from finite data records. The second is due to the fact that the higher-order cumulant functions inherently involve more computations than their second-order counterpart. A number of derivations quantifying the bias and variance of the HOS estimators of a sinusoidal signal in white Gaussian noise are given in Appendix A. This chapter explores the computational aspects of HOS implementations:

- The first part addresses complexity and proposes an algorithm for efficiently computing the 3^{rd}-order cumulant function with a reduced number of multiplications. The algorithm exploits the redundancy in the time samples to infer a factored expression with fewer multiplications.
- The second part deals with executing a DSP algorithm on a parallel architecture. A general scheduling and allocation model is proposed for mapping a set of operations on a configurable multi-unit architecture. The algorithm is based on branch-and-bound concepts with the use of a non-deterministic heuristic that accounts for the opportunity cost of resources and the scheduling urgency of the operations.

8.1 Efficient Computation of the Third Order Cumulant

8.1.1 Motivation

The third-order cumulant function (Eq 3.6) for a finite length sequence is computed as:

$$C_3[k, m] = \frac{1}{N}\sum_{n=0}^{N-1} s[n]\, s[n+k]\, s[n+m] \; . \qquad \text{(E 8.1)}$$

Computing the above for an N-point windowed sequence requires on the order of $K \cdot M \cdot N$ multiplications and additions (the division by N is discarded for the rest of this discussion), where N is the sequence length and K and M are the ranges for the two lags, respectively. The number of multiplications may be reduced by exploiting certain properties of $C_3[k, m]$. The following are observed:

- Most samples appear two or three times as multiplicands and can be factored accordingly.
- The product terms may be grouped into symmetry regions that repeat and can be factored.

The approach presented here extends a similar one that is used to compute the autocorrelation function with a reduced number of multiplications [Rab78]. The case of the third cumulant is however more involved given the presence of two lags.

8.1.2 The algorithm

Without loss of generality, it is assumed that $k < m$, and a new index is defined: $d = m - k$ to denote the difference between the two lags. Factoring the cumulant into an efficient form depends on the relation between the two lags, namely between the smaller lag (k) and the difference between the lags (d). Two cases are thus considered:

Case 1: k < d

The product terms required for computing $C_3[2, 8]$ are shown in Figure 8-1 (here $k = 2$, $m = 8$, $d = 6$). First, note that the terms may be divided into symmetry regions, each consisting of $m+d$ (in this case, 14) product terms. Thus the terms in each of these regions is first re-written in a factored way and the process replicated to the other regions.

In this example, it is observed that $s[8]$ and $s[9]$ occur in three terms and may be factored as:

$$s[8]\,\{s[0]\,s[2] + s[6]\,s[14] + s[10]\,s[16]\} + s[9]\,\{s[1]\,s[3] + s[7]\,s[15] + s[11]\,s[17]\}$$

The above expression thus eliminates lines 1, 2, 7, 8, 9, 10 (the lines with a check). The remaining samples, namely $s[10]$, $s[11]$, $s[12]$, $s[13]$, no longer occur in three product terms since some of these occurrences have already been factored in with $s[8]$ and $s[9]$. However they occur twice and may be factored as:

$$s[10]\{s[2]s[4]+s[12]s[18]\}+s[11]\{s[3]s[5]+s[13]s[19]\}+$$
$$s[12]\{s[4]s[6]+s[14]s[20]\}+s[13]\{s[5]s[7]+s[15]s[21]\}$$

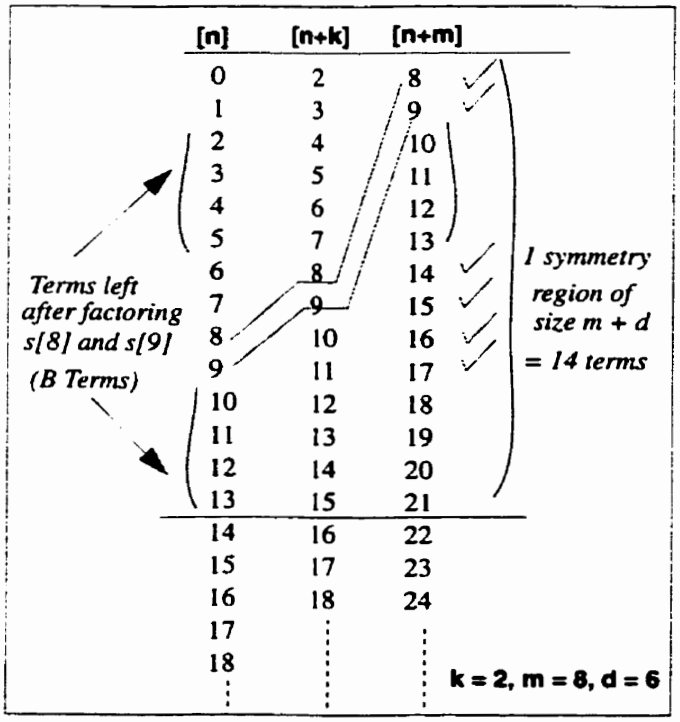
In general, for the case where $k < d$, each symmetry region may be expressed as: $A + B$, where A captures the products that are factored three times:

$$A = \sum_{i=0}^{k-1} s[i+m]\{s[i]s[i+k]+s[i+d]s[i+m+d]+s[i+k+m]s[i+2m]\}$$

and B the remaining products that are factored two times:

$$B = \sum_{i=k}^{d-1} s[i+m]\{s[i]s[i+k]+s[i+k+m]s[i+2m]\} .$$

Figure 8-1 Product terms for computing $C_3[2,8]$



In order to add all the symmetry regions, a periodic time index p is added and incremented by $(m+d)$ every period until all N samples have been used. This assumes that the total number of samples is a multiple of $(m+d)$; if this is not the case, the remaining product terms are computed in their original form. Thus, the total number of time samples used, N, is written as: $N = P(m+d) + r$, where P is an integer, and denotes the number of symmetry regions, and r is the number of residual product terms that cannot be factored. The cumulant is then written as a sum of three expressions:

$C_3[k, m] = A + B + R$ with:

$$A = \sum_{P} \sum_{k=0}^{k-1} s[i+m+p] \{ s[i+p] s[i+k+p] + s[i+d+p] s[i+m+d+p] + s[i+k+m+p] s[i+2m+p] \}$$

$$B = \sum_{P} \sum_{i=k}^{d-1} s[i+m+p] \{ s[i+p] s[i+k+p] + s[i+k+m+p] s[i+2m+p] \}$$

$$R = \sum_{i=P(m+d)}^{N-1} s[i] s[i+k] s[i+m]$$

where R includes the remaining unfactored product terms, whose number is less than the size of a symmetry region $(m+d)$. The index p starts from 0, and is incremented by $(m+d)$ each time, for a total of P times: $p = 0, \ldots, (m+d), 2(m+d), \ldots, (P-1)(m+d)$. The number of operations required for this case are given in Table 8-1 below:

Table 8-1 Number of operations for case 1

Term	Multiplies	Adds
A	4kP	P (3k-1)
B	3P(d-k)	P { 2(d-k) - 1 }
R	2{N - P(m+d)}	N - P(m+d)

Case 2: d < k

The expression developed for this case is similar to those of case 1, except for:

- The limits of the A & B summations.
- The size of the symmetry region.
- The first component of the B summation.

The product terms required for computing $C_3[6, 9]$ are shown in Figure 8-2 below (here, $k = 6$, $m = 9$, $d = 3$). As before, symmetry regions are identified, and the product terms may be divided in symmetry regions, in this case, of size $(k + m)$. Thus, the total number of time samples used, N, is written as: $N = P(m+k) + r$. Using the same reasoning as in case 1, the 3rd-order cumulant is written as a sum of three expressions:

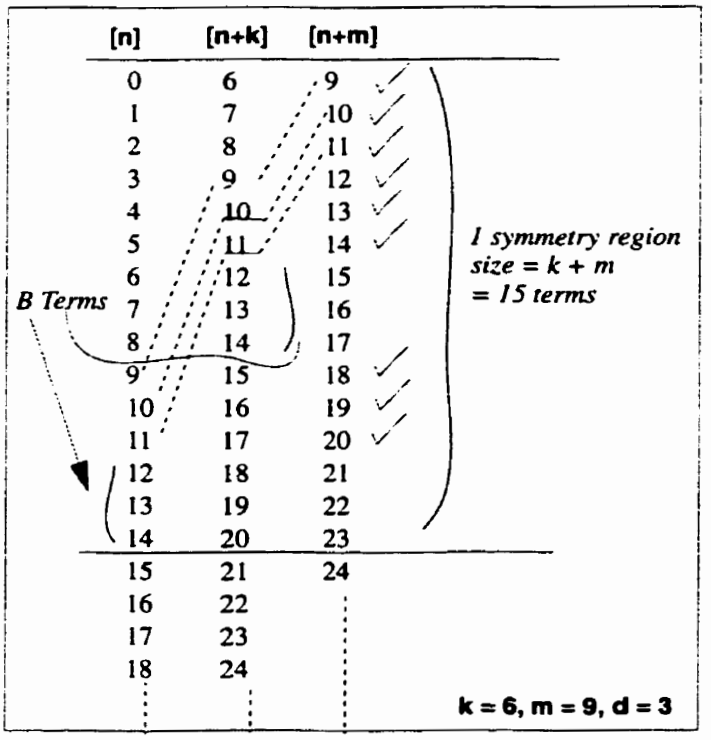
$C_3[k, m] = A + B + R$ with:

$$A = \sum_{p} \sum_{i=0}^{d-1} s[i+m+p] \{ s[i+p] s[i+k+p] + s[i+d+p] s[i+m+d+p] + s[i+k+m+p] s[i+2m+p] \}$$

$$B = \sum_{p} \sum_{i=d}^{k-1} s[i+m+p] \{ s[i+d+p] s[i+m+d+p] + s[i+k+m+p] s[i+2m+p] \}$$

$$R = \sum_{i=P(m+k)}^{N-1} s[i] s[i+k] s[i+m] .$$

Figure 8-2 Product terms for computing $C_3[6,9]$



Here the index p starts from 0, and is incremented by $(m+k)$ each time, for a total of P times. Thus $p = 0, \ldots, (m+k), 2(m+k), \ldots, (P-1)(m+k)$.

The number of operations required for this case are given in Table 8-2 below:

Table 8-2 Number of operations for case 2

Term	Multiplies	Adds
A	4dP	P (3d - 1)
B	3P(k-d)	P {2(k-d) -1}
R	2{N - P(m+k)}	N - P(m+k)

8.1.3 Comparative Results

Both cases result in a reduced number of multiplications (the number of additions is unchanged). The percentage savings in either case depend on how close the number of product terms is to an integer number of symmetry regions.

Table 8-3 Comparative results

$C_3[k,m]$ N / Case	Regular Computations		This Algorithm	
	Adds	Mul	Adds	Mul
$C_3[2,8]$ N=14 / 1	14	28	14	20
$C_3[2,8]$ N=200 /1	200	400	200	288
$C_3[6,9]$ N=15 /2	15	30	15	21
$C_3[40,50]$ N=200 /2	200	400	200	300
$C_3[20,30]$ N=300 /2	300	600	300	420
$C_3[20,30]$ N=150 /2	150	300	150	210
$C_3[20,30]$ N=200 /2	200	400	200	280

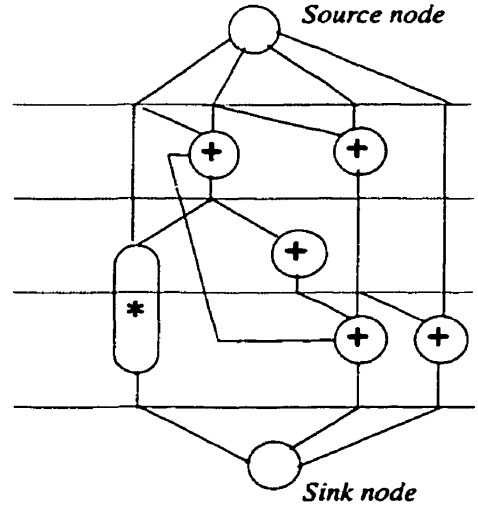
Table 8-3 illustrates some examples for various lags (thus cases 1 and 2) and various N values (thus various number of terms for the residual R). In applications where all valid combinations of (k, m) are considered (e.g., [Ran95]), the total number of multiplications is reduced by 20 to 25%. Considering that the total number of operations is on the order of 10^6 per analysis frame, this clearly results in significant savings. Finally, note that the exceptional case where both lags are zero cannot be factored and thus $C_3[0, 0]$ is evaluated using the regular form.

8.2 Mapping DSP Algorithms onto Configurable Architectures

The rapid advancement in VLSI and microprocessor technology has made single-chip parallel architectures economically available. Fixed and reconfigurable arrays containing a number of arithmetic and combinational modules have been proposed [Asc96][Che95][Mas91][Mol86]. The use of these arrays is becoming an attractive alternative to fixed ASICs for implementing iterative DSP algorithms. When mapping an algorithm onto such an array, one needs to optimally schedule and allocate the available hardware resources to the various operations in order to fully exploit the available parallelism and minimize the execution time of an iteration. This resource scheduling/allocation problem consists of determining when and on which unit to schedule a given task, while respecting precedence relations and resource constraints.

The general scheduling/allocation problem entails sequencing the operations of a control/data flow graph (Figure 8-3) into a correct order and using the available hardware resources that can perform these operations. Scheduling and allocation are interdependent tasks and should be performed simultaneously in order to achieve an optimal solution. This optimization problem may be *resource-constrained* or *time-constrained.*

Figure 8-3 A control / data flow graph



The former minimizes the number of control steps when the number of hardware resources is fixed and the latter minimize the number of resources when the number of control steps is fixed. The first problem is typically the case in mapping an algorithm on a distributed architecture, whereas the second is

more typical of a synthesis situation, where more hardware units may be synthesized as needed. Scheduling assigns each operator node to a control step that represents the controller state in which this operator will execute. Figure 8-3 shows a typical schedule where the control step boundaries are shown as horizontal lines. Scheduling fixes the order in which operators are implemented in a way that meets all dependencies specified by the edges of the graph. Hardware allocation assigns function units to execute the operations and storage units to store intermediate and final results. Allocation takes into account the number of available resources, the communication costs between units and buffers, and the possibility of resource substitution.

To solve the scheduling problem, both heuristic and exact scheduling algorithms have been used. In the early approaches, simple schemes such as ASAP or ALAP were suggested to minimize the schedule length while ignoring the hardware costs and timing constraints. Other approaches attempt to minimize hardware costs in the resulting allocation. These include greedy heuristics such as list scheduling [Goo89] and force-directed scheduling [Pau89], iterative transformational approaches such as simulated annealing, and exact approaches such as integer linear programming (ILP) [Lee89]. The ILP is attractive in providing an optimal solution but is an NP-hard problem in its basic form [Hwa91]. The typical execution time of an ILP scheduler are exponential, though special characteristics of the scheduling problem can be exploited to reduce the runtime.

The work presented here deals with the scheduling and allocation of a set of operations to a parallel architecture that contains a fixed number of hardware resources. The objective is to minimize the total execution time using the various available resources at each control step. The communication cost is not accounted for, in that it is assumed that the functional units may be connected in any way. The approach is a branch-and-bound search procedure that uses a non-deterministic heuristic at each time step to match the free resources to the operations that are ready for scheduling. The idea is partially inspired from a proposed approach to allocation in the context of operations management [Dre91]. The following set of requirements is captured in the model:

- There are several types of hardware resources and several instances of each type.
- A given operation may be executed on more than one hardware unit type.
- There is an order precedence relation between the various operations.
- There is a limited number of hardware resources.
- There is an upper limit on the execution time of the entire flow graph.

- The time to perform a given operation is known and depends on the hardware unit type used to execute it.

8.2.1 Problem Formulation

Multiple resource-constraint scheduling and allocation may be formulated as follows:

$$\text{Minimize } f_n \tag{E 8.2}$$

Subject to:

$$f_j - f_i \geq d_j \qquad (i, j) \text{ contained in } H \tag{E 8.3}$$

$$\sum_{S_t} r_{ik} \leq b_k \quad ; \quad t = 1, \ldots, f_n, k = 1, 2, \ldots, K; \tag{E 8.4}$$

where:

- f_i : Finish time of activity i (n being the last activity).
- H : Set of pairs of activities indicating a precedence relation.
- d_i : Processing time of activity i.
- r_{ik} : Amount of resource type k required by activity i.
- S_t : Set of activities in progress in time interval $[t-1, \ldots, t]$.
- b_k : Total number of units of resource type k.

The precedence constraint (Eq 8.3) indicates that an activity j can only be started if all predecessor activities are completed. Once started, activities run to completion (non-preemptive condition). The resource constraints (Eq 8.4) indicate that for each time period and for each resource type k, the resource amounts required by the activities in progress cannot exceed the resource availability. The total schedule duration is minimized by minimizing the finish time of the ending activity (n).

8.2.2 The Branch-and-bound Search Process

In a branch-and-bound search process [Sti78][Dem92], the nodes in the search tree correspond to partial schedules in which finish times have been temporarily assigned to a subset of the activities. These partial schedules (PS_m) are feasible and satisfy both precedence and resource constraints. They are built up starting at time 0 and proceed systematically throughout the search by adding at each decision point subsets of activities until a complete feasible schedule is obtained. The assignment is temporary since it may be changed later to resolve a resource conflict.

At each decision point *m*, the procedure identifies all activities that can be put in progress according to various rules. If it is impossible to schedule all eligible activities at time m, a resource conflict occurs. Such a conflict will produce a new branching in the solution tree. The branches describe ways to resolve the conflict by making decisions about which combinations of activities are to be delayed. The delaying set *D(p)* consists of all activities -either in progress or eligible- whose delay would resolve the current resource conflict at level *p* of the search tree. To resolve the resource conflict, various delaying strategies are considered at each level to determine which set of activities that may be delayed in favor of others [Dem92],[Sti78].

8.2.3 A Non-deterministic Allocation Heuristic

The allocation scheme described here follows the same general framework as the branch-and-bound approach. Instead of delaying strategies, a heuristic is used at each decision point to determine the matching of a candidate operation with an available resource. The key idea of this heuristic is as follows: At any point in time *m*, given the set of available resources (*AR*) and the set of candidate operations (*SC*) that meet precedence constraints, the criteria for matching operation *j* with resource *k* is a probabilistic decision based on quantifying the opportunity cost of *k* and the scheduling urgency of *j*:

- The opportunity cost of the matching pair (*j*, *k*) is the consequence on the overall schedule delay if *k* were not available. This depends on the availability of other resource types -if any- that can be used to execute operation *j*;
- The scheduling urgency of operation *j* depends on the position of *j* on the critical path, the number of immediate successors of *j*, and their position on the critical path.

The following sets are defined:

- *S0* = {operation *j* currently unscheduled}.
- *S1* = {operation *j* currently scheduled}.
- *AR*: the set of hardware resources currently available.
- $SC = \{ j \in S0 | ES_j \leq t, V_j \in S1 \}$. This is the set of unscheduled resources whose predecessors have been scheduled; V_j is the set of predecessors of operation *j*.

The following variables are used:

- NA_k: Number of hardware units of resource type k currently available.
- AD_k: Earliest availability date of any unit of resource type k.
- d_{jk}: Time required to perform operation j with unit of type k, and $\delta_{jk} = min\{d_{jk} | (k = 1, ..., K)\}$.
- ES_j: The earliest possible start time of operation j.
- LF_j: The latest possible finish time of operation j.
- r_{jk}: Number of units of type k required to perform operation j.
- b_k: Total number of hardware units of type k.
- J: Total number of operations.
- K: Total number of hardware unit types.
- t: Specific period $t = 0, 1, ..., T$.
- T: Imposed time limit to finish all operations.

8.2.3.1 Computing the matching criteria

At each decision instant, given the set of candidates operations SC and the set of available resources AR, a set of 'probability weights' is computed:

$$\gamma_{jk} = [\underbrace{min(d_{jq})}_{q \in AR,\, q \neq k} - d_{jk}] \cdot cp_j \quad , \quad \text{for all } j \in SC \text{ and } k \in AR. \qquad \text{(E 8.5)}$$

The bracketed term -denoted by μ_{jk}- is the delaying consequence on the schedule that is incurred if resource k were not available and operation j were to be scheduled on another resource type (q). Consider for example a multiplication that can be scheduled on either a multiplier (MUL) or a look-up table (LUT). The delaying costs for the LUT and the multiplier are respectively (assuming that the execution time for a multiplier is 1 clock and that of a LUT access is 3 clocks):

- $\mu_{j\,LUT} = \min(1) - 3 = -2$ and $\mu_{j\,MUL} = \min(3) - 1 = 2$

Now, consider two operations i and j, where j can be scheduled on the multiplier only and i can be scheduled on either:

- $\mu_{i\,LUT} = -2$ and $\mu_{i\,MUL} = 2$
- $\mu_{jLUT} = min(1) - \infty = -\infty$

- $\mu_{jMUL} = \infty - 3 = \infty$

As expected, the delaying cost of j on the multiplier is the highest since it cannot be executed on any other resource type. An allocation decision based on the delaying cost should favor the allocation of the multiplier to operation j. The second term in (Eq 8.42) is the critical path weight of operation j, cp_j. The weight of an operation depends on its relative position on the critical path, the number of immediate successors and their relative positions:

$$cp_j = \left[\frac{z}{z_j} + \sum_{S_j} \frac{z}{z_{s_j}}\right]^{\alpha} \qquad \text{(E 8.6)}$$

- z is the total length of the critical path.
- z_j is the length of the critical path up to operation j.
- S_j is the set of successors of j.
- α is a tuning coefficient < 1.

The set μ_{jk} is first normalized by finding a lower bound:

$$\mu_{min} = \min \{\mu_{jk} \mid \text{for all } j \in SC \text{ and } k \in AR\}$$

and a new 'probability weight' is then computed:

$$\gamma_{jk} = (\mu_{jk} - \mu_{min}) \cdot cp_j . \qquad \text{(E 8.7)}$$

This probabilistic matching measure resolves the two cases of multiple matching possibilities:

- A given operation may be scheduled on more than one of the available resource types (a multiplication may be scheduled on a multiplier or a look-up table).
- In tight resource times, a given resource has to be allocated to one operation over another.

8.2.4 The Scheduling Algorithm

A critical path analysis is first performed to determine the earliest possible starting time (ES_j) and the latest possible finish time (LF_j) of each operation j. More specifically, starting with task 1, ES_j is computed by a recursive relation:

$ES_1 = 0$.

$ES_j = max\{(ES_h + \delta_h) \mid h$ belongs to the set of immediate predecessors of task $j\}$.

Similarly, to find LF_j start with the time limit T by which all tasks must be completed and perform a critical path analysis using the execution times d_m of all immediate successors m of task j and working backward in time. Starting with the last operation (denoted by the index J) in the flow graph, the recursion is used:

$$LF_J = T.$$

$$LF_j = min\ \{(LF_m + d_m) \mid m \text{ belongs to the set of immediate successors of task } j\}.$$

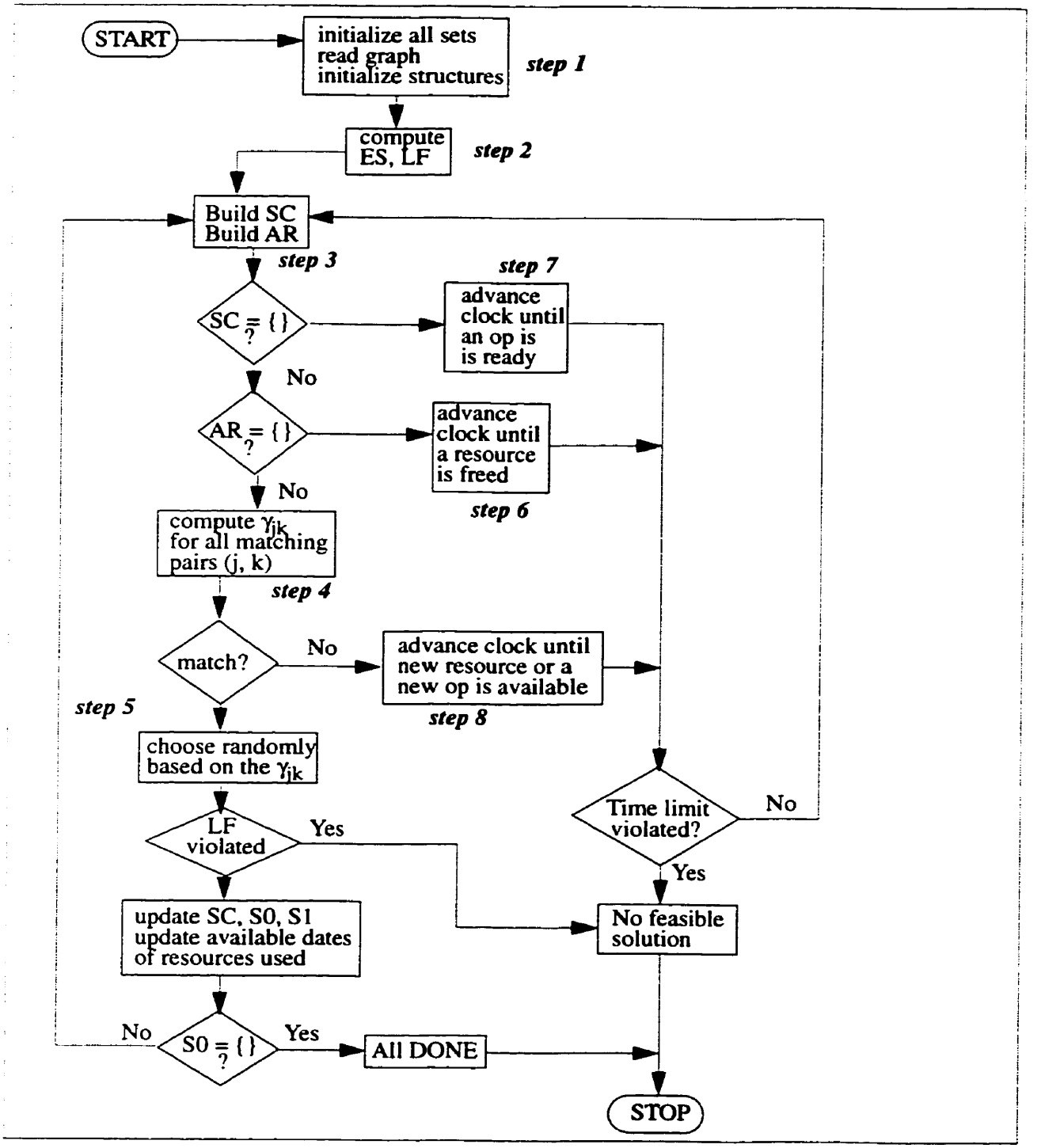
The interval $[ES_j, LF_j]$ thus represents a bound on the maximum time interval within which operation j can be scheduled without violating the precedence and time limit constraints.

At every iteration, all operations whose earliest starting time (ES) is less than the current time and whose predecessors are already scheduled are placed in a "candidates" set (SC). All resource types whose availability date is less than the current time are placed in an "available resource" set AR (Figure 8-4).

An attempt is made to match a candidate operation with an available resource. If no such match is possible, the clock is advanced until a new operation or a new resource type becomes available (which ever comes first) and a match is reattempted. When at least one match is possible, the matching criteria are computed for all potential pairs (j, k) using (Eq 8.44). The allocation decision is then determined by choosing randomly a pair (i, j) based on the weight γ_{jk}. The chosen operation is moved from the unscheduled set ($S0$) to the scheduled set ($S1$). The critical path analysis is performed again for all successors of j, and the availability date of the chosen resource k is re-evaluated.

When no more resources are available, the clock is advanced until at least one instance of any resource type k is free. When the candidates set SC gets empty, the clock is advanced until at least one unscheduled operation could get scheduled according to ES and precedence constraints. In a sense, the algorithm follows the general framework of branch and bound methods discussed earlier. It simultaneously decides about operation sequencing (which operation precedes others) and resource assignment. The algorithm builds precedence and resource feasible partial schedules; "partial" in the sense that not all operations have currently been scheduled. Scheduling a new operation is equivalent to augmenting the partial feasible solution.

Figure 8-4 Flow of control of the allocation algorithm



8.2.5 Results and Discussion

8.2.5.1 Data used

To illustrate the allocation model, the following matrix operation is used: $BA^{-1}X$, where A and B are size-2 matrices and X is 2-point vector. Thus,

$$BA^{-1}X = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \frac{a_{22}}{D} & \frac{-a_{12}}{D} \\ \frac{-a_{21}}{D} & \frac{a_{11}}{D} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{b_{11}a_{22}x_1}{D} + \frac{-b_{11}a_{12}x_2}{D} + \frac{-b_{12}a_{21}x_1}{D} + \frac{b_{12}a_{11}x_2}{D} \\ \frac{b_{21}a_{22}x_1}{D} + \frac{-b_{21}a_{12}x_2}{D} + \frac{-b_{22}a_{21}x_1}{D} + \frac{b_{22}a_{11}x_2}{D} \end{bmatrix}$$

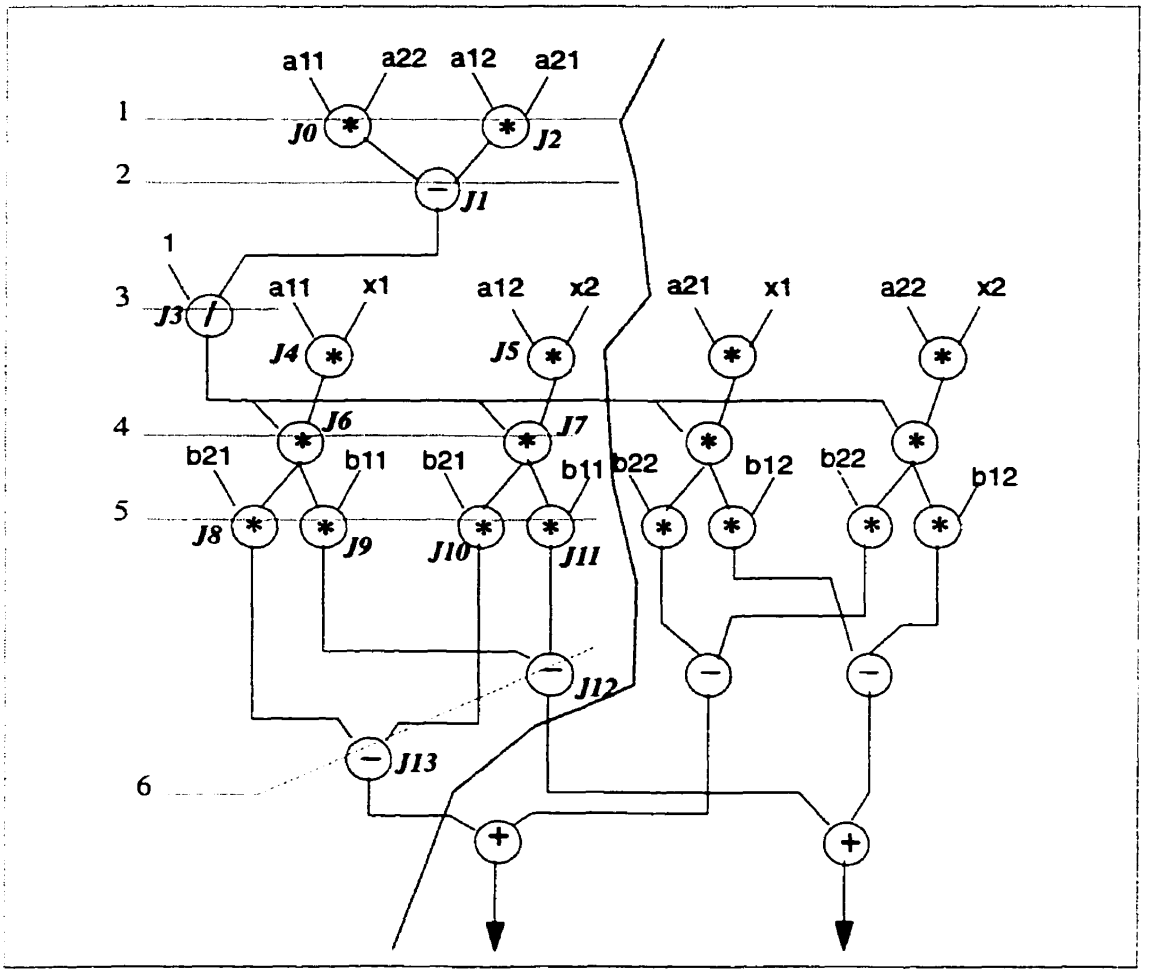
with $D = a_{11}a_{22} - a_{12}a_{21}$. The flow graph of the operations is shown in Figure 8-5 below. Due to its symmetry, only the left half of the graph is considered. The various critical path positions are also shown and numbered.

The time limit for the entire graph was set to 14 clock cycles. There were 14 operations, and 4 types of hardware resources were provided (Table 8-4).

Table 8-4 Hardware resources available

Resource	Type	Nb units	Execution Time
R0	MUL	2	2 clock cycles
R1	Adder	2	1 clock cycle
R2	Divider	1	2 clock cycles
R3	LUT	1	3 clock cycles

Figure 8-5 Flow graph for the matrix operation $BA^{-1}X$



8.2.5.2 Scheduling results

The critical path analysis was performed on the graph to determine the earliest starting and latest finishing times of each operation (prefix *R* denotes a resource and prefix *J* an operation). The scheduling/allocation algorithm was run using the above data. The results are shown in Figure 8-6. The algorithm took 20 iterations to find the complete schedule. To find a relative comparison, the problem was formulated using a basic ILP formulation. Precedence relations, resource constraints and concurrency of operations was modeled and required 50 equations. The resulting schedule had the same total execution time (Figure 8-7) as the branch-and-bound approach; however it required 19,000 pivoting iterations to generate the optimum solution. In this simple example, the two algorithms generate similar

schedules and clearly in a more complex problem, the branch and bound approach cannot generate the optimum solution that an ILP model produces. However, the complexity of the ILP problem becomes quickly prohibitive, particularly if resource substitution is to be factored in. For this reason, the proposed approach is attractive for the kind of DSP algorithms required for the applications presented in this work.

Figure 8-6 Scheduling results using the branch-and-bound model

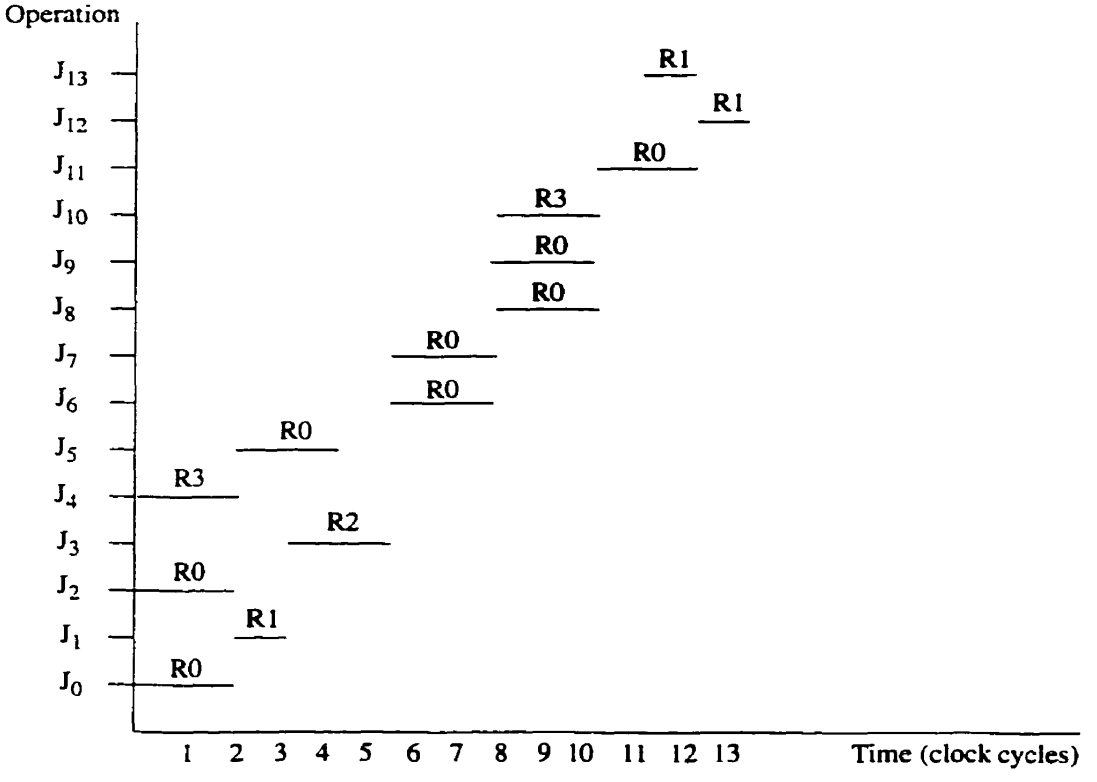
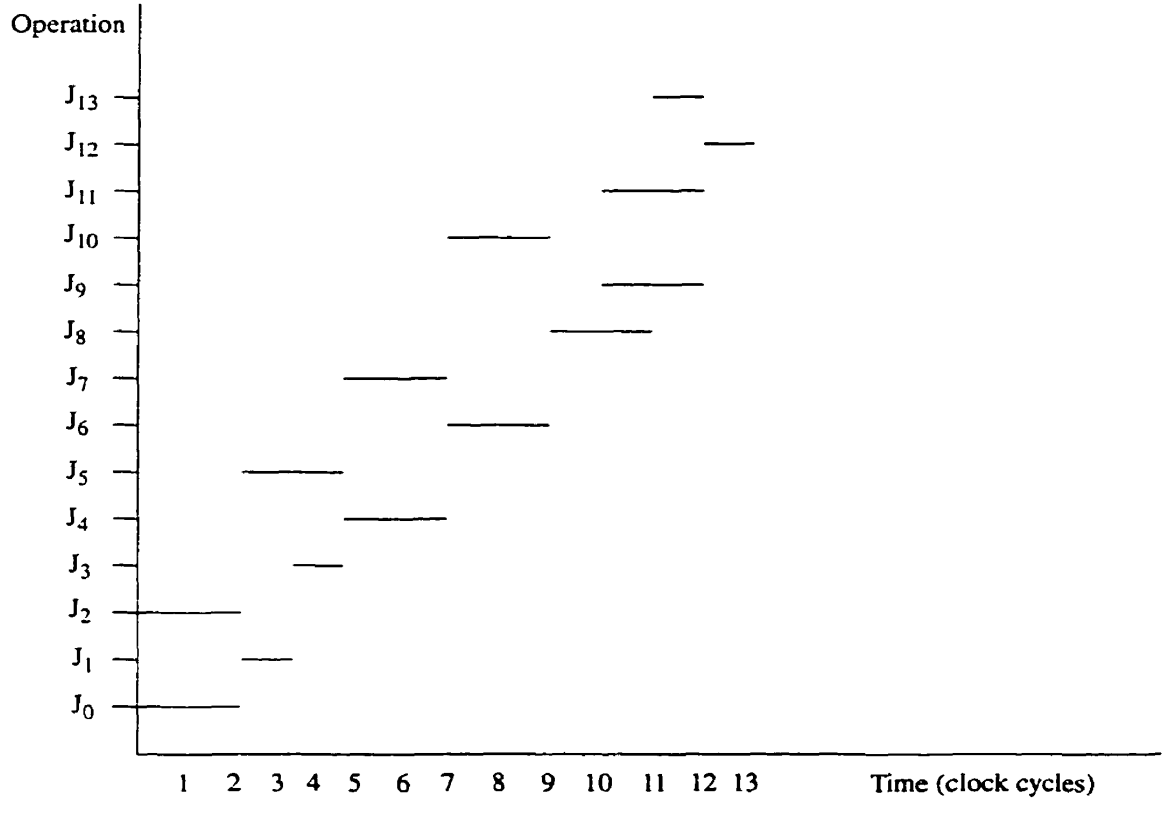


Figure 8-7 Scheduling results using Integer Linear Programming



8.3 Conclusion

This chapter addressed some of the computational issues in HOS-based applications. A new algorithm was proposed for efficiently computing the 3^{rd}-order cumulant function. The concept extends similar approaches used for the autocorrelation function in that it exploits the redundancy in the product terms to infer a factored expression with fewer multiplications.

In the second part of this chapter, a new scheduling and allocation scheme was proposed for the general problem of mapping a set of parallel operations on a configurable multi-unit architecture. The concept of resource substitution was introduced and the algorithm is based on branch-and-bound concepts with a non-deterministic heuristic that accounts for the opportunity cost of the resources and the scheduling urgency of the operations. The proposed scheme is a trade-off between finding an optimal solution, such as ILP methods, and keeping the complexity of the problem manageable and easy to formulate for large flow graphs. A typical matrix operation is used to compare the resulting schedule with that of the ILP and to illustrate the effectiveness of the heuristic. Some of the improvements to consider include using additional metrics to improve the decision criterion and accounting for the communication costs between units.

CHAPTER 9 Conclusion and Future Work

9.1 Conclusion

The objective of this thesis was to exploit the HOC properties of speech in the aim of finding new algorithms for speech enhancement and for robust voice activity detection. To achieve this objective, it was necessary to take an exploratory detour into the peculiarities of the HOC of speech and some of the general properties of HOS and their relation with their second order counterparts. Unlike the reported work on the use of HOC for speech, the approach taken here is more formal, whereby the HOC properties of speech are first established analytically, then verified using simulations with actual signals, and finally used for estimating specific speech parameters for the two applications considered.

To establish the analytical framework, two speech domains were considered, namely a narrow subband representation and the LPC residual signal. These domains were chosen because they allow a mathematically tractable model and because they have inherent peculiarities that yield useful characteristics of the HOC.

In both representations, the analytical expression of short-term speech was based on the sinusoidal model proposed in [McA86] and appropriately modified to suit the domain considered. The argument for using a linear deterministic model is twofold: it is realistically accurate for short term voiced segments, which are more perceptually important, and it offers a mathematically manageable base for deriving the expressions of the higher-order cumulants in terms of vital speech attributes.

HOC of subbanded speech

It was found that the 3rd-order cumulant of subbanded speech is identically zero, except in rare situations, which suggests that this cumulant is not very useful in this context. The 4th-order cumulant of voiced speech on the other hand is non-zero and may be expressed in terms of speech parameters, such as harmonic amplitudes, frequencies, damping factor, and sometimes energy. Therefore, it is conceivable to estimate these parameters using the value of the 4th-order cumulant at few lags.

The properties thus derived analytically and verified by simulation prove to be quite interesting: The 4th-order statistics of speech allow detecting the presence of speech harmonics and provide an upper and lower bound on the speech energy, which in turn provide bounds on the noise energy present in a given band. These findings are the basis for the approach used in the context of speech enhancement, where the 2nd-order statistics of the speech signal are estimated from the 4th-order cumulant of the noisy speech.

The simulations carried out in Chapter 4 refuted the claim that unvoiced speech can be modeled as a harmonic process, and suggested that it is more likely Gaussian given its zero skewness and kurtosis. The simulations also verified that the sinusoidal model assumed is in general valid for voiced speech, though not appropriate for transitional segments, which may be better represented by an exponentially decaying sinusoid.

Speech Enhancement

An optimal filter approach and subband representation are used as the basis for speech enhancement. The key idea is to use the 4th-order statistics of the noisy speech to estimate the required parameters for the enhancement filters, namely the SNR, the autocorrelation of speech and the probability of speech presence. Enhancement is carried out in the time domain, using the p most recent samples in each band. By making a white assumption about the noise, the problem is formulated using symmetric matrix algebra, resulting in an easily solvable matrix system. Finding the filter coefficients requires the speech and noise energies as well as the autocorrelation of speech. The probability of speech presence in each band is combined with the filtering to further attenuate non-speech bands. This probability is quantified by the kurtosis of each frame and the estimated noise energy. Using the computed variance of this estimator, it is possible to compute this probability using the *erfc* function.

To properly smooth the variance of the HOS estimators and the parameters derived from them, an inter-frame smoothing was adopted. For the SNR, the scheme proposed in [Eph84] is used and a simi-

lar one formulated for the speech autocorrelation. Finally, frequency masking is factored in by modelling the auditory filter shapes and combining the noise and speech energy in each auditory filter prior to computing the SNR.

Informal subjective listening and examination of the spectrograms showed that the resulting algorithm is effective on typical noises encountered in mobile telephony such as street, office and fan noise. This finding is not surprising, as these noise types contain a significant Gaussian component, being generated by a large number of independent sources. The algorithm does not however eliminate the non-Gaussian components in these noises, such as dominant conversations, as these processes are impulsive and do not have zero HOS.

The performance is compared to the TIA standard [IS127] for noise reduction. The results show that the HOS algorithm is better at preserving the harmonic structure of the speech and results in less speech distortion. This result is an important one and clearly demonstrates that 4^{th}-order statistics are effective in isolating bands containing speech harmonics and preventing overattenuation of these bands or their use as noise bands for noise estimation. The algorithm also results in more overall reduction of the noise, but that comes at the cost of more noise artifacts, particularly at very low SNR where the variance of the HOS estimators start to cause large errors in the estimation of the noise and the identification of harmonic bands.

HOC of the LPC Residual

The LPC residual of steady voiced speech is far from being Gaussian as shown analytically by the HOC derivations and verified by simulation. Moreover, the horizontal slices of the 3^{rd} and 4^{th} order cumulants have a zero-phase characteristic for steady voiced speech, and when normalized have a magnitude that is independent of the signal energy. These features suggest that these cumulant slices can be used in a similar way as the autocorrelation function for such applications as pitch estimation and voicing detection.

The skewness and kurtosis of steady voiced speech may be expressed in terms of the number of harmonics (i.e., related to the pitch) and speech energy. When normalized, they are independent of speech energy and as such may be used as a voicing detector. The normalization however results in the degradation of the detection capability in low SNR conditions, as the new metrics can now be expressed in terms of the number of harmonics and SNR. In addition, these two metrics degrade in effectiveness as

voiced speech becomes non-stationary. On the other hand, the DC component of the horizontal cumulant slice of voiced speech is non-zero in both steady and non-stationary conditions.

The simulations carried out in Chapter 6 verified the validity of the derivations and the underlying model for steady voiced segments. The simulations on sustained unvoiced speech revealed that it cannot be modeled as a harmonic process as suggested by the sinusoidal model, but has a Gaussian-like nature. As a result, detection and quantification of this type of speech cannot be done with HOS.

Voice Activity Detection

The proposed VAD algorithm combines HOS measures with SNR to classify speech and noise frames and determine whether a speech frame is voiced. The variance of the estimators of the 3^{rd} and 4^{th} order statistics is used to yield a likelihood measure for noise frames, and a voicing condition for speech frames is derived based on the relation between the skewness and kurtosis of voiced speech. The algorithm is designed as a finite 2-state machine. Its performance is quantified in various noise conditions using the probability of correct and false classification and compared to those of the G.729B VAD [Ben97].

The performance in noise of the two algorithms showed the HOS-based VAD has overall comparable performance to the G.729B. Its probability of false classification is lower in low SNR and Gaussian-like noise, but higher in speech-like noises and high SNR. It is however based on more analytical ground, is conceptually simpler and uses a smaller parameter set, therefore it is easier to tune. It is also appropriate to use in conjunction with speech coders where such parameters as the LSF's, required by the G.729B, are not available.

HOS Generalities

When dealing with sinusoidal or periodic signals, such as speech, 2D slices of the higher cumulants are often used and, in the case of a deterministic signal, their Fourier transform may be expressed in terms of the Fourier transform of the underlying signal, as was shown in Chapter 3. In the 3^{rd}-order case, the Fourier transform of the horizontal slice may be computed from the bispectrum points. Since this bispectrum is expressed in terms of the Fourier transform of the underlying signal, this transform may be recovered from the bispectrum points. New schemes were proposed for Fourier magnitude recovery from the bispectrum. These schemes are a good compromise between using all the available information and finding a computationally manageable method, that will not break down if some of these points are zero.

Another insight in Chapter 3 is the use of the horizontal slice of the 4^{th}-order cumulant and the significance of the DC component of that slice. In the case of a flat spectrum signal, this DC component may be written in terms of the signal energy and the bandwidth.

While there is no direct way to estimate the signal power spectrum from its higher cumulants, it was shown that the geometric mean of the power spectrum may be estimated by considering the product of bispectrum magnitudes along a diagonal line in the bifrequency plane.

Implementation Aspects of HOS

Implementation issues related to using HOS include the effect of finite length segments and the complexity of computing cumulants. For instance, it is shown that the estimators of the higher moments are unbiased for any random process, but when computing the higher moments of a sinusoid, a bias term is produced whenever the length of the segment is not an integer number of signal periods. Moreover, in the case of a random process, the estimator of the 4^{th}-order statistic is biased due to the squaring of the second moment term. In the case of a white Gaussian process, this bias may be quantified and removed. In the case of a deterministic sinusoid, the bias term may be reduced by averaging the moments over a few overlapping frames prior to computing the cumulants.

The variance of the higher moments depends on the pdf and the correlation of the underlying process. In the case of white Gaussian noise, the variances may be expressed in terms of the variance of the underlying process. Similarly, the variance of the skewness and kurtosis may be expressed in terms of the process variance. These important results are used in the context of enhancement and voicing detection for determining the probability of a frame consisting only of noise.

Implementing higher cumulant functions inherently involve more computations. In the case of the 3^{rd}-order cumulant, the redundancy in the product terms may be exploited to result in an efficient formulation with a reduced number of multiplications. The idea of the proposed algorithm is to exploit the fact that samples appear two or three times as multiplicands and can be factored accordingly. The other observation is that the product terms may be grouped into symmetry regions that repeat and can also be factored. Depending on the two lags, the savings in multiplications are about 20 to 25% and in typical real time applications, this results in significant savings of computing power.

The last part in Chapter 8 addressed the problem of mapping a DSP algorithm onto a configurable architecture. The relevance of this problem to the context of an HOS-based application is that algorithms based on HOS are highly parallelizable and good candidates for this kind of architectures in

real-time applications, such as mobile telephony. The proposed allocation model builds on the concept of branch and bound methods whereby at various points of the scheduling process, matching decisions have to be done. A non-deterministic matching criterion was proposed to resolve the problem where a given operation may be executed on more than one resource and the case where a resource has to be assigned to one operation over another.

Overall...

This thesis has provided fresh insights into the use of HOC for speech analysis. The exploratory part revealed important relations and derivations of the HOC of subbanded and LPC filtered speech. These findings may be used in a variety of applications. Two of these were addressed in detail and it was demonstrated that in spite of the practical limitations of using HOC and the approximate nature of the speech model assumed, effective applications are possible. By making use of only HOC measures, the performance of the proposed algorithms is shown to be comparable, and even better in some respects, to the current standards. As this is the first iteration of this type of work, it clearly demonstrates the promising potential of HOC in yielding algorithms that would surpass the current state of the art. The work however does not claim these cumulants to be superior in and of themselves to 2^{nd}-order approaches, but rather that they provide additional information about the signal that is immune to the presence of noise, which make them particularly effective in applications designed for low SNR conditions. Clearly, successful algorithms are those that can combine the two approaches and harness the advantages of both.

9.2 Proposed Future Work

There is room to further investigate and refine the ideas presented in this thesis. The following is a list of some of the improvements to consider:

- In the subband representation, the expression of the higher cumulants may be extended to the case of larger bands containing three or four harmonics. This case is more challenging than the two sinusoid case since it is more difficult to find a general expression for the higher cumulants that is independent of the number of harmonics in the band and independent of the relation between the harmonics.
- The variance of the HOS estimators was quantified in terms of the variance of the underlying process in the case of a white Gaussian process. It may be possible to extend the derivations to the case where the process is a mixture of speech and noise. The case will have to be considered in the different speech domains that are used.
- The modelling of speech adopted here may need to be expanded. A probabilistic approach may have to be investigated to find possible distributions of speech samples in the subband and LPC residual domains considered.
- In the LPC residual, an improved model that captures non-stationary speech needs to be investigated and the derivations for the higher-order cumulants need to be extended accordingly. The case is more complex when the speech model contains time-varying or random amplitude and frequency components, since the expression for the cumulants requires the knowledge of the higher-order correlation of this random amplitude process.
- In the VAD application, it is worth investigating the use of other metrics and whether combining them with HOS measures can lead to more accurate classification.
- Other applications of the theoretical findings may be investigated. For example the problem of pitch estimation using the slices of the 3^{rd} and 4^{th} order cumulants and whether some of the traditional approaches ([Mar72], [Wis76], [Sec83]) that use the autocorrelation functions may be extended to the higher cumulants.
- The algorithm for efficiently computing the 3^{rd}-order cumulant function can be extended to the 4^{th}-order cumulant. The case is more challenging as the expression for that cumulant involves more lags as well as both the 2^{nd} and 4^{th} order moment functions.

APPENDIX A Bias and Variance of the HOS Estimators

When computing the higher-order statistics of mixed signal and noise, only finite data records are available; as a result, both probabilistic errors and deterministic errors are present. The former are due to the bias and variance of the estimator of a random process. The later are due to the size of the data frame relative to the period of the signal, in the case of a harmonic signal.

A.1 Definitions

Given a zero-mean, discrete random or mixed process, $x(n)$, the estimators of the 2nd, 3rd and 4th moments are defined as:

$$M_{kx} = \frac{1}{N}\sum_{n=0}^{N-1} [x(n)]^k \text{ estimator for } E[\{x(n)\}^k], \qquad \text{(E A.1)}$$

for $k = 2, 3$ and 4. Let the true skewness and kurtosis of $x(n)$ be denoted by SK and KU respectively. The estimators for the skewness and kurtosis are respectively:

$$\hat{SK} = M_{3x} \qquad \text{(E A.2)}$$

$$\hat{KU} = M_{4x} - 3(M_{2x})^2. \qquad \text{(E A.3)}$$

In the case where $x(n)$ consists of both signal and Gaussian noise, the true variance of the Gaussian noise is denoted by υ_g and the true variance of the signal is denoted by υ_s.

A.2 Bias of the HOS Estimators of a Sinusoid in Noise

A.2.1 Case 1: noise only {x(n) = g(n)}

• **Proposition 1.** *The estimator for the skewness (Eq A.2) is unbiased whereas that of the kurtosis (Eq A.3) is only asymptotically unbiased. For the case of a white Gaussian process, this estimator may be unbiased for any segment size N by introducing an artificial bias in the first term; thus a new unbiased estimator is proposed for the kurtosis:*

$$\hat{KU}_U = \left(1 + \frac{2}{N}\right) M_{4g} - 3\,(M_{2g})^2 \qquad \text{(E A.4)}$$

•**Proof:** Assuming that $g(n)$ is a white Gaussian process with zero mean, then all its odd-power moments are zero. The 2nd and 4th moments are denoted by υ_g and δ_g respectively. Moreover, the higher moments of $g(n)$ may be expressed in terms of the variance by evaluating the moment integral, $E[g^m] = \int_{-\infty}^{\infty} g^m \frac{1}{\sqrt{2\pi\upsilon}} e^{-g^2/2\upsilon} dg$ for $m = 4$, 6, and 8; namely:

$$E[\{g(n)\}^4] = 3\,(\upsilon_g)^2 \qquad \text{(E A.5)}$$

$$E[\{g(n)\}^6] = 15\,(\upsilon_g)^3 \qquad \text{(E A.6)}$$

$$E[\{g(n)\}^8] = 105\,(\upsilon_g)^4. \qquad \text{(E A.7)}$$

The true skewness and kurtosis of $g(n)$ are then: $SK = 0$ and $KU = 0$.

The estimators for the higher moments are all unbiased since for any process $g(n)$:

$$E[M_{kg}] = \left(\frac{1}{N}\right) E\left[\sum_{n=0}^{N-1} [g(n)]^k\right] = \frac{1}{N}\sum_{n=0}^{N-1} E[g^k(n)] = E[g^k(n)]. \qquad \text{(E A.8)}$$

Therefore: $E[M_{2g}] = \upsilon_g$, $E[M_{3g}] = 0$, $E[M_{4g}] = 3\,(\upsilon_g)^2$. (E A.9)

The estimator of the skewness is unbiased; but the same is not true for the estimator of the kurtosis:

$$E[\hat{KU}] = E[M_{4g} - 3\,(M_{2g})^2] = E[M_{4g}] - 3E[(M_{2g})^2]. \qquad \text{(E A.10)}$$

First it is to note that:

$$E[(M_{2g})^2] = E\left[\left(\frac{1}{N}\sum_{i=0}^{N-1} g^2(i)\right)^2\right] = \frac{1}{N^2} E\left[\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} g^2(i)\, g^2(j)\right]$$

$$= \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[g^2(i)\, g^2(j)].$$

In the above double sum, there are N terms where $i = j$ and $(N^2\text{-}N)$ terms where $i \neq j$ Thus,

$$E[(M_{2g})^2] = \frac{1}{N^2}\{NE[g^4(i)] + (N^2 - N)E[g^2(i)]E[g^2(j)]\}$$

$$= \frac{1}{N^2}\{3N\upsilon_g^2 + (N^2 - N)\upsilon_g^2\}$$

$$E[(M_{2g})^2] = \upsilon_g^2\left(\frac{2}{N} + 1\right). \qquad \text{(E A.11)}$$

substituting Eq A.11 into Eq A.10,

$$E[\hat{KU}] = E[M_{4g}] - 3\left(\frac{2\upsilon_g^2}{N} + \upsilon_g^2\right)$$

and using Eq A.8 and Eq A.5,

$$= E[x^4(n)] - 3\left(\frac{2\upsilon_g^2}{N} + \upsilon_g^2\right) = 3\upsilon_g^2 - 3\left(\frac{2\upsilon_g^2}{N} + \upsilon_g^2\right)$$

$$E[\hat{KU}] = \frac{-6}{N}\upsilon_g^2 \neq KU. \qquad \text{(E A.12)}$$

The estimator for the kurtosis of Gaussian noise is only 'asymptotically unbiased' and for a small number of points, the bias terms is significant. One way to unbias the estimator is to introduce a factor that is a function of N and that will cause a bias in the first term of Eq A.10 which will cancel the bias of the second term, thus a new estimator for the kurtosis is defined as:

$$\hat{KU}_U = \left(1 + \frac{2}{N}\right)M_{4g} - 3(M_{2g})^2$$

and this estimator is unbiased for any N since:

$$E[\hat{KU}_U] = \left(1 + \frac{2}{N}\right)(3\upsilon_g^2) - 3\upsilon_g^2\left(\frac{2}{N} + 1\right) = 0 = KU.$$

A.2.2 Case 2: sinusoidal signal only {x(n) = s(n)}

- **Proposition 2.** *When s(n) consists of a deterministic sinusoid, the estimators for the skewness and kurtosis are biased whenever the segment size N is not an integer number of signal periods. The bias term in these estimators may be significantly reduced if the estimators for the 2nd, 3rd and 4th moments are computed for a few overlapping segments and averaged prior to using them for computing the skewness and kurtosis.*

• **Proof:** Let $s(n)$ be a single sinusoid of deterministic amplitude, frequency and phase:

$$s(n) = a\cos(nTw + \theta) \qquad \text{(E A.13)}$$

where T is the sampling period. In Section 4.3.2, it was found that the true skewness and kurtosis of this signal are:

$$SK = 0 \quad \text{and} \quad KU = -1.5\upsilon_s^2 \qquad \text{(E A.14)}$$

where υ_S is the average signal energy. In the derivations of Chapter 4 it was assumed that the analysis frame is made of an integer number of signal periods. When the segment size is arbitrary, a bias term is introduced in the HOS estimators. This can be seen by considering the expression for the higher moments of $s(n)$, defined as:

$$M_{ms} = E[s^m(n)] = \frac{1}{N}\sum_{n=0}^{N-1}(a \cdot \cos[nTw + \theta])^m \qquad \text{for } m = 2, 3 \text{ and } 4.$$

The above is evaluated by first noting the identities:

$$\sum_{n=0}^{N-1} e^{j(2knTw+\theta)} = \frac{1 - e^{j2kTwN}}{1 - e^{j2kTw}} e^{j\theta} = \frac{\sin(TkwN)}{\sin(Tkw)} e^{j[wkT(N-1)+\theta]}$$

$$\sum_{n=0}^{N-1} \cos(2knTw + \theta) = \frac{2\sin(TkwN)}{\sin(Tkw)} \cos[wkT(N-1) + \theta]$$

for real values of k. Thus the *second* moment is:

$$M_{2s} = \frac{a^2}{2}\frac{1}{N}\sum_{n=0}^{N-1}[\cos 2(wnT + \theta) + 1]$$

$$M_{2s} = \frac{a^2}{2}\left[\frac{2\sin(TwN)\cos(Tw[N-1] + \theta)}{N\sin(Tw)} + 1\right]. \qquad \text{(E A.15)}$$

The *third* moment is:

$$M_{3s} = \frac{1}{N}\sum_{n=0}^{N-1}(a \cdot \cos[nTw + \theta])^3$$

$$= \frac{a^3}{4}\frac{1}{N}\sum_{n=0}^{N-1}\{\cos[3(wnT + \theta)] + 3\cos(wnT + \theta)\}$$

$$M_{3s} = \frac{a^3}{4}\left[\frac{2\sin(1.5TwN)\cos(1.5Tw[N-1] + 3\theta)}{N\sin(1.5Tw)} + 6\frac{\sin(TwN)\cos(Tw[N-1] + \theta)}{N\sin(Tw)}\right] \qquad \text{(E A.16)}$$

Finally, the *fourth* moment is:

$$M_{4s} = \frac{1}{N}\sum_{n=0}^{N-1} (a \cdot \cos[nTw + \theta])^4$$

$$= \frac{a^4}{16}\frac{1}{N}\sum_{n=0}^{N-1} \{\cos[4(wnT + \theta)] + 4\cos(wnT + \theta) + 6\}$$

$$M_{4s} = \frac{a^4}{16}\left[\frac{2\sin(2TwN)\cos(2Tw[N-1] + 4\theta)}{N\sin(2Tw)} + 8\frac{\sin(TwN)\cos(Tw[N-1] + 2\theta)}{N\sin(Tw)} + 6\right] \quad \text{(E A.17)}$$

In the above expressions (Eq A.15, Eq A.16 and Eq A.17), the trigonometric terms vanish whenever N is a multiple of the signal period (i.e. $N = 2k\pi/Tw$). Due to the *sinc(x)* function the bias factor is bounded by $(1/N)$. If the effect of these terms is removed, then the expressions for the higher moments simplifies to the ones found in Chapter 4: $M_{2s} = a^2/2$, $M_{3s} = 0$, and $M_{4s} = 3a^4/8$.

If the above estimators are computed for a number of overlapping segments K and then averaged, it is equivalent to averaging out the effect of the phase term θ, since each segment will generate a different phase. It is easy to see that if the bias term in the above estimators is integrated over the phase range $[-\pi, \pi]$ (assuming a uniformly distributed phase), then the entire bias term vanishes since:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \cos[knTw + m\theta]\, d\theta = 0.$$

It is to note here that the above statement is true only when an infinite number of segments is used. In reality, using a few segments will reduce the bias significantly. Another point to note is that if the modified kurtosis estimator that was proposed in the previous section (Eq A.4) is used, then the kurtosis of a sinusoid will still be a function of the signal energy (as in Eq A.14) but an extra term involving the segment size is now in place:

$$\hat{KU}_U = \left(1 + \frac{2}{N}\right)M_{4s} - 3(M_{2s})^2 = \left(1 + \frac{2}{N}\right)\left(\frac{3}{2}\upsilon_s^2\right) - 3\upsilon_s^2$$

$$\boxed{\hat{KU}_U = \left(\frac{3}{N} - \frac{3}{2}\right)\upsilon_s^2} \quad \text{(E A.18)}$$

This does not cause a problem, as long as this term is accounted for when, for example, the kurtosis is used to estimate the signal energy.

A.2.3 Case 3: sinusoid and Gaussian noise {x(n) = s(n) + g(n)}

- **Proposition 3.** *When x(n) consists of a deterministic sinusoid in Gaussian noise, then the estimators for the skewness and kurtosis will contain unwanted bias terms, due to both the deterministic signal period as well as to the noise energy. If the estimators for the moments are averaged over overlapping segments as explained in the section above, and if the modified estimator for the kurtosis is used (Eq A.4), then the estimators for the skewness and kurtosis will not have unwanted bias terms and the kurtosis thus computed can be expressed in terms of the signal energy and the segment size, as given by Eq A.18.*

- **Proof:** Let $x(n)$ consist of both a deterministic sinusoid and white Gaussian noise: $x(n) = s(n) + g(n)$. In addition, it is assumed that both signal and noise are zero-mean and that they are statistically independent. The true higher moments of $x(n)$ are given by:

$$E[x^2(n)] = E[s^2(n) + g^2(n) + 2s(n)g(n)]$$

$$E[x^2(n)] = E[s^2(n)] + E[g^2(n)]. \qquad \text{(E A.19)}$$

$$E[x^3(n)] = E[s^3(n) + g^3(n) + 3s^2(n)g(n) + 3s(n)g^2(n)]$$

$$= E[s^3(n)] + E[g^3(n)]$$

$$E[x^3(n)] = 0. \qquad \text{(E A.20)}$$

$$E[x^4(n)] = E[s^4(n) + g^4(n) + 6s^2(n)g^2(n) + 4s^3(n)g(n) + 4s(n)g^3(n)]$$

$$= E[s^4(n)] + E[g^4(n)] + 6E[s^2(n)]E[g^2(n)]$$

$$E[x^4(n)] = \frac{3}{2}\upsilon_s^2 + 3\upsilon_g{}^2 + 6\upsilon_g\upsilon_s. \qquad \text{(E A.21)}$$

Thus the true skewness and kurtosis are given by:

$$SK = E[x^3(n)] = 0 \qquad \text{(E A.22)}$$

$$KU = E[x^4(n)] - 3(E[x^2(n)])^2$$

$$= E[s^4(n)] - 3(E[s^2(n)])^2 = -1.5(E[s^2(n)])^2$$

$$KU = -1.5\upsilon_s^2. \qquad \text{(E A.23)}$$

Since the estimators for the higher moments are unbiased, it follows that the estimator for the skewness is unbiased: $E[\hat{SK}] = E[x^3(n)] = 0 = SK$.

If the modified estimator for the kurtosis is used, and if the estimators for the moments are averaged over overlapping segment first prior to computing the kurtosis, then this estimator will be free of bias terms due to noise energy and will be expressed in terms of speech energy only, as desired:

$$E[\hat{KU}_U] = E[M_{4x}] - 3E[(M_{2x})^2]. \qquad \text{(E A.24)}$$

First, the expected value of the squared second moment is:

$$E[(M_{2x})^2] = \frac{1}{N^2}E\left[\left(\sum_{i=0}^{N-1} x^2(i)\right)^2\right]$$

$$= \frac{1}{N^2}E\left[\left(\sum_{i=0}^{N-1} [s^2(i) + g^2(i) + 2s(i)g(i)]\right)^2\right]$$

$$= \frac{1}{N^2}E\left[\left(\sum_{i=0}^{N-1} s^2(i) + \sum_{i=0}^{N-1} g^2(i) + 2\sum_{i=0}^{N-1} s(i)g(i)\right)^2\right]$$

$$= \frac{1}{N^2}E\left[\begin{array}{c}\left(\sum_{i=0}^{N-1} s^2(i)\right)^2 + \left(\sum_{i=0}^{N-1} g^2(i)\right)^2 + 4\left(\sum_{i=0}^{N-1} s(i)g(i)\right)^2 + \\ 2\sum_{i=0}^{N-1} s^2(i)\sum_{i=0}^{N-1} g^2(i) + 4\sum_{i=0}^{N-1} s^2(i)\sum_{i=0}^{N-1} s(i)g(i) + 4\sum_{i=0}^{N-1} g^2(i)\sum_{i=0}^{N-1} s(i)g(i)\end{array}\right]$$

$$= \left[\begin{array}{c}\left(\frac{1}{N}\sum_{i=0}^{N-1} s^2(i)\right)^2 + \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[g^2(i)g^2(j)] + \frac{4}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[s(i)s(j)g(i)g(j)] \\ + \frac{2}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[s^2(i)g^2(j)]\end{array}\right]$$

The first term is simply the square of the signal energy υ_s^2; the second term was evaluated in (Eq A.11) to be: $\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} g^2(i)g^2(j) = \upsilon_g^2\left(1 + \frac{2}{N}\right)$;

the third term is only non-zero when $i=j$ and can be written in terms of speech and noise energies:

$$\frac{4}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[s(i)s(j)g(i)g(j)] = \frac{4N}{N^2}E[s^2(i)]E[g^2(i)] = \frac{4\upsilon_g\upsilon_s}{N},$$

and the last term is simply the product of speech and noise energies:

$$\frac{2}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E[s^2(i)g^2(j)] = 2\upsilon_g\upsilon_s.$$

Thus the expected value of the squared second moment becomes:

$$E[(M_{2x})^2] = \upsilon_s^2 + \upsilon_g^2\left(1+\frac{2}{N}\right) + \frac{4\upsilon_g^2\upsilon_s^2}{N} + 2\upsilon_g^2\upsilon_s^2 \qquad \text{(E A.25)}$$

and by substituting Eq A.25 into Eq A.24, the expected value of the estimator of the kurtosis is:

$$E[\hat{KU}_U] = \left(1+\frac{2}{N}\right)E[x^4(n)] - 3\left[\upsilon_s^2 + \upsilon_g^2\left(1+\frac{2}{N}\right) + \frac{4\upsilon_g^2\upsilon_s^2}{N} + 2\upsilon_g^2\upsilon_s^2\right]$$

and using Eq A.21,

$$= \left(1+\frac{2}{N}\right)\left[\frac{3}{2}\upsilon_s^2 + 3\upsilon_g^2 + 6\upsilon_g\upsilon_s\right] - 3\left[\upsilon_s^2 + \upsilon_g^2\left(1+\frac{2}{N}\right) + \frac{4\upsilon_g\upsilon_s}{N} + 2\upsilon_g\upsilon_s\right]$$

$$\boxed{E[\hat{KU}_U] = \left(\frac{3}{N}-\frac{3}{2}\right)\upsilon_s^2}$$

A.3 Variance of the HOS Estimators of a White Gaussian Process

The estimators for the higher moments (Eq A.1) of any random process including Gaussian processes are unbiased, as was mentioned in the previous section. Their variance however depends on the pdf and correlation of the process.

• **Proposition 4.** *For the case of a white Gaussian process g(n) with zero mean, the variance of the* M_{2g}, M_{3g} *and* M_{4g} *estimators may be expressed in terms of the process variance* υ_g *as:*

$$\boxed{Var[M_{2g}] = \frac{2\upsilon_g^2}{N}}, \boxed{Var[M_{3g}] = \frac{15\upsilon_g^3}{N}} \text{ and } \boxed{Var[M_{4g}] = \frac{96\upsilon_g^4}{N}}. \qquad \text{(E A.26)}$$

• **Proof:** The higher moments of *g(n)* may be expressed in terms of the variance of the process as shown in Eq A.5, Eq A.6 and Eq A.7. In addition, all odd-power moments are zero, thus:

1. Variance of M_{2g}

$$Var[M_{2g}] = E[(M_{2g}-\upsilon_g)^2] = E[M_{2g}^2] - \upsilon_g^2$$

$$= E\left[\frac{1}{N^2}\sum_i\sum_j g^2(i)\,g^2(j)\right] - \upsilon_g^2 = \frac{1}{N^2}\sum_i\sum_j E[g^2(i)\,g^2(j)] - \upsilon_g^2.$$

In the above double sum, there are N terms where $i = j$ and (N^2-N) terms where $i \neq j$ Thus,

$$Var[M_{2g}] = \frac{1}{N^2}\{NE[g^4(i)] + (N^2-N)E[g^2(i)]E[g^2(j)]\} - \upsilon_g^2$$

and using Eq A.5, $Var[M_{2g}] = \frac{1}{N^2}\{3N\upsilon_g^2 + (N^2-N)\upsilon_g^2\} - \upsilon_g^2 = \frac{2\upsilon_g^2}{N}$.

2. Variance of M_{3g}

For a Gaussian process, the actual skewness is zero, thus:

$$Var[M_{3g}] = E[(M_{3g}-0)^2] = E[M_{3g}^2]$$

$$Var[M_{3g}] = E\left[\frac{1}{N^2}\sum_i\sum_j g^3(i)\,g^3(j)\right] = \frac{1}{N^2}\sum_i\sum_j E[g^3(i)\,g^3(j)].$$

There are N terms where $i = j$ and $(N^2\text{-}N)$ terms where $i \neq j$, thus,

$$Var[M_{3g}] = \frac{1}{N^2}\{NE[g^6(i)] + (N^2-N)E[g^3(i)]E[g^3(j)]\}$$

and using Eq A.6, $Var[M_{3g}] = \frac{1}{N^2}\{15N\upsilon_g^3 + (N^2-N)0\} = \frac{15\upsilon_g^3}{N}$.

3. Variance of M_{4g}

$$Var[M_{4g}] = E[(M_{4g}-\delta_g)^2] = E[M_{4g}^2] - \delta_g^2$$

$$= E\left[\frac{1}{N^2}\sum_i\sum_j g^4(i)\,g^4(j)\right] - \delta_g^2 = \frac{1}{N^2}\sum_i\sum_j E[g^4(i)\,g^4(j)] - \delta_g^2$$

where δ_g is the true 4th moment and is given (using Eq A.5) by: $\delta_g = E[g^4(j)] = 3\upsilon_g^2$.
In the above double sum, there are N terms where $i = j$ and $(N^2\text{-}N)$ terms where $i \neq j$, thus:

$$Var[M_{4g}] = \frac{1}{N^2}\{NE[g^8(i)] + (N^2-N)E[g^4(i)]E[g^4(j)]\} - \delta_g^2$$

and using Eq A.7 and Eq A.5

$$Var[M_{4g}] = \frac{1}{N^2}\{105N\upsilon_g^4 + (N^2-N)\delta_g^2\} - \delta_g^2$$

$$= \frac{1}{N}\{105\upsilon_g^4 - \delta_g^2\} = \frac{96\upsilon_g^4}{N}.$$

- **Proposition 5.** *For the case of a white Gaussian process g(n) with zero mean, the variance of the estimator of the kurtosis (Eq A.4) may be expressed in terms of the process variance* υ_g *as:*

$$\boxed{Var[\hat{KU}_U] = \frac{3\upsilon^4}{N}\left(104 + \frac{452}{N} + \frac{596}{N^2}\right)} \qquad \text{(E A.27)}$$

•**Proof:** The unbiased estimator of the kurtosis is (Eq A.4): $\hat{KU}_U = \left(1 + \frac{2}{N}\right)M_{4g} - 3\,(M_{2g})^2$

Therefore its variance is:

$$Var[\hat{KU}_U] = \left(1 + \frac{2}{N}\right)^2 Var[M_{4g}] + 9Var[(M_{2g})^2] + 6\left(1 + \frac{2}{N}\right)Cov[M_{4g}(M_{2g})^2] \qquad \text{(E A.28)}$$

The first term was computed in Proposition 4 as: $Var[M_{4g}] = \frac{96\upsilon_g^4}{N}$. (E A.29)

The second term in Eq A.28 is computed as follows:

$$Var[(M_{2g})^2] = E[(M_{2g})^4] - (E[(M_{2g})^2])^2 \qquad \text{(E A.30)}$$

$$E[(M_{2g})^4] = \frac{1}{N^4}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} E[g_i^2 \cdot g_j^2 \cdot g_k^2 \cdot g_m^2] .$$

In the above sums, there are:

$[N\cdot(N^2-2N+1)]$	terms where	$i=j$	only
$[N\cdot(N^2-2N+1)]$	terms where	$i=k$	only
$[N\cdot(N^2-2N+1)]$	terms where	$i=m$	only
$[N\cdot(N^2-2N+1)]$	terms where	$j=k$	only
$[N\cdot(N^2-2N+1)]$	terms where	$j=m$	only
$[N\cdot(N^2-2N+1)]$	terms where	$k=m$	only
$[N\cdot(N-1)]$	terms where	$i=j=k$	only
$[N\cdot(N-1)]$	terms where	$i=j=m$	only
$[N\cdot(N-1)]$	terms where	$i=k=m$	only
$[N\cdot(N-1)]$	terms where	$j=k=m$	only
N	terms where	$i=j=k=m$	
$[N^4-6N^3+8N^2-3N]$	terms where	$i\neq j\neq k\neq m$	

Therefore $E[(M_{2g})^4] = \frac{1}{N^4}\left[\begin{array}{c} 6(N^3-12N^2+6N)\,E[g^4]\,(E[g^2])^2 + (4N^2-4N)\,E[g^6]\,E[g^2] \\ +NE[g^8] + (N^4-6N^3+8N^2-3N)\,(E[g^2])^4 \end{array}\right]$,

and using Eq A.5, Eq A.6 and Eq A.7 for the higher moments:

$$E[(M_{2g})^4] = \upsilon^4\left(1 + \frac{12}{N} + \frac{32}{N^2} + \frac{60}{N^3}\right) . \qquad \text{(E A.31)}$$

The second term in Eq A.30 is given (from Eq A.11) by:

$$(E[(M_{2g})^2])^2 = \upsilon^4\left(1+\frac{4}{N}+\frac{4}{N^2}\right). \qquad \text{(E A.32)}$$

Substituting Eq A.32 and Eq A.31 into Eq A.30 yields:

$$Var[(M_{2g})^2] = \frac{\upsilon^4}{N}\left(8+\frac{28}{N}+\frac{60}{N^2}\right). \qquad \text{(E A.33)}$$

The last term in Eq 8.42 is given by:

$$Cov[M_{4g}(M_{2g})^2] = E[M_{4g}(M_{2g})^2] - E[M_{4g}]E[(M_{2g})^2] \qquad \text{(E A.34)}$$

where the first term may be evaluated as follows:

$$E[M_{4g}(M_{2g})^2] = \frac{1}{N^3}E\left[\left(\sum_{i=0}^{N-1} g_i^4\right)\left(\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} g_j^2 \cdot g_k^2\right)\right] = \frac{1}{N^3}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} E[g_i^4 \cdot g_j^2 \cdot g_k^2].$$

In the above sum, there are:

$[N\cdot(N-1)]$ *terms where* $(i=j\neq k)$
$[N\cdot(N-1)]$ *terms where* $(i=k\neq j)$
$[N\cdot(N-1)]$ *terms where* $(j=k\neq i)$
N *terms where* $(i=j=k)$
$[N^3-3N^2+2N]$ *terms where* $(i\neq j\neq k)$

Therefore, $E[M_{4g}(M_{2g})^2] = \frac{1}{N^3}\left[\begin{array}{l}2(N^2-N)E[g^6]E[g^2]+(N^2-N)(E[g^4])^2\\ +NE[g^8]+(N^3-3N^2+2N)E[g^4](E[g^2])^2\end{array}\right]$

$$E[M_{4g}(M_{2g})^2] = 3\upsilon^4\{1+\frac{10}{N}+\frac{24}{N^2}\}. \qquad \text{(E A.35)}$$

Substituting Eq A.35, Eq A.9 and Eq A.11 into Eq A.34 yields:

$$Cov[M_{4g}(M_{2g})^2] = 3\upsilon^4\left(\frac{8}{N}+\frac{24}{N^2}\right). \qquad \text{(E A.36)}$$

Finally substituting Eq A.36, Eq A.33 and Eq A.29 into Eq A.28 yields:

$$Var[\hat{KU}_U] = \left(1+\frac{2}{N}\right)^2\left(\frac{96\upsilon^4}{N}\right)+9\frac{\upsilon^4}{N}\{8+\frac{28}{N}+\frac{60}{N^2}\}+18\upsilon^4\left(1+\frac{2}{N}\right)\left(\frac{8}{N}+\frac{24}{N^2}\right)$$

and after rearranging terms, $Var[\hat{KU}_U] = \frac{3\upsilon^4}{N}\left[104+\frac{452}{N}+\frac{596}{N^2}\right]$.

References

[Ami91] M. Amin, "A Frequency-domain LMS Comb Filter", *IEEE Trans. on Circuits and Systems*, Vol 38, No 12, Dec 1991, pp. 1573 - 1576.

[Ars95] L. Arslan, "New Methods for Adaptive Noise Suppression", *Proc. ICASSP* 1995, pp. 812 - 815.

[Asc96] G. Ascia, "A Reconfigurable Parallel Architecture for a Fuzzy Processor", *Information Sciences*, Vol. 88, Jan 1996, pp. 299 - 315.

[Bei79] M. Beirouti, "Enhancement of Speech Corrupted by Acoustic Noise", *Proc. ICASSP* 1979, pp. 208 - 211.

[Ben97] A. Benyassine, E. Shlomot, H. Su, "ITU-T Recommendation G.729, Annex B, A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications", *IEEE Communications Magazine*, pp. 64-72, Sept. 1997.

[Bol79] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No 2, April 1979, pp. 113-120.

[Bril94] D. Brillinger, "Some Basic Aspects and Uses of Higher-order Spectra", *Signal Processing*, Vol. 36, No 3, April 1994, pp. 239 - 250.

[Cap94] O. Cappe, "Elimination of the Musical Noise Phenomena with the Ephraim and Malah Noise Suppressor", *IEEE trans. on Speech and Audio Processing*, Vol. 2, No. 2, April 1994, pp. 345 - 349.

[Che91] Y. Cheng, D.O'Shaughnessy, "Speech Enhancement Based Conceptually on Auditory Evidence", *IEEE trans. on Signal Processing*, Vol. 39, No. 9, Sept. 1991, pp. 1943-1954.

[Che95] J. Cheng, "A Reconfigurable High-speed Optoelectronic Interconnect Technology for Multi-processor Computers", *SPIE*, Vol 2481, 1995, pp. 2 - 12.

[Cox81] R. Cox and D. Malah, "A Technique for Perceptually Reducing Periodically Structured Noise in Speech", *Proc ICASSP* 1981, pp. 1089 - 1092.

[Dem92] E. Demeulemeester and W. Herroelen, "A Branch-and-Bound procedure For the Multiple Resource-Constrained Project Scheduling Problem", *Management Science* Vol 38, No. 12 Dec 1992, pp. 1803- 1818.

[Don95] D. Donoho and I. Johnstone, "Adapting to Unknown Smoothness via Wavelet Shrinkage", *Jrnl Acoustical Society of America*, Vol. 90, 1995, pp. 1200-1223.

[Dre91] A. Drexl, "Scheduling of Project Networks by Job Assignment", *Management Science*. Vol 37, No. 12, Dec 1991, pp. 1590-1602.

[Eph84] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", *IEEE Trans. on Speech and Audio Processing*, Vol. ASSP-32, No 6, Dec. 1984, pp. 1109- 1121.

[Fac94] J. Fackrell and S. McLaughlin, "The Higher-order Statistics of Speech Signals", *IEE Colloquium on Techniques for Speech Signal Processing and their Applications*, June 1994. pp. 7.1 - 7.6.

[Fac96] J. Fackrell, "Bispectral Analysis of Speech Signals", *Ph.D. Thesis*, University of Edinburgh, 1996.

[Fal93] A. Falaschi and I. Tidei, "Speech Innovation Characterization by Higher-Order Moments", *Visual Representation of Speech Signals*. Martin Cooke, Steve Beet, and Malcolm Crawford (eds.), 1993 by John Wiley & Sons Ltd.

[Fre89] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The Voice Activity Detector for the Pan European Digital Cellular Mobile Telephone Service", *Proc. ICASSP*, May 1989, pp. 369-372.

[Ful93] R. Fulchiero and A. Spanias, "Speech Enhancement Using the Bispectrum", *Proc ICASSP* 1993, vol 4, pp. 488 - 491.

[Gab88] G. Gabor and Z. Gyorfi, "On the Higher Order Distributions of Speech Signals", *IEEE Trans. on ASSP*. Vol. 36, No. 4, April 1988, pp. 602 - 603.

[Gan97] S. Gannot, D. Burshtein, E. Weinstein, "Iterative-batch and Sequential Algorithms for Single Microphone Speech Enhancement", *Proc. ICASSP* 1997, pp. 1215 - 1218.

[Goo89] G. Goossens, "Loop Optimization in Register-Transfer Scheduling for DSP Systems", *Proc. 26th DAC*, June 1989, pp. 826 - 831.

[Gra93] Joseph Graf, "Dynamic Time Warping Comb Filter for the Enhancement of Speech Degraded by White Gaussian Noise", *Proc. ICASSP*, 1993, pp. 339 - 342.

[Hai93] J. A. Haigh and J.S. Mason, "Robust Voice Activity Detection Using Cepstral Features", *IEEE TENCON* 1993, pp. 321-324.

[Hir95] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition", *Proc. ICASSP* 1995, pp. 153 - 156.

[Hoe97] R. Hoeldrich and M. Lorber, "Broadband Noise Reduction Based on Spectral Subtraction", *Proc. ICSPAT* 1997, pp. 265 - 269.

[Hoy94] J. D. Hoyt and H. Wechsler, "Detection of Human Speech in Structured Noise" in *Proc. ICASSP*, May 1994, pp. 237-240.

[Hwa91] C.T. Hwang, "A Formal Approach to the Scheduling Problem in High Level Synthesis", *IEEE trans. CAD*, vol. 10, April 1991, pp. 464 - 475

[Koi92] R. Koilpillai and P. Vaidyanathan, "Cosine-Modulated Filter Banks Satisfying Perfect Reconstruction", *IEEE Trans. Signal Processing*, Vol. 40, no. 4, April 1992, pp. 770 - 783.

[Koo89] B. Koo, J. Gibson, and S. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding", *Proc. ICASSP* 1989, pp. 349 - 352.

[Lee89] J. H. Lee, "A New Integer Linear Programming Formulation for the Scheduling Problem in Data Path Synthesis", *Digest of technical papers for the Int Conf. of Computer Aided Design* 1989, pp. 20-23.

[Leo89] A. Leon-Garcia. *Probability and Random Process for Electrical Engineering*. Addison-Wesley 1989. Chap 7, section 4, pp. 405 - 414.

[Lim78] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-26, No. 4, Aug. 1978, pp. 354 - 358.

[Mal82] D. Malah and R. Cox, "A Generalized Comb Filtering Technique for Speech Enhancement", *Proc ICASSP 1982*, pp. 160 - 163.

[Mal99] D. Malah, R. Cox, A. Accardi, "Tracking Speech-Presence Uncertainty To Improve Speech Enhancement in non-Stationary Noise Environments", *Proc. ICASSP* 1999, pp. 789-792.

[Mal192] S. Mallat and W. Hwang, "Singularity Detection and Processing with Wavelets", *IEEE trans. on Info. Theory*. 1992, Vol. IT-38, pp. 617-643.

[Mar90] J. Marron, P. Sanchez, R. Sullivan, "Unwrapping Algorithm for Least-Squares Phase Recovery From the Modulo-2π Bispectrum Phase", in *Jrnl Opt. Soc. of Ame. A*, Vol. 7, No. 1, Jan. 1990, p 14.

[Mar72] J. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", *IEEE Trans. on Audio and Electroacoustic*, Vol AU-20, No. 5, Dec 72., pp. 367 - 377.

[Mar93] R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech signals", *Eurospeech* 1993, pp. 1093 - 1096.

[Mas91] G. Masera *et al*, "A Microprogrammable Parallel Architecture for DSP", *ICCS* June 1991, pp. 824 - 827.

[Mat84] T. Matsuoka and T. Ulrych, "Phase Estimation Using the Bispectrum", *Proc IEEE*, Vol. 72, Oct 1984, pp. 1403 - 1411.

[McA80] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-decision Noise Suppression Filter", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No.2, April 1980, pp. 137 - 145.

[McA86] R. McAulay and T. Quatieri, "Speech Analysis/synthesis Based on a Sinusoidal Representation", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No.4, Aug 1986, p 744 - 754.

[Men91] J. Mendel, "Tutorial on Higher-order Statistics in Signal Processing and System Theory: Theoretical Results and Some Applications", *Proc IEEE*, Vol. 79, No. 3, March 1991, pp. 278 - 305.

[Mol86] D. Moldovan and J. Fortes, "Partioning and Mapping Algorithms into Fixed Size Systolic Arrays" *IEEE Trans. on Computers*, Vol 35, No. 1, Jan 1986, pp. 1 - 12.

[Moo81] B. Moore and B. Glasberg, "Auditory Filter Shapes Derived in Simultaneous and Forward Masking". *Jrnl Acoustical Society of America.*, Vol. 70 No. 4, Oct. 1981, pp. 1003 - 1014.

[Moo83] B. Moore and B. Glasberg, "Suggested Formulae for Calculating Auditory-filter Bandwidths and Excitation Patterns", *Jrnl Acoustical Society of America.*, Vol. 74, No. 3, Sep. 1983, pp. 750- 753.

[Mor92] A. Moreno and J. Fonollosa, "Pitch Determination of Noisy Speech Using Higher Order Statistics", *Proc ICASSP 1992*, vol 1, pp. 133- 136.

[Nik87] C. Nikias and M. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework", *Proc. IEEE*, vol. 75, No. 7, July 1987, pp. 869-891.

[Nik93] C. Nikias and J. Mendel, "Signal Processing with Higher-Order Statistics", *IEEE Signal Processing*, July 1993, pp. 10 - 38.

[O'Sh87] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley. 1987.

[O'Sh89] D. O'Shaughnessy, "Enhancing Speech Degraded by Additive Noise or Interfering Speakers", *IEEE Communications Magazine*, Feb 1989, pp. 46-52.

[Pal87] K. Paliwal & A. Basu, "A Speech Enhancement Method Based on Kalman Filtering", *Proc. ICASSP* 1987, pp. 177 - 180.

[Pal91] K. Paliwal and M. Sondhi, "Recognition of Noisy Speech Using Cumulant-based Linear prediction Analysis", *Proc. ICASSP 1991*, pp. 429 - 432.

[Pau89] Paulin, P. and J. Knight, "Force Directed Scheduling for the Behavioral Synthesis of AISCs", *IEEE trans. CAD*, Vol. 8, no. 6, June 1989, pp. 661 - 679.

[Qua86] T. Quatieri and R. McAulay, "Speech Transformations Based on a Sinusoidal Representation", *IEEE Trans on Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 6, Dec 1986. pp. 1449-1464.

[Rab77] L. Rabiner and M.R. Sambur, "Voiced-Unvoiced-Silence Detection Using the Itakura LPC Distance Measure", *Proc. ICASSP* 1977 pp. 323-326.

[Rab78] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ, Prentice-Hall, 1978.

[Ram79] V. Ramamoorty and T. Ericson, "Speech Coding Based on a Composite-Gaussian Source Model", *Proc. ICASSP* 1979, pp. 534 - 537.

[Ram80] V. Ramamoorty, "Voice/Unvoice Detection Based on a Composite-Gaussian Source Model", *Proc. ICASSP* 1980, pp. 57 - 60.

[Ran95] M. Rangoussi and G. Carayannis, "Higher-order Statistics Based Gaussianity Test Applied to On-line Speech Processing", *Asilomar Conf. Record*, 1995, p 303.

[Rui95] D. Ruiz, M. Carrion, A. Gallego and A. Medouri, "Parameter Estimation of Exponentially Damped Sinusoids Using a Higher Correlation-based Method", *IEEE Trans. on ASSP*, Vol. 43, No. 11, Nov. 95, pp. 2665- 2667.

[Sal94] J. Salavedra, E. Masgrau, A. Moreno, J. Estarellas, X. Jove, "Robust Coefficients of a Higher Order AR Modelling in a Speech Enahncement System Using Parametrized Wiener Filtering", *Proc. ICASSP* 1994, pp. 69 - 72.

[Sca96] P. Scalart and J. Vieira-Filho, "Speech Enhancement Based on Apriori Signal to Noise Estimation", *Proc. ICASSP* 1996, pp. 629 - 632.

[Sch97] I. Schick and H. Krim, "Robust Wavelet Thresholding for Noise Suppression", *Proc. ICASSP* 1997, pp. 3421 - 3424.

[Sec83] B. Secrest and G. Doddington, "An Integrated Pitch Tracking Algorithm for Speech Systems", *Proc. ICASSP* 1983, pp. 1352-1355.

[See88] S. Seetharaman and M. Jernigan, "Speech Signal Reconstruction Based on Higher Order Spectra", *Proc ICASSP* 1988, pp. 703-706.

[Sor97] P. Sorqvist, P. Handel, B. Ottersten, "Kalman Filtering for Low Distortion Speech Enhancement in Mobile Communication", Proc. *ICASSP* 1997, pp. 1219 - 1222.

[Sti78] Stinson, "Multiple Resource-Constrained Scheduling Using Branch-and-Bound", *AIIE Transactions*, Vol 10, No. 3, 1978, pp. 252-259.

[Sun90] G. Sundaramoorthy, M. Raghuveer, and S. Dianat, "Bispectral Reconstruction of Signals and Noise. Amplitude Reconstruction Issues", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 38, No 7, July 1990, pp. 1297 - 1306.

[Swa91] A. Swami and J. Mendel, "Cumulant-based Approach to the Harmonic Retrieval and Related Problems", *IEEE Trans. on ASSP*, Vol. 39, No. 5, May 1991, pp 1099 - 1109.

[Tea83] H. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract" in R.G. Daniloff, Ed., *Speech Sciences: Recent Advances*. San Diego, CA. College-Hall press. pp. 73-109, 1983.

[Tea90] H. Teager and S. Teager, "Evidence for Non-linear Sound Production in the Vocal Tract" in *Speech Production and Speech Modelling* (W. J. Hardcastle and A. Marchal, eds.), pp. 241-261, Kluwer Academic Publishers, 1990.

[Tso93] D. Tsoukalas, M. Paraskevas, J. Mourjopoulos, "Speech Enhancement Using Psychoacoustic Criteria", *Proc. ICASSP* 1993, pp. 359-362.

[Tuc92] R. Tucker, "Voice Activity Detection Using a Periodicity Measure", *IEE Proceeding-I*, Vol. 139, No. 4, pp. 377-380, August 1992

[Van68] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley 1968, pp. 54 - 56, 198 - 206.

[Var85] P. Vary, "Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits", *Signal Processing*, Vol.8 No. 4, July 1985, pp 387 - 400.

[Vee89] D. Veeneman and B. Mazor, "A Fully Adaptive Comb Filter for Enhancing Block-coded Speech", *IEEE trans ASSP*, Jun 1989, pp. 955 - 957.

[Vir95] N. Virag, "Speech enhancement based on masking properties of the auditory system", *Proc. ICASSP* 1995 pp. 796 - 799.

[Wan93] F. Wang, P. Kabal, and D. O'Shaughnessy, "Frequency Domain Adaptive Postfiltering for Enhancement of Noisy Speech", *Speech Communication*, Vol. 12, No 1, March 1993, pp. 41- 56.

[Wel85] B. Wells, "Voiced/Unvoiced Decision Based on the Bispectrum", *Proc ICASSP 1985*, March 1985, pp. 1589 - 1592.

[Wis76] J. Wise, J. Carpiro, and T. Parks, "Maximum Likelihood Pitch Estimation", *IEEE Trans. on ASSP*, vol. 24, Oct. 1976, pp. 418 - 423.

[Yan93] J. Yang, "Frequency-Domain Noise Suppression Approaches in Mobile Telephone Systems", *Proc. ICASSP* 1993, pp. 363-366.

[PN-3292] TIA Document, PN-3292, "Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems", Jan. 1996.

[IS127] TIA/EIA, "Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems", Interim Standard, Jan 1997.

Publications

- Note: All papers by Elias Nemer, Rafik Goubran and Samy Mahmoud.

Published

[1] "SNR Estimation of Speech Signals Using Subbands and Fourth Order Statistics", *IEEE Signal Processing Letters, Vol. 6, No. 7,* July 1999, pp. 171-174.

[2] "The Third-order Statistics of Speech Signals with Application to Reliable Pitch Estimation", *IEEE Statistical Signal and Array Processing workshop,* Sept. 1998, pp. 427-430.

[3] "Noise Estimation Using Higher Order Statistics for Improved Speech Enhancement", *Proc. IASTED Signal and Image Processing,* 1998, pp. 539-543.

[4] "Pitch Estimation and Voicing Detection Using Third-Order Statistics", *Proc. ICSPAT* 1997, pp. 1668 - 1673.

[5] "An Efficient Algorithm for Computing the Triple correlation", *Proc. CCECE,* May 1997, pp. 215 - 218.

[6] "A Non-deterministic Scheduling and Allocation Model for Mapping Algorithms on Configurable Architectures", *Proc. CCECE,* 1997, pp. 19 - 22.

[7] "Speech Enhancement Using Higher Order Statistics", *Proc. ICSPAT,* 1996, pp. 1645 - 1650.

Accepted for Publication

[8] "The 4^{th}-Order Cumulant of Speech Signals with Application to Voice Activity Detection", *EUROSPEECH* 1999.

[9] "Speech Enhancement Using Higher Order Cumulants and Time-domain Optimal Filters", *EUROSPEECH* 1999.

In Review

[10] "Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain", *IEEE Transactions on Speech and Audio Processing.*

[11] "Speech Enhancement Using Fourth-Order Cumulants and Optimal Filters in the Subband Domain", *Speech Communication.*