LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Perceptually Motivated Audio Time-Frequency Analysis

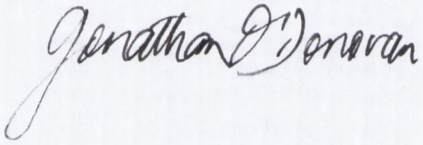
Jonathan O'Donovan,
Department of Electronic and Electrical Engineering,
Trinity College Dublin, Ireland.

Thesis submitted for the degree of Ph.D.
October, 2000

TRINITY COLLEGE
0 2NOV 2001
LIBRARY DUBLIN
THESIS
6482

Declaration

I, Jonathan James Patrick O'Donovan, declare that this thesis is my own work, and that it has not been submitted for a degree at any other university or institution. I agree that the Library may lend or copy this thesis upon request.

Signed : 

Date : 13/9/2001

Summary

Fourier analysis has long been an indispensable tool for the investigation of signals whose frequency content does not change with time. It has likewise been of fundamental importance in engineering for the study of linear time-invariant systems. Clearly, however, most natural and man-made signals *do* have a time-varying frequency content. Fourier analysis is unfortunately inadequate for the analysis of such signals, as it cannot tell us how the signal's frequency content changes with time—it can determine which frequency components are present in the signal, but not when they occur. In order to overcome the limitations of Fourier analysis, therefore, a *joint time-frequency* signal description is required.

The most widely adopted time-frequency signal analysis technique is the spectrogram, developed over 50 years ago for the analysis of speech. In recent years, however, there has been much research into more complex time-frequency analysis techniques for a wide variety of applications including acoustics, speech and music analysis, machine monitoring, biomedical signal analysis, radar and economic data analysis. The ever increasing computing power of the personal computer and more widespread time-frequency analysis software have also facilitated an increased awareness of the potential advantages of time-frequency analysis techniques.

Audio time-frequency analysis has for many years provided researchers with a mathematical approach to gaining a better understanding of the physical characteristics of acoustic signals such as speech or music. The field of *psychoacoustics* on the other hand, is concerned with understanding the workings of the human auditory system and how the auditory sensations we perceive relate to such acoustic signals. With the advent of telecommunications, perceptually-based audio compression schemes, hearing-aid technology and digital audio and digital signal processing in general, there has been a recent surge of interest in auditory modelling. While our knowledge of the human auditory system is not yet complete, much of its behaviour is now well understood.

One of the most important features of the auditory system is that of *auditory masking*, which refers to a situation where a relatively weak signal component is rendered inaudible (masked) by a relatively strong signal component if the two are either temporally or spectrally close. Masking is divided into *temporal* masking and *spectral* masking; both are related to the ear's limited temporal and spectral resolutions respectively. A large body of

work in the psychoacoustic literature has been devoted to modelling the auditory resolutions of the ear. Despite this work and the fact that auditory masking has a crucial effect on our perception of sound, there has been relatively little attempt to incorporate auditory resolutions in joint time-frequency signal analysis techniques. Some attempts using linear time-frequency techniques such as the spectrogram or wavelet transform have been made. However, these have generally been designed with perceptual compression schemes in mind and do not strive to accurately achieve the auditory resolutions specified in the psychoacoustic literature. Also, linear time-frequency techniques are inherently limited by the fact that independent control of time and frequency resolution is not possible.

This thesis is primarily concerned with the design and validation of a new *bilinear* time frequency distribution, which simultaneously (and independently) incorporates the temporal and spectral resolutions of the auditory system. The distribution, dubbed the *'EarWig'* distribution (EWD), is based on the smoothed Wigner distribution and is a member of *Cohen's class* of bilinear time-frequency distributions. Cohen class distributions are well suited to the task of modelling auditory temporal and spectral resolution as they allow independent control of distribution time and frequency resolutions. To the best of the author's knowledge, however, no previous successful attempt has been made at modelling auditory resolution using a member of Cohen's class. This is perhaps due to the difficulties normally associated with this class which include excessive memory requirements and high computational complexity. Furthermore, the presence of cross-term interference and clutter can confuse the interpretation and limit the useful dynamic range of members of the class. This thesis, however, describes how these difficulties may all be overcome to achieve a joint time-frequency distribution which very accurately incorporates the time and frequency resolutions specified in the psychoacoustic literature. The EWD does not suffer from cross-term interference and has a very wide dynamic range. This thesis also describes the application of the EWD to two important areas of audio research, namely 3-D surround sound design and artificial reverberation analysis. These application areas not only highlight the benefit and utility of incorporating auditory resolutions in audio time-frequency analysis but also the elegance of the mathematical signal-analytic framework offered by Cohen's class.

Acknowledgements

First of all, I would like to express my heartfelt gratitude to Dr. Dermot Furlong for providing me with the opportunity of conducting research in the fascinating areas of audio, acoustics, signal processing and psychoacoustics. I have been very fortunate indeed to be able to thus combine interests in both engineering and music. The years spent as a post-graduate student at Trinity College Dublin have been a wonderful learning experience for me both academically and personally. Dr. Furlong's friendship and supervision, constant good humour and enthusiasm for all things audio-related has been an inspiration to me and has ensured that this research has always been both challenging and rewarding.

Much thanks are due to the staff and students of the Department of Electronic and Electrical Engineering in Trinity College for creating such a pleasant working environment. In particular, I would like to thank Professor Boland and Dr. Foley for their support and interest, Dr. Liam Dowling and Dr. Anthony Quinn for their signal processing teaching and Dr. Anil Kokaram for his much appreciated help with LaTeX. I also would like to thank Conor Nolan and Robbie Dempsie for all manner of technical advice and Nora Moore for her help with colour printing. Finally, many thanks are due to the students of the Department for their friendship and good laughs and I apologise for my reclusive behaviour during these last few months !

The provision of a studentship from Silicon Software Systems (S3) was crucial to allowing me to pursue this work. The support of S3 is greatly appreciated and I thank them wholeheartedly for their generosity and patience. I would also like to express my gratitude to the many researchers who, in the true spirit of scientific learning, freely answered queries of mine relating to their work and shared with me their knowledge and expertise. Without this help, much of this work would not have been possible.

Finally and most importantly, no amount of words can sufficiently express the gratitude I feel for the unconditional love and support provided to me by my parents Patrick O'Donovan and Gobnait Scally. If anything, this thesis really represents the fruits of their labours, not mine. I also thank my brothers Rory and Shane and my sister Deirdre for their love, companionship, thoughtfulness and musical abilities !

To all of you, thank you ! May all your efforts be rewarded with every happiness and success and each day of your lives be blessed !

Contents

1 Introduction **14**

1.1 Thesis Outline 18

2 Joint Time-Frequency Analysis **21**

2.1 The Spectrogram 22

2.1.1 Spectrogram Time-Frequency Resolution 22

2.2 The Wigner Distribution 24

2.2.1 Wigner Distribution Time-Frequency Resolution 25

2.2.2 Wigner Distribution Properties 27

2.2.3 The Cross-Wigner Distribution 30

2.2.4 The Pseudo-Wigner Distribution 33

2.2.5 Inter-relationships of the WDF, TCF, CAF and SCF 35

2.3 Cohen's Class 36

2.3.1 Separable Smoothing 38

3 The Auditory System and Masking **41**

3.1 Overview of the Auditory System 42

3.2 Spectral Masking 45

3.2.1 The Power Spectrum Model 46

3.2.2 The Notched-Noise Method 49

3.2.3 The Critical-Band Rate Scale and ERB Scale 53

3.2.4 Auditory Filter Shape 55

3.2.5 Excitation Patterns 67

3.3 Temporal Masking 69

3.3.1 Temporal Resolution 70

3.3.2 Measuring Temporal Resolution 70

3.3.3 The Temporal Window Model 72

3.4 Review of Perceptually Motivated TF Analysis 74
3.4.1 Wavelet Based Models 75
3.4.2 Spectrogram Based Models 77
3.4.3 Auditory Modelling using the Smoothed Wigner Distribution 78

4 EarWig Distribution Design **80**
4.1 The Discrete Wigner Distribution 80
4.1.1 Temporal Correlation Function 82
4.1.2 Implementation Issues 85
4.1.3 Discrete Pseudo-Wigner Distribution 87
4.1.4 The Discrete Smoothed WD 91
4.1.5 Frequency Dependent Kernel Design 94
4.2 The EarWig Distribution 96
4.2.1 Frequency Smoothing Window Design 97
4.2.2 Excitation Pattern Generation 99
4.2.3 Temporal Smoothing Window Design 103
4.3 Modelling the Outer-Middle Ear 108
4.4 Computational Cost of the EWD 112
4.4.1 Memory Requirements for EWD Calculation 115
4.4.2 Real Time EWD Analysis 116
4.4.3 Comparison of EWD and Filterbank-Based Auditory Models 117
4.5 EWD Computation Platform 119

5 Model Validation **120**
5.1 EWD Properties 120
5.2 Spectral and Temporal Masking : Model Accuracy 121
5.2.1 Spectral Resolution Accuracy 122
5.2.2 Temporal Resolution Accuracy 135
5.2.3 Comparison of EWD and Spectrogram 138
5.3 Cross-Term Interference 143
5.3.1 Frequency Response of the Temporal Window 145
5.4 Effective Dynamic Range 146
5.5 Psychoacoustic Limitations of the EarWig Distribution 154
5.5.1 Spectral Resolution 157
5.5.2 Temporal Resolution 157

6 HRTF Smoothing **161**

6.1 Auditory Localisation and Localisation Cues 163

6.1.1 Head Related Transfer Functions 165

6.2 Obtaining a Generic HRTF Set 166

6.2.1 PCA HRTF Smoothing 170

6.3 EWD Derived HRTF Smoothing 177

6.3.1 Empirical Comparison of PCA and EWD Smoothing 183

6.3.2 Subjective Comparison of PCA and EWD Smoothing 190

6.3.3 Comparison of 'Raw' and EWD Averages 197

7 Reverberation Analysis **207**

7.1 Natural Reverberation 208

7.2 Reverberation Algorithms 210

7.2.1 Simulating Early Reverberation 212

7.2.2 Simulating Late Reverberation 214

7.2.3 Late Reverberation Algorithms 215

7.3 EWD Analysis of Reverberation 220

7.3.1 Vibraphone Analysis 229

7.3.2 Spectral Analysis of Partial Amplitude Modulation 237

7.3.3 Analysis of the Effects of Comb Filter Delays 240

7.4 Spectrogram Reverberation Analysis 248

8 Summary and Conclusions **252**

8.1 Speaker, Mic. and Equalisation Impulse Response Analysis 257

8.2 Musical Instrument Analysis and Synthesis 258

8.3 Automatic Artificial Reverberator Design 259

8.4 Virtual Auditory Display Robustness 260

8.5 Speech Processing and Auditory Scene Analysis 260

8.6 A Final Note 261

List of Figures

2.1 Waveform, energy spectral density and spectrogram of a chirp signal. 23

2.2 Waveform, spectrum and WD of a chirp signal. 26

2.3 Waveform, energy spectral density and Wigner distribution of a signal consisting of three Gaussian atoms. 32

2.4 Waveform, energy spectral density and pseudo-Wigner distribution of a signal consisting of three Gaussian atoms. 34

2.5 Inter-relationship between the WDF, CAF, TCF and SCF domains for the WD and smoothing kernalEach arrow represents a (forward) Fourier transform. 36

2.6 WDF (a), PWDF (b), SPWDF (c) and spectrogram (d) of a signal consisting of three Gaussian atoms. 39

3.1 Outer, middle and inner ear structure [108]. 42

3.2 Cross section of the cochlear duct [176]. 43

3.3 Displacement of the basilar membrane at an instant in time [108]. 44

3.4 The threshold of a 2 kHz sinusoidal signal plotted as a function of the bandwidth of a noise masker centred at 2 kHz [108]. 47

3.5 Power spectra of masking noise, signal and rectangular filter used by Fletcher to estimate the critical bandwidth CB 48

3.6 Illustration of Patterson's notched-noise experiment. The vertical line at centre frequency f is the signal. The notch is placed symmetrically around the signal; the upper and lower noise edges are separated from the signal by frequency deviation Δf. The triangular region represents the auditory filter shape and the shaded areas represent the noise passing through this filter. . 50

3.7 Illustration of a typical auditory filter shape derived using Patterson's notched-noise method. The centre frequency of the filter is 1 kHz [108]. 51

3.8 A summary of ERB measurements made by several researchers—the solid curve plots a fit to this data [108]. The dashed line plots Zwicker's [193] earlier measurements of the critical bandwidth, which is seen to flatten for frequencies below 500 Hz. 52
3.9 Representation of masker, tone and hypothetical auditory filter shape in Patterson's masking experiment [129] to determine auditory filter shape. 57
3.10 Tone plus notched-noise masker. (a) Tone plus lowpass noise (b) Use of off-frequency listening to improve the signal to noise ratio (c) Upper noise band used to limit off-frequency listening [130]. 59
3.11 The bold and faint solid curves plot the the λ-asymmetric and symmetric rounded-exponential filters, respectively [135]. Also shown are the average, fixed symmetric filters from [130] (bold dashed curves) and [129] (faint dashed curves). The filters are plotted on both linear and logarithmic ordinates. . . 62
3.12 (a) The individual excitation patterns of first 10 harmonics of a 300 Hz fundamental. (b) The total excitation pattern. 68
3.13 Non-simultaneous masking. 70
3.14 The 12 temporal window shapes derived from the average data for the four subjects [151]. The abscissa denotes time in ms relative to the centre of the window shape and the ordinate denotes the relative weighting in dB. 73

4.1 Real signal $s[n]$ with $L = 4$. 83
4.2 Calculation of $g[n, l]$ at $n = -3$, $n = -2$ and $n = -1$ for a real signal, $s[n]$, with $L = 4$. 84
4.3 Diamond-shaped TCF region $g[n, l]$ with $L = 4$. 85
4.4 Rotation, re-numbering and zeropadding of TCF, $g[n, l]$, with $L = 4$. 86
4.5 TCF cropping for DPWD. 88
4.6 DWD of a chirp signal. 89
4.7 DPWD of a chirp signal. 89
4.8 Energy spectral density of a two sinusoid signal windowed by a hanning window for the FFT, (a), and PWDF (b). Energy spectral density of the same two sinusoid signal windowed by a gamma window for the FFT, (c), and PWDF, (d). 92
4.9 Operations required to calculate a SPWDF frequency slice : (A) : TCF block calculation, (B) : TCF time smoothing (C) : TCF frequency smoothing and frequency transformation. 93

4.10 Operations required to calculate a SPWDF frequency slice with frequency dependent time and frequency resolution : (A) : TCF block calculation, (B) : TCF time smoothing (C) : TCF frequency smoothing. 95
4.11 Waveforms (left) and frequency responses (right) for the gammatone filter (top panels (a) and (b)), baseband gammatone filter (middle panels (c) and (d)) and gamma window (bottom panels (e) and (f)). 100
4.12 Upper panel : Intersection of sinusoid spectrum and roex filter shapes giving the excitation for 5 nominal frequencies A, B, C, D and E. Lower panel: The resulting overall excitation pattern. 101
4.13 A comparison of the calculation of one excitation level for the EWD, (a) and method of Glasberg and Moore, (b). The filter shape is symmetric. The two methods are shown to give the same result *i.e.* $b = a$. 102
4.14 A comparison of the calculation of one excitation level for the EWD, (a) and method of Glasberg and Moore, (b). The filter shape is asymmetric, having a shallower lower slope than upper slope. The two methods are shown to give the same result *i.e.* $b = a$. 104
4.15 Interpolated temporal window shape parameters for the design of a temporal window of arbitrary frequency : (a) total peak duration, T_p, (b) total skirt duration, T_s, (c) skirt level, w and (d) window asymmetry, a. 105
4.16 (a) Plan view of the frequency dependent temporal window shape at 20 centre frequencies logarithmically spaced between 100 Hz and 10 kHz. (b) The seven temporal windows used for the EWD smoothing kernel. 107
4.17 Flowchart detailing the overall EWD algorithm. 109
4.18 Threshold of audibility (threshold in quiet) curve [194]. 110
4.19 Outer-middle ear filter response. The values of R, from bottom to top are : 0.95, 0.96, 0.97, 0.98 and 0.99. 111
4.20 Normalised TCF block size for temporal window dynamic ranges of 10 to 100 dB (increasing in steps of 10 dB from bottom to top). The abscissa denotes the minimum frequency analysed. 114

5.1 (a) EWD of a 150 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the symmetric roex(p) shape (red curve) and the EWD (blue curve). $MSE = 3.71.10^{-4}, PED = 2.40.10^{-3}$. 126
5.2 (a) EWD of a 1200 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the roex(p) shape (red curve) and the EWD (blue curve). $MSE = 2.89.10^{-4}, PED = 3.11.10^{-3}$. 127

5.3 (a) EWD of a 6000 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the roex(p) shape (red curve) and the EWD (blue curve). $MSE = 4.04.10^{-4}, PED = 3.32.10^{-3}$. 128
5.4 (a) EWD of a complex tone consisting of the first 10 harmonics of a 200 Hz fundamental. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) filter shape (red curve) and the EWD (blue curve). $MSE = 7.51.10^{-3}, PED = 1.52.10^{-2}$. 130
5.5 Excitation patterns generated by the asymmetric roex(p) filter shape (black curves), symmetric roex(p) filter shape (red curves) and EWD (blue curves) for the first 10 harmonics of a 200 Hz fundamental for four component levels of 30, 50, 70 and 90 dB SPL. 132
5.6 (a) EWD 2 kHz signal tone plus noise masker. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) shape (red curve) and the EWD (blue curve). $MSE = 1.32.10^{-3}, PED = 1.04.10^{-2}$ . . 134
5.7 Harmonic levels of /l/ vowel sound, from fundamental at 125 Hz to 32nd harmonic at 4 kHz. (data taken from [111]). 136
5.8 (a) EWD of a /l/ vowel sound, fundamental 125 Hz. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) shape (red curve), symmetric roex(p) shape (black curve) and the EWD (blue curve) for stimulus levels of 50 and 80 dB. 137
5.9 (a) Waveform and (b) frequency response of the 20th order FIR lowpass filter used to test EWD temporal resolution. 138
5.10 (a) EWD of the 20th order FIR filter for 7 analysis frequencies, using all 7 temporal windows. (b) A comparison of the EWD output and the actual temporal window shape. The curves plotted represent temporal window 1 to 7 (top to bottom), with the graphs normalised and offset by 10 dB for clarity. 139
5.11 EWD of the first 10 harmonics of a 200 Hz fundamental. 140
5.12 Spectrogram of the first 10 harmonics of a 200 Hz fundamental, using the 1000 Hz temporal window as the analysis window. 141
5.13 Spectrogram of the first 10 harmonics of a 200 Hz fundamental, using the 1000 Hz gammatone window as the analysis window. 142
5.14 EWD mesh plot illustrating the cross-term interference between a sinusoid of 250 Hz and a linear chirp signal, sweeping from 250 Hz to 4 kHz. 144
5.15 Frequency response of all 7 temporal windows. 146

5.16 Constant-Q distribution of a vibraphone note F#4 (hard mallet attack), showing the first 40 dB of the distribution. 149
5.17 EWD of a vibraphone note F#4 (hard mallet attack), showing the first 40 dB of the distribution. 150
5.18 Constant-Q distribution of a vibraphone note F#4 (hard mallet attack). The full dynamic range of the distribution is shown. 151
5.19 EWD of a vibraphone note F#4 (hard mallet attack). The full dynamic range of the distribution is shown. 152
5.20 Constant-Q distribution of a flutter-tongued flute attack for the note C5. The first 60 dB of the distribution is shown. 155
5.21 EWD of a flutter-tongued flute attack for the note C5. The full dynamic range (140 dB) of the distribution is shown. 156

6.1 HRTF magnitude responses for subject AFW : (a) left ear and (b) right ear 167
6.2 HRTF magnitude responses for subject SOW : (a) left ear and (b) right ear 168
6.3 HRTF magnitude responses for subject SOU : (a) left ear and (b) right ear 169
6.4 (a) Raw DTF data and (b) PCA-smoothed data (1 PC) for the left ear of subject SOW. 172
6.5 PCA-smoothed data for the left ear of subject SOW. (a) 2 PCs (b) 3 PCs. . 173
6.6 PCA-smoothed data for the left ear of subject SOW. (a) 4 PCs (b) 5 PCs. . 174
6.7 PCA-smoothed data for the left ear of subject SOW. (a) 6 PCs (b) 7 PCs. . 175
6.8 A typical HRIR measurement (subject SOU). 178
6.9 EWD of the HRIR shown in Figure 6.8. The EWD is calculated at $t = 0\ ms$, $t = 3\ ms$ and $t = 6\ ms$ at 58 bins between 400-11025Hz. 179
6.10 (a) A comparison of the original HRTF response (red curve) and the EWD-smoothed HRTF response (blue curve) illustrating the high frequency bias introduced by EWD smoothing at high frequencies. (b) A comparison of the corresponding DTFs, with the original DTF (red) and the smoothed DTF (blue). (c) A comparison of the unsmoothed HRTF (red) and the smoothed DTF, with the AAM information re-inserted (blue). 180
6.11 (a) Raw DTF data and (b) EWD-smoothed data for the left ear of subject SOW. 182
6.12 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 0^o and 30^o respectively. 184

6.13 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 60^o and 90^o respectively. 185
6.14 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 120^o and 150^o respectively. 186
6.15 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 180^o and 210^o respectively. 187
6.16 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 240^o and 270^o respectively. 188
6.17 A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 300^o and 330^o respectively. 189
6.18 Virtual auditory display for a two loudspeaker setup. 191
6.19 (a) A comparison of the desired β_{2spk} filter magnitude and phase response (red curves) and the IIR fits achieved by the filter fitting process (blue curves). The green curves plot the difference. (b) The same scenario, now presented for the $[\alpha/\beta]_{2spk}$ filter. 193
6.20 Experimental setup used for comparison of PCA and EWD-smoothed HRTFs. 194
6.21 (a) Average 'raw' DTFs for the left ear of subject SOW. (b) Average EWD-smoothed DTFs for the same ear. $PED = 8.34.10^{-3}$, $MSE = 8.27.10^{-4}$ . . 198
6.22 (a) Magnitude response of the $[\alpha/\beta]_{2spk}$ filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data. 199
6.23 (a) Magnitude response of the β_{2spk} filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data. 201

6.24 (a) Magnitude response of the $[\alpha/\beta]_{2spk} \times \beta_{2spk}$ filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data. 202
6.25 A comparison of the average 'raw' (dashed curves) and EWD-smoothed (solid curves) HRTF filter specifications common to all azimuth positions : (a) $HRTF_L(-30)$, (b) $HRTF_R(30)$, (c) $HRTF_L(30)$ and (d) $HRTF_R(-30)$. . 204

7.1 An echogram of a typical room. 209
7.2 Schroeder's filter structure implementing early echoes and late reverberation [60] 213
7.3 Processing of early reflections by absorptive and localisation filters [60]. . . 214
7.4 Filter structure, impulse response, pole-zero diagram and spectrum magnitude of a single comb filter [60]. 216
7.5 Filter structure, impulse response, pole-zero diagram and spectrum magnitude of a single allpass filter [60], [168]. 216
7.6 Schroeder's reverberator [60] 217
7.7 A comparison of (a) real room response and (b) Schroeder reveberator response. 221
7.8 A comparison of EWDs of (a) real room response and (b) Schroeder reveberator response. Amplitude unit is relative excitation level (dBs.) 222
7.9 (a) Speech signal *'... changed in eighteen seventy five.'* (b) EWD of same signal. Amplitude unit is relative excitation level (dBs.) 223
7.10 (a) EWD of speech signal reverberated with (a) real room response and (b) Schroeder response. Amplitude unit is relative excitation level (dBs.) 224
7.11 Partial amplitudes of the synthetic signal. Amplitude unit is relative excitation level (dBs.) 227
7.12 A comparison of EWDs of (a) the unreverberated vibraphone signal and (b) the vibraphone signal reverberated with the 3-second, non-prime Schroeder reverb. Amplitude unit is relative excitation level (dBs.) 230
7.13 The temporal marginals for the vibraphone fundamental, f_0 (370 Hz). . . . 231
7.14 The temporal marginals for vibraphone partial, $4f_0$ (1480 Hz). 232
7.15 The temporal marginals for the vibraphone partial, $9.5f_0$ (3515 Hz). 233
7.16 Modulation spectra of the vibraphone partial at $4f_0$ (1480 Hz) for the 2 second non-prime (green), prime (blue) and real (red) reverberations. 238
7.17 Modulation spectra of the vibraphone partial at $4f_0$ (1480 Hz) for the 3 second non-prime (green), prime (blue) and real (red) reverberations. 239

7.18 Schroeder's reverberator with additional variable-gain, 'non-prime' comb filter section. 242
7.19 The temporal marginals for the fundamental of the vibraphone signal reverberated with the 2 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1. 244
7.20 The temporal marginals for the partial at $4f_0$ of the vibraphone signal reverberated with the 2 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1. 245
7.21 The temporal marginals for the fundamental of the vibraphone signal reverberated with the 3 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1. 246
7.22 The temporal marginals for the partial at $4f_0$ of the vibraphone signal reverberated with the 3 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1. 247
7.23 EWD of an unreverberated violin signal. Amplitude is in dBSPL. 249
7.24 The spectrograms of an unreverberated violin signal using (a) the temporal window shape centred at 1000 Hz and (b) the gamma window centered at 1000 Hz as the analysis window. Amplitude is in dBSPL. 250

List of Tables

2.1 Cohen's class properties and corresponding kernel constraints 38

3.1 Construction of the critical-band rate scale with z indicating the critical-band rate, f_c being centre frequency, and f_l, f_u and Δf_G being the lower frequency, upper frequency and bandwidth of each critical band. Data taken from [194]. 54

4.1 Values of averaged parameters T_p, T_s, w and a for frequency dependent temporal window model. 105
4.2 Start, centre and end frequencies for each of the 7 temporal windows. . . . 108
4.3 Values of parameters N_{Tmax}, N_{Fmax}, N_{Tav}, N_{Fav}, N_{Tmaxa} and T_{step} for a particular case of real time EWD analysis. 116
4.4 Total number of multiplications per second for TCF calculation, time smoothing and frequency smoothing/transformation for the particular example given. 117

5.1 Cohen's class properties and corresponding kernel constraints. A tick indicates that a property is satisfied by the EWD. 121
5.2 MSE and PED values for a comparison of the EWD with the symmetric and asymmetric filter models for stimuli levels of 30, 50, 70 and 90 dBSPL. . . . 131
5.3 MSE and PED values for a comparison of the EWD with the symmetric and asymmetric filter models for stimuli levels of 30, 50, 70 and 90 dBSPL. . . . 135

6.1 Horizontal and vertical localisation cues listed in order of importance. . . . 165
6.2 Percentage of variance preserved in DTF data after PCA. 176
6.3 Subjective comparison of PCA and EWD models for all orders. 'L' denotes localisation and 'N' denotes naturalness. 196

7.1 Audible degradation noted in a variety of musical signals. 'G' denotes graininess, 'F' denotes fluttering and 'M' denotes a metallic timbre. The number of each letter denotes the presence of a weak, moderate or strong degradation. 225

7.2 Details of the samples taken from the MUMS database where f_0 is the fundamental frequency of the particular note. The synthetic signal was generated in MATLAB and is included for completeness. It has a note value of approximately $G\natural 3$. 225

7.3 Integration regions for the partials at f_0, $4f_0$ and $9.5f_0$ of the vibraphone signal. f_c is the partial frequency and the integration is performed over the region f_{min} to f_{max}. 235

7.4 Feedback delays for the prime, non-prime and variable-gain, non-prime comb-filter sections of the Schroeder reverberator. Delay time is in ms. 241

Chapter 1

Introduction

Arguably one of the most important scientific instruments ever devised is the oscilloscope which was invented in 1897 by German physicist Karl Braun. This device remains to this day one of the most useful tools available to the electronic engineer and was crucial to the development of many technologies, including television. The oscilloscope also revolutionised the study of musical sound. Prior to its invention, it was impossible to visualise the rapid acoustic vibrations produced by musical instruments. However, in combination with a microphone, the oscilloscope provided a method of examining the waveforms generated by musical instruments. This allowed musical acousticians to study and compare the periodic waveforms produced by the steady state portions of different musical notes. Using harmonic analysis it was thus possible to characterise to a certain degree the perception of an instrument's timbre by its harmonic spectrum.

Jean-Baptiste Fourier formulated the mathematical theory behind harmonic analysis in 1807, well before the invention of the oscilloscope. Fourier's new mathematics was originally intended to solve a problem concerning heat flow and temperature discontinuities. However, his work also provided the ground-breaking mathematical tools necessary for the harmonic or spectral analysis of signals. Fourier analysis is useful for studying the steady state acoustic waveforms produced by the oscilloscope because the frequency content of such signals can be assumed to be constant with time. Such signals are termed *stationary*. It is clear, however, that the majority of signals present in nature do change with time and are thus *non-stationary*. For example, musical tones are in fact inherently non-stationary because of the transient effects of their onset, decay and also other effects such as vibrato and tremolo. Even the execution of a simple melody on any given musical instrument produces a highly non-stationary signal. If several musicians perform simultaneously and the complexity of the music increases, our hope of gaining any useful understanding of the acoustic properties

of the resulting music using either the 'scope or spectral analysis rapidly dwindles to zero.

If we consider only the human senses of vision and hearing, we are continuously presented with a vast array of colours and sounds which change with time in a highly unpredictable manner. Fourier analysis is not appropriate for studying such signals. The spectrum of a speech signal, for example, can tell us what frequency components are present overall in the signal, but not when and for how long they occurred. Discounting the phase response, Fourier analysis cannot differentiate between the speech signal 'upsidedown' and 'downsideup'. Therefore, if we wish to gain a proper understanding of such signals and determine their physical properties, something better than mere spectral analysis is necessary. What is required is a three dimensional representation of how a signal's energy is distributed in time and frequency *i.e.* a *joint time-frequency (TF) signal representation.* Such a representation, or *time-frequency distribution*, is a far more satisfactory method of analysis than spectral analysis as it can tell us the signal frequencies present at any given time, or conversely, the time intervals (if any) over which a given frequency is present.

The practical use of time-frequency analysis began in earnest in the 1940's, with the development of the *short-time Fourier transform* or *spectrogram* for the study of speech signals. Spectrogram analysis is a simple extension of Fourier analysis which seeks to determine the frequency content of a signal at a given instant in time by performing a spectral analysis of a short section of the signal centred around this time. This signal section is extracted by means of a multiplication of the signal with a (much shorter) *window function.* However, the success of such analysis is limited by the fact that the time and frequency resolutions achieved depend on the characteristics of the window used. If the window length is short, temporal resolution is improved at the expense of frequency resolution. Conversely, if the window is long, then frequency resolution is improved at the expense of temporal resolution. The spectrogram therefore suffers from an inherent time resolution versus frequency resolution tradeoff. The time and frequency resolutions of the spectrogram cannot be independently set.

Time-frequency analysis is very useful for visualising and studying how the *physical properties* of a signal change with time. However, for the specific case of audio signal analysis, these properties tell us little about the *auditory perception* of a signal. The study of auditory perception is the domain of *psychoacoustics*, which is concerned with the workings of the human auditory system and with gaining a better understanding of the link between the physical characteristics of acoustic signals and the sensations which they produce. Psychoacoustics involves the investigation of the anatomy and physiology of the ear and the identification of the underlying mechanisms leading to auditory perception. While there is

still considerable uncertainty concerning some of the components of the auditory system, much of its behaviour is now well understood. Perhaps the most useful feature of auditory processing for psychoacoustics is *masking*. Masking refers to a situation where a stimulus (sometimes referred to as the *signal*) becomes inaudible in the presence of another, usually stronger, stimulus (sometimes referred to as the *masker*) if both stimuli are temporally or spectrally close. Masking is inherently related to the limited spectral and temporal resolutions of the auditory system—these resolutions have been the subject of much ongoing research in the psychoacoustic literature and are now well established. The *frequency resolution* or *spectral resolution* of the human auditory system is markedly frequency dependent, broadening with increasing frequency, and also depends on a number of other factors including stimuli level. The ear's spectral resolution is caused by the frequency selectivity of the basilar membrane and the neurons situated along its length. Two signal components close in frequency may not be resolvable by the ear, resulting in (spectral) masking of the weaker component. Typically, the spectral resolution of the ear is modelled by a bank of overlapping bandpass filters known as an *auditory filterbank*, whose individual filters model as closely as possible the spectral resolution of the ear at each filter centre frequency. The overall response to a stimulus for such a filterbank is known as an *excitation pattern*. Excitation patterns are an important indication of the degree to which components of a stimulus are resolved internally by the auditory system. *Temporal masking* refers to the fact that a masking signal may produce masking both before and after its onset and decay. The causes of temporal masking are less well understood than those of spectral masking, but it is thought that temporal masking may be due to a more central limitation of the auditory system.

A great deal of study of auditory modelling has been carried out in the psychoacoustic literature. These models generally consist of combinations of physiologically based models of individual components of the overall auditory system. Models of auditory masking have also been employed in subband or transform coding audio compression schemes. Such coding schemes split a signal into a number of subbands and utilise auditory masking models to estimate the masking threshold within each subband. From this, the number of bits necessary to code each subband without perceptible degradation in audio quality may be determined. Despite this research, relatively little use to date has been made of joint time-frequency signal analysis in auditory modelling. If both temporal and spectral masking are to be simultaneously modelled using joint time-frequency techniques, independent control of the time and frequency resolutions of the distribution is necessary. Because of the Heisenberg uncertainty principle, which states that the time and frequency extent of any given

signal cannot be made arbitrarily narrow, such independence is not possible using linear time-frequency analysis techniques such as the spectrogram[1] or wavelet transform. However, an important bilinear class of time-frequency distributions, known as *Cohen's class*, does allow the temporal and spectral resolutions to be independently set. Cohen's class is expressed as the convolution of a fundamental distribution known as the *Wigner distribution* with a *smoothing kernel*. The smoothing kernel is used to control the time and frequency resolutions of the class as is also often used to reduce the undesirable *cross-term interference* inherent to bilinear time-frequency techniques. To the best of the author's knowledge, Cohen's class has not previously been used in auditory modelling, despite its many advantages which are described later this work. To date, audio signal analysis using Cohen's class has been geared towards finding a compromise between high time and frequency resolution and good cross-term suppression, rather than incorporating auditory resolution. Cohen's class distributions have also traditionally been very expensive in terms of computation time and memory usage and this has limited their applicability to the analysis of long signals.

The main focus of this thesis is the design and evaluation of a new member of Cohen's class which accurately incorporates the temporal and spectral resolutions of the auditory system. This is a departure from the normal kernel design approach of simply striving for a compromise between high time-frequency resolution and good cross-term suppression. This present work is carried out in order to attain a much more accurate reflection of auditory perception than achieved using conventional audio time-frequency analysis techniques. The approach outlined here also has advantages over typical psychoacoustic models of the auditory system because the analytic framework provided by Cohen's class may be utilised in a consistent manner to extract meaningful distribution information. It is thus the intention that this new modelling approach form a link or compromise between standard TF signal analysis techniques, designed to extract signal properties, and physiologically-based auditory modelling, designed to investigate auditory signal perception.

The distribution designed in this work is dubbed the *'EarWig'* distribution (EWD) as it is a smoothed Wigner distribution (*i.e.* a member of Cohen's class) which incorporates the resolutions of the human ear. The EWD is designed to generate the temporally evolving excitation pattern of a given audio signal and thus to simultaneously reflect both spectral and temporal masking. Unlike most Cohen class distributions, the EWD effectively does not have a limit on its dynamic range as cross-term suppression is very high ensuring that

[1] Strictly speaking, the spectrogram is a member of Cohen's class and may be expressed as the convolution of the Wigner distribution of the signal with the Wigner distribution of the window function. However, the spectrogram nevertheless suffers from the time versus frequency resolution trade-off flaw and is therefore often referred to as linear rather than bilinear.

such interference is virtually non-existent for practically all signals encountered. This is important as the human auditory system exhibits a dynamic range in excess of 100 dB for most of the audible frequency range. The EWD is shown to very accurately achieve the resolutions specified by the models of spectral and temporal masking adopted in the design of its kernel. The EWD is thus a highly useful audio signal analysis tool since by incorporating spectral and temporal masking it only registers unmasked *i.e.* audible signal information. As a demonstration of its utility, this work applies the EWD to two major areas of audio research, namely 3-D surround-sound system design and the investigation of audible degradation in artificially generated reverberation. For the case of surround-sound system design, the EWD is demonstrated to be a very effective method of reducing inter-subject variability in head related transfer function measurements (HRTFs). For artificial reverberator design, the EWD is shown to be an excellent tool for the identification and characterization of undesirable artifacts of synthetic reverberation. The EWD is thus demonstrated to be a very useful, perceptually motivated time-frequency signal analysis tool.

1.1 Thesis Outline

Chapter 2 and Chapter 3 are designed to give the reader a thorough understanding of the areas of time-frequency analysis, psychoacoustics and masking. Chapter 2 is concerned with time-frequency analysis and initially introduces the spectrogram and discusses the limitations of this distribution. The properties of the Wigner distribution and Cohen's class are then described in considerable detail. As the workings of the human auditory system will be unfamiliar ground to most readers coming from a non-psychoacoustic background, Chapter 3 presents an extensive review of the areas of psychoacoustics relevant to this thesis, namely spectral and temporal masking. An overview of the peripheral auditory system is initially presented, to familiarise the reader with the most important components of the ear. Spectral and temporal masking are then described in detail and a historical review of models of these phenomena is given, both for the sake of completeness and also to provide the reader with a clear understanding of the nature of masking and of the psychoacoustic issues and terminology relevant to its modelling.

Chapter 4 proceeds with a detailed explanation of the EWD and its design. Initially, the implementation of a discrete version of Cohen's class is explained in detail. Here, it is demonstrated how the psychoacoustic models describing auditory temporal and spectral resolution may be incorporated into the design of a frequency dependent, Cohen's class

smoothing kernel. Issues pertaining to efficiency of implementation and memory usage are also described and the computational cost of the EWD is calculated. Chapter 5 evaluates the EWD and demonstrates that the spectral and temporal resolutions of the EWD successfully match those adopted from the psychoacoustic literature. It is shown that the EWD is effectively a temporally evolving excitation pattern. Furthermore, the cross-term suppression and dynamic range of the EWD are investigated and discussed.

In order to highlight the utility of the EWD as a joint time-frequency model of both temporal and spectral masking, Chapters 6 and 7 describe the application of the EWD to 3-D surround-sound system design and artificial reverberation analysis respectively. In a real listening environment, sound may approach us from any direction, with the perception of each sound being accompanied by a sense of its location. Normal stereo reproduction systems, however, can only produce sounds which appear to emanate from a point along the arc joining the two loudspeakers. The aim of surround sound technology is to re-create the full three-dimensional spatial hearing we experience in real listening. This may be achieved by replicating the localisation cues which the auditory system uses to determine the location of a sound source. Such localisation cues are summarised by a given listener's head-related transfer functions (HRTFs), which encapsulate for each source location the modifications sound undergoes as it interacts with the listener's body. However, since no two people are physically identical, there is considerable inter-subject variability in HRTF measurements in general. This complicates the task of designing a 3-D surround sound system which is effective for all listeners. In this work, the EWD is demonstrated to be very well suited to the reduction of such inter-subject variability by removing spectrally masked, and therefore redundant, detail from HRTF measurements.

Chapter 7 describes the application of the EWD to the analysis of real and synthetic reverberation. Reverberation is important as it affects our perception of a sound's timbre, loudness and provides also the spatial impression of an acoustic space. When sound reaches our ears in a reverberant environment such as a cathedral, it does so not only via a direct path from the sound source but also via paths involving reflections off the surfaces of the space such as the walls, floor and ceiling. The term *reverberation* refers to the collection of these reflections. The acoustical qualities or 'sound' of a performance space, such as a theatre or concert hall, are characterised by its reverberant response. Without reverberation, music sounds dry and lifeless. However, an excessive amount of reverberation can cause it to sound 'washed out'. Excessive reverberation can also make speech unintelligible in spaces designed for speech such as theatres, lecture halls or churches. A great deal of attention is therefore required in the design of acoustic spaces in order to ensure an appropriate quality

of reverberation.

For the case of music or speech recorded in a dry studio environment, reverberation is provided artificially using electronic reverberators typically implemented using digital filtering algorithms on DSPs. Such algorithms aim to generate realistic sounding reverberation by re-creating the perceptually relevant characteristics of a real or imagined performance space. However, there is still disagreement over which characteristics of reverberation are perceptually relevant and which are irrelevant. Therefore, a good deal of the tuning of artificial reverberators must be carried out by trial and error before a pleasant and natural sounding reverberation effect is arrived at. Since the EWD is a perceptually motivated time-frequency signal analysis tool, Chapter 7 explores the use of the EWD in analysing both real and synthetic reverberant responses. This is carried out in an attempt at gaining a better indication of the perceptually relevant characteristics of reverberation. Furthermore, a close examination of the EWDs of synthetic reverberant responses is conducted in order to achieve a better characterization of the unpleasant degradation artifacts sometimes present in artificial reverberation.

Chapter 2

Joint Time-Frequency Analysis

This chapter presents a discussion of an important group of joint time-frequency distributions known as Cohen's class. This class of distribution is very well known and has already been applied with great success to an extensive range of time-frequency signal analysis problems including sonar and radar analysis, biological signal analysis, machine monitoring [11], the processing of seismic signals [19], the analysis of loudspeaker impulse responses [80] and musical signal analysis [148]. To gain a clear understanding of time-frequency analysis the well known short-time Fourier transform is initially introduced. The Wigner distribution is then described and its mathematical properties are listed and explained. The set of all possible linear transformations of the Wigner distribution, known as Cohen's class, is also defined and we show how manipulation of a smoothing kernel is used to control distribution behaviour. An important feature of Cohen's class is that it allows independent control of both time and frequency resolution. However, a disadvantage of the class is that it suffers from bilinear *cross-terms*, which complicate its interpretation. However, these cross-terms may be removed by smoothing as described later in this chapter. Chapter 3 describes spectral and temporal masking, two auditory phenomena which exist as a consequence of the ear's limited spectral and temporal resolutions respectively. Independent control of a given time-frequency distribution's time and frequency resolutions is very useful if the spectral and temporal masking inherent to the auditory system are to be simultaneously modelled. This chapter therefore introduces the concept of the *separable smoothing kernel* which is used to independently control the time and frequency resolutions of a Cohen's class distribution and also to suppress unwanted cross-term interference. The use of such a separable kernel also allows a very computationally efficient Cohen's class algorithm to be employed.

2.1 The Short-Time Fourier Transform and Spectrogram

The *spectrogram* [33] is a popular method of time-frequency analysis, initially developed during the 1940's for the purpose of speech analysis. The concept is simple: to study the signal properties at a certain time t, we extract a short section of the signal centred around this time and calculate its spectrum. In practice, this is performed by multiplying the signal by a window, $h(t)$, centred at time, t, and calculating the Fourier transform of the resulting modified signal. This is known as the *short-time Fourier transform (STFT)* and may be expressed as :

$$S(t,\omega) = \int e^{-j\omega\tau} s(\tau) h(\tau - t) d\tau \ , \tag{2.1}$$

where $S(t,\omega)$ is the STFT at time t and frequency ω, $e^{-j\omega\tau}$ is the complex exponential of frequency ω and $s(t)$ and $h(t)$ are the signal and window functions, respectively, at time t.

The energy density spectrum, P_{SP}, at time, t, is the absolute magnitude squared of the STFT:

$$P_{SP}(t,\omega) = |S(t,\omega)|^2 \tag{2.2}$$

Spectrograms are composed of a collection of energy density spectrums taken at regular time intervals and plotted together on a three-dimensional time-frequency grid. Figure 2.1 presents the waveform, energy spectral density and spectrogram of a linear chirp signal. The chirp is represented in time-frequency by a broad peak centred on the chirp's instantaneous frequency. The advantages of this type of signal representation are clear; we can tell at any particular time the frequency content of the signal.

2.1.1 Spectrogram Time-Frequency Resolution

In order to calculate the *temporal resolution* of the STFT we let $s(t)$ be a Dirac impulse [9]:

$$s(t) = \delta(t - t_0) \ . \tag{2.3}$$

From Equation 2.1 and the sifting theorem, the resulting STFT $S(t,\omega)$ becomes :

$$S(t,\omega) = e^{-j\omega t_0} h(t_0 - t) \tag{2.4}$$

If $h(t)$ is symetric about $t = 0$, Equation 2.4 may be re-written as :

$$s(t) = \delta(t - t_0) \Rightarrow S(t,\omega) = e^{-j\omega t_0} h(t - t_0) \tag{2.5}$$

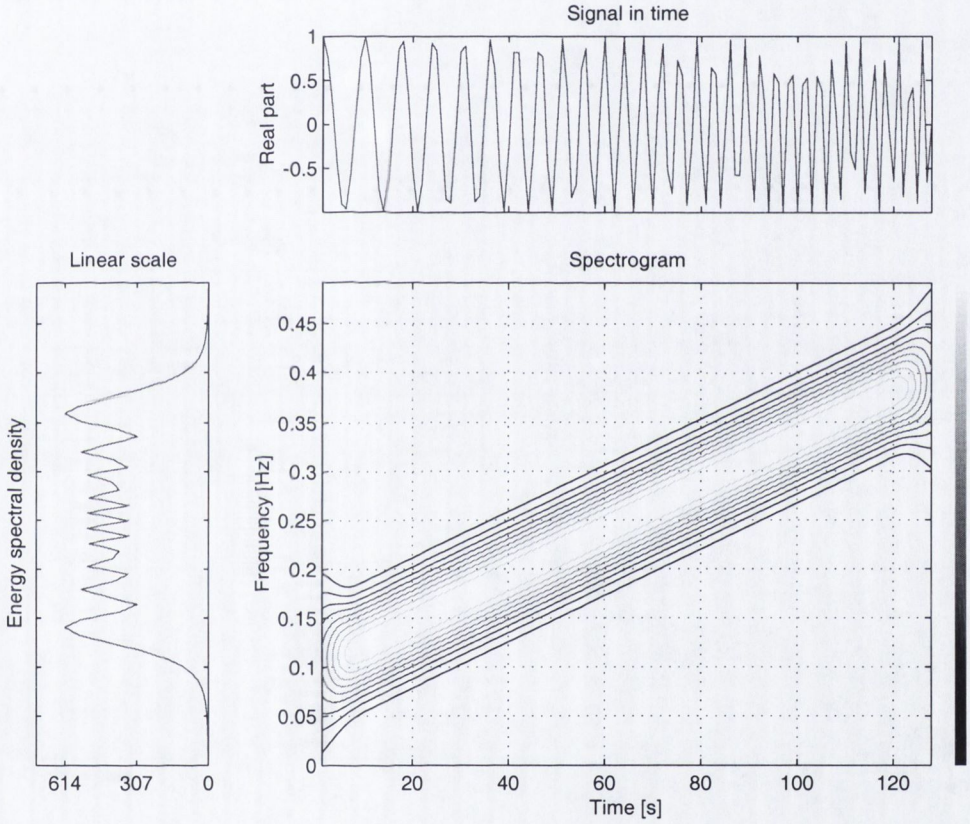


Figure 2.1: Waveform, energy spectral density and spectrogram of a chirp signal.

The spectrogram for this signal is therefore

$$P_{SP}(t,\omega) = |h(t-t_0)|^2 \ . \tag{2.6}$$

It is clear from Eq. 2.5 that the temporal resolution of the STFT is proportional to the effective length of the window function, $h(t)$ and from Eq. 2.4 that the temporal resolution of the spectrogram is proportional to the effective length of $|h(t)|^2$. We can calculate the *frequency resolution* in a similar manner by letting $s(t)$ be a complex sinusoid:

$$s(t) = e^{j\omega_0 t} \Rightarrow S(t,\omega) = e^{-j\omega t} H(\omega - \omega_0) \tag{2.7}$$

In this case, the spectrogram for this signal is

$$P_{SP}(t,\omega) = |H(\omega - \omega_0)|^2 \ . \tag{2.8}$$

Thus, the frequency resolutions of the STFT and spectrogram are governed by the effective bandwidth of $H(\omega)$ and $|H(\omega)|^2$ respectively, where $H(\omega)$ is the Fourier transform of window $h(t)$. Because of the uncertainty principle, a window's duration and bandwidth cannot both be made arbitrarily narrow [33]. Therefore, the STFT suffers from an inherent temporal resolution versus frequency resolution *trade-off*. By shortening the window length, temporal resolution may be increased; this reduces the frequency resolution and vice versa. It is thus impossible to independently set time and frequency resolutions. Effective spectrogram analysis requires that the signal be stationary for the duration of the window. If this is not the case for a given signal, the window can be shortened. However, sustained shortening of the window implies increasingly poor frequency resolution - the signal spectral detail becomes more and more blurred until we are effectively using the signal to study the window spectrum [31] !

Clearly, it would be useful to identify joint time-frequency representations which would not suffer from resolution trade-off limitations. Section 2.3 describes a generalised class of bilinear time-frequency representations, known as *Cohen's class* [33] which is free from the resolution trade off limitations of the spectrogram. The *Wigner distribution*, described in the following section, is the most fundamental member of this class.

2.2 The Wigner Distribution

The concept of the *Wigner distribution* (WD) was introduced by Eugene Wigner [188] in 1932 in the context of quantum mechanics as a means of representing the joint distribution of position and momentum. In 1948 Ville [181] recognised a partial formal mathematical

correspondence between quantum mechanics and signal analysis and applied the WD to communication theory. However, it was not until a set of three papers by Claasen and Mecklenbräuker [27], [28], [29] that the WD received serious attention in the field of digital signal processing.

The WD may be expressed in terms of its signal, $s(t)$, or its spectrum, $S(\omega)$, as follows[1] :

$$W(t,\omega) = \int_{-\infty}^{\infty} s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2})\, e^{-j\tau\omega}\, d\tau \tag{2.9}$$

$$= \int_{-\infty}^{\infty} S^*(\omega-\frac{\theta}{2})\, S(\omega+\frac{\theta}{2})\, e^{jt\theta}\, d\theta \tag{2.10}$$

From this formulation it can be seen that the WD is the Fourier transform of the (deterministic) autocorrelation function, $s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2})$, also known as the *temporal correlation function* (TCF). In a dual manner, the WD may also be expressed as the inverse Fourier transform of the *spectral correlation function* (SCF), $S^*(\omega-\frac{\theta}{2})\, S(\omega+\frac{\theta}{2})$. As an example, Figure 2.2 portrays the WD of the linear chirp signal analysed earlier using the spectrogram in Figure 2.1. From the plot we see that the WD provides superior time and frequency resolution to that of the spectrogram—the signal energy is almost completely localised along the instantaneous frequency of the chirp.

2.2.1 Wigner Distribution Time-Frequency Resolution

The superior time and frequency resolution of the WD over the spectrogram is explained here as follows. Earlier in Section 2.1.1 we saw that the time and frequency resolution of the spectrogram was dependent on the window shape and frequency response, respectively. This, however, is not true for the WD. Consider the case where the signal is perfectly localised in time *i.e.* $s(t)$ is a Dirac impulse :

$$W(t,\omega) = \int_{-\infty}^{\infty} s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2}) e^{-j\tau\omega}\, d\tau \quad ; \quad s(t) = \delta(t-t_0) \tag{2.11}$$

This gives

$$W(t,\omega) = \int_{-\infty}^{\infty} \delta^*(t-\frac{\tau}{2}-t_0)\, \delta(t+\frac{\tau}{2}-t_0) e^{-j\tau\omega}\, d\tau \tag{2.12}$$

Therefore,

$$W(t,\omega) = \delta(t-t_0) \quad \forall\ \omega \tag{2.13}$$

[1]The integration limits indicate a region from $-\infty$ to $+\infty$. For simplicity, these limits will be omitted in future expressions for the WD and are implied unless otherwise stated.

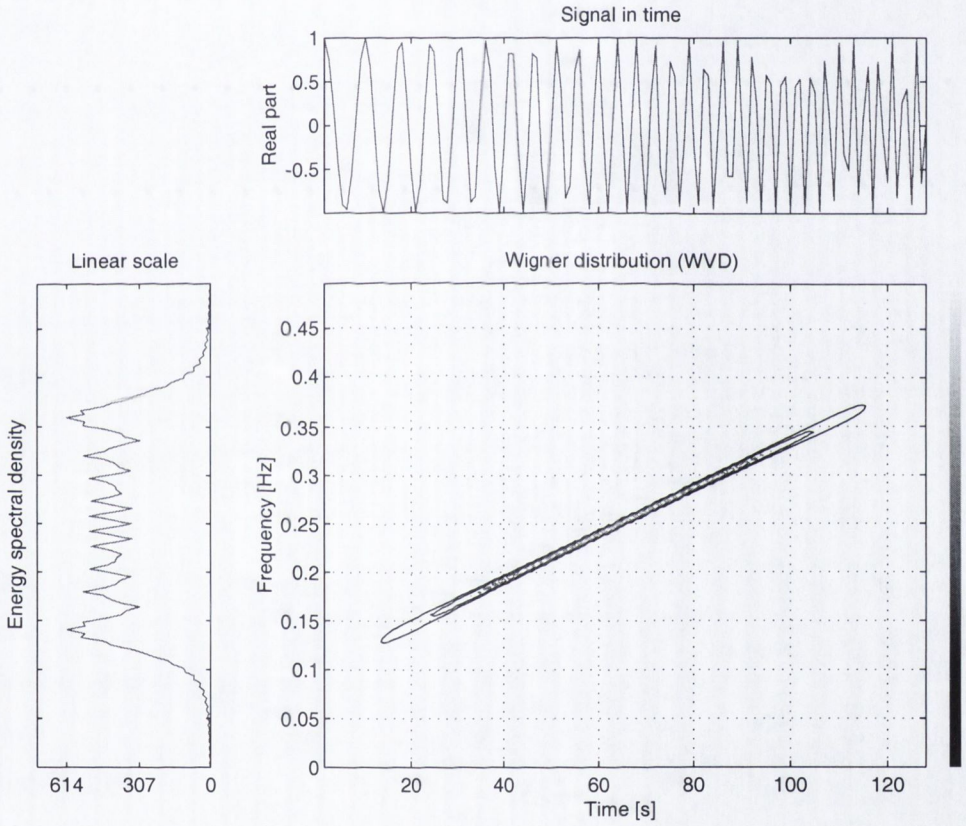


Figure 2.2: Waveform, spectrum and WD of a chirp signal.

Here, it is clear that $W(t,\omega)$ is perfectly localised in time, being equal to unity at $t = 0$ for all ω and zero everywhere else. Therefore, the WD has the same temporal resolution as the signal itself and is not compormised by the resolution of an analysis window, as is the case for the spectrogram.

In a dual manner, we can examine the WD for a signal which is perfectly localised in frequency, *i.e.* a complex sinusoid. So, for this signal,

$$W(t,\omega) = \int_{-\infty}^{\infty} s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2}) e^{-j\tau\omega}\, d\tau \quad ; \quad s(t) = e^{j\omega_0 t} \tag{2.14}$$

This gives

$$\begin{aligned} W(t,\omega) &= \int_{-\infty}^{\infty} e^{j\omega_0(t+\tau/2)} e^{-j\omega_0(t-\tau/2)} e^{-j\tau\omega}\, d\tau \\ W(t,\omega) &= \int_{-\infty}^{\infty} e^{j\omega_0\tau} e^{-j\tau\omega}\, d\tau \\ W(t,\omega) &= \int_{-\infty}^{\infty} e^{j(\omega-\omega_0)\tau}\, d\tau \end{aligned}$$

Therefore,

$$W(t,\omega) = \delta(\omega-\omega_0) \quad \forall\, t \tag{2.15}$$

Again, unlike the spectrogram, whose frequency resolution is limited by the Fourier transform of the analysis window, the frequency resolution of the Wigner distribution is equal to that of the signal's spectrum. Here, both signal and WD are perfectly localised in frequency.

Time-frequency analysis is useful because it provides us with an intuitive 3-D representation of where a signal's energy content is concentrated at any given time. It can provide us knowledge not apparent from either a signal's spectrum or time waveform. Furthermore, time-frequency analysis may be employed to deduce information about a signal, for example its instantaneous frequency or group delay. In order for time-frequency analysis to provide us with such information, the particular time-frequency distribution in question must satisfy a number of mathematical rules relating it to the signal and/or its properties. The Wigner distribution satisfies a number of such useful mathematical properties which are summarised in the following section.

2.2.2 Wigner Distribution Properties

1. *Reality*: The WD of any signal s (real or complex) is real :

$$W^*(t,\omega) = W(t,\omega) \ . \tag{2.16}$$

2. *Evenness*: The WD of any signal s (real or complex) is even in frequency :

$$W(t,-\omega) = W(t,\omega) \ . \tag{2.17}$$

3. *Energy Conservation*: The integration of the WD of signal s over the time-frequency plane yields the total signal energy E :

$$E = \int\int W(t,\omega)\, d\omega\, dt = \int |s(t)|^2\, dt = \int |S(\omega)|^2 d\omega \ . \tag{2.18}$$

4. *Time Marginal*: Integration of the WD of s with respect to frequency yields the instantaneous power $|s(t)|^2$:

$$\int W(t,\omega)\, d\omega = |s(t)|^2 \ . \tag{2.19}$$

5. *Frequency Marginal*: Integration of the WD of s with respect to time yields the energy spectral density $|S(\omega)|^2$:

$$\int W(t,\omega)\, dt = |S(\omega)|^2 \ . \tag{2.20}$$

Properties 4 and 5 thus state that the Wigner distribution satisfies both time and frequency marginals.

6. *Time Shift Covariance*: The WD preserves time shifts of signal s i.e. the WD is time shift covariant :

$$\text{If} \quad s(t) \Rightarrow s(t-t_0) \quad \text{then} \quad W(t,\omega) \Rightarrow W(t-t_0,\, \omega) \ . \tag{2.21}$$

7. *Frequency Shift Covariance*: The WD preserves modulations of signal, s, i.e. the WD is frequency covariant :

$$\text{If} \quad s(t) \Rightarrow e^{jw_0 t} s(t) \quad \text{then} \quad W(t,\omega) \Rightarrow W(t,\, \omega-\omega_0) \ . \tag{2.22}$$

8. *Finite Time Support*: If the signal $s(t)$ is zero outside a finite time interval, then its WD will be correspondingly time limited :

$$\text{If} \quad s(t) = 0, \quad |t| > T \quad \text{then} \quad W(t,\omega) = 0, \quad |t| > T \; . \tag{2.23}$$

9. *Finite Frequency Support*: If the signal s is of finite bandwidth, then its WD will be bandlimited to exactly the same frequency interval :

$$\text{if} \quad s(t) = 0, \quad |\omega| > B \quad \text{then} \quad W(t,\omega) = 0, \quad |\omega| > B \; . \tag{2.24}$$

10. *Group Delay*: The group delay of a signal s can be obtained by calculating the first order moment in time of the WD of the analytic signal [67], s_a, where s_a is the analytic signal associated with s :

$$t_g(\omega) = \frac{\int t\, W_{s_a}(t,\omega)\, dt}{\int W_{s_a}(t,\omega)\, dt} \; . \tag{2.25}$$

11. *Instantaneous Frequency*: Similarly, calculation of the first order moment in frequency of the WD of the analytic signal, s_a, yields the instantaneous frequency of signal, s :

$$\Omega(t) = \frac{\int \omega\, W_{s_a}(t,\omega)\, d\omega}{\int W_{s_a}(t,\omega)\, d\omega} \; . \tag{2.26}$$

12. *Convolution Covariance*: If y is the convolution of x and h then the Wigner distribution of y is the convolution in time of the WD of x with the WD of h :

$$y(t) = \int h(t-\tau)\, x(\tau)\, d\tau \quad \Rightarrow \quad W_y(t,\omega) = \int W_h(t-\tau,\omega) W_x(\tau,\omega)\, d\tau \; . \tag{2.27}$$

13. *Modulation Covariance*: In a similar manner, if y is the modulation of x and h then the Wigner distribution of y is the convolution in frequency of the WD of x with the WD of h :

$$y(t) = h(t)\, x(t) \quad \Rightarrow \quad W_y(t,\omega) = \int W_h(t,\omega-\xi) W_x(t,\xi)\, d\xi \; . \tag{2.28}$$

It is thus clear that the Wigner distribution neatly encapsulates many useful signal measures, including the total energy, energy spectral density, instantaneous power, group delay and instantaneous frequency. Furthermore, signal operations such as time or frequency shifts, convolutions and modulations are mirrored by the WD in an intuitive manner. Overall it is readily apparent that the WD is a powerful and elegant signal analysis framework. However, these advantages do not come without a price. The following section details the

main disadvantage typically associated with the WD, namely cross-term interference. This interference causes non-zero values to appear in the WD of a particular signal at unexpected locations in time-frequency. It will be shown, however, that with appropriate smoothing, this interference can be suppressed, albeit with some weakening of the properties discussed above.

2.2.3 The Cross-Wigner Distribution

Any given signal may be arbitrarily broken up into two (or more) components :

$$s(t) = s_1(t) + s_2(t) \ . \tag{2.29}$$

Because of the bilinear nature of the WD, the resulting WD of $s(t)$ is

$$W(t,\omega) = W_{11}(t,\omega) + W_{22}(t,\omega) + W_{12}(t,\omega) + W_{21}(t,\omega) \ , \tag{2.30}$$

where $W_{11}(t,\omega)$ and $W_{22}(t,\omega)$ are the *auto Wigner distributions* of $s_1(t)$ and $s_2(t)$, respectively and $W_{12}(t,\omega)$ and $W_{21}(t,\omega)$ are known as the *cross-Wigner distributions* of $s_1(t)$ and $s_2(t)$. These may be expressed as :

$$\begin{aligned}
W_{12}(t,\omega) &= \int s_1^*(t-\frac{\tau}{2})\, s_2(t+\frac{\tau}{2})\, e^{-j\tau\omega}\, d\tau && (2.31)\\
&= \int S_1^*(\omega-\frac{\theta}{2})\, S_2(\omega+\frac{\theta}{2})\, e^{jt\theta}\, d\theta \quad \text{and} && (2.32)\\
W_{21}(t,\omega) &= \int s_2^*(t-\frac{\tau}{2})\, s_1(t+\frac{\tau}{2})\, e^{-j\tau\omega}\, d\tau && (2.33)\\
&= \int S_2^*(\omega-\frac{\theta}{2})\, S_1(\omega+\frac{\theta}{2})\, e^{jt\theta}\, d\theta \ . && (2.34)\\
& && (2.35)
\end{aligned}$$

The cross WD is complex valued. However, it can be shown that $W_{12} = W_{21}^*$. Therefore, $W_{12} + W_{21}^*$ is real. Consequently, from Equation 2.30,

$$W(t,\omega) = W_{11}(t,\omega) + W_{22}(t,\omega) + 2Re\{W_{12}(t,\omega)\} \ . \tag{2.36}$$

From this expression we see that the Wigner distribution of $s_1(t) + s_2(t)$ is not just the sum of the Wigner distributions of the individual signals $s_1(t)$ and $s_2(t)$, but contains an additional term $2Re\{W_{12}(t,\omega)\}$. This term is commonly referred to as an *interference term* or *cross-term*. Such interference can often make interpretation of the Wigner distributions difficult. As an example, consider the (auto) Wigner distribution of the sum of two sinusoids :

$$s(t) = s_1(t) + s_2(t) = A_1 e^{j\omega_1 t} + A_2 e^{j\omega_2 t} \ . \tag{2.37}$$

From Equations 2.36 and 2.15 we have

$$\begin{aligned} W(t,\omega) &= W_{11}(t,\omega) + W_{22}(t,\omega) + 2Re\{W_{12}(t,\omega)\} \\ W_{11} &= |A_1|^2 \delta(\omega - \omega_1) \qquad (2.38) \\ W_{22} &= |A_1|^2 \delta(\omega - \omega_2) \qquad (2.39) \\ W_{12} &= \int_{-\infty}^{\infty} A_1 e^{j\omega_1(t+\tau/2)} A_2^* e^{-j\omega_2(t-\tau/2)} e^{-j\omega\tau}\, d\tau \\ &= \int_{-\infty}^{\infty} A_1 A_2^* e^{j(\omega_1 t + \omega_1\tau/2 - \omega_2 t + \omega_2\tau/2)} e^{-j\omega\tau}\, d\tau \\ &= A_1 A_2^* e^{jt(\omega_1-\omega_2)} \int_{-\infty}^{\infty} e^{-j\tau(\omega - \frac{\omega_1+\omega_2}{2})}\, d\tau \\ &= A_1 A_2^* e^{jt(\omega_1-\omega_2)} \delta(\omega - \frac{\omega_1+\omega_2}{2}) \end{aligned} \tag{2.40}$$

This results in the Wigner distribution

$$W(t,\omega) = |A_1|^2\, \delta(\omega-\omega_1) + |A_2|^2\, \delta(\omega-\omega_2) + 2|A_1 A_2|\, \delta(\omega - \frac{1}{2}(\omega_1+\omega_2)) \cos\big((\omega_1-\omega_2)t\big) \ . \tag{2.41}$$

The terms at ω_1 and ω_2 are known as the auto-terms. The cross-term at $(\omega_1 + \omega_2)/2$ oscillates sinusoidally with time at rate $(\omega_1 - \omega_2)$. This interference is present despite the fact that no signal energy whatsoever exists at $(\omega_1 + \omega_2)/2$. Furthermore, the cross-term interference introduces both positive and negative values; this is counter-intuitive to the notion that the WD be interpreted as a distribution of signal energy is be strictly positive.

Figure 2.3 portrays the waveform, Fourier transform magnitude and Wigner distribution of a signal comprised of three time-shifted and sinusoidally modulated Gaussian impulses (also known as Gaussian atoms [10]). The equation for this signal is :

$$\begin{aligned} s(t) &= \exp\left(-\left(\frac{t-t_0}{T}\right)^2 \pi\right) \exp(j2\pi f_0 t) \\ &+ exp\left(-\left(\frac{t-t_1}{T}\right)^2 \pi\right) \exp(j2\pi f_1 t) \\ &+ exp\left(-\left(\frac{t-t_2}{T}\right)^2 \pi\right) \exp(j2\pi f_2 t) \end{aligned} \tag{2.42}$$

The plots were generated using the Time-Frequency Toolbox for Matlab [9] where the paramaters were set as follows :

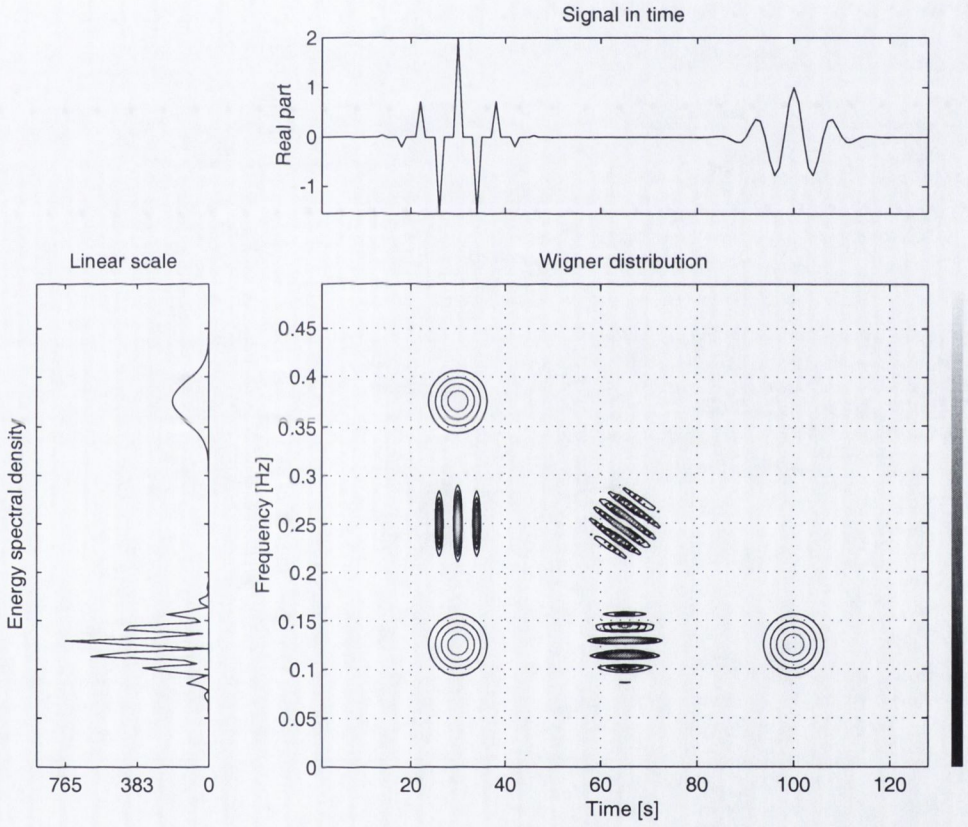


Figure 2.3: Waveform, energy spectral density and Wigner distribution of a signal consisting of three Gaussian atoms.

$$(t_0, f_0) = (30, 0.125), \quad (t_1, f_1) = (100, 0.125) \quad \text{and} \quad (t_2, f_2) = (30, 0.375)$$

From the plot we see that the auto terms appear at time-frequency locations (t_0, f_0), (t_1, f_1) and (t_2, f_2) as expected. In addition, however, three cross-terms located midway between the auto-terms are also present. The following simple geometric laws govern cross-term behaviour:

- Any two auto terms which are disjoint (separated) in the time-frequency plane interfere to create a cross-term located at their geometrical midpoint.
- The cross-term oscillates at a frequency proportional to the distance between the two auto-terms.
- The direction of oscillation is perpendicular to the line joining the two auto-terms.
- The cross-term magnitude is independent of the degree of separation between the auto-terms for the (un-smoothed) Wigner distribution.
- For the *smoothed Wigner distribution*, to be discussed in Section 2.3, cross-terms amplitude may be affected by smoothing. The amplitude of these terms will therefore depend on the type of smoothing kernel and the cross-term oscillation frequency. For example, smoothing kernels of a lowpass type are generally employed to reduce cross-term interference. Therefore, cross-terms become increasingly suppressed with increasing auto-term separation and cross-term oscillation frequency. However, for kernels of a high-pass type, cross-term interference increases with increasing auto-term separation [72].

2.2.4 The Pseudo-Wigner Distribution

From Equations 2.9 and 2.10 we see that in order to calculate the Wigner distribution, we need to calculate the temporal correlation function $s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2})$, for $\tau = -\infty$ to $\tau = +\infty$ or the spectral correlation function $S^*(\omega - \frac{\theta}{2})\, S(\omega + \frac{\theta}{2})$ from $\theta = -\infty$ to $\theta = +\infty$. This is only possible analytically and not numerically. Considering Equation 2.9, we may overcome this limitation by emphasising the TCF around t. This is achieved by multiplying the TCF by finite length window function $h(\tau)$, which is peaked around $\tau = 0$. We thus arrive at a new distribution, with WD-like properties, called the pseudo-Wigner distribution (PWD) :

$$W_{PS}(t, \omega) = \int h(\tau)\, s^*(t - \frac{\tau}{2})\, s(t + \frac{\tau}{2})\, e^{-j\tau\omega}\, d\tau \tag{2.43}$$

By emphasising $s^*(t-\frac{\tau}{2})\,s(t+\frac{\tau}{2})$ around $\tau = 0$, we alleviate somewhat the highly non-local nature of the WD [33]. Cross-term interference oscillating parallel to the frequency axis is also reduced since multiplication of the TCF by $h(\tau)$ amounts to a frequency smoothing of the WD [28]:

$$W_{PS}(t,\omega) = \int H(\omega-\xi)W(t,\xi)\,d\xi \tag{2.44}$$

Unfortunately, however, some of the useful properties of the WD such as the frequency marginal and the instantaneous frequency property are weakened once frequency smoothing takes place. As an example, Figure 2.4 presents the PWD of the signal presented earlier for the WD in Figure 2.3. Cross term interference oscillating parallel to the frequency axis has been reduced at the expense of compromising frequency resolution slightly[2].

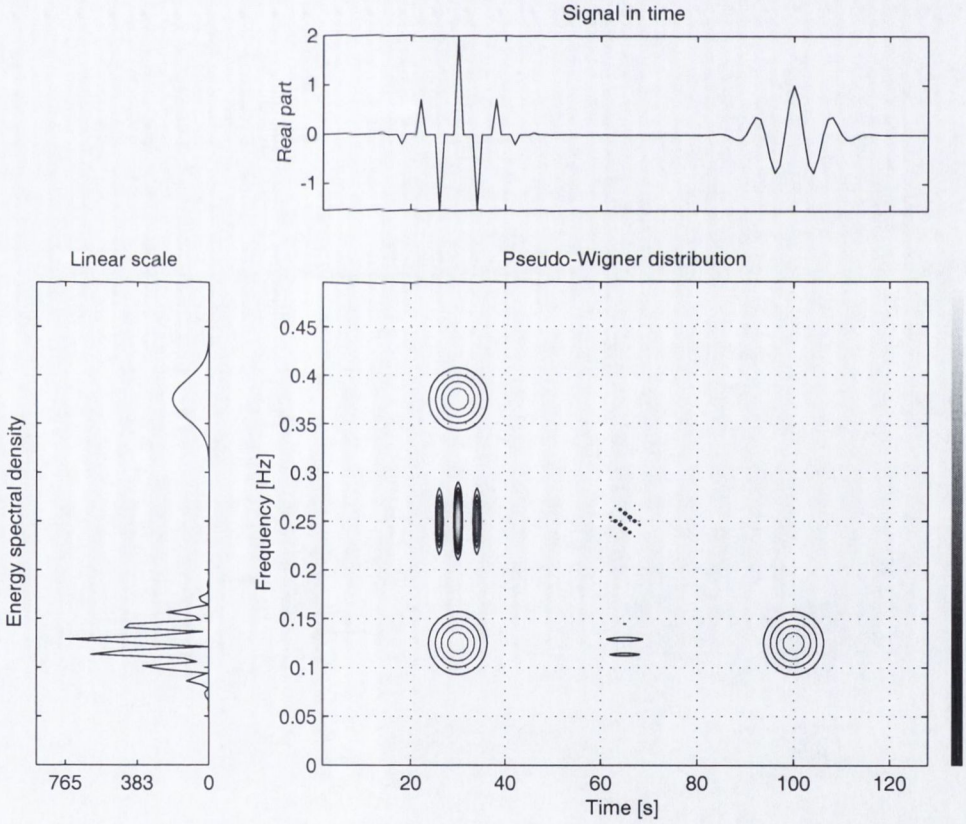


Figure 2.4: Waveform, energy spectral density and pseudo-Wigner distribution of a signal consisting of three Gaussian atoms.

[2]This slight compromise of frequency resolution may be difficult to see in this instance, because the distrubution amplitude is plotted in linear units.

2.2.5 Inter-relationships of the Wigner Distribution, Temporal Correlation Function, Complex Ambiguity Function and Spectral Correlation function

From Equation 2.9, the Wigner distribution may be expressed as the Fourier transform with respect to τ of the two variable function

$$R(t,\tau) = s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2}) \;, \tag{2.45}$$

known as the temporal correlation function (TCF), where t represents time and τ represents relative time :

$$W(t,\omega) = \int R(t,\tau)\, e^{-j\tau\omega}\, d\tau \;. \tag{2.46}$$

From Equation 2.10, the Wigner distribution may also be arrived at from the inverse Fourier transform with respect to relative frequency, θ, of the two variable function

$$A(\omega,\theta) = S^*(\omega-\frac{\theta}{2})\, S(\omega+\frac{\theta}{2}) \;, \tag{2.47}$$

know as the spectral correlation function (SCF), where ω represents frequency and θ represents relative frequency :

$$W(t,\omega) = \int A(\omega,\theta)\, e^{jt\theta}\, d\theta \;. \tag{2.48}$$

Thus, the TCF and SCF form a two-dimensional Fourier transform pair since

$$\int\int s^*(t-\frac{\tau}{2}) s(t+\frac{\tau}{2})\, e^{-j\tau\omega} e^{-jt\theta}\, d\tau\, dt = \int W(t,\omega) e^{-jt\theta}\, dt = A(\omega,\theta) \;. \tag{2.49}$$

Furthermore, the two dimensional Fourier transform of the Wigner distribution yields a two dimensional function of relative time, τ, and relative frequency, θ, known as the *complex ambiguity function* (CAF), $\Gamma(\tau,\theta)$. The CAF may also be expressed as the one-dimensional Fourier transform with respect to time, t, of the TCF, or in a dual manner, the one-dimensional Fourier transform with respect to frequency, ω, of the SCF as follows :

$$\Gamma(\tau,\theta) = \int s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2})\, e^{-jt\theta}\, dt \tag{2.50}$$

$$= \int S^*(\omega-\frac{\theta}{2})\, S(\omega+\frac{\theta}{2})\, e^{j\omega\tau}\, d\omega \;. \tag{2.51}$$

The WDF, $W(t,\omega)$, CAF, $\Gamma(\tau,\theta)$, TCF, $R(t,\tau)$ and SCF, $A(\omega,\theta)$, thus comprise a family of four two-dimensional functions which are linearly related via one or two-dimensional

Fourier transform pairs [155]. A graphical representation of these relationships is presented in Figure 2.5. Each arrow represents a (one-dimensional) Fourier transform pair, the direction of each arrow indicating a forward Fourier transform. It is shown in Chapter 4 that this interrelationship may be exploited to devise a very efficient algorithm for calculating the WD on a digital computer.

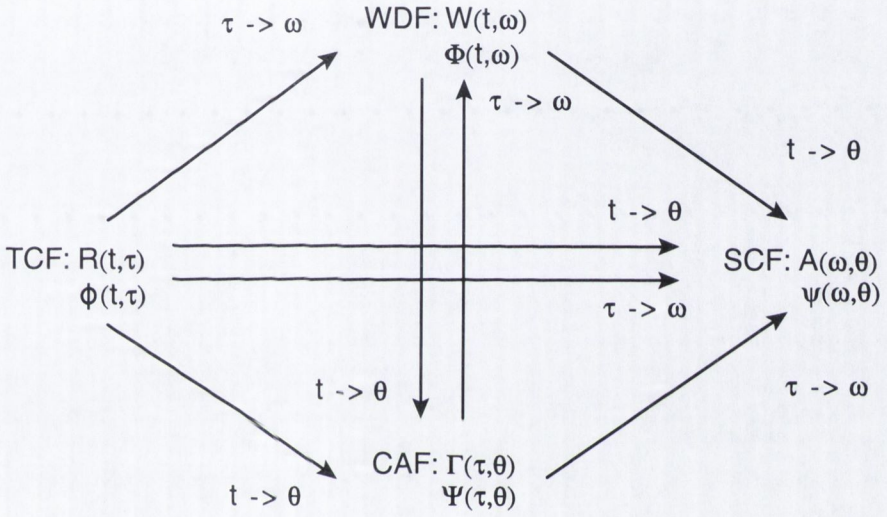


Figure 2.5: Inter-relationship between the WDF, CAF, TCF and SCF domains for the WD and smoothing kernalEach arrow represents a (forward) Fourier transform.

2.3 Cohen's Class

Section 2.2 introduced the Wigner distribution and discussed its important properties. On its own, the Wigner distribution is not necessarily a very useful time-frequency analysis tool, since it suffers from serious cross-term interference. It has been shown that interference oscillating parallel to the frequency axis may be suppressed in the case of the pseudo Wigner distribution using a frequency smoothing window. In a dual manner, cross-term interference oscillating parallel to the time axis may be suppressed using a time smoothing window. Since there are an infinite number of smoothing functions available, there are actually an infinite number of time frequency distributions in existence, expressed as a combination of the smoothing functions and Wigner distribution itself. These time and frequency smoothing functions are often expressed as a single function of both time and frequency, called a

smoothing kernel. The combination of this kernel and the Wigner distribution is generally termed a *Cohen's class* distribution, described below.

An infinite number of time-frequency representations may be generated from Cohen's class of bilinear time-frequency representations [30], [33], expressed as :

$$C(t,\omega) = \int\int\int s^*(u-\frac{\tau}{2})\, s(u+\frac{\tau}{2})\, \Psi(\theta,\tau) e^{-j\theta t - j\tau\omega + j\theta u} du\, d\tau\, d\theta \ , \qquad (2.52)$$

where $\Psi(\theta,\tau)$ is known as the *smoothing kernel* (expressed in the CAF domain). Aided by Figure 2.5 we can see that in this expression the TCF is transformed into the CAF domain, multiplied by the smoothing kernel, $\Psi(\theta,\tau)$, and then transformed into the WDF domain. Because all four domains are equivalent, we can also express $C(t,\omega)$ as

$$C(t,\omega) = \int\int W(u,\xi)\Phi(t-u,\omega-\xi)\, du d\xi \ , \qquad (2.53)$$

where $\Phi(t,\omega)$ is the same *smoothing kernel* (expressed in the WDF domain). The smoothing kernel is denoted $\phi(\tau,\theta)$ in the TCF domain and $\psi(\omega,\theta)$ in the SCF domain, as shown in Figure 2.5.

From Equation 2.53 we see that Cohen's class is the convolution of the Wigner distribution, $W(t,\omega)$, and the smoothing kernel, $\Phi(t,\omega)$. Thus, a member of Cohen's class may sometimes be referred to as a *smoothed pseudo-Wigner distribution* (SPWD). The kernel determines the properties of $C(t,\omega)$ and may depend on frequency, time or even the signal s[3] .

Cohen's class was first proposed in 1966 in [30] and its importance lies in the fact that it demonstrates that a variety of different time-frequency distributions proposed since the spectrogram and Wigner distribution are in fact related via linear transformation. The Page [126], Rihaczek [160] and Margenau-Hill [101] distributions are examples of apparently different time-frequency representations subsequently shown to be linearly related by Cohen. There are several clear advantages to Cohen's unified approach. Firstly, the properties of a given time-frequency representation, once re-expressed in terms of Cohen's class can readily be determined through simple examination of the kernel. Secondly, the unified approach clarifies the relationship between existing time-frequency representations such as the spectrogram and Wigner distribution and simplifies the design of new representations employing more advanced kernel functions [21], [11]. Thirdly, if we wish to design a time-frequency representation with certain desired properties, we merely convert these distribution properties into kernel constraints. For example, if we wish to design a distribution which satisfies the time marginal exactly then we may only choose kernels which

[3]The kernel is sometimes expressed (in the TCF domain) as $\phi(\theta,\tau;t,\omega,s)$ to denote this.

satisfy $\Psi(0,\theta) = 1$. Table 2.1 presents a list of the kernel constraints that must be satisfied to ensure the corresponding distribution properties described earlier in Section 2.2.2. As stated at the begining of this chapter, such a unified approach has been very effective in the design of a multitude of Cohen's class members for such diverse applications as loudspeaker analysis [80], musical signal analysis [148], audio filter analysis [158], speech recognition, sonar and radar echo analysis, the study of muscle and cardiac vibrations and machine monitoring [11].

Property	Kernel Constraint
P_1 : Reality	$\Psi^*(-\tau,-\theta) = \Psi(\tau,\theta)$
P_2 : Evenness	$\Psi(\tau,\theta)$ is real
P_3 : Energy Conservation	$\Psi(0,0) = 1$
P_4 : Time Marginal	$\Psi(0,\theta) = 1$
P_5 : Frequency Marginal	$\Psi(\tau,0) = 1$
P_6 : Time Shift Covariance	$\Psi(\tau,\theta)$ is not dependant on time t
P_7 : Frequency Shift Covariance	$\Psi(\tau,\theta)$ is not dependant on frequency f
P_8 : Finite Time Support	$\phi(t,\tau) = 0, \quad \lvert t/\tau \rvert > 1/2$
P_9 : Finite Frequency Support	$\psi(\omega,\theta) = 0, \quad \lvert \omega/\theta \rvert > 1/2$
P_{10} : Group Delay	$\Psi(\tau,0) = 1$ and $\frac{\delta}{\delta\theta}\,\Psi(\tau,\theta)\big\rvert_{\theta=0} = 0$
P_{11} : Instantaneous Frequency	$\Psi(0,\theta) = 1$ and $\frac{\delta}{\delta\tau}\,\Psi(\tau,\theta)\big\rvert_{\tau=0} = 0$
P_{12} : Convolution Covariance	$\Psi(\tau_1+\tau_2,\theta) = \Psi(\tau_1,\theta)\Psi(\tau_2,\theta)$
P_{13} : Modulation Covariance	$\Psi(\tau,\theta_1+\theta_2) = \Psi(\tau,\theta_1)\Psi(\tau,\theta_2)$

Table 2.1: Cohen's class properties and corresponding kernel constraints

2.3.1 Separable Smoothing

We have shown above that any member of Cohen's class is simply a convolution of $W(t,\omega)$ with a smoothing kernel, $\Phi(t,\omega)$. The kernel controls the time and frequency resolutions of the distribution and also determines the degree of *cross-term suppression.* Interestingly, we can re-express $\Phi(t,\omega)$ as a separable kernel $h_t(t)H_\omega(\omega)$ where $h_t(t)$ is a time smoothing window and $H_\omega(\omega)$ is the Fourier transform of the frequency smoothing window, $h_\omega(\tau)$. This extra degree of freedom allows us to independently control time and frequency resolution. Previously, we saw that for the spectrogram we were forced to compromise between time and frequency resolution. With a separable kernel, however, the compromise is between joint time-frequency resolution and the degree of cross-term suppression. An

increase in time/frequency smoothing leads to a decrease in cross-term interference parallel to the time/frequency axis and a corresponding decrease in time/frequency resolution. Furthermore, we have independent control of the type and degree of time and frequency smoothing. For example, we may decide to smooth only in frequency, as is the case for the pseudo-Wigner distribution. Figure 2.6 portrays the SPWD of the signal $s_g(t)$ (Eq. 2.42) along with its WDF, PWDF and spectrogram. For comparison purposes, the level of the auto-terms has been normalised to unity and the contour lines maintained at the same level for each of the four plots. The time-frequency resolution of the SPWDF can be seen to be slightly inferior to that of the WDF. However, cross-terms have been reduced and the time and frequency resolution of the SPWDF is superior to that of the spectrogram.

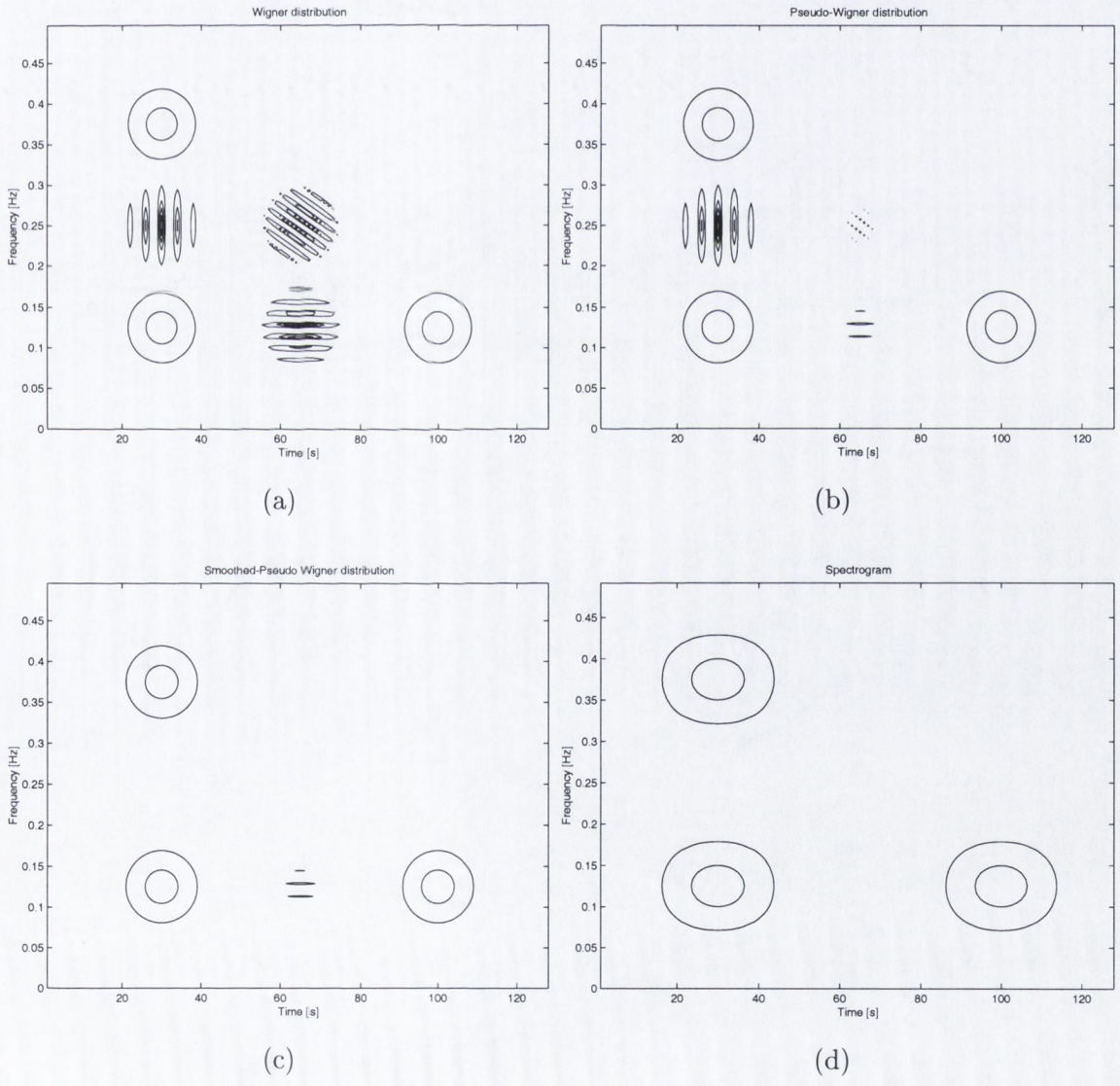


Figure 2.6: WDF (a), PWDF (b), SPWDF (c) and spectrogram (d) of a signal consisting of three Gaussian atoms.

Separable Smoothing in the Temporal Correlation Domain

We can re-express Cohen's class to show the operation of the separable smoothing kernel in the temporal correlation function (TCF) domain as follows :

$$C(t,\omega) = \int h_\omega(\tau) \int h_t(u-t) s^*(u-\frac{\tau}{2})\, s(u+\frac{\tau}{2})\, du\, e^{-j\omega\tau} d\tau \tag{2.54}$$

Since, by Fourier's theory, convolution (smoothing) in the frequency domain implies multiplication in the time domain, the frequency smoothing of $h_\omega(\tau)$ in Equation 2.54 is expressed as a multiplication along the τ axis. This temporal multiplication is converted to frequency smoothing once the expression is transformed into the Wigner domain via a Fourier transform in τ. However, the temporal smoothing of h_t is not affected by a Fourier transform in τ. Therefore, the temporal smoothing of h_t remains a convolution in the WDF domain. As we will show in Chapter 4, significant computational savings can be achieved if the time-frequency smoothing is performed in the TCF domain as in Eq. 2.54 above rather than in the WDF, SCF or CAF domains. However, we must first deduce what smoothing functions are required to accurately model the temporal and spectral resolutions of the ear. This is the subject of the following chapter.

Chapter 3

The Auditory System and Masking

This chapter presents a summary and explanation of the development of useful yet simple models of the spectral and temporal resolutions of the auditory system. The term *resolution* refers to the ability of the ear to resolve or discriminate between separate auditory events. The resolutions of the auditory system are of great importance in the psychoacoustic literature as they are fundamentally linked to *auditory masking*. Auditory masking occurs when a relatively weak signal component is rendered inaudible (masked) by a relatively strong signal component (masker) if both are either temporally or spectrally close. The classic example of auditory masking often given is that of two people having a conversation by the side of a road. A truck passes; if they continue to talk at the same level as previously then their speech may be rendered inaudible. In such a circumstance the noise of the truck is said to mask the speech signal. Masking is divided into temporal and spectral (critical-band) masking; both are related to the ear's limited temporal and spectral resolution respectively.

Much research in psychoacoustics has been concerned with achieving accurate models of auditory resolution. By allowing us to predict the response of the ear to stimuli, such models provide us with a better understanding of the relationship between sound and sensation. However, to date, relatively little use has been made of actual joint time-frequency techniques in auditory modelling. No use has been made of the smoothed Wigner distribution, despite the fact that it affords independent control of its temporal and spectral resolutions and thus provides us with a method of simultaneously and independently modelling both spectral and temporal masking. This chapter begins with a short overview of the functioning of the human auditory system and continues with a discussion of spectral masking and a historical review of the concepts of the auditory filter and auditory filter shape. Popular auditory filter shapes are reviewed. Temporal resolution is subsequently discussed and a review of research investigating the ear's temporal acuity or resolution is

given. A model of temporal resolution known as the 'temporal window model' is presented. Finally a review of auditory modelling in general is presented, with particular emphasis on auditory models incorporating joint time-frequency techniques. The advantages of using the smoothed Wigner distribution rather than linear time-frequency analysis techniques for auditory modelling are also outlined.

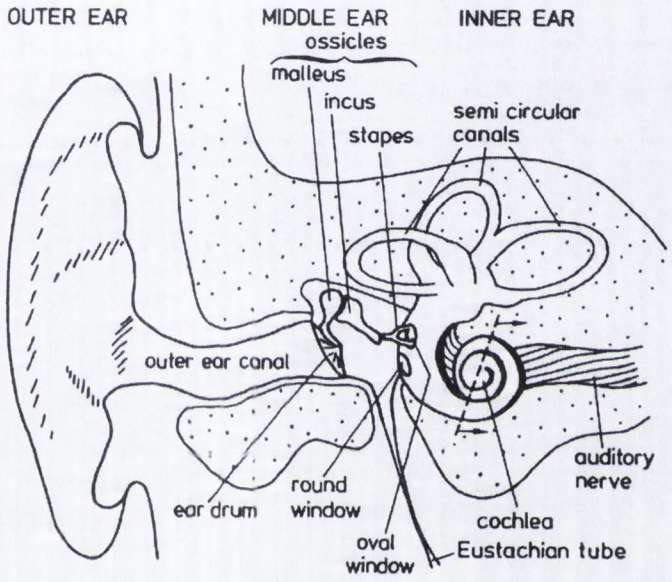


Figure 3.1: Outer, middle and inner ear structure [108].

3.1 Overview of the Auditory System

There is much similarity in the peripheral auditory system of most mammals. This system can be divided into three main sections : the outer ear, the middle ear and the inner ear [108]. Figure 3.1 shows the overall structure of the human peripheral auditory system. The pinna is the cartilaginous tissue on the outside of the head that is usually associated with the ear. The pinna and outer ear canal (also known as the auditory canal or meatus) compose the outer ear. Sound is reflected and diffracted initially by the pinna causing spectral modifications which are important for our ability to localise sound sources. The sound then travels down the outer ear canal (meatus) and causes the tympanic membrane or eardrum to vibrate. The eardrum forms the boundary between the outer and middle ear. In

the middle ear, sound is mechanically transmitted by three small bones, the malleus, incus and stapes (collectively known as the ossicular chain or ossicles), to a membrane covered opening in the inner ear, or cochlea. The opening is called the oval window. Movement of the oval window sets up pressure waves in the cochlear fluid which are relieved by a sympathetic motion of the round window. The primary function of the middle ear is thus the efficient transfer of sound from the air to the fluids of the cochlea.

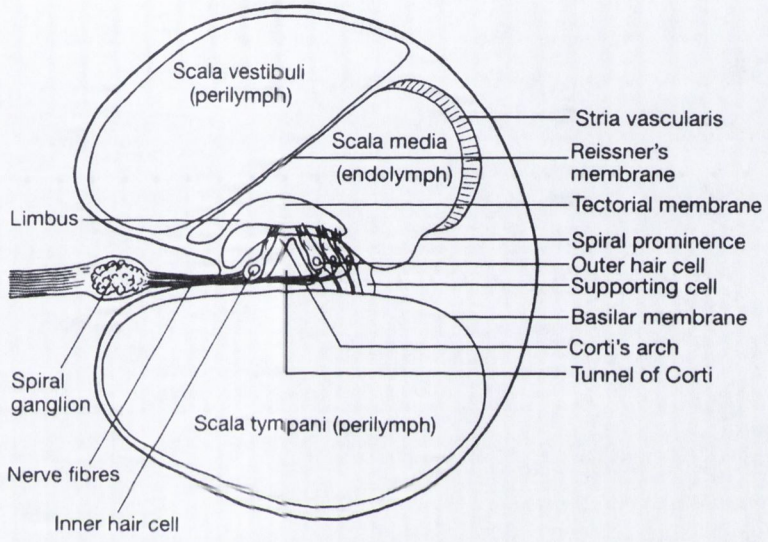


Figure 3.2: Cross section of the cochlear duct [176].

The cochlea is the most important component of the peripheral auditory system. This is shown in Figure 3.2, which presents a cross section of the cochlear duct. Reissner's membrane and the basilar membrane (BM) divide the rigid, fluid-filled cochlea along its entire length. The base refers to the start of the cochlea, where the oval window occurs, whereas the apex refers to its end. The cochlea narrows from base to apex. As stated previously, the ossicular chain moves the oval window in response to a sound, which sets up a pressure difference across the BM, causing the BM to move. For the case of a sinusoidal stimulus, a travelling wave is set up which propagates along the BM from base to apex. Figure 3.3 depicts the displacement of the basilar membrane at two particular instants in time in response to a 200 Hz sinusoid. The overall envelope of the response is also shown and exhibits a gradual increase up to a peak about 28 mm from the stapes, after which the amplitude envelope decays rapidly. Because of the mechanical properties of the BM, the position of the peak of the travelling wave depends on the frequency of stimulation. High frequencies produce a point of maximum displacement near the oval window, with little movement beyond this point. Low frequencies cause movement of the BM along much of its

length, up to the point of maximum displacement, after which the envelope amplitude falls quickly. The response of the BM to two individual sinusoids depends on their frequency separation. If they are sufficiently far apart, the envelope of the travelling wave produced will exhibit two peaks corresponding to the two stimulus frequencies. For a smaller separation, some points on the BM respond to both tones, with a displacement following the complex waveform produced by the interference of the two tones. For even smaller separations, there are no longer two distinct peaks in the envelope of the BM motion. Instead, there is a single, broad peak. In this instance, the BM fails to resolve the two sinusoidal components. This has important consequences for the spectral masking properties of the auditory system, to be discussed in Section 3.2.

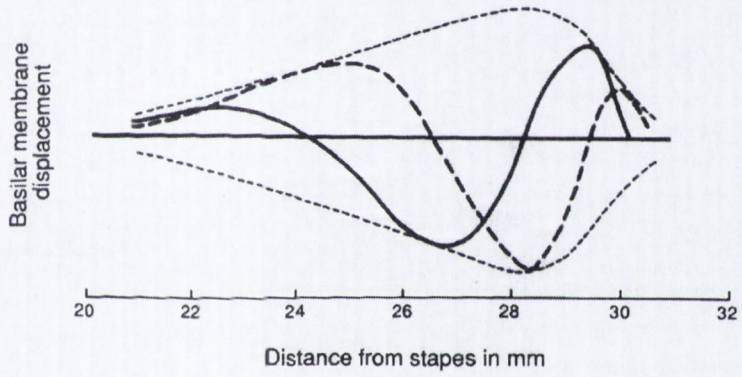


Figure 3.3: Displacement of the basilar membrane at an instant in time [108].

In order for the motion of the BM to be converted into the neural signals of the auditory nervous system, a transduction process must take place. This occurs in the organ of Corti as follows. Between the basilar membrane and tectorial membrane are two groups of hair cells known as the inner and outer hair cells. The two groups are divided by an arch known as the tunnel of Corti, the inner hair cells being on the side closest to the inside of the cochlea, and the outer hair cells being on the side closest to the outside of the cochlea. When the BM moves in response to a sound, a shearing motion is set up between it and the tectorial membrane. This has the effect of causing the hairs at the end of the hair cells to move. This causes an ionic exchange in the inner hair cells which ultimately leads to the firing of a nerve impulse. The strength of this nerve impulse (or action potential) is constant at a few millivolts. Greater stimuli levels cause an increase in the nerve firing rate. The nerve fibres themselves also display a marked frequency selectivity, each nerve firing best to a limited range of frequencies. A given nerve's tuning curve plots its firing threshold as a function of frequency. This threshold reaches a minimum at the nerve fibre's characteristic

frequency. Remarkably, the nerve fibres are positioned along the length of the cochlea in a very organised fashion so that each fibre's characteristic frequency matches closely the frequency causing maximal displacement of the BM at that position. Furthermore, temporal information about the signal is also contained in the temporal pattern of the nerve firing. This is because nerves tend to fire only at particular phases of the stimulus; the firing is said to be 'phase locked' with the stimulus. The upper frequency limit of this effect is about 4 kHz. There is general agreement that this phase locking is important for our ability to localise sounds as localisation depends on our being able to compare temporal information from both ears. However, there is still considerable disagreement over how important the effect of phase locking is to auditory masking.

While the overview presented above is a greatly simplified one, it serves to give an adequate introduction to the overall functioning of the peripheral auditory system. This thesis is primarily concerned with the phenomenon of auditory masking, comprising temporal and spectral masking. The phenomenon of spectral masking is a result of the limited frequency resolving capability of the auditory system and is strongly linked to both the basilar membrane motion and the neural activity of the inner ear. Spectral masking is reviewed in detail in Section 3.2. The causes of temporal masking are less well understood, especially backward masking. However, it is thought that a more central limitation or 'sluggishness' of the auditory system may be a cause. Temporal masking is reviewed in detail in Section 3.3 of the thesis.

3.2 Spectral Masking

One of the most fundamental concepts in auditory theory is that the ear behaves as a frequency analyser [108]. This frequency analysis occurs in the cochlea and is due to the motion of the basilar membrane [191]. In 1940, Fletcher [57] proposed that the frequency analytic behaviour of the ear be likened to a bank of continuously overlapping bandpass filters. Fletcher's filterbank concept is still very much in use today. The filterbank is commonly referred to as the *auditory filterbank* and individual filters are referred to as *auditory filters*. Glasberg and Moore describe the auditory filter as follows [63] : "The shape of the auditory filter at a given centre frequency can be thought of as a weighting function which is applied to the power spectrum of a sound to determine the effective magnitude of the output of the filter."

The ability of the ear to resolve, or distinguish between, spectral components is limited. A component can be masked (made inaudible) or partially masked (causing a decrease in

loudness) by another component which is close in frequency. This is known as *simultaneous masking* or spectral masking when the masking component (*masker*) and masked component (*signal*) are both present at the same time. Non-simultaneous masking or *Temporal masking* describes masking produced by a masker both before and after its duration and is discussed further in Section 3.3.

Spectral masking reflects the ability or inability of the ear to resolve spectral components. In other words, it reflects the ear's *frequency resolution.* By studying masking, it is possible to characterise the frequency resolution of the auditory system and to thus determine the shape and bandwidth of the auditory filter.

3.2.1 The Power Spectrum Model of Masking

In [57], Fletcher performed an experiment which measured the threshold of audibility[1] of a sinusoidal signal in the presence of a bandpass noise masker. The masker was centred on the signal frequency. The noise power density was held constant during the experiment and its bandwidth was varied. The total power of the noise masker was therefore directly proportional to its bandwidth. Fletcher found that the signal threshold increased with increasing noise bandwidth only up to a certain bandwidth, which he termed the *critical bandwidth* (CB). He noted that beyond the critical bandwidth, the signal threshold did not increase significantly. Several researchers have repeated this experiment and found similar results [65], [70]. As an example, Figure 3.4 presents data from a study conducted by Moore et al. [108]. The threshold of the signal increases for increasing masker bandwidth up to 400Hz, after which it remains constant.

In order to explain these interesting results, Fletcher proposed that the frequency selectivity of the ear be modelled as a bank of band pass filters called *critical-band filters.* The total noise power passing through a critical-band filter centred on the signal frequency therefore increases as the noise bandwidth increases only as long as the noise bandwidth is less than the bandwidth of the critical-band filter. The *power spectrum model* of masking, based on Fletcher's observations, depends on the following assumptions :

- The frequency selectivity of the ear can be modelled as a bank of linear overlapping band pass filters, known as critical-band filters.
- Only one critical-band filter, centred on the signal frequency, is used by the auditory system when attempting to detect a tone in the presence of masking noise. Thus, the

[1] In this work, all references to threshold refer to threshold of audibility.

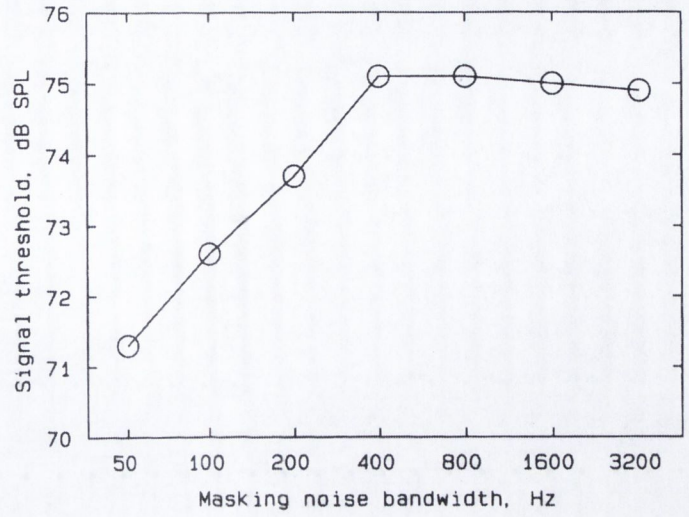


Figure 3.4: The threshold of a 2 kHz sinusoidal signal plotted as a function of the bandwidth of a noise masker centred at 2 kHz [108].

effect of the filter is to pass the signal while attenuating noise outside the bandwidth of the filter.

- Only noise components passed by the critical-band filter produce masking.
- The signal masking threshold corresponds to a ratio K between signal power and noise power at the output of the critical-band filter. Furthermore, this output is determined by the long-term power spectra of signal and noise—short term temporal fluctuations of the masker are ignored, as is signal phase.

In order to calculate the bandwidth of the auditory filter, or critical bandwidth, Fletcher carried out a second experiment which measured the threshold (of audibility) of a tone as a function of frequency in the presence of a white noise masker. Fletcher assumed that the filter shape could be approximated by a rectangular shape. All components within the passband of such a filter are passed equally, while those outside the passband are completely removed. By using the assumptions of the power spectrum model, Fletcher derived the following equation which relates critical bandwidth, CB, to noise power density, N_0, and signal power, P_s, at threshold :

$$P_s = K(CB \times N_0) \ . \tag{3.1}$$

Thus,

$$CB = P_s/(K \times N_0) \ . \tag{3.2}$$

Figure 3.5 illustrates the noise, filter and signal power spectra. The total noise power passing through the rectangular filter is $CB \times N_0$. Thus, the power of the signal at threshold depends on the signal-to-noise ratio, K, multiplied by the noise power, $CB \times N_0$. In order to determine the value of CB, Fletcher made one final assumption—he assumed the value of K to be 1. Thus the value of CB was easily estimated as P_s/N_0. This ratio is now known as the *critical ratio* (CR).

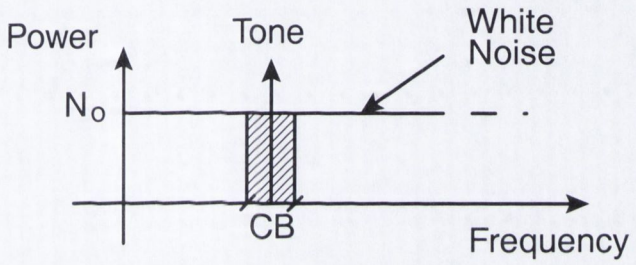


Figure 3.5: Power spectra of masking noise, signal and rectangular filter used by Fletcher to estimate the critical bandwidth CB

Subsequent studies [166], [177], [134], [114] show that a value of 0.4 for K is a better estimate, and also that K varies with centre frequency. Furthermore, none of Fletcher's assumptions for the power spectrum model are completely true. It is now known that the auditory filters are non-linear [111], although the frequency selectively of the ear can adequately be modelled using filters that are piece-wise linear. Signal detection may also be influenced by noise occurring outside the passband of an auditory filter centred on the signal frequency [68]. Also, information from several auditory filters may be combined to assist in signal detection [174], [25]. The auditory filter may not always be centred on the signal frequency—if the signal stimulates a large frequency region a higher signal-to-noise ratio may exist in a frequency region offset from that of the signal. This is known as *off-frequency listening.* Finally, the temporal fluctuations inherent to noise maskers (especially for narrow-band noise) can influence the masking they produce—subjects can 'listen to the dips' in the temporal masker to aid signal detection [109].

Despite these deviations, the power spectrum model has proved very useful in deter-

mining the shape of the auditory filter. This is all the more true if the experiment is deliberately designed so as not to strongly violate the assumptions of this model. The next section describes such an experiment.

3.2.2 The Notched-Noise Method

In order to limit the effect of off-frequency listening on the estimation of the auditory filter shape, Patterson [130] proposed a masking experiment known as the *notched-noise method.* This method utilises a noise masker with a spectral notch centred on the signal frequency. This effectively removes the advantage of off-frequency listening and the auditory filter with the highest signal to noise ratio can be assumed to be centred on the signal frequency. Furthermore, there is little advantage in combining the outputs of several auditory filters distributed over a range of frequencies [130], [134].

The notched-noise method is useful because it distinguishes between the frequency selectivity and detection efficiency of the auditory system [138], thus correcting a flaw in Fletcher's original experiment. Figure 3.6 illustrates the experiment. The signal frequency is represented on the graph as the vertical line at frequency f. The two rectangular areas on either side represent the noise. The auditory filter is represented by a triangular shape. The notch is placed symmetrically around the signal; each edge is separated by Δf from the signal centre frequency. Noise passing through the auditory filter is shown as two shaded areas; the total noise power passing through the auditory filter is proportional to this area. As the spectral notch widens, less noise passes through the auditory filter, causing a reduction in the signal threshold. This experiment assumes that the shape of the auditory filter is symmetric on a linear frequency scale. This is a reasonable assumption at least for the peak of the filter shape and for moderate sound levels [108] (asymmetric filter shapes are discussed further in Section 3.2.4).

According to the power-spectrum model of masking, the power P_s of the signal at threshold may be estimated as :

$$P_s = P_s = KN_0 \int_0^{f_c-\Delta f} W(f)\,df \quad + \quad KN_0 \int_{f_c+\Delta f}^{\infty} W(f)\,df\ , \tag{3.3}$$

where N_0 is the power spectral density of the noise in its passbands, K is the signal-to-noise ratio at threshold, f_c is the centre frequency of the signal, Δf is the frequency deviation and $W(f)$ is the auditory filter shape to be derived. This equation gives a relationship between the integral of the auditory filter shape, $W(f)$, and signal power at threshold, P_s, for a notch width of $2\Delta f$. Therefore, the shape of the auditory filter at $f + \Delta f$ can be calculated from the differentiation of the function relating the signal threshold to the notch

width, $2\Delta f$ [107] (see Section 3.2.4, Equation 3.9).

In this particular case, where a symmetric auditory filter is considered, the value of each half of the integrand in Equation 3.3 is equal.

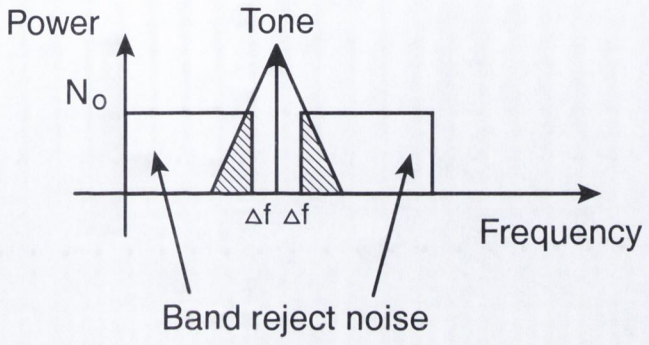


Figure 3.6: Illustration of Patterson's notched-noise experiment. The vertical line at centre frequency f is the signal. The notch is placed symmetrically around the signal; the upper and lower noise edges are separated from the signal by frequency deviation Δf. The triangular region represents the auditory filter shape and the shaded areas represent the noise passing through this filter.

In order not to violate the assumptions of the power spectrum model of masking, the type of masker used must be carefully considered. There are several reasons for using the notched-noise masker described above. As previously stated, placing the notch around the signal frequency is an effective way of reducing off-frequency listening. It is reasonable to ask, however, whether two sinusoidal maskers placed symmetrically around the signal would have served just as well. This is not the case, unfortunately, since in general, sinusoidal maskers interact with the signal to produce *beats* [194], [108], [109]; such beats are much less obvious for noise maskers. Beats form an additional cue to the presence of the signal. If the (sinusoidal) masker and signal frequencies are close, such beats are perceived as a periodic fluctuation in loudness. As the separation between masker and signal increases, the beat rate increases, causing a sensation of roughness or dissonance [154]. Therefore, these beats provide the listener with a cue to masker-signal separation, and therefore to masker frequency in the power spectrum model, for which the signal frequency is fixed. As such, the presence of beats violates the assumption of the power spectrum model that the threshold corresponds to a constant signal-to-masker ratio at the output of the auditory

filter [106], [14].

In addition to choosing noise maskers over sinusoidal maskers, broadband noise is preferred to narrowband noise. This is because narrowband noise has relatively slow temporal fluctuations compared to broadband noise. Such slow temporal fluctuations can interact with the sinusoidal signal in much the same was as described above for sinusoidal maskers, resulting in audible cues. Therefore, for the reasons outlined above, maskers consisting of broadband noise with a spectral notch centred at the signal frequency are used to best satisfy the assumptions of the power spectrum model.

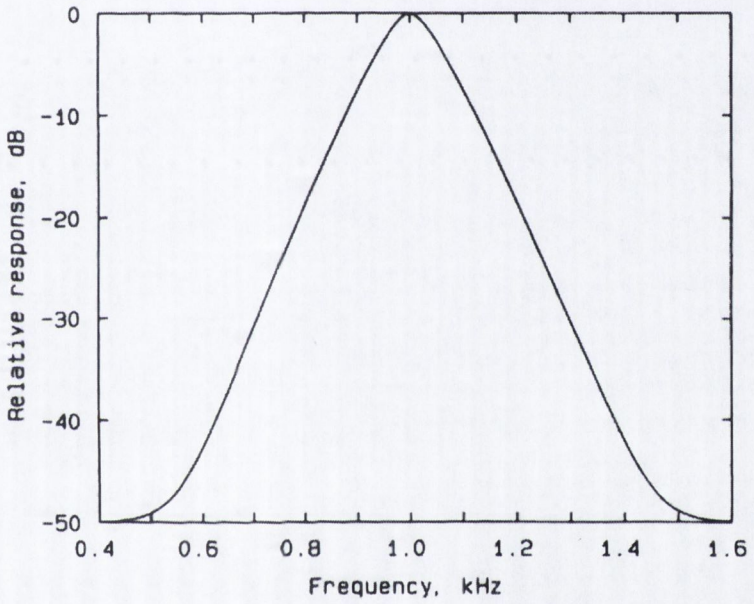


Figure 3.7: Illustration of a typical auditory filter shape derived using Patterson's notched-noise method. The centre frequency of the filter is 1 kHz [108].

Figure 3.7 graphs a typical auditory filter shape derived using the notched-noise method. The filter is symmetric and has a rounded top and steep skirts. The *roex* family of filter shapes, to be discussed in Section 3.2.4, are often employed to characterise auditory filter shape using a small number of free parameters. The *equivalent rectangular bandwidth* (ERB) [107] is a useful summary measure of auditory filter shape. Put simply, the equivalent rectangular bandwidth is a measure of the bandwidth of an idealised bandpass filter with equal peak level and passing equal total power to the auditory filter.

Figure 3.8 presents an estimation of ERB by several researchers, using Patterson's method. There is a good deal of agreement between the different sets of results. The solid line plots the following function relating ERB to centre frequency, f, (in kHz) :

$$ERB = 24.7(4.37f + 1) \tag{3.4}$$

This equation can be seen to fit the ERB data well. For comparison, an estimation of Zwicker's 'classical' critical bandwidth [195] is also plotted. The critical bandwidth can be seen to flatten off for frequencies below 500 Hz, unlike the ERB, which continues to decrease below this frequency. ERB is currently thought to be a more accurate measure of auditory filter bandwidth that the traditional value of critical bandwidth. This is because the effectiveness of the detector system following the auditory filter varies with centre frequency; therefore indirect measurements of the critical ratio may be in error, leading to an incorrect measure of critical bandwidth [110], [108].

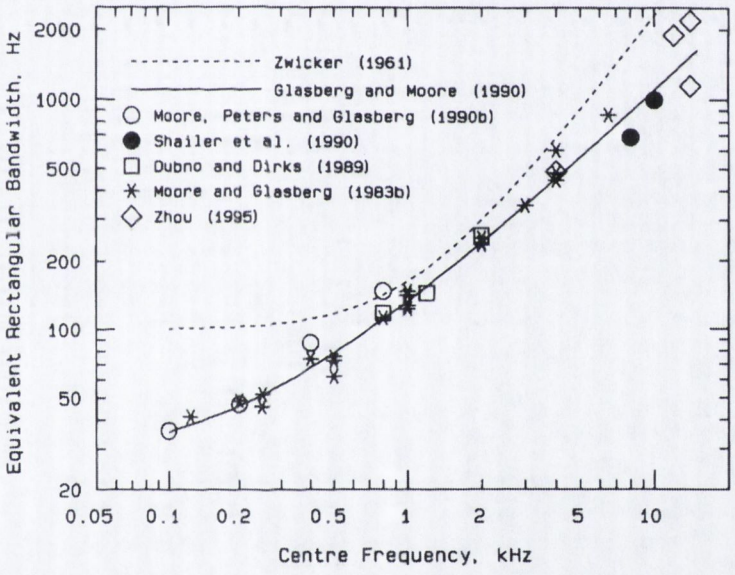


Figure 3.8: A summary of ERB measurements made by several researchers—the solid curve plots a fit to this data [108]. The dashed line plots Zwicker's [193] earlier measurements of the critical bandwidth, which is seen to flatten for frequencies below 500 Hz.

The power spectrum model of masking, outlined above, is the most widely used method of determining the shape of the auditory filter. Modifications to this procedure allowing for non-symmetric filter shape, level dependent filter skirts, off-frequency listening and the transmission characteristics of the outer ear, middle ear and transducers are summarised in Section 3.2.4. For completeness, we also mention two other methods of determining the auditory filter shape: the *rippled-noise method* [107] and the measurement of *psychophysical tuning curves* (PTCs) [194]. Briefly, the rippled-noise method measures signal threshold

in the presence of *comb-filtered noise*. The filter shape can be calculated using the method described above for the notched-noise method or by expressing the filter shape in terms of a Fourier series. The filter shapes derived are quite similar to those derived using the notched method, but have a slightly broader top. The method is however unsuited to measuring the auditory filter shape over a wide dynamic range [107]. The measurement of psychophysical tuning curves measures the masker level (usually a narrow band noise) necessary to just mask a sinusoidal signal of fixed frequency and level. The signal level is typically very low (say 10 dB SPL); the method depends on the assumption that only one auditory filter is involved in its detection. The method also assumes that at threshold, the masking noise produces a constant output from the auditory filter—enough to just mask the signal. Each PTC is a plot of this masking threshold versus frequency. Since the method also assumes that the auditory filter is linear, inverting the PTC yields an estimation of the filter shape. Unfortunately, there is strong evidence to show that off-frequency listening can result in an overestimation of the masking threshold [107], since the masking noise is not centred around the signal. Furthermore, the assumption that the auditory filter is linear leads to an underestimation and overestimation of the lower and upper auditory filter skirts, respectively. Therefore, in general, the notched-noise method is generally thought to be the most effective method of determining the auditory filter shape.

3.2.3 Critical-Band Rate Scale and Equivalent Rectangular Bandwidth Scale

Because auditory filter bandwidth increases with frequency it is sometimes preferable to plot psychoacoustic data on a logarithmic rather than linear frequency scale, the logarithmic scale being more perceptually relevant. A further improvement on the logarithmic scale is the *critical-band rate* scale which relates critical bandwidth to frequency. The scale is constructed by adding one critical-band to the next, letting the upper frequency limit of one critical-band equal the lower frequency limit of the next critical band. In such a manner, the intersections of adjacent critical bands denotes a point on the critical-band rate scale. The unit of the scale is the Bark, named after Heinrich Barkausen [69]. Table 3.1 demonstrates the construction of the scale. It takes 24 Barks to cover the frequency range 0 to 16 kHz. The relationship between Barks, z and frequency in kHz, f, can be approximated by the following equation [107] :

$$z = [26.8/(1 + 1.96/f)] - 0.53 \ . \tag{3.5}$$

The critical-band rate scale is derived from the value of critical bandwidth. The cor-

z	f_l, f_u	f_c	z	Δf_G	z	f_l, f_u	f_c	z	Δf_G
Bark	Hz	Hz	Bark	Hz	Bark	Hz	Hz	Bark	Hz
0	0				12	1720			
		50	0.5	100			1850	12.5	280
1	100				13	2000			
		150	1.5	100			2150	13.5	320
2	200				14	2320			
		250	2.5	100			2500	14.5	380
3	300				15	2700			
		350	3.5	100			2900	15.5	450
4	400				16	3150			
		450	4.5	110			3400	16.5	550
5	510				17	3700			
		570	5.5	120			4000	17.5	700
6	630				18	4400			
		700	6.5	140			4800	18.5	900
7	770				19	5300			
		840	7.5	150			5800	19.5	1100
8	920				20	6400			
		1000	8.5	160			7000	20.5	1300
9	1080				21	7700			
		1170	9.5	190			8500	21.5	1800
10	1270				22	9500			
		1370	10.5	210			10500	22.5	2500
11	1480				23	12000			
		1600	11.5	240			13500	23.5	3500
12	1720				24	15500			
		1850	12.5	280					

Table 3.1: Construction of the critical-band rate scale with z indicating the critical-band rate, f_c being centre frequency, and f_l, f_u and Δf_G being the lower frequency, upper frequency and bandwidth of each critical band. Data taken from [194].

responding rate for the value of ERB is the ERB rate E. A formula relating number of ERB's, E to frequency f in kHz, derived from Equation 3.4 is

$$E = 21.4 \log_{10}(4.37f + 1) \ . \tag{3.6}$$

The ERB scale is often used as a frequency scale for plotting the output of the entire auditory filterbank to a given stimulus. Such representations are known as *excitation patterns* and are discussed in Section 3.2.5. It is easy to confuse equivalent rectangular bandwidth, ERB, at a given frequency, f, with the number of ERBs, E, corresponding to that frequency. The equivalent rectangular bandwidth refers to the bandwidth of the auditory filter at frequency, f, whereas the number of ERBs relates frequency, f, to its corresponding position on the ERB scale.

3.2.4 Auditory Filter Shape

This section presents a brief historical review of early research into determination of auditory filter shape. This is included to highlight several different models of auditory spectral resolution and to bring to the attention of readers not familiar with psychoacoustics several issues relevant to auditory filter shape including off-frequency listening, filter asymmetry and filter level dependency. Four current important filter shapes are then discussed : the *rounded exponential* or *roex* family, the *gammatone* filter (GTF), the *gammachirp* filter (GCF) and the *all-pole gammatone* filter (APGF).

Critical Bandwidth and the Rectangular Filter

As described in Section 3.2.1, Fletcher's early work on critical bandwidth and masking led to an important theory which persists to this day, known as the power-spectrum model of masking. Fletcher proposed that the ear be modelled as a bank of continuously overlapping bandpass filters with bandwidth equal to the critical bandwidth[2]. In order to estimate the auditory filter bandwidth using this model (Equation 3.2), Fletcher assumed a simple rectangular filter shape[3]. Fletcher argued that, at threshold, the signal power and noise power passing through such a filter were equal; from this he deduced that the signal-to-noise ratio P_s/N_0 gave a measure of the critical bandwidth. He concluded that the width

[2]The idea of using a filterbank to model auditory frequency selectivity was actually suggested as early as 1863 by Helmholtz [71]

[3]Fletcher was, of course, well aware that the 'true' auditory filter shape was highly unlikely to be rectangular, however this shape was a convenient one; at the time Fletcher was more interested in basilar membrane excitation and was not overly concerned with determining the exact auditory filter shape.

of the auditory filter (critical bandwidth) increased in proportion to increasing frequency. Overall, Fletcher found the above model very useful for accurately predicting the threshold of sinusoidal signals masked by noise. Similar experiments to Fletcher's were conducted by a number of researchers at the time [65], [70] and are in good agreement with his results.

The Universal Resonance and Gaussian Filter

In order to investigate a more plausible filter shape than the rectangular filter, Schafer *et al.* [165] predicted threshold curves using both the rectangular filter shape and the *universal-resonance* filter shape and compared these to threshold data from a masking experiment. One of the limitations of the rectangular filter is that is has unity gain within its passband and complete attenuation outside its passband. Therefore, threshold curves predicted using this model of auditory filtering show discontinuities. As the noise bandwidth inside the passband of a rectangular filter increases, predicted threshold increases linearly. Once, the noise bandwidth exceeds the filter bandwidth, however, the threshold remains constant. There is therefore a discontinuity at the point where the noise bandwidth equals the filter bandwidth. Threshold curves predicted using the universal resonance filter, however, do not suffer from this discontinuity. Such a breakpoint was not present in the experimental data of Schafer et al.; it was therefore concluded that the universal resonance filter was a more realistic filter shape than the rectangular shape.

Webster *et al.* [184] also compared the rectangular and universal resonance filter shapes. They measured sinusoidal signal threshold in the range 400 to 1400 Hz in the presence of bandstop filtered white noise, with the bandstop edges at 600 and 1200 Hz. They found the universal resonance filter shape to be superior to the rectangular shape in predicting the results of this masking experiment. In order to obtain a more accurate estimation of the critical bandwidth, Swets *et al.* [175] assumed one of three shapes for the auditory filter : universal resonance, rectangular and Gaussian. They measured the threshold of a 1000 Hz tone in the presence of a broadband noise masker of variable width centred on the tone. The noise was generated by passing white noise through a universal resonance filter with a 1000 Hz centre frequency and 3 dB bandwidth which varied from 12 to 90 Hz. Critical bandwidth estimations were 91, 41 and 95 Hz for the rectangular, Gaussian, and universal resonance filter shapes respectively. They concluded that the estimated value of critical bandwidth depended strongly on the assumption of a particular auditory shape.

The Symmetric Filter

In 1974, Patterson [129] proposed a method of determining not only auditory filter bandwidth, but auditory filter shape. Unlike earlier researchers, no initial assumptions were made about this shape. The experiment Patterson used measured the threshold of a sinusoid of constant frequency in the presence of lowpass noise[4]. The cutoff of the noise was increased from far below the signal frequency to just above it. Figure 3.9 presents the long-term power spectrum of signal and noise and also the hypothetical filter shape. The shaded area represents the noise power passed by the auditory filter.

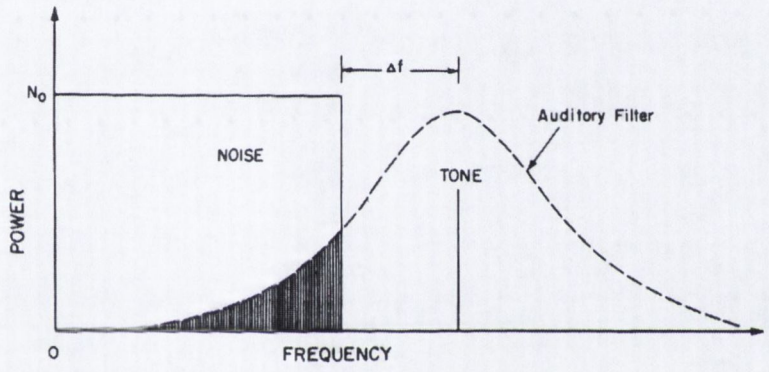


Figure 3.9: Representation of masker, tone and hypothetical auditory filter shape in Patterson's masking experiment [129] to determine auditory filter shape.

Following the power-spectrum model of masking, Patterson assumed that the sinusoid reached threshold once the signal power and noise power passed by the auditory filter reached some constant proportion K :

$$P_s = K \int_0^\infty N(f)W(f)df \ , \tag{3.7}$$

where P_s is the signal power, $N(f)$ is the noise power-spectrum and $W(f)$[5] is the auditory filter power-spectrum. As stated earlier, the cutoff of the lowpass noise was very sharp.

[4]As special efforts were made to ensure the noise cutoff was very sharp, the actual noise used in the experiment was bandpass. However, the lower edge of the noise was remote enough from the signal (and the auditory filter) to ensure that the masking threshold was not influenced.

[5]Originally, Patterson uses the notation $|H(f)|^2$ instead of $W(f)$ to describe the amplitude squared of the auditory filter transfer function. Subsequently, following others, the term $W(f)$ is here used to describe the auditory filter in terms of a weighting function acting on the energy or power spectrum of the stimulus. For clarity therefore, we use $W(f)$ throughout.

Therefore Eq. 3.7 reduces to :

$$P_s = K\,N_0 \int_0^{f_c} W(f)df \ , \tag{3.8}$$

f_c being the the cutoff frequency. Patterson pointed out that the lower half of the auditory filter shape, $W(f)$, could be calculated by simply differentiating Eq. 3.8 :

$$\frac{dP_s}{df_c} = K\,N_0 \frac{d}{df_c}\Big[\int_0^{f_c} W(f)df\Big] \quad = \quad K\,N_0\,W(f_c) \ . \tag{3.9}$$

In order to determine the upper half of the auditory filter, the experiment was repeated exactly as described above, but with a highpass noise masker. Patterson found that, unexpectedly, the resulting auditory filter shape was symmetric on a linear frequency scale—he referred to the shape as the *symmetric filter*. This shape has a simple expression[6] :

$$W(\Delta f/f_0) = \frac{1}{[(\Delta f/\alpha f_0)^2 + 1]^2} \ , \tag{3.10}$$

where Δf is the frequency deviation from filter centre frequency and α is a frequency dependent parameter which controls the bandwidth of the filter.

Further research into auditory filter shape includes the work of Margolis and Small [102] who performed a similar experiment to Patterson, measuring the detectability of a tone of constant power and frequency as a function of the cutoff of digitally generated noise. Their results agree closely with those of Patterson, although they are slightly less symmetric. Also, Houtgast [75], [76] estimated the shape of the auditory filter by measuring the threshold of a tone of fixed frequency and level in the presence of a continuous, rippled (comb filtered) noise masker[7]. Houtgast varied the delay time and polarity of the comb filtered noise and measured the noise level at which the tone was just audible. By representing the auditory filter shape as a Fourier series, and using a masker with a cosinusoidal or sinusoidal spectrum, Houtgast was able to derive the auditory filter shape from the masking data.

Off-Frequency Listening and Filter Asymmetry

The filter shapes determined by Patterson and by Margolis and Small show a relatively flat section about the centre frequency and quite steep skirts. However, one of their assumptions

[6] In [98], Lyon points out that this is also the expression for the response of an order 2 gammatone filter, which is discussed shortly.

[7] Rippled noise is generated by adding delayed copy of a white noise signal to itself. When the delayed noise has the same polarity as the original, the rippled noise spectrum is cosinusoidal, having first peak at 0 Hz subsequent peaks every $1/T$ Hz, where T is the delay. This is sometimes referred to as *cosine+*. In a dual manner, when the delayed noise has the opposite polarity as the original, the rippled noise spectrum is sinusoidal, having a dip at 0 Hz and first peak at $1/2T$ Hz. This is sometimes referred to as *cosine-*.

is that the auditory filter remains centred on the signal throughout the experiment. In other words, off-frequency listening is ruled out. However, for the particular combination of lowpass noise masker and signal shown in Figure 3.9, and because the peaks of their derived filter shapes are relatively flat, a significant increase in signal-to-noise ratio can be achieved by centering the filter somewhat to the upper frequency side of the tone. This meant that the bandwidths of the filter shapes they derived were too small.

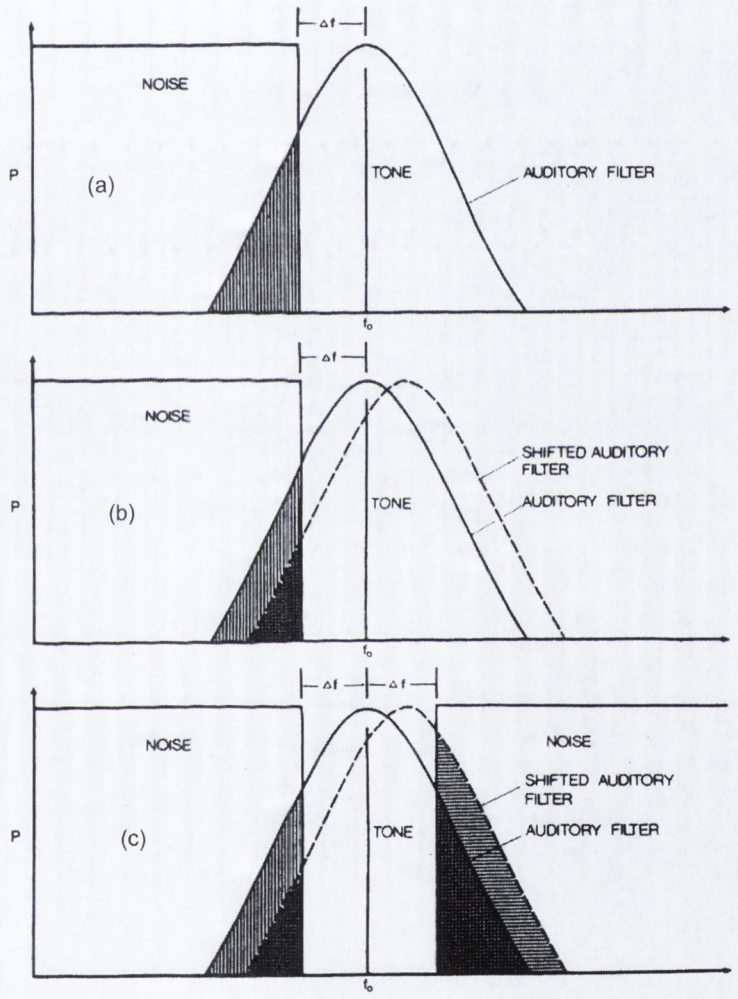


Figure 3.10: Tone plus notched-noise masker. (a) Tone plus lowpass noise (b) Use of off-frequency listening to improve the signal to noise ratio (c) Upper noise band used to limit off-frequency listening [130].

In order to limit the extent of off-frequency listening, Patterson [130] suggested the use of a notched-noise masker and assumed that the filter shape was symmetrical. The masker

and signal power spectrum are portrayed in Figure 3.10 (c); the signal tone is centred at f_0 and placed inside the notch. The cutoffs of the lowpass and highpass sections of the noise are placed equidistant from the tone. Figure 3.10 (a) and (b) illustrate off-frequency listening; it is clear that utilising a filter placed above the tone results in an improvement in signal-to-masker ratio. However, as Figure 3.10 (c) demonstrates, this improvement is effectively lost if a notched-noise masker is used and the filter is reasonably symmetric. The advantage of this approach is that filter position is restricted to a narrow region about the tone frequency; hence the assumption that off-frequency listening does not occur is a reasonable one—the filter can be assumed *fixed.* Patterson noted that this method yielded filters with similar skirt slopes to those of [129] and [102]. However, the bandwidths were wider and compared well with those of Houtgast. Weber [183] also used the notched-noise masker/symmetric filter model to investigate the influence of masker level on filter bandwidth and noted an increase in filter bandwidth with level. Patterson and Henning [132] reversed the roles of tone and noise—they measured the threshold of a narrowband noise signal centred between two tone maskers. The resulting filter shape is very close to that yielded by the notched-noise masker of [130].

The assumption that the filter shape is symmetrical is in reality only valid at moderate stimulus levels. For increasingly higher levels, the slope of the lower skirt decreases, resulting in the well known *upward spread of masking.* If the auditory filter is asymmetric, then off-frequency listening is likely to occur, even for notched-noise maskers placed symmetrically about the signal. Thus, there is an important *interaction* between filter asymmetry and off-frequency listening—neither effect can be studied in isolation. To investigate the degree of auditory filter asymmetry, Patterson and Nimmo-Smith [136] proposed a new notched-noise masker experiment similar to his earlier model discussed above [130]. The experiment was designed to promote off-frequency listening; the notched-noise was *not* placed symmetrically about the signal tone; either the lower or upper noise band edge was closer to the signal. The signal frequency was fixed at 2 kHz. The effect of the asymmetrical masker in each case was to force the auditory filter to move away from the signal tone towards the centre of the notch. The filter was assumed to shift to a centre frequency maximising the signal-to-noise ratio at its output; this was referred to as the "max(S/N)" assumption. This shift towards the centre was larger for the former case (signal near lower noise band), because the lower skirt of the auditory filter was shallower than the upper skirt. For the duration of the experiment, the noise bandwidth and level were fixed and only the location of the noise bands was shifted. In a similar manner to previously, Patterson and Nimmo-Smith measured the threshold of tone audibility as the noise bands were shifted. To provide a best

fit to the data, they present strong arguments in favour of using a filter shape produced by the product of two back-to-back decaying exponentials and a cubic polynomial. The exponentials are responsible for the overall decay of the filter shape, whereas the cubic polynomial produces a more rounded filter top and also provides the shallower filter tails by temporarily counteracting the exponential decay. The filter shape may be expressed as

$$W(f) = [1 + \lambda g + a_2(\lambda g)^2 + a_3(\lambda g)^3]e^{-\lambda g} \ . \tag{3.11}$$

The filter parameters a_2, a_3 and λ are different for the lower and upper halves of the filter shape. Parameter g is normalised frequency ($g = |f - f_c|/f_0$), the exponential factor λ controls the slope of the steep filter skirts. The quadratic term is responsible for rounding the top of the filter and the cubic term generates the more shallow filter tails. Patterson and Nimmo-Smith examined the effect on filter asymmetry of allowing different combinations of filter parameters to vary. They concluded that the best and most convenient fit was achieved by simply allowing the parameter λ to be higher for the upper half of the filter shape. The filter is thus λ-*asymmetric* and the upper half can be considered a compressed version of the lower half of the filter shape. Patterson and Nimmo-Smith's main conclusions were that by allowing for off-frequency listening, a much better fit to the masking data was achieved. The rounded-exponential shape was also a convenient filter shape analytically as it was easy to calculate its integral. The experiment yielded filter shapes having a flat top, steep skirts (typically 100 dB/octave) and more shallow tails (30-50 dB/octave) located about 30 dB below the filter top. The 3-dB bandwidth of the filter was approximately equal to 11% of the centre frequency. Figure 3.11 compares the asymmetric rounded-exponential filter shape (bold solid curves) with a rounded-exponential filter shape (faint solid curves) derived for the case of symmetric masker and fixed, symmetric filter. The filter shapes are plotted on both linear and logarithmic ordinates. The lower skirt and tail of the asymmetric filter can be seen to have shallower slopes than those of the fixed, symmetric filter shape. In addition, the figure plots average, symmetric filter shapes at 2 kHz derived in [130] for the fixed filter, notched-noise experiment (bold dashed curves) and the average, symmetric filter shapes derived in [129] for the fixed filter, lowpass noise experiment (faint dashed curves). For the former, the filter bandwidth is slightly less than that of the λ-asymmetric filter shape, indicating that the fixed filter assumption is reasonable for symmetric, notched-noise maskers if the auditory filter can be assumed symmetric. For the latter, however, the filter is notably narrower than all others. The fixed filter assumption in this case is invalid and leads to an underestimation of filter bandwidth. Overall the authors conclude that the λ-asymmetric filter shape provides the best fit to the asymmetric notched noise data.

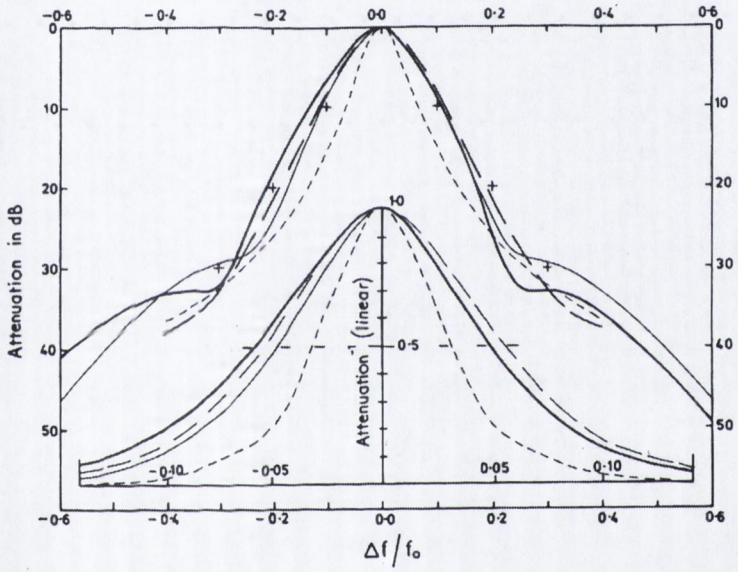


Figure 3.11: The bold and faint solid curves plot the the λ-asymmetric and symmetric rounded-exponential filters, respectively [135]. Also shown are the average, fixed symmetric filters from [130] (bold dashed curves) and [129] (faint dashed curves). The filters are plotted on both linear and logarithmic ordinates.

An examination of the evolution of auditory filter shape from Fletcher's original rectangular filter and critical bandwidth measure, proposed in 1940, to Patterson and Nimmo-Smith's λ-asymmetric filter shape, proposed in 1980, shows a clear pattern of development. Over this period, the assumptions of filter symmetry and lack of off-frequency listening are dropped as the experiments to determine filter shape become more sophisticated. This leads to the development of increasingly realistic and accurate filter models. Since then, with the advent of digital signal processing, mobile communications and perceptually based audio compression, there there has been a significant increase in the amount of research into auditory modelling and perception, and a corresponding increase in the number of auditory filter models. It is therefore convenient to present in isolation several current, well known filterbank models which are relevant to this work, namely the *rounded exponential* or *roex* family of filter shapes, the *gammatone* (GTF) filter, the *gammachirp* filter (GCF) and the *all-pole gammatone* filter (APGF).

The Roex Family of Filter Shapes

In [138], Patterson, Nimmo-Smith, Weber and Milroy suggest a family of filter expressions approximating the auditory filter shape using a small number of parameters. Each has the form of a decaying exponential with a rounded top; the family of expressions is therefore referred to as the *roex* family. The simplest member, the *roex(p)* shape, is a minor modification of the rounded-exponential filter of [136] and is expressed as :

$$roex(p) = (1 + pg)\exp(-pg) \quad , \tag{3.12}$$

where g is normalised frequency ($g = |f - f_c|/f_0$), and p controls both the filter bandwidth and filter skirt slope. The ERB of this filter is $4f_c/p$ for the case of a symmetric filter. For asymmetric filters, p has values p_l and p_u on the lower and upper frequency sides of the filter, controlling the slopes of the lower and upper filter skirts respectively. The ERB is $[2f_c/p_l + 2f_c/p_u]$ for such filters. The indefinite integral of the roex(p) filter can easily be derived, hence the filter form is convenient for predicting masking thresholds.

The *roex(p,w,t)* shape may be expressed as follows :

$$roex(p, w, t) = (1 - w)(1 + pg)\exp(-pg) + w(1 + tg)\exp(-tg) \; . \tag{3.13}$$

This shape is useful in masking experiments covering a wide range of levels, where auditory filter shape with shallow tails are needed to account for masking data. The filter shape is the sum of two rounded exponentials, the second term $w(1+tg)\exp(-tg)$ gives the shallower

filter tail; w determines the relative weight of the filter tail and t determines its slope. For asymmetric filter shapes, the parameters p and w may vary for the upper and lower halves of the filter shape.

As a convenient approximation to the $roex(p, w, t)$ shape, Patterson *et al.* [138] introduce the $roex(p, r)$ shape, which may be expressed as :

$$roex(p, r) = (1 - r)(1 + pg)\exp(-pg) + r \ . \tag{3.14}$$

The term r introduces a flat tail or dynamic range limit, and is intended to approximate the very shallow tail of the auditory filter outside its passband. It is noted by Patterson that the parameters of the $roex(p, r)$ shape are more stable than those of the $roex(p, w, t)$ shape.

The roex family has been extensively used to model auditory filtering. Moore and Glasberg [111] use the roex model to characterise auditory filter level dependency and present a convenient method for calculating the level dependent excitation patterns of arbitrary signals. This work is subsequently refined by Glasberg and Moore in [63] and results in a better fit to masking data for very low and high frequencies and for hearing-impaired subjects. Rosen and Stuart [163] use the roex shape to address the question as to whether signal or masker level should be kept constant in notched-noise experiments investigating auditory filter level dependency. In contrast to Glasberg and Moore [63], they present convincing arguments in favour of keeping the signal tone constant and varying the masker level instead. More, recently, the roex filter has been used by Fagelson and Champlin [52] to investigate non-linearities in auditory filtering for neighbouring centre frequencies and by Rosen, Baker and Darling [164] to derive level-dependent auditory filter shapes using a variable level, notched-noise masker.

The Gammatone Filter

The *gammatone* filter (GTF) was originally introduced as early as 1960 by Flanagan [56] to model the basilar membrane. It was also used by Johannesma [83], de Boer [41], [42] and Carney and Yin [26] to characterise "revcor" measurements of auditory nerve fibre impulse responses. An excellent discussion of the history of the gammatone filter is presented by Lyon [98] and by Cooke [35].

The gammatone filter may be expressed as :

$$g(t) = at^{n-1}\exp(-2\pi bt)cos(2\pi f_c t + \phi) \qquad t > 0 \tag{3.15}$$

The equation is the product of a "gamma" envelope and a cosine "tone"; frequency, f_c, sets the filter centre frequency with the phase term, ϕ, being usually set to zero for convenience. The parameter b controls the bandwidth of the filter as it sets impulse response length, and n is the filter order—it determines the slope of the filter skirts. The equivalent rectangular bandwidth (ERB) of the gammatone filter is $0.982b$. In [69], Hartmann summarises the advantages of this filter :

- Unlike the roex filter shape, which only describes the energy spectral density of the auditory filter shape, the gammatone filter is expressed in the form of a causal impulse response. The filter is therefore realizable, and is actually minimum phase [137]. Furthermore, a highly efficient IIR representation of the filter has been derived by Slaney [173].
- Gammatone filters of order 4 have an amplitude response which matches very well with that of the symmetric *roex(p)* filter shape of equal ERB.
- The gammatone filter impulse response corresponds well to "revcor" measurements of auditory nerve fibre impulse responses [83], [41], [42], [26]. It therefore provides a link between physiological and psychoacoustic data [143].

The gammatone filterbank is an array of gammatone filters designed to simulate the spectral processing of the auditory system. Gammatone filterbanks have been extensively used in auditory modelling by many researchers including Patterson [137], [133], [167] in the 'Auditory Image Model', Meddis and Hewitt [103] for perceptual modelling and Cooke [35] for auditory scene analysis.

The Gammachirp Filter

The *gammachirp* filter was introduced recently by Irino [78] as an alternative to the gammatone filter. The filter is derived as the function satisfying minimum uncertainty in joint time-scale representations [32]. A scaled signal is simply a stretched or compressed version of an original signal. When the speed of a signal is varied, for example, using a variable speed dictaphone, the pitch of the sound source changes but the perceived source does not [78]. For example, after a scale change, a voice is generally still recognisable as a voice, a violin as violin and so on. In a mathematically analogous manner, the scale transform, $S(p)$, of a signal $s(t)$ is invariant to scale changes (such as a compression in time by factor a below), apart from a change of magnitude :

$$\text{If} \quad s(t) \Rightarrow S(p), \quad \text{then} \quad s(at) \Rightarrow a^{-p}S(p) \tag{3.16}$$

Irino postulates that the role of the cochlea is to produce a time-frequency representation of sound suitable for subsequent processing by a scale transform. Scale transforms are useful as the scale transform of a signal can be viewed as representing the pitch and timbre of the signal [78], with the pitch corresponding to the magnitude, a^{-p} and the timbre corresponding to the invariant distribution $S(p)$. He argues that the filtering occurring in the cochlea be considered optimal for this scale representation and derives the gammachirp function to satisfy this optimality criterion.

The gammachirp response consists of the product of a gamma envelope and a chirp and may be expressed as :

$$g_c(t) = at^{n-1}\exp(-2\pi bt)cos(2\pi f_c t + c\ln t + \phi) \qquad t > 0 \tag{3.17}$$

The expression is almost identical to the expression for the gammatone filter, except for the inclusion of the log-time phase term, $c\ln t$. This term controls filter asymmetry. In [78], Irino finds that the gammachirp provides an excellent fit to notched-noise masking experiments and has the advantage that its impulse response is defined, unlike the roex family of filter shapes. An efficient IIR implementation of the gammachirp filter is also developed by Irino [79], composed of an IIR gammatone filterbank followed by an IIR asymmetric compensation filterbank. The time-varying asymmetric filterbank is controlled by the overall filterbank output; this models auditory filter level dependency. Since the inverse IIR gammachirp filterbank is stable, the filterbank and its inverse can be configured as an analysis/synthesis system.

The All-Pole Gammatone Filter

The *all-pole gammatone filter* (GTF) was originally introduced by Slaney [173] as an efficient approximation of the gammatone filter. The Laplace transform of the gammatone filter may be expressed as follows [99], [96] where ω replaces $2\pi f_c$:

$$H_{GT} = \frac{e^{j\phi}[s + b + i\omega_r]^N + e^{-i\phi}[s + b - i\omega_r]^N}{[(s + b)^2 + \omega_r^2]^N} \quad . \tag{3.18}$$

Unfortunately, the roots of the numerator (zeros) fall on the real axis in the s plane; their location depends on the parameters of the filter in a complicated way which makes modelling of level dependent gain using the filter difficult. The properties of the poles, however are much easier to compute as the denominator simply consists of a complex conjugate set of order-N poles. Lyon therefore proposes the all-pole gammatone filter, not simply as an approximation of the gammatone filter, but as a useful filter in its own right. The all-pole

gammatone filter (APGF) is arrived at by removing the zeros of the gammatone filter. Its Laplace transform is :

$$H_{AP} = \frac{b^2 + \omega_r^2}{[(s+b)^2 + \omega_r^2]^N} \quad . \tag{3.19}$$

The GTF amplitude spectrum is (almost) symmetric on a linear frequency scale. In contrast, the amplitude spectrum of the APGF is flat for low frequencies. The APGF does not have the gammatone's simple time-domain description, but properties such as gain, delay and bandwidth can be analytically determined for the APGF since it is simply the Nth power of a simple pole-pair filter [99]. The *differentiated all pole gammatone filter* (DAPGF) has the same poles as the GTF and APGF, but has a single zero at D.C. :

$$H_D = \frac{Ks}{[(s+b)^2 + \omega_r^2]^N} \quad . \tag{3.20}$$

The properties of the DAPGF thus lie somewhere between those of the GTF and APGF. In [98] [100], Lyon also demonstrates how an all-pole filter cascade (APFC) can be used as an efficient auditory filterbank.

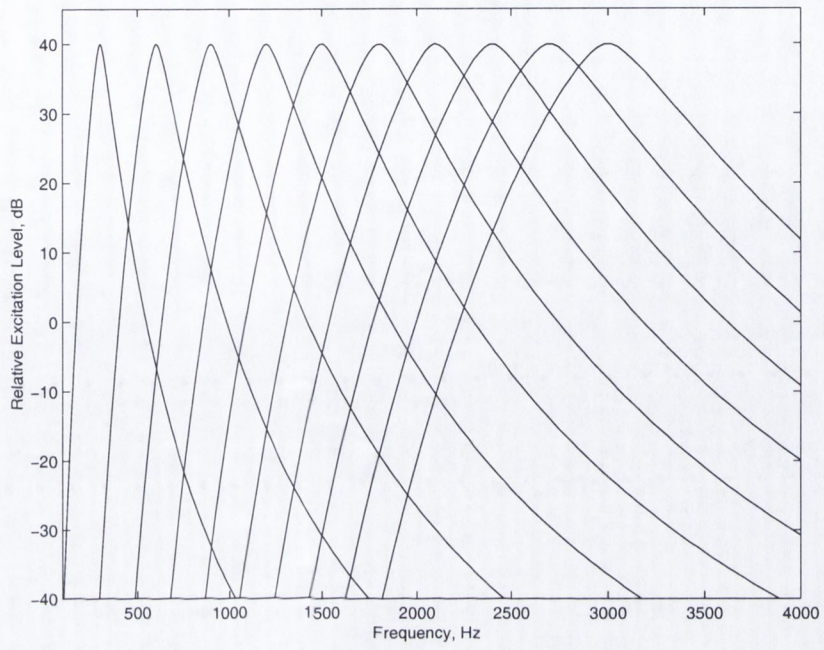
3.2.5 Excitation Patterns

In the previous section, the shape of the auditory filter was discussed in some depth. The response of an auditory filter centred at frequency, f_c, or ERB rate, E, to a particular stimulus models the ear's response to that stimulus at a single value of frequency or ERB rate. The ear's *overall* response to a given stimulus can therefore be modelled as the output of a continuous bank of auditory filters. The plot of this filterbank output with respect to ERB rate, frequency or *place*[8] is known as an *excitation pattern*, and is an indication of the internal auditory representation of this stimulus. This internal representation is important as it illustrates the degree to which a sound's spectral components are resolved by the ear. Furthermore, excitation patterns can be used to predict spectral masking; *masked audiograms* or *masking patterns*[9] and excitation patterns are closely related [194], [108], [105].

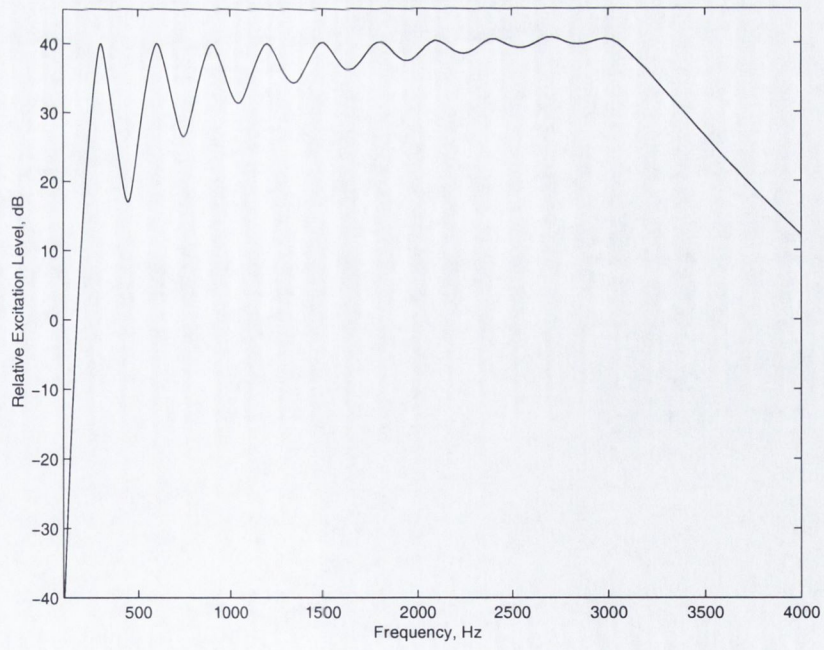
In [110] and [111], Moore and Glasberg present a simple method for calculating excitation patterns of signals composed of sinusoidal components; software implementing this code is freely available [8]. The excitation pattern of each component is calculated and these are then summed to give the signal's overall excitation pattern. This method is explained in more detail in Chapter 4, Section 4.2.2. As an example, Figure 3.12 (a) plots

[8]Place theory maps characteristic frequency to distance along the length of the basilar membrane.

[9]A masked audiogram or masking pattern is a graph of the amount of masking in dBs produced by a masker, with respect to frequency.



(a)



(b)

Figure 3.12: (a) The individual excitation patterns of first 10 harmonics of a 300 Hz fundamental. (b) The total excitation pattern.

the individual excitation patterns of the first 10 harmonics of a 300 Hz fundamental, each at 40 dB SPL. Figure 3.12 (b) plots the overall excitation pattern. The excitation pattern is composed of ripples, the peaks of which represent each harmonic. The peak-to-valley ratio decreases with increasing frequency. It thus becomes progressively more difficult for the ear to resolve harmonics as the harmonic number increases. In fact, the ear can rarely resolve more than the first 5-7 harmonics of any given musical note [154]; excitation patterns predict this behaviour very accurately. Note that the excitation pattern of a single tone graphed on a linear frequency axis is not symmetric, but shows a more shallow upper slope. This is known as the *upward spread of masking* and occurs because the auditory filter bandwidth increases with frequency. The upward spread of masking is even more apparent for level dependent filters at high levels because the low frequency slope of the auditory filter becomes increasingly shallow with increasing level.

3.3 Temporal Masking

Simultaneous masking, or *spectral masking* refers to a situation where signal threshold is increased by a masker which occurs at the same time as the signal [194]. Section 3.2 has attempted to fully explain this phenomenon by reviewing the concepts of the auditory filter and excitation pattern. *Non-simultaneous masking*, or *temporal masking*, which takes place when a masker produces masking outside of its duration, is now considered. When a masker raises the threshold of a signal preceding the masker in time, *backward masking* or *pre-masking* is said to occur. Similarly, *forward masking* or *post-masking* occurs when the masker precedes the signal in time. The effect of forward masking is more dominant than that of backward masking. Forward masking can persist for up to 100-200 ms after the cessation of the masker [107]. There is still some doubt concerning the extent of backwards masking. In [107], Moore describes the main properties of forward masking. These properties are summarised below :

- The closer the signal is to the masker, the greater the effect of forward masking.
- The rate of recovery from forward masking increases as the masker level increases.
- Unlike simultaneous masking, increments in masker level do not produce corresponding increases in forward masking. This effect is known as the non-linear growth of forward masking and is signal dependent. For low signal levels (below about 35 dB-SPL), signal thresholds increase at a much slower rate than masker level. For signal levels greater than this value, the masking function steepens, becoming close to unity

for high signal levels [152].

- Forward masking increases with increasing masker duration, at least up to masker durations of 200 ms.

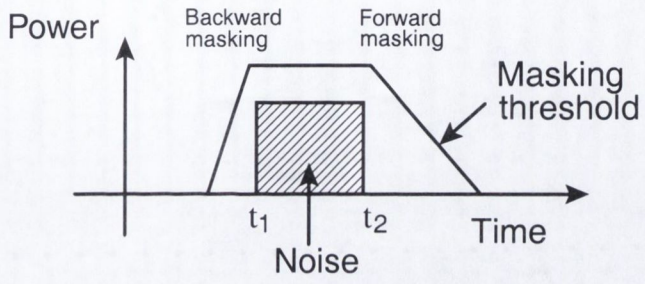


Figure 3.13: Non-simultaneous masking.

Figure 3.13 illustrates the effect of simultaneous and non-simultaneous masking on the threshold of a signal in the presence of a masker. The masker is limited to the time interval t_1 to t_2. However, the effects of the masker extend outside this interval.

3.3.1 Temporal Resolution

Auditory *temporal resolution* (also known as *temporal acuity*) refers to the extent to which the ear is able to accurately perceive a sound's temporal detail[10]. The *fine structure* of a sound refers to its rapid temporal variations; slower temporal changes in amplitude give a sound's *envelope*. As an example, for an amplitude modulated sinusoid, $s(t) = A(t)sin(\omega t)$, fine structure would refer to the periodic oscillations occurring at ω radians per second, whereas $A(t)$ would give the signal envelope. As explained by Plack [150], temporal resolution usually refers to the ability of the ear to follow changes in a sound's envelope only.

3.3.2 Measuring Temporal Resolution

One of the main problems encountered in experiments measuring temporal resolution is that it is often difficult to design experiments which eliminate spectral cues and which test temporal resolution alone. If a detection task involves discriminating between two stimuli with different temporal shape, for example a single click versus two closely spaced clicks, differences in the spectra of the stimuli can provide strong detection cues. Thus,

[10]These terms should not be confused with *temporal integration* or *summation*, which refer to the ability of the ear to combine information over a period of time in order to improve signal detection [107].

such experiments do not measure temporal resolution alone [150], [108]. However, by using stimuli whose spectra do not change, for example the discrimination of time reversed signals, or by masking any spectral differences that do occur, a more accurate measure of temporal resolution can be arrived at.

Gap Detection Experiments

The discrimination of two stimuli, one of which is interrupted by a short gap, is often used as a detection task in gap detection studies. An example would be the detection of a gap in *broadband noise*. If a broadband noise source is interrupted by a short gap, its *spectrum* is hardly distinguishable from that of an uninterrupted broadband noise. Therefore, no spectral cues are present. Using this fact, Plomp [153] found that the minimum detectable gap in broadband noise was about 2-3 ms and that this gap was relatively level independent. Many other studies have been conducted to investigate the effect of other stimulus parameters on gap detection, a good summary of which is presented in [107]. The main conclusions drawn are that changes in noise stimulus level, total noise duration and temporal position of the gap do not appreciably influence gap detection thresholds.

For *narrowband noise*, gap detection is often aided by spectral cues caused by the gap, known as *spectral splatter* [69]. This splatter can be masked using appropriate maskers. However, special care is needed to ensure that the maskers do not also mask the gap itself and thus interfere with the detection task [108]! Furthermore, random temporal fluctuations in noise envelopes can often be confused with gaps, an effect which becomes more pronounced as the noise bandwidth decreases. Gap detection experiments such as those which have been carried out by Eddins [50], Eddins and Green [49], Shailer and Moore [171], [171] and Glasberg and Moore [64] conclude that the gap threshold drops with increasing noise bandwidth and that the noise centre frequency has little effect on gap detection.

A study of the detection of gaps in *sinusoidal stimuli* has also been conducted by Shailer and Moore [170]. They introduced phase shifts at the gap and found that their results could be explained by *ringing* of the auditory filter. For stimuli with no phase shift, however, they found that ringing only affected gap detection in sinusoids at very low frequencies.

Other experiments investigating temporal acuity include work carried out by Ronken [162], who asked his subjects to discriminate between a pair of clicks of different amplitude, and a time reversed copy of this stimulus. Since time-reversal of a signal only changes its phase spectrum and not its magnitude spectrum, this detection task was not aided by spectral cues. A general method of generating signals of equal magnitude spectra but differing phase spectra was proposed by Huffman [77]. Huffman sequences were used as stimuli by Pat-

terson and Green [127] to measure temporal acuity; they found that discrimination of two such waveforms failed when the waveform duration fell below about 2 ms.

Finally, Viemeister [179] used the term *temporal modulation transfer function* (TMTF) to characterise the sensitivity of the auditory system to the modulation of a stimulus at various modulation rates. He found that the sensitivity to modulation decreased as the modulation rate increased, thus, the TMTF had the form of the magnitude response of a lowpass filter [150].

3.3.3 The Temporal Window Model of Temporal Resolution

One of the difficulties concerning the majority of studies presented in the previous section is that temporal resolution is characterised by a single measure of time. However, as pointed out by Plack [150], "There is a serious problem with all the single-value measures of temporal resolution ... Take for example the gap detection task. If temporal resolution is to be regarded as some form of temporal 'blurring', analogous to the loss of spatial resolution caused by an out of focus microscope, then the internal representation of a gap in a stimulus will be a smooth dip. The shorter the gap the shallower the dip will be. It is clear then that the gap detection task involves two components; a temporal resolution component which determines the depth of the dip and an intensity discrimination component which determines the smallest dip that can be detected. Hence the gap detection experiment cannot be regarded as a measure of temporal resolution *per se*".

As an exact analog of Fletcher's critical-band concept, Penner [142] and Penner and Cudahy [141] introduced the notion of the *critical masking interval* and attempted to measure its duration, which was found to vary between 6 and 30 ms. Later work by Penner [140] attempted to explain the data in terms of a temporal weighting function or *temporal window* with sloping skirts, which varied considerably with masker duration. Festen and Plomp [54], [55] used a symmetrical exponential or Gaussian shape to describe temporal window shape. However, such a symmetric window shape was deemed unlikely as forward masking is clearly dominant over backward masking; this was also noted by Pollack and Robinson [161].

A more thorough investigation of temporal window shape was conducted by Moore, Glasberg, Plack and Biswas [112] and Plack and Moore [151]. They used an approach similar in concept to the notched noise experiments of Patterson [135] to measure the shape of the temporal window function. The threshold of a brief sinusoidal burst presented in a temporal gap in noise was measured as a function of the noise gap duration. In subsequent work, Plack and Moore [151] measured the signal thresholds of four subjects for four frequencies and

three noise levels. *Off-time listening*, an exact temporal analog of off-frequency listening, was taken into account as the signal was presented both symmetrically and asymmetrically in the gap. The data is fitted to produce a temporal window shape which may be expressed as the sum of two rounded exponential functions [138]. The window is in two halves, one before $t = 0$ and one after $t = 0$. Together these two halves model forward and backward masking, respectively. Each half of the window has the form

$$W(t) = (1 - w)(1 + 2t/T_p)exp(-2t/T_p) + w(1 + 2t/T_s)exp(-2t/T_s) \ , \tag{3.21}$$

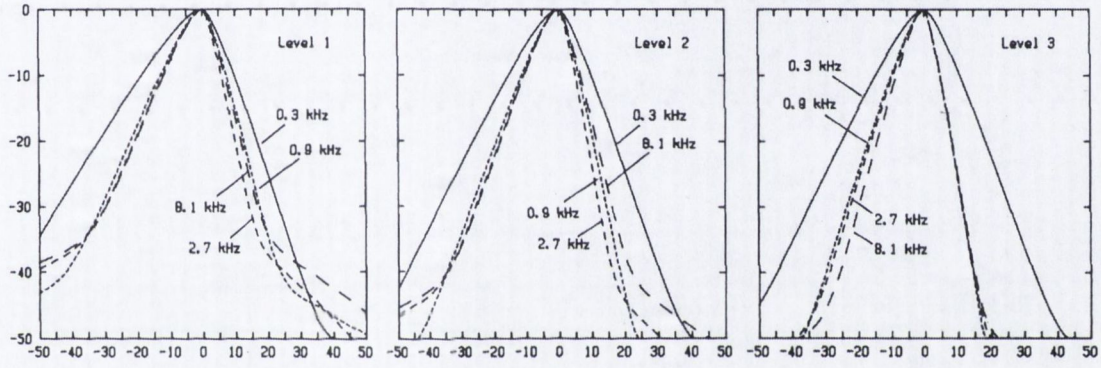


Figure 3.14: The 12 temporal window shapes derived from the average data for the four subjects [151]. The abscissa denotes time in ms relative to the centre of the window shape and the ordinate denotes the relative weighting in dB.

where t is time relative to the window centre, parameters T_p and T_s control the slope of the central part and skirt of the window shape, respectively, and w controls the relative importance of the these two components. The *equivalent rectangular duration*,[11] (ERD), is approximately equal to the sum of T_p for both sides of the window, with an overall average of about 8 ms across frequency and level. For the form given in Equation 4.17, the shape has a total of six free parameters. However, Moore *et al.* found that this number could be reduced to four without significantly worsening the goodness of fit by keeping the ratio of T_p/T_s and the value of w equal on either side of the window. The temporal window function is intended as a running averager of stimulus *energy* and is non-causal so as to model the temporal spread of masking both forwards (forward masking) and backwards (backward

[11]The equivalent rectangular duration (ERD) is defined as the duration of a rectangular window shape having peak amplitude and total area equal to that of the temporal window shape. It is an exact analogy of the equivalent rectangular bandwidth (ERB) measure used for auditory filter shape.

masking) in time. The temporal window shape sharpens with increasing frequency and level. This can be seen from Figure 3.14, which presents all 12 temporal windows calculated from the average data for the the four subjects across frequency and level. There are some limitations to modelling temporal resolution in such a manner which are discussed in Chapter 5, Section 5.5. However, a considerable amount of temporal masking data can be predicted using this simple model of temporal resolution and it has been employed by several researchers [2], [74], [178] to model temporal masking. It is also noted that a more refined model of temporal masking which also accounts for the nonlinear growth of forward masking has recently been proposed by Plack and Oxenham [152], who employ a simulated basilar membrane response followed by a sliding temporal window to model nonlinear forward masking effects.

3.4 Review of Perceptually Motivated Joint Time-Frequency Analysis

A great deal of research in the psychoacoustic literature has concerned itself with modelling the behaviour of elements or components of the auditory system. Over the years, psychoacousticians have gradually formed a more complete picture of the overall system by piecing together knowledge of such components. However, there are still relatively few models of the entire system, and assumptions still have to be made about the behaviour of certain parts of such systems. An in-depth overview of auditory modelling can be found in [124]. Briefly, models of the auditory system can broadly be divided into those modelling *impaired* and *unimpaired* hearing. *Macromechanical* models focus on modelling the fluid dynamics of the cochlea. In general, these describe the motion of the basilar membrane using either *classical* [4], [180], [43], [39], [45] or *non-classical* fluid mechanical models [5], [44]; both of which seek to determine the impedance of the basilar membrane and its response to a given stimulus. *Micromechanical* auditory models study the behaviour of the basilar membrane, tectorial membrane and organ of Corti, all of which lie within the cochlea [5]. Unlike micromechanical and macromechanical models, which aim to replicate the exact physiological behaviour of auditory system components, *hybrid models* generally try to characterise such component's input-output function only. For example, auditory frequency selectivity is often modelled using a parallel or cascade filterbank structure [143], [128], as opposed to complicated basilar membrane fluid mechanics. A subsequent simple nonlinearity and smoothing stage often follows the filter structure to model hair cell response and temporal masking respectively [150].

The most important hearing impairments modelled are those of *frequency smearing* and *loudness recruitment* [144], [82]. Frequency smearing describes a degradation in the frequency selectivity of the auditory system, caused by a broadening of basilar membrane tuning and a reduction in outer hair cell function [113]. Auditory filtershapes obtained from subjects suffering from frequency smearing are often much broader than those obtained from normal listeners [108]. Loudness recruitment is caused by damage to the outer hair cells of the cochlea causing a reduction in the compressive nature of the basilar membrane. This results in an unusually rapid and uncomfortable increase in loudness for certain sounds. Models of loudness recruitment can be found in [58], [94].

Despite the large body of work in the psychoacoustic literature, and the wide variety of auditory models in existence, relatively little use has been made of *joint time-frequency analysis* in auditory modelling. Of course, it is unreasonable to expect all features of auditory processing to be replicable using time-frequency signal analysis alone. However, audio time-frequency analysis could certainly benefit from a closer modelling of certain important characteristics of the auditory system. Perhaps the most notable characteristic worth incorporating is spectral and temporal masking or simultaneous and non-simultaneous masking. As Yang notes in [190] : "Auditory approaches ... often involve complex, multistage, nonlinear transformations which make analytical treatments intractable. This has made assessing performance improvements and their underlying causes very difficult and uncertain, and almost totally dependent on extensive experimentation. Therefore, in order to fully realise the potential benefits of applying auditory processing, it is essential that a viable analytical approach be developed to characterise the signal representations generated by such processing." Later, in this thesis, it is demonstrated how a smoothed Wigner distribution, with its useful mathematical properties, can provide a useful method of generating such auditory signal representations.

3.4.1 Wavelet Based Models

The wavelet transform is a relatively new development in the field of signal processing that has quickly found application in several important areas including image and video processing and compression, and audio compression [46], [172], [159]. Wavelet analysis has seen some application in perceptual coding of speech as its inherently constant Q-factor properties approximate in some sense the frequency resolution of the auditory system. Some recent examples of the use of wavelets in auditory filter modelling and audio compression may be found in [190], [36], [40], [13].

In general, auditory models based on wavelet analysis seek to achieve an auditory-like

frequency resolution. However, a close match to auditory filter shape, as detailed in the psychoacoustic literature, is not generally achieved. In general, the aim of such schemes is to achieve an efficient wavelet filterbank structure with auditory-like properties suitable for incorporation in efficient audio compression schemes. Auditory frequency selectivity, as represented by the excitation pattern, is therefore not accurately achieved. Furthermore, in wavelet analysis, time and frequency resolution cannot be independently set. Even if a form of wavelet analysis which accurately models frequency masking is designed (by closely matching wavelet frequency responses to auditory filter frequency responses), its temporal resolution is set. It is therefore highly unlikely that this temporal resolution will by chance equal that of of a psychoacoustically appropriate smoothing function characterising temporal masking. In addition, wavelet analysis is linear, whereas auditory filtering is nonlinear. Finally, the wavelet transform is inherently constant-Q for all frequencies, whereas auditory filtering is approximately constant-Q for high frequencies, and approximately constant-bandwidth for low frequencies[12]. One possible method of mirroring auditory frequency resolution using wavelet analysis is the use of wavelet packet analysis [3], [159]. However, there are serious difficulties to be overcome in accurately modelling spectral and temporal masking using wavelet analysis techniques and in general such techniques, while being adequate for compression schemes, are not accurate enough for serious psychoacoustic auditory modelling.

One example of the use of the wavelet transform is presented in [2], [3]. Here, a model of the auditory system generating joint time-frequency excitation patterns is proposed by Agerkvist which uses wavelet analysis as a preprocessing stage. This stage uses a generalised orthogonal wavelet packet analysis scheme, with wavelet bandwidths approximating auditory filter ERB. The output of this wavelet analysis is treated as a time-frequency image. Auditory filter level dependency and temporal masking are modelled by smearing the image in frequency and time to generate a joint time-frequency excitation pattern. The overall model is intended for the evaluation of perceptually coded audio. Degradation is deemed detectable if the difference between the time-frequency excitation pattern of the original and coded signal differ by more than a critical ratio value at any point in the TF plane. Unfortunately, in this model the effective bandwidth of the wavelet analysis only approximates the ERB; it is not a smooth function of frequency, but increases stepwise with frequency due to the discrete dilations inherent to the discrete wavelet transform [159]. In a dual manner, temporal resolution decreases stepwise with increasing frequency. The accuracy

[12]There is still uncertainty over the exact value of auditory filter bandwidth, but whatever the final outcome, it is highly unlikely to be exactly constant-Q across the audible frequency range.

of modelling temporal masking and auditory filter nonlinearity using graphical smoothing of the TF images is also questionable. Furthermore, the time-frequency resolution tradeoff inherent to the wavelet transform meant that it was not possible to properly model auditory temporal resolution. Agerkvist states that : "If the model is to be improved, then a bilinear method such as the Wigner distribution should be used as no linear method can provide better resolution in one domain without sacrificing the resolution in the other." He also explains that : "Bilinear methods were not considered as they are slow and require massive amounts of computer memory, even for relatively short signals." In the following chapters, it is clearly demonstrated that the bilinear smoothed Wigner distribution does indeed offer the solution to simultaneously, accurately and independently modelling temporal and spectral masking. Furthermore, it is clearly demonstrated how the high computation and memory requirements of such methods may be overcome.

3.4.2 Spectrogram Based Models

Spectrograms have also been used to model the auditory system. *Perceptually-based dynamic spectrograms* proposed in [7] utilise time differentiation, Bark-spaced filterbank processing, equal loudness compensation and compression to emphasise perceptually important features and increase robustness in noisy conditions. In [116], a perceptually based derivative of the spectrogram called the *FTT-spectrogram* using frequency dependent analysis windows is proposed. Another similar scheme is proposed by Jeong and Ih [81] known as the *variable frequency resolution short-time Fourier transform* (VFR-STFT). This scheme is based on the *variable frequency resolution fast Fourier transform* proposed in [1]. In this scheme, the Fourier transform of windowed signal sections are calculated and stored. The signal is then downsampled by a factor of two. This is repeated a number of times, yielding partial spectra of increasing frequency resolution. Since the length of the window is the same at all iterations, the spectral resolution is effectively doubled after each iteration. The VFR-FFT is composed of these partial spectra. Unfortunately, the VFR-FFT scheme is unsuited for incorporation with the STFT because a lack of window overlap causes non-uniform temporal resolution behaviour [81]. The VFR-STFT is therefore proposed to overcome this problem by allowing for some window overlap. As for the wavelet-based auditory model of Agerkvist [3], the bandwidth of the VFR-STFT does not vary gradually with frequency, but increases stepwise after each successive downsampling of the signal. Temporal resolution behaves in a similar manner, decreasing stepwise with increasing frequency. Again, auditory filter shapes are not accurately modelled—only a crude approximation to ERB is achieved. No attempt at modelling the temporal resolution of the ear is made in

this model.

3.4.3 Auditory Modelling using the Smoothed Wigner Distribution

In the author's opinion, attempts to date at accurately modelling auditory temporal and spectral resolution using linear time-frequency analysis methods such as the spectrogram or wavelet transform have been unavoidably limited by the time versus frequency resolution trade off inherent to these methods. Auditory-like frequency resolution has been achieved in the work reviewed above, but the inflexibility of such methods limits accuracy. The majority of these methods have been geared more towards audio compression schemes than psychoacoustic auditory modelling *per se*. As Agerkvist points out, since it is not possible to independently set time and frequency resolution using linear time-frequency analysis methods, the accuracy of all such attempts at incorporating auditory resolution will of necessity be compromised. To the best of the author's knowledge, no use at all has been made of the smoothed Wigner distribution for auditory modelling, despite the fact that it allows independent control over time and frequency resolutions. Furthermore, its elegant mathematical properties, outlined in Chapter 2, suggest that it would provide a means for the analytical modelling approach proposed by Yang [190] described above.

There may be several reasons why researchers in the past have shied away from using the smoothed Wigner distribution for auditory modelling. Firstly, the Wigner distribution is computationally expensive; it is an order-n^2 algorithm, where n is signal length[13]. Therefore it has been perceived as being very computationally expensive for long signals. In comparison, linear methods such as the spectrogram and wavelet transform are order-n. Some researchers may have been misguided in believing that the entire TCF needed to be computed before transformation into the WDF domain, resulting in a requirement for much memory space. The presence of cross-term interference has also limited the use of the Wigner distribution; traditionally Wigner distribution research has striven to attain a compromise between cross-term suppression and high resolution. Unfortunately, in the past, smoothed Wigner distributions have typically been plotted on a linear, rather than logarithmic magnitude axis. Cross-term interference, noise and general clutter can often *seem* well suppressed if the distribution is plotted on a linear scale. However, such undesirable artifacts can often completely swamp the very same distribution if the log magnitude is plotted. Linear scales are of little use in filter design and in audio research in general due to their limited dynamic range. Therefore, there may have been some doubts concerning the

[13]Once TCF windowing is applied however, for the pseudo and smoothed Wigner distributions, the algorithm becomes order-n.

useful dynamic range of the smoothed Wigner distribution, especially when cross-term interference is considered. There has also been relatively little research using smoothed Wigner distributions with non-uniform frequency spacings, as for example, bins spaced according to the logarithmic or ERB scales. This is because the FFT algorithm has generally been employed to transform from the TCF domain into the WDF domain, resulting in a linear bin spacing[14]. Finally, perhaps due to the relative complexity of the Wigner distribution algorithm, there has been relatively little attempt to produce distributions with variable frequency resolution; traditionally, for computational efficiency, Wigner distributions have been computed using the FFT algorithm, resulting in a constant frequency resolution.

In the following chapters, the design of a smoothed Wigner distribution with frequency dependent time and frequency resolutions is described. These resolutions exactly match those of the ear as described in the psychoacoustic literature across a wide dynamic range. As such, the distribution is a temporally varying excitation pattern which simultaneously models both spectral and temporal masking. The distribution is also computationally efficient and does not suffer from the excessive memory requirements normally associated with the WD for long signals. The distribution frequency bin spacing are at the discretion of the user. Furthermore, it will be shown that the time and frequency smoothing of this distribution is sufficient to attenuate practically all cross-terms below auto-term excitation level. The distribution is also free from windowing artifacts and has a very wide dynamic range.

[14]The Bertrand P_k distributions and the Q distribution of Altes are notable exceptions [21], [156], [6].

Chapter 4

EarWig Distribution Design

This chapter details the design of a time-frequency distribution which simultaneously models both spectral and temporal masking, known as the EarWig distribution (EWD) [119], [120], [121]. The implementation of the EWD on a digital computer is also tackled. Initially, this chapter introduces the theory behind the discrete Wigner distribution, leading to discrete forms of the pseudo-Wigner distribution and smoothed pseudo-Wigner distribution. Issues pertaining to the implementation of this distribution are then discussed, including signal sampling requirements, computational efficiency and memory usage. It is shown how computation of the discrete pseudo-Wigner distribution from the temporal correlation function domain leads to significant savings in terms of computation time and memory space. Computation time is markedly reduced because knowledge of the smoothing window allows us to significantly relax the temporal sampling requirement of the EWD to be calculated. Furthermore, memory limitations are no longer a problem since the TCF may be computed in stages 'as required'.

After this discussion of the discrete smoothed Wigner distribution, the EarWig distribution is described in some detail. The design approach of its frequency dependent separable kernel function is fully explained. It is demonstrated that the EarWig distribution is effectively equivalent to a temporally evolving auditory excitation pattern. Finally, the computational load and memory requirements of the EarWig distribution are estimated. In the following chapter, this approach to modelling temporal and spectral masking will be assessed in terms of resolution accuracy, cross-term interference and dynamic range.

4.1 The Discrete Wigner Distribution

In order to compute the Wigner distribution on a digital computer, a discrete form of the distribution is required. The derivation of the *discrete* WD from the continuous WD

is presented in detail in [155] and in [28], and is similar to the derivation of the discrete Fourier transform from the continuous Fourier transform. It is reproduced in abbreviated form here. Starting with the expression for the (continuous) WD we have :

$$W(t,\omega) = \int s^*(t-\frac{\tau}{2})\, s(t+\frac{\tau}{2})\, e^{-j\tau\omega}\, d\tau \tag{4.1}$$

Instead of evaluating $W(t,\omega)$ using the continuous-time signal, $s(t)$, we evaluate the distribution using discrete signal, $s(nT)$, where T is the sampling interval. Hence, time, t, is replaced with nT and relative time, τ, is replaced with $2lT$, where n and l represent discrete time index and discrete relative time index respectively. The expression is also multiplied by $2T$ to preserve total energy. We therefore arrive at the expression for the *discrete-time* WD :

$$W(nT,\omega) = 2T \sum_{l=-\infty}^{\infty} s^*((n-l)T)s((n+l)T)\, e^{-j2Tl\omega}\ . \tag{4.2}$$

Because τ is set to $2lT$, the sampling frequency, f_s, of the discrete time signal, $s(nT)$, must exceed *twice* the Nyquist rate. In other words $T = 1/f_s$, where $f_s \geq 4f_c$, f_c is being the highest signal frequency present in the signal[1]. Normalising frequency by setting $T = 1$ for a unit sample period we obtain :

$$W(n,\theta) = 2 \sum_{l=-\infty}^{\infty} s^*[n-l]s[n+l]\, e^{-j2l\theta}\ . \tag{4.3}$$

We note that $W(n,\theta)$ is a function of discrete time, n, but continuous normalised frequency θ. By proceeding in an exactly analogous manner to the derivation of the discrete Fourier transform from the discrete-time Fourier transform, we can arrive at an expression for the the discrete-time, discrete-frequency Wigner distribution (or simply discrete Wigner

[1] An excellent discussion of discrete WD sampling issues is presented by Nuttall [117], [118]. He demonstrates that the usual signal sampling frequency of $f_s \geq 2f_c$ is sufficient for exact alias-free reconstruction of the original *continuous* WD of the *continuous* signal at all time and frequency locations, if all available TCF information is utilised. He points out that it is the WD, and not the signal, which needs to be computed twice as finely in time and frequency as the usual Nyquist-rate sampled discrete signal and discrete spectrum respectively. He also presents an algorithm for generating the discrete WD from a continuous signal sampled at the Nyquist rate. For a discrete signal N samples in length, the resulting discrete WD time-frequency grid is $2N \times 2N$ samples. The algorithm does, however, involve an intrinsic upsampling of the discrete signal. Nuttall's discrete WD algorithm does not therefore result in a reduction in computation over the usual discrete WD procedure of sampling the signal at twice the Nyquist rate. Therefore, in this work, the usual discrete WD procedure is used. Also, Nuttall points out that the use of the analytic signal halves the number of multiplications necessary for TCF calculation. However, as pointed out in Section 4.4, the resulting multiplications become complex, effectively cancelling this computational advantage.

distribution) [155]. Specifically. for a discrete sequence of length $2L-1$, we may express the discrete WD as

$$W(n, k\pi/M) = \sum_{l=-L+1}^{L-1} s^*[n-l]s[n+l]\, e^{-j2lk\pi/M} \;, \tag{4.4}$$

where n is discrete-time, $k\pi/M$ is discrete frequency (often abbreviated to k), M is the number of samples in the frequency domain and the factor of 2 outside the summation is omitted for simplicity. As is the case for the DFT, M should be greater than or equal to the number of samples $2L-1$, to avoid aliasing. Also, because of the factor $2l$, the discrete WD is periodic in frequency with period π, unlike the DFT, which repeats every 2π. Thus, as previously stated, the signal sampling rate for the discrete Wigner distribution must be at least twice the Nyquist rate.

The properties of the continuous, discrete-time and discrete Wigner distributions are discussed at length in [27], [28], [29]. In general, with the exception of frequency periodicity, the properties of the discrete-time WD have a direct correspondence with those of the continuous WD discussed in Section 2.2.2. The properties of the discrete-time, discrete-frequency Wigner distribution are closely related to those of the discrete-time Wigner distribution (and thus Wigner distribution), with the exception of discrete frequency variable $k\pi/M$. In conclusion, if the discrete sequence, $s[n]$, represents the values of the continuous time signal, $s(t)$, sampled at twice the Nyquist frequency, then its discrete WD, $W[n,k]$, provides an excellent estimation of the continuous WD, $W(t,\omega)$, of $s(t)$ sampled at regular intervals in the time-frequency plane.

4.1.1 Temporal Correlation Function

We will denote the discrete temporal correlation function (TCF) as a two dimensional discrete function $g[n,l]$, where $g[n,l] = s^*[n-l]s[n+l]$. The function $g[n,l]$ is most easily visualised as a product of the signal $s[l]$ time shifted by n samples with the time reversed signal $s^*[-l]$, also shifted in the opposite direction by n samples. Figure 4.2 graphically portrays the calculation of $g[n,l]$ at $n=-3$, $n=-2$ and $n=-1$ for signal, $s[n]$, of odd length $2L-1$, with $L=4$. The signal $s[n]$ is graphed in Figure 4.1. For simplicity's sake, $s[n]$ is chosen to be real. Figure 4.2 demonstrates that as n increases from $-L+1$ to $L-1$, $s[n-l]$ shifts to the right and $s[n+l]$ shifts to the left. Overall, the non-zero samples of the sampled TCF, $g[n,l]$, form a diamond shaped region in the discrete time-

discrete relative time domain[2]. A plot of this diamond shaped region for this particular signal length is shown in Figure 4.3. The function $g[n,l]$ is even in l for real signals; it shows Hermitian symmetry for complex signals. Therefore, the number of multiplications required to calculate $g[n,l]$ may be halved by calculating the function for either positive or negative l only and transposing the result.

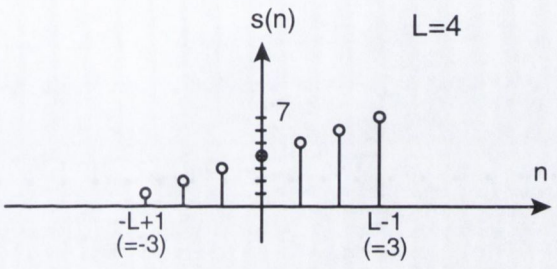


Figure 4.1: Real signal $s[n]$ with $L = 4$.

It is worthwhile pointing out at this point that a definition of the sum of the TCF along n is very similar to the definition of the (unscaled) autocorrelation function, $\hat{\phi}_{xx}[l]$, from random signal processing theory [123] :

$$\hat{\phi}_{xx}[l] = \sum_{n=-L+1}^{L-1-|k|} s[n]s[n+l] \; , \tag{4.5}$$

where real signal $s[n]$ in non-zero from $-L+1$ to $L-1$. The TCF of $s[n]$, summed along n, may be defined as

$$G[l] = \sum_{n=-L+1+|k|}^{L-1-|k|} s[n-l]s[n+l] \; . \tag{4.6}$$

The only difference is that for the TCF, both signals slide in time, whereas for the autocorrelation function, only one signal slides; $G[l]$ therefore contains only the even values of $\hat{\phi}_{xx}[l]$. However, if, as required by the discrete WD, the signal is upsampled by a factor of 2 prior to TCF calculation, both TCF sum, $G[l]$, and autocorrelation, $\hat{\phi}_{xx}[l]$, give exactly the same result. For the autocorrelation sequence, transforming into the frequency domain via the Fourier transform yields the *power spectrum* of the (stationary) random process of which

[2]The non-zero values of the *spectral* correlation function also form a diamond shaped region in the frequency-relative frequency domain.

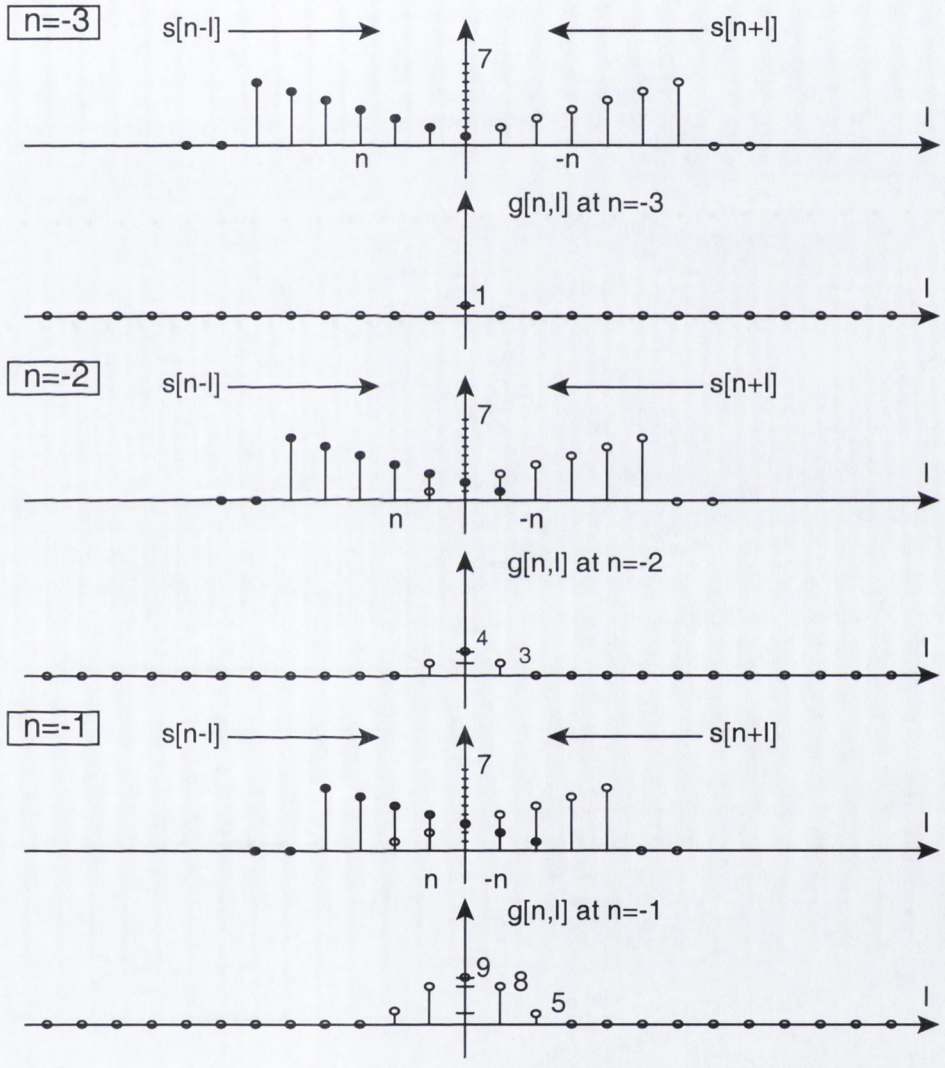


Figure 4.2: Calculation of $g[n,l]$ at $n=-3$, $n=-2$ and $n=-1$ for a real signal, $s[n]$, with $L=4$.

$s[n]$ is a sample sequence. For the TCF summed across n, the Fourier transform yields the frequency marginal of the WD, which is the *energy spectral density* of deterministic signal $s[n]$. With these concepts in mind, the WD can be thought of as distributing the energy of the energy spectral density of deterministic signal, $s[n]$, in the time-frequency plane in a meaningful, temporally evolving manner[3].

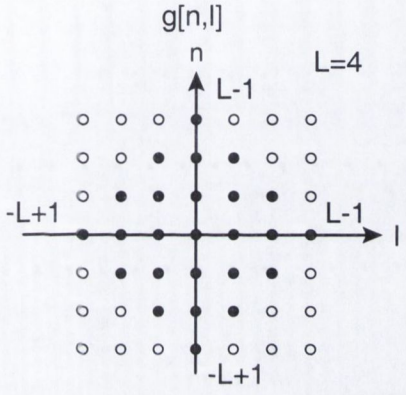


Figure 4.3: Diamond-shaped TCF region $g[n,l]$ with $L = 4$.

4.1.2 Implementation Issues

Section 4.1.1 described the calculation of the TCF, $g[n,l]$, for a signal of odd length, $2L-1$. As described in Section 4.1, the discrete WD may be calculated by performing a DFT in l on $g[n,l]$ at each time index n. We can take advantage of highly efficient DFT algorithms by zeropadding the TCF, $g[n,l]$, in l to a highly composite length. More specifically, for the popular Cooley and Tukey radix-2 FFT algorithm, we can zero pad $g[n,l]$ in l by one zero and choose L to be a power of 2 [80]. Figure 4.4 (a) and (b) portray this step, for the case of $L = 4$. Unfortunately, DFT/FFT routines are generally defined for causal signals; for a signal of length $2L$ the integration limits would be defined from $L = 0$ to $2L - 1$. Unfortunately, for the TCF, the data exists from $-L + 1$ to $L - 1$. However, a DFT/FFT of length $2L$ is periodic with period $2L$, therefore any values in $g[n,l]$ may be shifted in l by an integer multiple of $2L$ samples without affecting the result[4]. Therefore, the samples

[3]In this work only deterministic signal analysis is considered. However, we also note that the WD has been used to as a tool to characterise the spectral content of *nonstationary* random processes with time [73].

[4]If more frequency samples are required, a larger FFT may be performed by extra zeropadding. However, the zeros should always be placed from $l = L$ to $l = (N-1) - (L-1)$ where N is the FFT length and the signal is of length $2L - 1$.

of $g[n,l]$ for negative l from $-L+1$ to -1 are shifted to $L+1$ to $2L-1$. This is portrayed in Figure 4.4 (c). This shifting is sometimes referred to a 'rotating and renumbering' [80]. A radix-2 FFT may be performed on the data over the region $L=0$ to $L=2L-1$ to give the required result. Furthermore, the discrete Wigner distribution is always real, whereas the output of the FFT routine is complex. This redundancy can therefore be exploited by efficient implementations of the discrete Wigner distribution making use of the fast Hartley transform [23] which are presented in [139], [189].

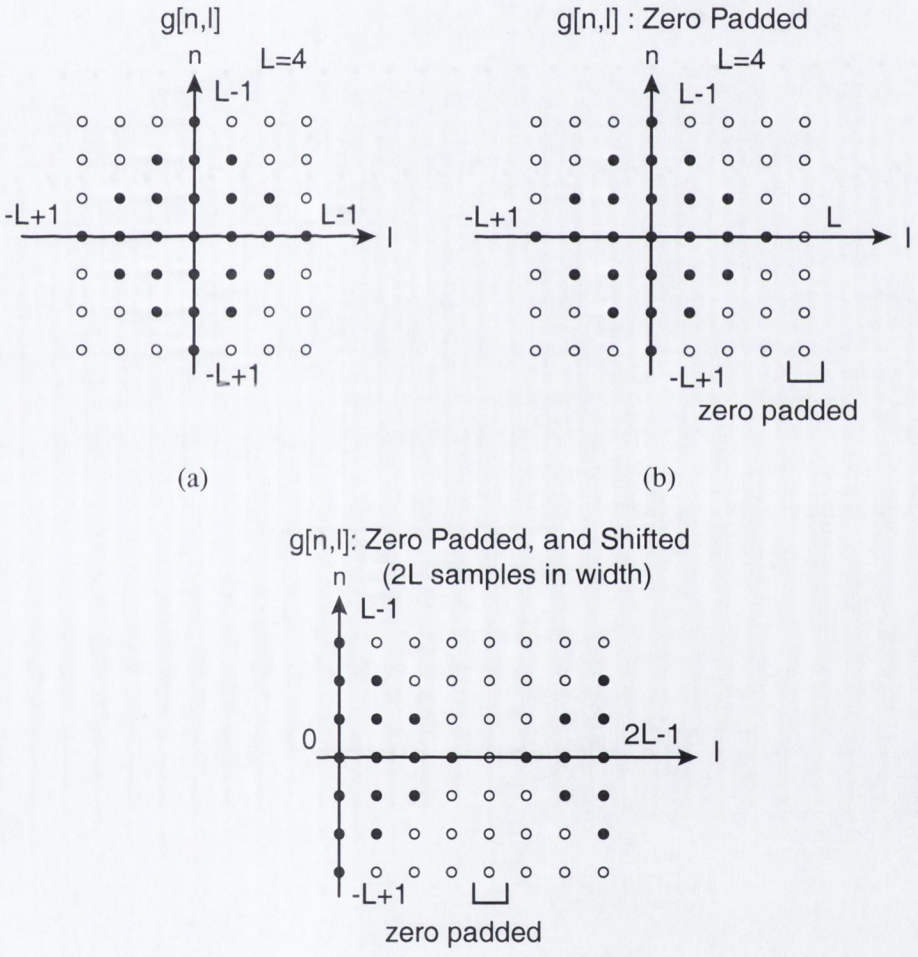


Figure 4.4: Rotation, re-numbering and zeropadding of TCF, $g[n,l]$, with $L=4$.

4.1.3 Discrete Pseudo-Wigner Distribution

The discrete version of the continuous-time pseudo-Wigner distribution may be arrived at by multiplying the TCF $g[n,l]$ by a discrete frequency smoothing window of finite length. The choice of window depends on the application; typically windows are chosen to exhibit properties such as good frequency resolution and sidelobe attenuation. The discrete pseudo-WD (DPWD) may be expressed as

$$W_{DP}(n,k\pi/M) = \sum_{l=-P+1}^{P-1} h[l]s^*[n-l]s[n+l]\,e^{-j2lk\pi/M}\ , \tag{4.7}$$

where $h[l]$ is the smoothing window and is non-zero from $-P+1$ to $P-1$. P is typically chosen so that $P << L$ *i.e.* the window length is much shorter than the signal length. Figure 4.5 demonstrates the cropping action of $h[l]$ on TCF, $g[n,l]$. After multiplication with the window function, the TCF data may be zeropadded, shifted and Fourier transformed in exactly the same manner as for the discrete WD. In order to demonstrate the effect of this windowing, Figures 4.6 and 4.7 compare the discrete WD and discrete pseudo WD for a chirp signal. It can be seen that for the discrete WD, the distribution reaches a peak at time $t = 65$. This point corresponds to the centre of its TCF, *i.e.* $n = 0$, which is where the non-zero values of the TCF reach their maximum width in relative time. Thus the peak amplitude of the chirp is proportional to the width of the non-zero samples of the TCF. However, for the discrete pseudo-WD, the distribution reaches peak amplitude once the effective TCF width equals the window length. Therefore, for the discrete pseudo-WD shown in Figure 4.7, the distribution is at peak amplitude for most of its duration.

It is clear from Figures 4.6 and 4.7 that the frequency resolution of the DPWD is less than that of the DWD. The frequency smoothing introduced by window, $h[l]$, results in a WD which may, for the continuous time definition, be expressed as the convolution of the WD with the window Fourier transform, $H(\omega)$:

$$W_{PS}(t,\omega) = \int H(\omega-\xi)W(t,\xi)\,d\xi\ . \tag{4.8}$$

Sometimes in the literature [18], [27], [28], the frequency smoothing window, $h[l]$, is referred to as being both even and real, to ensure a real pseudo-WD with symmetric frequency smoothing respectively[5]. Also, the frequency smoothing window, $h[l]$, is frequently applied prior to TCF calculation, resulting in the following definition for the discrete pseudo-WD :

[5]It may be desirable to design a real pseudo WD with asymmetric frequency smoothing, for example if we wish to model auditory filter asymmetry, as described in Section 3.2.4. For this purpose, a Hermitian symmetric, complex window may be used. This is discussed in more detail in Section 4.2

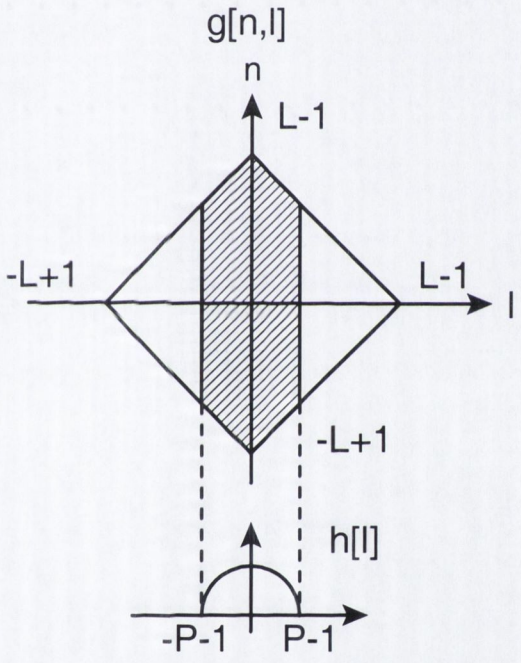


Figure 4.5: TCF cropping for DPWD.

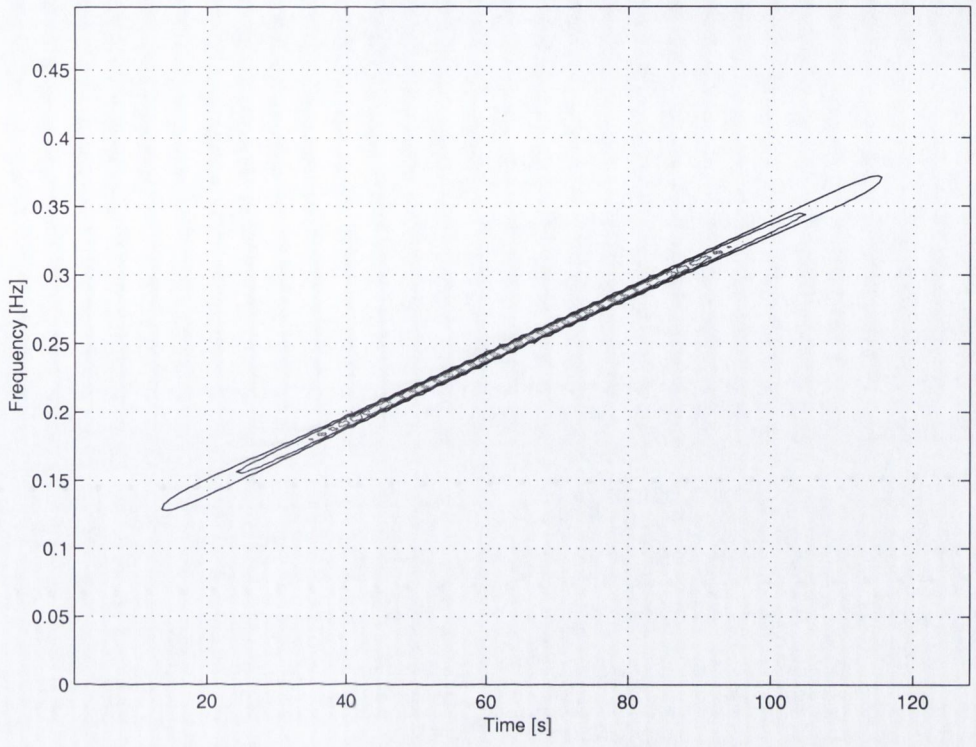


Figure 4.6: DWD of a chirp signal.

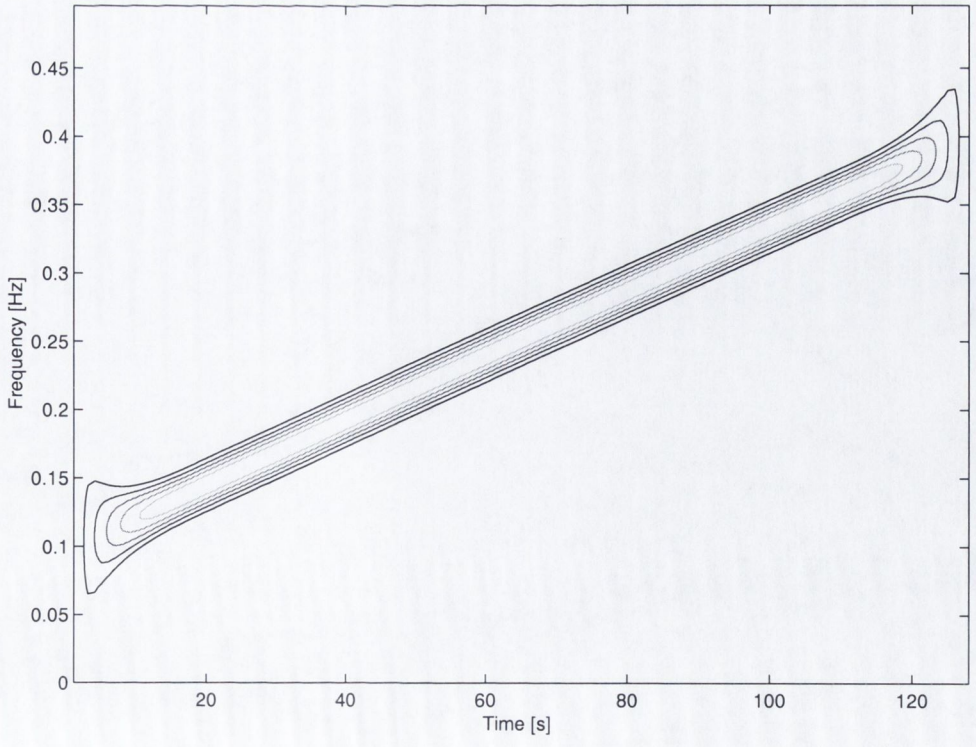


Figure 4.7: DPWD of a chirp signal.

$$W_{PS}(n, k\pi/M) = \sum_{l=-P+1}^{P-1} h^*[-l]h[l]s^*[n-l]s[n+l]\, e^{-j2lk\pi/M} \ . \tag{4.9}$$

Since $h[l]$ is often assumed to be real and even; the abbreviation $|h[l]|^2$ is therefore commonly substituted for $h^*[-l]h[l]$. Thus, as pointed out by Pielemeier [147], the Fourier transform of the effective smoothing window is now equal to the convolution of the Fourier transform of $h[l]$ with itself, since multiplication in the time domain implies convolution in the frequency domain. Unfortunately, this halves the frequency resolution of a WD computed using such a window. Pielemeier points out [147], [145] that if the correct pseudo-WD sidelobe attenuation is to be achieved, $h[l]$ should be *autocorrelated*, not multiplied, with itself prior to multiplication with TCF $g[n, l]$. This can be explained as follows. In Equation 4.8, we see that the frequency resolution of the pseudo-WD computed with frequency smoothing window, $h[l]$, is equal to the convolution in frequency of the WD with the Fourier transform, $H(\omega)$, of the window. However, $W(t, \omega)$ is in units of *Joules* $Hz^{-1}s^{-1}$, *i.e. energy spectral density* per unit time, whereas $H(\omega)$ is in units of *Volts* Hz^{-1} or *Fourier transform magnitude*. Therefore, in order to achieve the correct smoothing window performance, $H(\omega)$ should appear in squared form in Equation 4.8. More specifically, we wish to have

$$W_{PS}(t, \omega) = \int H_f(\omega - \xi)W(t, \xi)\, d\xi \ , \tag{4.10}$$

where $H_f(\omega) = |H(w)|^2$, $H_f(\omega)$ being the Fourier transform of window $h_f[l]$. To achieve this, Pielemeier proposes the use of (unscaled) autocorrelation of $h[l]$, or equivalently, convolution of $h[l]$ with its time reversed complex conjugate to achieve window $h_f[l]$:

$$h_f[l] = h[l] \otimes h^*[-l] \ , \tag{4.11}$$

where $\otimes$ denotes convolution.

Thus, to design the correct smoothing window $h_f[l]$, a window $h[l]$ exhibiting the desired spectral properties is first specified and then autocorrelated to give $h_f[l]$, which is then used to window the TCF. It is worth noting that the length of $h_f[l]$ is twice that of $h[l]$. Also, as we shall see in Section 4.2, the procedure is applicable to complex, non-Hermitian-symmetric windows as the autocorrelation procedure guarantees Hermitian symmetry of $h_f[l]$, thus ensuring the reality of the resulting pseudo-WD.

As an example, Figure 4.8 (a) plots the energy spectral density of a signal consisting of two complex sinusoids at normalised frequencies $\pi/8$ and $3\pi/8$. The signal was windowed with a hamming window and then zeropadded prior to FFT calculation. The resulting energy spectral density calculated as $20log_{10}(|FFT|)$, where $|FFT|$ is the magnitude of the

FFT values displayed up to the normalised frequency of $\pi/2$. The two main lobes of the signal frequencies are clearly visible along with the sidelobes. The peak sidelobe level is 41 dB below the main lobe, as expected for the hamming window function. Figure 4.8 (b) displays the log magnitude of one frequency slice of the DPWD of exactly the same signal, where the log magnitude is calculated using $10log_{10}(|DPWD|)$. The frequency window used was generated by autocorrelating and normalising the hamming window used above prior to FFT calculation. The DPWD exhibits an extra lobe at $\pi/4$ corresponding to cross-term interference between the signal frequencies (no time smoothing was carried out to reduce cross-term interference oscillating parallel to the time axis). Apart from this difference, the resolutions match exactly with both plots displaying precisely the same mainlobe and sidelobe characteristics. Therefore, using the window design methodology of Pielemeier, the exact DPWD frequency resolution may be achieved.

Figures 4.8 (c) and 4.8 (d) portray exactly the same scenario, except for a different window function. The particular window used was a *gamma window*. The gamma window shape is discussed in detail in Section 4.2.1, suffice to say here that gamma windows are based on the gammatone auditory filter described in Chapter 3. Again, apart from the cross-term interference at $\pi/2$, the FFT and DPWD resolutions match exactly for all frequencies. We also point out at this point that the gamma window does not exhibit the strong lobing behaviour of the hamming window and is smooth across the entire frequency range. This will be shown later to be important for ensuring that the EWD is clutter-free over a wide dynamic range.

4.1.4 The Discrete Smoothed Wigner Distribution

The equation for the discrete smoothed pseudo-WD (DSPWV) can be given as :

$$DSPWV[n,k] = \sum_{l=-P+1}^{P-1} h_F[l] \sum_{m=-T+1}^{T-1} h_T[m]g[n-m,l]\, e^{-j2lk\pi/M} \,. \tag{4.12}$$

The only difference between the DSPWV and the DPWV is that the TCF $g[n,l]$ is smoothed in the time direction by time smoothing window $h_T[n]$ prior to FFT calculation. The properties of this distribution are exactly analogous to the properties of the continuous smoothed pseudo-WD (SPWD) described in Chapter 2. The kernel is separable, and is expressed $h_F[l]h_T[n]$ in the TCF domain. Again, as Pielemeier points out, autocorrelation of window $h_T[n]$ should be performed in a similar manner to that of frequency smoothing window $h_F[l]$ in order to achieve the desired resolution.

The entire operation necessary to compute the DSPWV at a particular time $n = N$ is

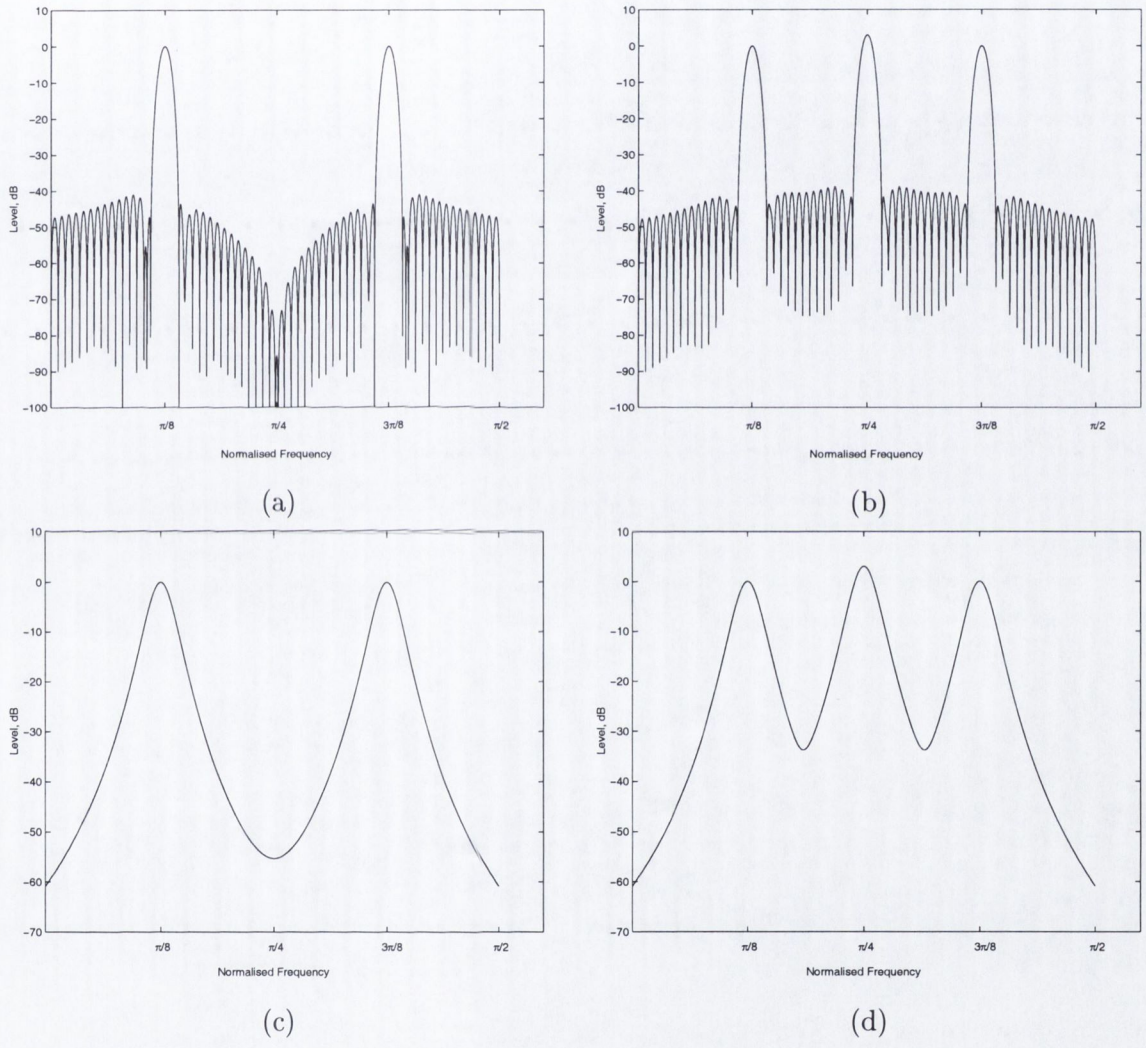


Figure 4.8: Energy spectral density of a two sinusoid signal windowed by a hanning window for the FFT, (a), and PWDF (b). Energy spectral density of the same two sinusoid signal windowed by a gamma window for the FFT, (c), and PWDF, (d).

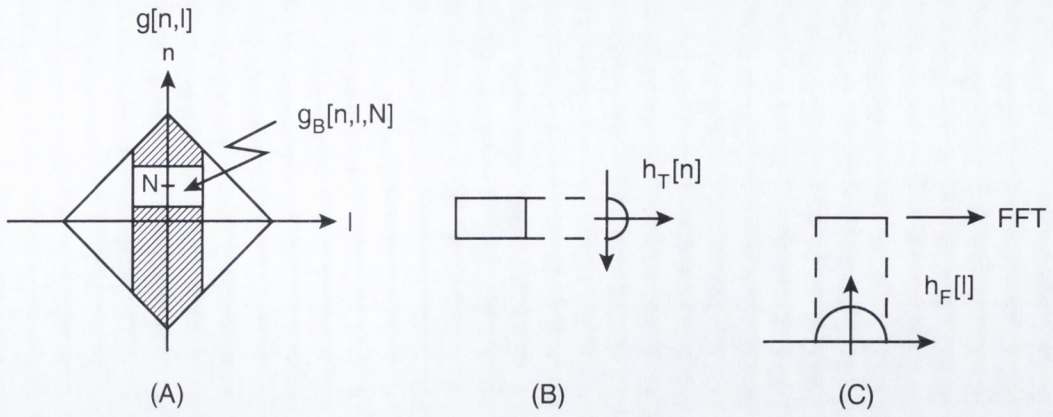


Figure 4.9: Operations required to calculate a SPWDF frequency slice : (A) : TCF block calculation, (B) : TCF time smoothing (C) : TCF frequency smoothing and frequency transformation.

graphically portrayed in Figure 4.9, where the operations of zeropadding and rotating and renumbering are omitted for simplicity. A *TCF block*, $g_B[n, l, N]$, of width equal to the width of the frequency smoothing window $h_F[l]$ and length equal to the length of the time smoothing window $h_T[n]$, centred around $[n, l] = [N, 0]$ is first calculated (Step A). This is then smoothed in time by computing a weighted sum along n for each l of the TCF block using time smoothing window $h_T[n]$. This operation results in a temporally smoothed TCF *slice* (Step B). This slice is then windowed using frequency smoothing window $h_F[l]$ (Step C). The result is zeropadded to the FFT length and rotated and renumbered. Calculating the FFT of this results in a frequency slice of the DSPWV at time $n = N$.

It can be seen that using this approach, the TCF need only be computed in stages as needed. Thus, a large amount of memory is no longer required to compute the WDF of very long signal lengths (*i.e.* a few seconds of audio samples at 44.1 kHz). The TCF is simply computed 'as needed'. This is important as otherwise the TCF memory requirements typically place a limit on the maximum signal length analysable using the DSPWD. Furthermore, as Pielemeier points out in [148], performing the time smoothing prior to FFT calculation results in a great reduction in the total number of FFTs needed. Since the time smoothing window is typically lowpass, time smoothing results in a reduction in the spatial bandwidth of the final DSPWD in the time direction. If the time smoothing window suppresses all spatial frequencies above ω_c, then each DSPWD slice need only be computed every $1/2\omega_c$ seconds without incurring any loss of distribution detail. However, if the time

smoothing is performed after FFT calculation, an FFT must be performed for every value of n for correct time smoothing, in other words, at the signal sampling rate. This simple observation, generally completely overlooked in the Wigner distribution literature, generally results in huge computational savings! Also, the memory needed to store the resulting distribution is significantly reduced since the TCF is computed in stages, as required.

4.1.5 Frequency Dependent Kernel Design

For the purposes of designing a joint time frequency distribution modelling auditory masking we wish to design a smoothing kernel whose temporal and spectral resolutions depend on frequency. The kernel is assumed to be time and signal independent. The exact design of this smoothing kernel is detailed in Section 4.2, however suffice to say at this point that, in general, frequency smoothing changes much more rapidly with frequency than temporal smoothing. Furthermore, we cannot use the FFT to transform from the TCF to WDF domain, since we may often want to compute the distribution on a non-linear frequency scale, for example, a logarithmic or ERB scale. As auditory filters overlap strongly, the number of frequency channels required is usually relatively small, for example 100 - 150 channels is sufficient to cover the range 20 - 20 kHz.

The frequency smoothing required at each frequency increases with increasing frequency. So, if the temporal resolution was constant across frequency then for each time, N, we would first compute the TCF block $g_B[n, l, N]$, and smooth it in time using constant time smoothing window, $h_T[n]$, yielding a time smoothed TCF strip. Then, for each frequency channel C, we would multiply this TCF strip by a different frequency smoothing window $h_F[l, C]$ and then compute the DFT value at that particular, single frequency. This would yield the required distribution values at each channel at time $n = N$, ie. a slice of the distribution along frequency at time $n = N$. However, if we wish to have variable spectral *and* temporal resolution, a different amount of temporal smoothing needs to be performed for each frequency channel. Therefore, at a given time, $n = N$, each channel, C, would normally require a TCF block, $g[n, l, N]$ to be both smoothed in n by window, $h_T[n, C]$, and multiplied in l by window, $h_F[l, C]$. This means that both time and frequency smoothing would need to be performed at every sample of the final TF distribution, resulting in a huge computational load. However, for our purposes, since auditory temporal resolution changes relatively slowly with frequency, applying a new temporal window, $h_T[n, C]$, at each channel, C, is not necessary. Typically, for the temporal window model of auditory temporal resolution discussed in Section 4.2, less than 10 temporal windows are sufficient to model the temporal resolution frequency dependency across the entire audible frequency

range. Each temporal window may be assigned to a certain frequency band. Thus, frequency channels lying within a given band use the same temporal smoothing and hence can make use of the same TCF strip for all frequencies in this band. In this manner, at a given time $n = N$, the number of time smoothing operations that need to be performed on the TCF block $g[n, l, N]$ is equal to the number of temporal windows used (10 or less), rather than the number of frequency channels (usually 100 or more).

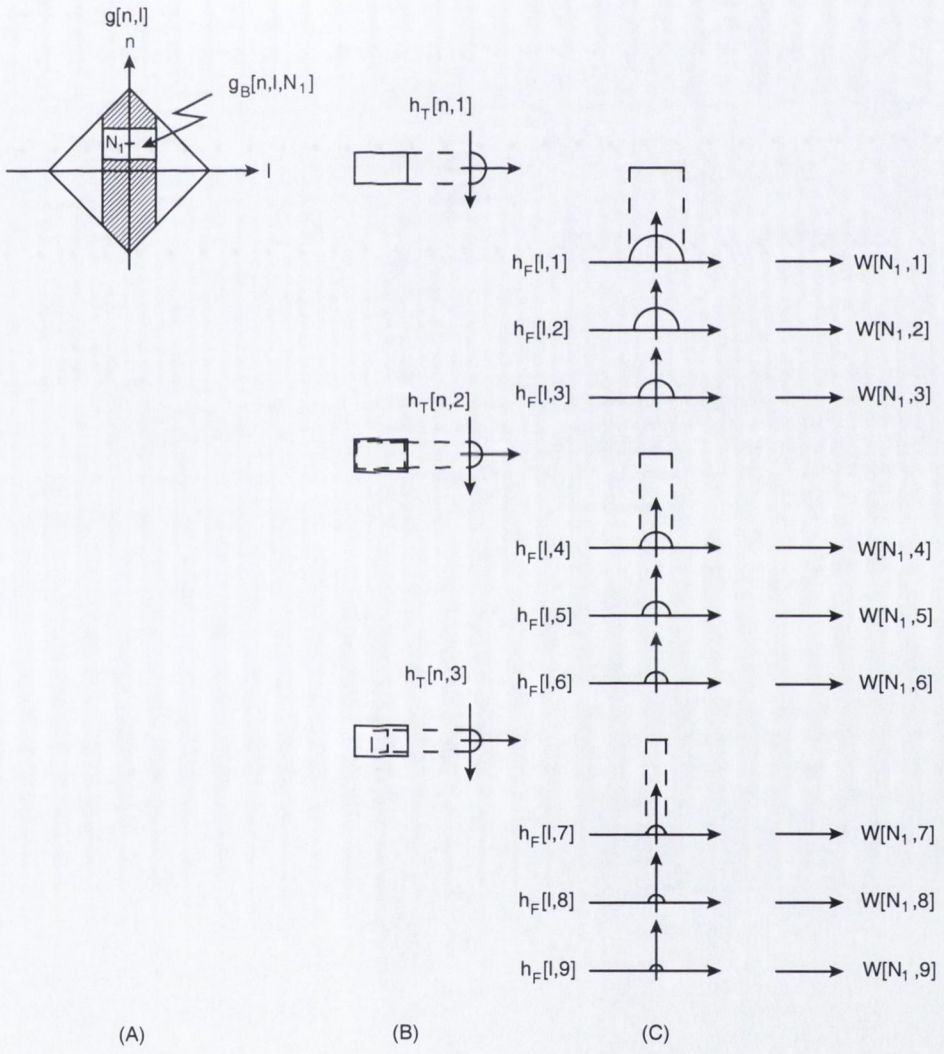


Figure 4.10: Operations required to calculate a SPWDF frequency slice with frequency dependent time and frequency resolution : (A) : TCF block calculation, (B) : TCF time smoothing (C) : TCF frequency smoothing.

The execution of the entire processes, at one time $n = N_1$ is graphically portrayed in Figure 4.10. In this instance, 3 temporal windows and 9 frequency channels are used, 3

channels being assigned to each window. Initially the TCF block $g[n, k, N_1]$ is computed, of width and length equal to the length of the longest frequency smoothing window and longest time smoothing window respectively (Step A). The longest frequency window $h_F[l, 1]$ occurs at the lowest frequency (channel 1) , since the auditory filter bandwidths are lowest for low frequencies. The temporal window model is also generally longest for low frequencies (see Section 3.3.3). This TCF block is then processed for each of the 3 temporal windows. For the lowest frequency temporal window (top), the TCF is first multiplied by $h_T[n, 1]$ and summed in n giving a TCF strip (Step B). This strip is then multiplied by each of the the frequency smoothing windows in the frequency band of this temporal window $h_F[l, 1]$, $h_F[l, 2]$ and $h_F[l, 3]$. A single frequency DFT calculation is performed after each frequency smoothing operation, yielding the value of $W[N_1, 1]$, $W[N_1, 2]$ and $W[N_1, 3]$; this is (Step C). Similar calculations are performed for the other 2 temporal window stages yielding $W[N_1, 4]$ to $W[N_1, 9]$. The overall procedure outlines the total processing required to calculate a slice in frequency of $W[n, l]$ at a particular instant in time $n = N_1$. For the next time $n = N_2$, a new TCF block $g[n, l, N_2]$ is computed and the whole process is repeated. In this manner, the required TF distribution is built up, frequency slice by frequency slice.

The value of $N_2 - N_1$ gives the time interval in samples at which the distribution is calculated. In general, there is a good deal of overlap between the current TCF block $g[n, l, N_{curr}]$ and the previous TCF block, $g[n, l, N_{curr-1}]$. This overlap can therefore be exploited to reduce TCF calculation cost. Furthermore, frequency smoothing windows can be pre-modulated and stored, so the frequency smoothing and frequency transformation steps in Step (C) can be performed simultaneously. Also, the Hermitian symmetry of the TCF and frequency smoothing window can be exploited to halve computation.

4.2 The EarWig Distribution

The previous sections have discussed an efficient method of calculating the discrete smoothed pseudo-Wigner distribution, and of achieving frequency dependant temporal and spectral resolution using frequency dependent windows $h_T[n, C]$ and $h_F[l, C]$. The following two sections discuss the design of separable smoothing kernel $h_{TA}[n, C]h_{FA}[l, C]$, where 'A' stands for 'auditory'. The kernel design is based on well known models of auditory spectral and temporal resolution. The resulting TF distribution, modelling both temporal and spectral masking, is dubbed the *'EarWig'* distribution (EWD).

4.2.1 Frequency Smoothing Window Design

Chapter 3 presented a historical review of the development of the auditory filter shape and described some important filter models. In order to model auditory frequency resolution using the DSPWD, we need to design a bank of lowpass, frequency-dependent frequency smoothing windows $h_{FA}[l, C]$ derived from one of these auditory filter shape specifications. As will become clear shortly, the frequency response of each frequency smoothing window must match exactly that of its corresponding auditory filter response, demodulated to DC. In other words, the smoothing window must be a baseband version of the auditory filter.

Section 3.2.4 presented four recent auditory filter shapes, namely the roex, gammatone, gammachirp and all-pole gammatone filter. As explained below, the most suitable of these for our purposes was the gammatone filter. The *roex* filter shape is specified only in terms of energy spectral density and has no phase description. Furthermore, it has not been possible to specify a phase function giving a satisfactory time domain response for this filter shape [98]. For window design, a time domain response is required, therefore no attempt was made to employ the roex filter shape in the design of frequency smoothing windows in this work.

The *all pole gammatone* has a simple frequency domain specification, but a complicated time domain one. As stated above, the design of smoothing windows requires a time domain response; one possible method of deriving such a response from the APGF might be to compute its windowed impulse response, generate the analytic version of this (FIR) response (to preserve asymmetries) and demodulate the result to DC. Another possible method would be to set the centre frequency, ω, to zero in the rational system transfer function and compute its windowed impulse response. However, in this work, these possible techniques were not explored, as the gammatone (and gammachirp) filter offered a more direct method of computing smoothing windows.

For the *gammatone* filter, the demodulated response can immediately be specified by omitting the 'tone' part of the expression, thus centering the filter shape at DC. This is referred to by Cooke as a *gamma response* or *lowpass gammatone filter* [35] and by Slaney as a *baseband gammatone filter* [173]. The gamma response may be expressed as :

$$g_W(t) = at^{n-1} \exp(-2\pi bt) \qquad t > 0 \; . \tag{4.13}$$

Parameters b and n control the bandwidth and slope of the filter in exactly the same manner as for the gammatone filter function. Parameter a may be ignored as it is generally set to give a DC level of 0 dB. Approximations of the gamma response in terms of rational

z-domain transfer functions (giving IIR filters) using pole-mapping, the bilinear transform and the impulse invariant transform can be found in [35]. For the purpose of designing frequency smoothing windows, however, an FIR response is required. This can be readily be calculated by sampling and windowing the gamma response, ie. using the well known 'window method'[6] of FIR filter design. The resulting window is referred to here as a *gamma window*. The gamma window is a real function, after autocorrelation it becomes symmetric, ensuring a symmetrically smoothed and real DSPWD respectively.

In an exactly parallel manner, the *gammachirp* filter may be converted into a baseband gammachirp filter by simply setting centre frequency f_c to zero in its analytic expression :

$$g_c(t) = at^{n-1}\exp(-2\pi bt)\exp(j(c\ln t + \phi)) \qquad t > 0 \tag{4.14}$$

The analytic form of the filter shape is exactly the same as the real form, except that the term $\exp(j(2\pi f_c t + c\ln t + \phi))$ is used instead of $\cos(2\pi f_c t + c\ln t + \phi)$ (Equation 3.17). The analytic form of the filter is used so as to preserve the asymmetry of the response once demodulated to DC. Parameters b, n and c control filter bandwidth, slope and asymmetry; parameter a may again be ignored. As for the gamma window, an FIR baseband 'gammachirp window' may be designed using 'window method'. However, as is demonstrated in Section 4.2.2, if such a window is to be used, its asymmetric frequency response must be mirrored about DC. This is most easily achieved by taking the complex conjugate of the gammachirp window either before or after it has been autocorrelated and normalised.

In this work, level dependency was not incorporated in the design of the EWD smoothing kernel. Using a level independent filter model is not a serious limitation for the purpose of excitation pattern calculation because differences between level dependent and level independent filter models only become important for narrowband signals at high levels [7]. Excitation patterns for broadband stimuli remain very similar across a wide range of levels [111]—this is demonstrated in Chapter 5 which presents an extensive comparison of excitation patterns generated using the EWD and those generated using level dependent and independent filter models. Level independent filters have therefore been successfully used for excitation pattern calculation [134], [106]. The gammatone filter also has been used extensively as an auditory filter model [137], [133], [35], [173], [143], whereas the

[6]The window method is a commonly used FIR filter design technique. An IIR filter is first designed to meet a certain specification. The impulse response of this IIR filter is then multiplied by a finite length window and this result is then used as the time domain response of the new FIR filter.

[7]Level dependency is important for modelling temporal masking effects such as the nonlinear growth of masking, the additivity of masking and the masker overshoot effect. The limitations of using a level independent temporal masking model are discussed further in Section 5.5

gammachirp is a relatively new filter shape and has yet to find widespread appeal [98]. Therefore, the gammatone filter was deemed the most appropriate auditory filter model on which to base smoothing window design. However, the frequency smoothing window design techniques described above for the gammatone filter are directly applicable to the gammachirp window if EWD level dependency is desired.

Gamma Window Response

Figure 4.11 (a) plots the impulse response of the analytic version of the gammatone filter of centre frequency 1 kHz (the sampling frequency chosen was 5 kHz). The analytic version is used to eliminate spectral content for negative frequencies and thus better illustrate the demodulation process. The solid and dashed lines indicate the real and imaginary part of the response respectively. The response is a sinusoidal oscillation modulated by a gamma response. The impulse response is normalised to give a gain of 0 dB at the centre frequency. Figure 4.11 (b) plots the filter's frequency response for both negative and positive frequencies. The response is almost completely free from windowing artifacts such as sidelobes. Figures 4.11 (c) and (d) plots the impulse and frequency response of the baseband gammatone filter respectively. The impulse response is asymmetric and real. It is clear that its frequency response is simply a demodulated version of the gammatone filter. Finally, the waveform and frequency response of the gamma window function of same centre frequency is plotted in Figures 4.11 (e) and (f) respectively. The window response is real and gaussian-like in shape. It is twice the length of the baseband gammatone response and is symmetrically centred about zero time lag. Its frequency response is exactly equal to that of the basebase band gammatone filter except that the filter attenuation is doubled, thus ensuring the correct DSPWD frequency resolution.

4.2.2 Excitation Pattern Generation

In [110], a method for computing excitation patterns is presented which uses the level independent, symmetric $roex(p)$ filter shape. The expression for this filter shape is :

$$roex(p) = (1 + pg)\exp(-pg) \quad . \tag{4.15}$$

Their method is summarised here. The upper panel of Figure 4.12 presents a spectrum of a sinusoid at 1 kHz and several roex(p) filter shapes. The intersection of each filter shape with the sinusoid gives the excitation level caused by the signal at the filter centre frequency. For example, the filter of lowest centre frequency intersects the sinusoid at point a; this therefore gives the excitation level at centre frequency A. This process is repeated for

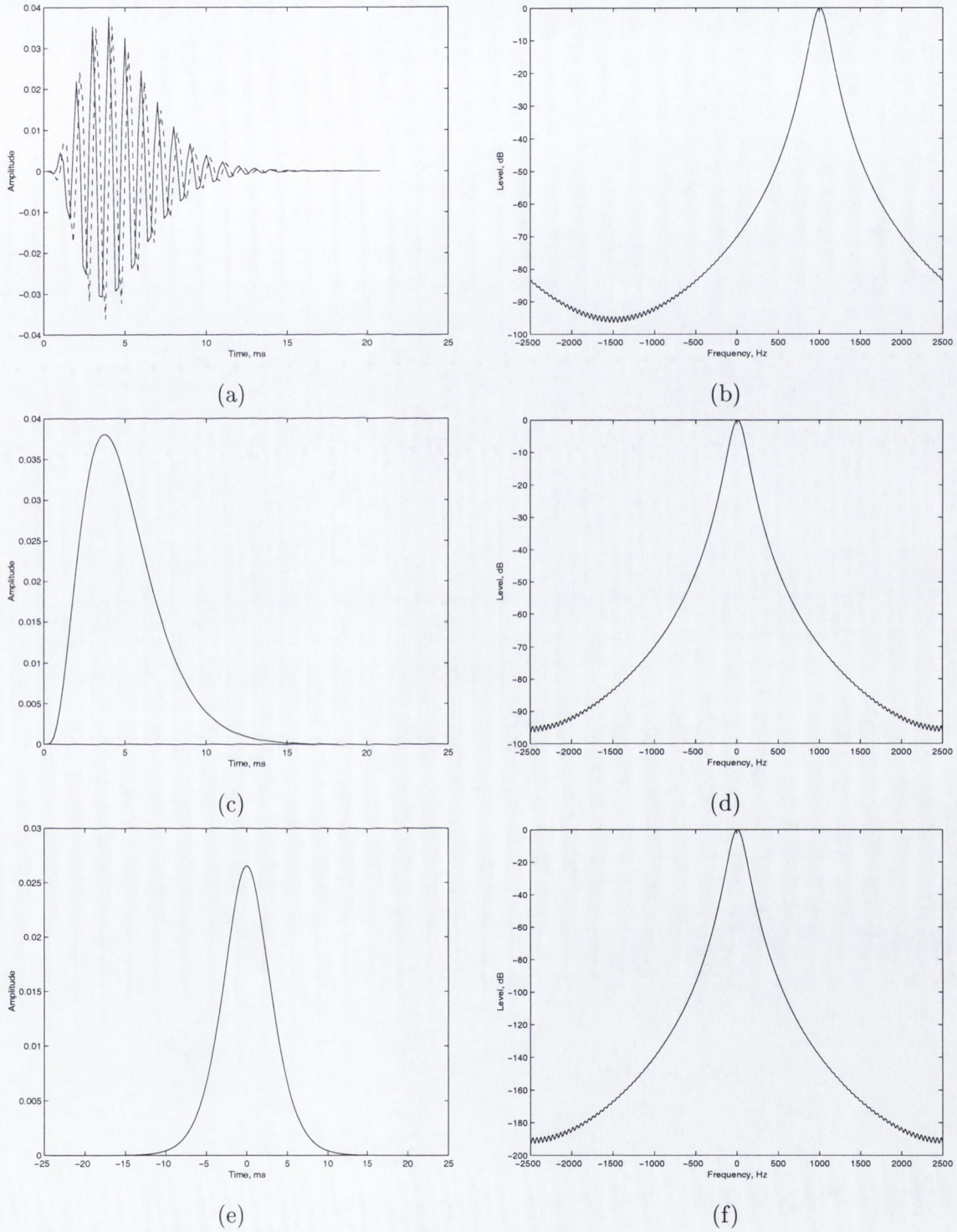


Figure 4.11: Waveforms (left) and frequency responses (right) for the gammatone filter (top panels (a) and (b)), baseband gammatone filter (middle panels (c) and (d)) and gamma window (bottom panels (e) and (f)).

centre frequencies B, C, D, and E. The lower panel of Figure 4.12 plots the overall excitation pattern, generated by calculating the excitation level at finely spaced intervals in frequency. This method is graphical and does not rely on signal processing techniques.

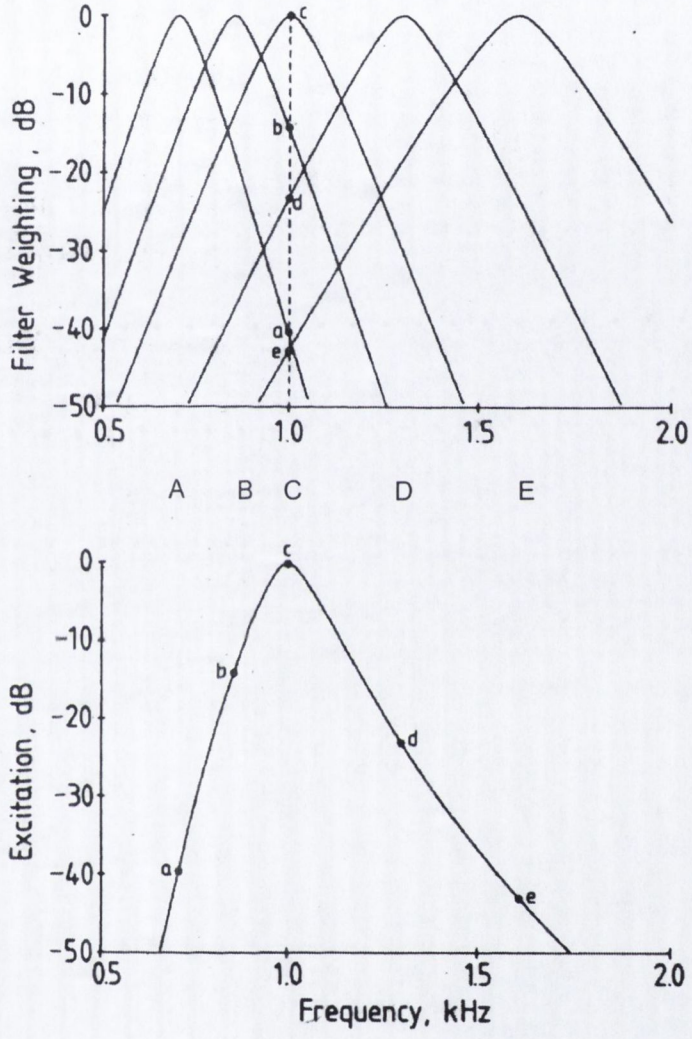


Figure 4.12: Upper panel : Intersection of sinusoid spectrum and roex filter shapes giving the excitation for 5 nominal frequencies A, B, C, D and E. Lower panel: The resulting overall excitation pattern.

Excitation pattern computation using the EWD relies on a different method than the one outlined above. However, both methods are exactly equivalent and give the same result. This is demonstrated in Figure 4.13 (a) and (b), which compare the computation of one excitation level for a particular frequency slice using the EWD and the method of Moore and Glasberg respectively. The stimulus in question is again a single sinusoid. Figure 4.13 (a) outlines the EWD method. Here, the curve is the spectrum of the gamma-

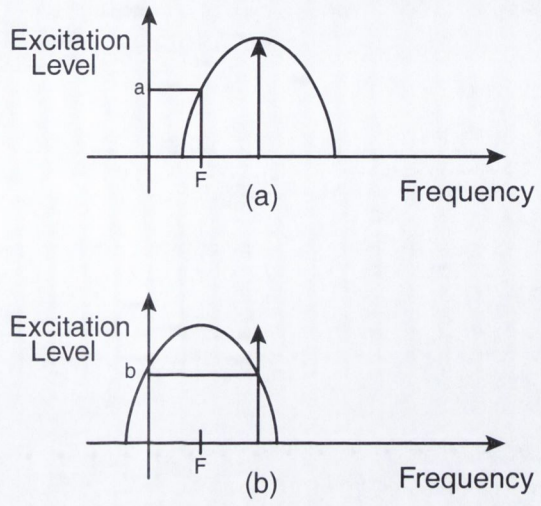


Figure 4.13: A comparison of the calculation of one excitation level for the EWD, (a) and method of Glasberg and Moore, (b). The filter shape is symmetric. The two methods are shown to give the same result *i.e.* $b = a$.

windowed sinusoid signal, given by the convolution of the sinusoid spectrum and the gamma window spectrum. The gamma window is designed from the corresponding gammatone filter of centre frequency F. The EWD method simply calculates the level of the signal's spectrum at frequency F, giving the excitation level, a, at this frequency. In Figure 4.13 (b), Moore and Glasberg's method is portrayed. As described earlier, this method calculates the intersection of the gammatone response centred on frequency F with the sinusoidal spectrum. This effectively determines the excitation at frequency F caused by the signal tone giving excitation level b at frequency F. However, as we have pointed out, the spectral shape of the gamma window for the EWD is just a demodulated replica of the gammatone filter. Hence, because the gammatone filter shape is symmetrical about its centre frequency, both methods yield exactly the same result *i.e.* b=a. The equivalence of the methods can be generalised to arbitrary signals, since, by Fourier's theory, all signals may be expressed as sums of sinusoidal components. We also point out again that the gammatone filter and symmetric roex(p) shapes are very similar. Hence, excitation patterns computed using the EWD method agree closely to those computed using Moore and Glasberg's method [110], as demonstrated in Chapter 5.

In this work, the gammatone filter was adopted as a suitable model of auditory frequency resolution. This filter shape is symmetrical about its centre frequency and is level

independent. If level dependency is to be incorporated in a more advanced EWD, then an asymmetrical filter shape must be used, as the slope of the lower skirt of the auditory filter decreases with increasing level, while the slope of the upper skirt increases with increasing level. Fortunately, the kernel design procedure outlined above for the gammatone filter may still be used to design window shapes having asymmetrical frequency responses (*i.e.* complex-valued window shapes). However, for the method of excitation pattern generation portrayed in Figure 4.13 to succeed, the magnitude of the frequency response of the resulting window must be mirrored about DC. This may be simply achieved by computing the complex conjugate of the window, either before or after autocorrelation and normalisation have been performed. These steps are summarised below, where $w[n]$ is the demodulated asymmetric filter response, $w_c[n]$ is the response after conjugation, $w_a[n]$ is the response after conjugation and autocorrelation and $h_{TA}[n]$ is the final normalised kernel smoothing window where N_F is the length of $w_a[n]$.

$$\begin{aligned} w_c[n] &= w^*[n] \\ w_a[n] &= w_c[n] \otimes w_c^*[-n] \\ h_{TA}[n] &= \frac{w_a[n]}{\sum_{n=1}^{n=N_F} |w_a[n]|} \end{aligned}$$

Figure 4.14 presents exactly the same situation as presented in Figure 4.13 for a symmetrical filter shape, now for an asymmetrical shape. The filter has a more shallow slope for its lower skirt than its upper skirt. It is clear that once the window response is flipped about DC, the EWD method of excitation pattern generation and the method of Glasberg and Moore are again equivalent, *i.e.* $b = a$.

4.2.3 Temporal Smoothing Window Design

In Chapter 3, temporal masking was discussed and a model of temporal resolution proposed by Plack and Moore [151] was described. This model proposes that the forward and backward spread of masking be modelled by a window shape acting as a running averager of stimulus energy. The shape of the window is non-causal and has longer duration for negative time, ensuring that once applied, the forward spread of masking is greater than the backward spread. Each side of the window may be expressed as the weighted sum of two rounded exponentials as follows :

$$W_b(t) = (1-w)(1+2t/T_{pb})\exp(-2t/T_{pb}) + w(1+2t/T_{sb})\exp(-2t/T_{sb}) \ , \quad (4.16)$$

$$W_a(t) = (1-w)(1+2t/T_{pa})\exp(-2t/T_{pa}) + w(1+2t/T_{sa})\exp(-2t/T_{sa}) \ . \quad (4.17)$$

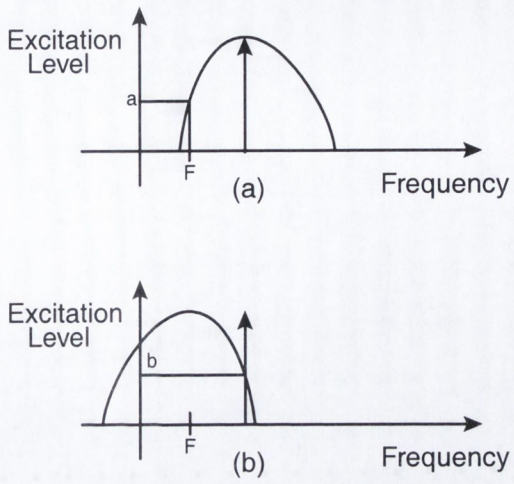


Figure 4.14: A comparison of the calculation of one excitation level for the EWD, (a) and method of Glasberg and Moore, (b). The filter shape is asymmetric, having a shallower lower slope than upper slope. The two methods are shown to give the same result *i.e.* $b = a$.

Here, $W_b(t)$ and $W_a(t)$ gives the window shape before and after the window centre at $t = 0$, i.e. for negative and positive time. Parameter t is (absolute) time relative to the window centre; T_{pb} and T_{sb} control the shape of the window peak and skirt for negative time respectively with T_{pa} and T_{sa} having the corresponding meaning for positive time. Parameter w controls the relative level of skirt to peak for both sides of the window. In the study by Plack and Moore [151], twelve averaged temporal window shapes were derived by fitting this window shape to masking data for three levels and four centre frequencies. Figure 3.14 in Chapter 3 plots these windows. The temporal window shape is intended as a smoother of excitation level and is therefore ideally suited as an EWD kernel time smoothing function. Furthermore, it is lowpass in nature and therefore suppresses cross-term interference oscillating parallel to the time axis; this is discussed in greater detail in Chapter 5. As the window shape is expressed in units of energy, it does not require autocorrelation and can be applied directly as an EWD time smoothing function or window. Thus, the asymmetric nature of the temporal window is preserved.

For the purpose of EWD kernel design, a *level independent*, frequency dependent time smoothing window is required. Since Plack and Moore derive a temporal window shape at centre frequencies of 300, 900, 2700 and 8100 Hz only, interpolation of their data is necessary to generate a temporal window shape of arbitrary centre frequency. Also, for each centre frequency, they derive three window shapes corresponding to the three stimuli

Frequency (Hz)	T_p (ms)	T_s (ms)	w (dB)	a
300	12.7	70.7	-58.3	1.6
900	9.3	68.3	-47.0	1.8
2700	8.1	46.0	-39.0	2.1
8100	7.5	48.0	-39.7	1.9

Table 4.1: Values of averaged parameters T_p, T_s, w and a for frequency dependent temporal window model.

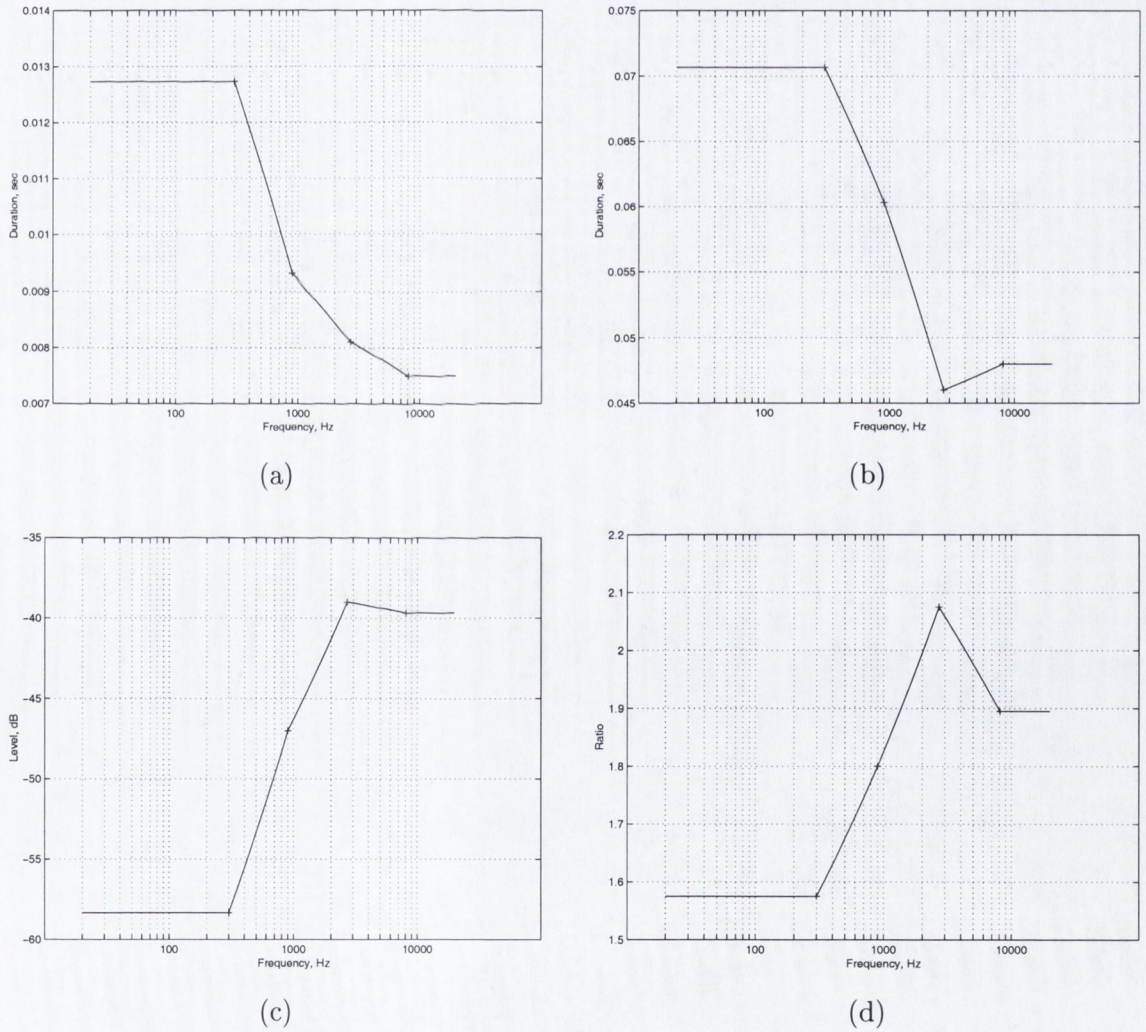
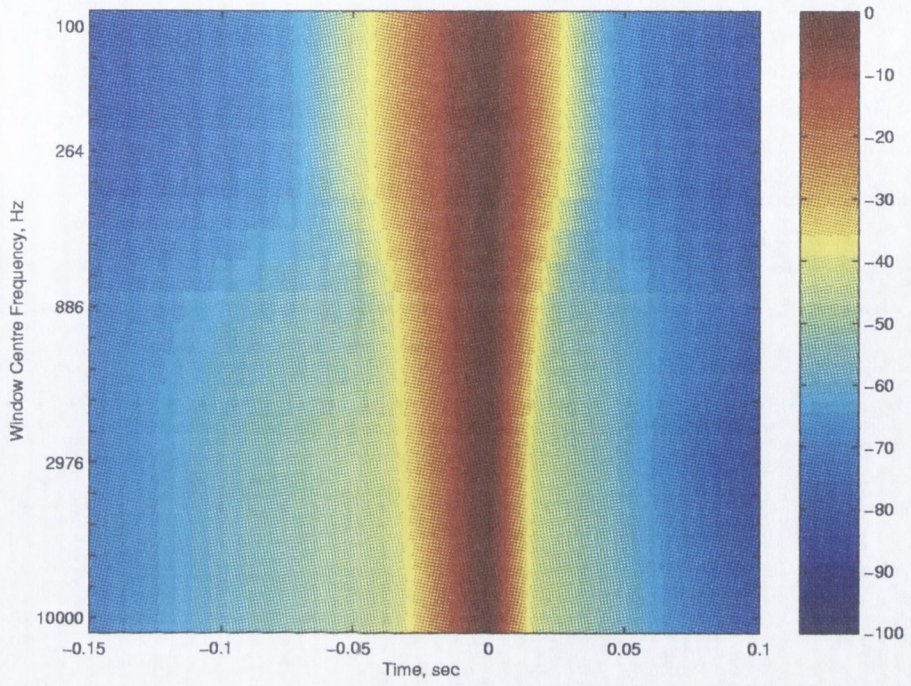


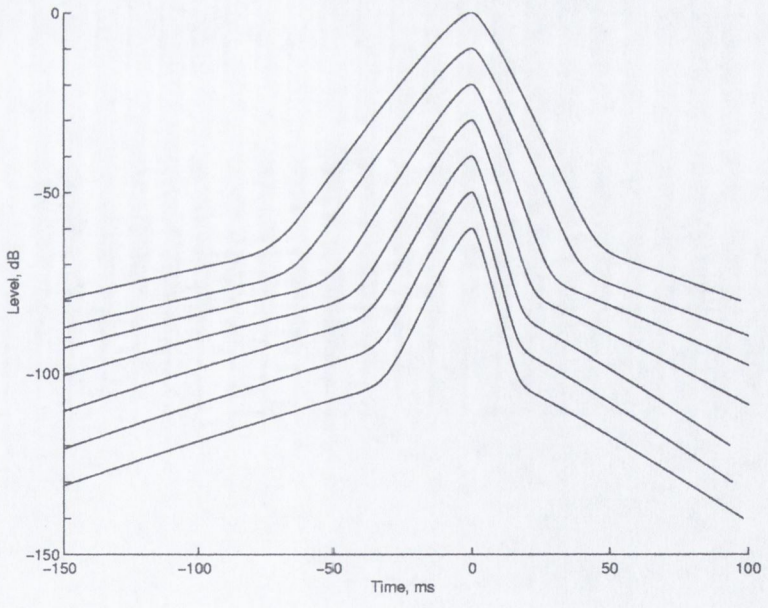
Figure 4.15: Interpolated temporal window shape parameters for the design of a temporal window of arbitrary frequency : (a) total peak duration, T_p, (b) total skirt duration, T_s, (c) skirt level, w and (d) window asymmetry, a.

levels used. However, level dependency is not incorporated in the design of the EWD. Therefore, in order to design a level independent window shape, the data of Plack and Moore is averaged across level. Table 4.1 presents this averaged data. Parameter T_p is the total of the peak time constants ($T_p = T_{pb} + T_{pa}$) with T_s being the total of the skirt time constants ($T_s = T_{sb} + T_{sa}$). Parameter w is averaged across level and ranges between about -60 and -40 dB. The asymmetry of window shape is reflected by the values of T_{pb}/T_{pa} and T_{sb}/T_{sa}, which agree closely at each centre frequency and level; they are therefore averaged across level to give parameter a, characterising overall window asymmetry for each of the four centre frequencies. In order to design a frequency dependent window shape, the values of T_p, T_s, w and a were interpolated with respect to a logarithmic frequency scale. Graphs of the interpolated data are presented in Figure 4.15 (a), (b), (c) and (d); note that the data is assumed to remain constant below 300 Hz and above 8100 Hz, respectively since these frequencies were the limits of Plack and Moore's study. This is a reasonable assumption above 8100 Hz. However it may not be true below 300 Hz since the response of the auditory filter shape may still increase with decreasing frequency [149]. Figure 4.16 (a) presents a surface plot of temporal window shapes derived from this interpolated data for 20 centre frequencies logarithmically spaced between 100 Hz and 10 kHz. The window shape varies smoothly with frequency, hence the interpolation of the temporal window data was deemed successful. Forward masking is more dominant for all frequencies. The peak of the window shape narrows with increasing frequency reflecting increasing temporal resolution. Interestingly, the more gradually sloped skirt has a greater effect at higher frequency; this behaviour is due to an increase in parameter w with frequency, w determining the level of the skirt relative to the peak.

The temporal window shape changes relatively slowly with frequency. For the purposes of EWD analysis, it was therefore decided that a total of 7 temporal windows was sufficient to cover the nominal audible frequency range 20 Hz - 20 kHz, with each temporal window covering a specific band of frequencies. Table 4.2 gives the frequency band of each of these temporal windows. The parameter 'Window F_C' gives the centre frequency of the temporal window assigned to that particular band. For all but the first and last window, this figure is equal to the logarithmic midpoint of 'Band Start' and 'Band End'. The temporal window shape was fixed for frequencies below 300 Hz and above 8.1 kHz because, again, these frequencies were the limits of Plack and Moore's study. The 7 temporal windows are plotted in Figure 4.16 (b), with the temporal window number increasing from top to bottom. The temporal windows are plotted at offsets of 10 dB for comparison purposes. It is clear that the frequency bands are appropriately chosen, as the window shape changes quite evenly from



(a)



(b)

Figure 4.16: (a) Plan view of the frequency dependent temporal window shape at 20 centre frequencies logarithmically spaced between 100 Hz and 10 kHz. (b) The seven temporal windows used for the EWD smoothing kernel.

window to window. Earlier on, in Section 4.1.5 an algorithm achieving frequency dependent temporal and spectral resolution was described which used a frequency dependent kernel. Figure 4.10 graphically portrays this approach. This approach is used for the EWD, its frequency dependent kernel being based on auditory temporal and spectral resolution. The overall algorithm is portrayed in flowchart form in Figure 4.17.

Window Number :	1	2	3	4	5	6	7
Band Start :	0	300	533	949	1687	3000	8100
Band End :	300	533	949	1687	3000	8100	20000
Window F_c :	300	400	711	1265	2250	4930	8100

Table 4.2: Start, centre and end frequencies for each of the 7 temporal windows.

4.3 Modelling the Outer-Middle Ear

Before sound reaches the cochlea it must first be transmitted through the outer and middle ear. These two sections of the peripheral auditory system modify incoming sound in a frequency dependent manner causing amplification of some frequencies and attenuation of others. These effects are most clearly reflected by the threshold of audibility curve, which plots the minimum sound pressure level at which a particular frequency is audible. This curve is plotted in Figure 4.18. From the diagram it is clear that the auditory system is not very sensitive to very low or very high frequencies. It is much more responsive to mid-range frequencies, these frequencies being important for speech production. The resonant effect of the external meatus (outer ear) is principally responsible for the mid-range boost at around 3-4 kHz.

The middle ear plays a vital role in converting the air-borne sound of the outer ear to the fluid-borne vibrations of the cochlea. The transmission of sound from air to water is very inefficient. This is because the impedance of water is about 3750 times that of air. Therefore, without the middle ear, the impedance mismatch between air and sound would mean that 99.9 % of the sound impinging on the eardrum would be reflected. Because of the mass, acoustic resistance and acoustic compliance of the ossicular chain, high and low frequency sounds are attenuated as they pass through the chain. This is reflected in the absolute threshold curve, which displays low sensitivities to these frequencies.

The frequency dependent gain of the combined outer and middle ear are best simulated by an appropriate transfer function modelling its overall input-output behaviour. The effects

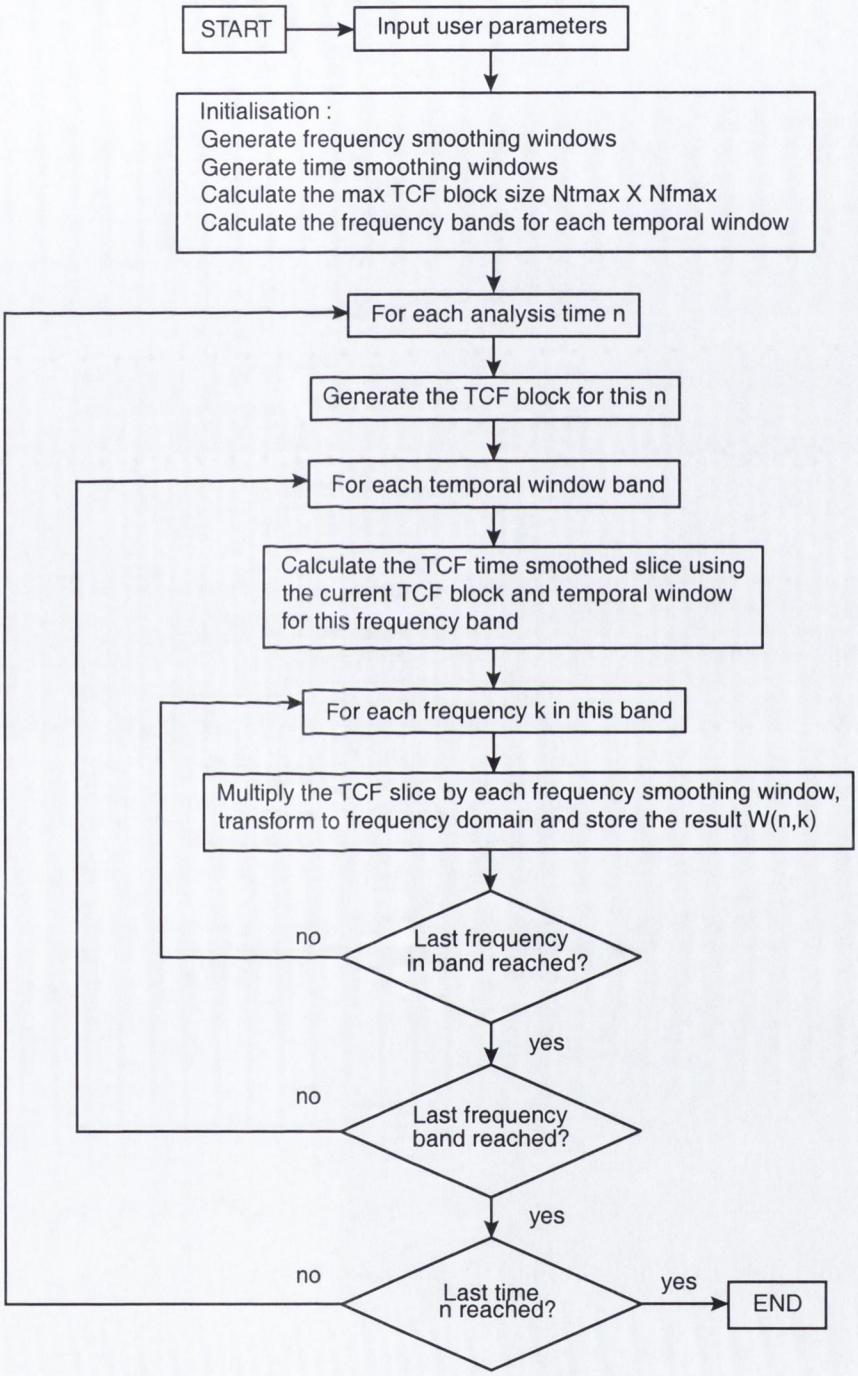


Figure 4.17: Flowchart detailing the overall EWD algorithm.

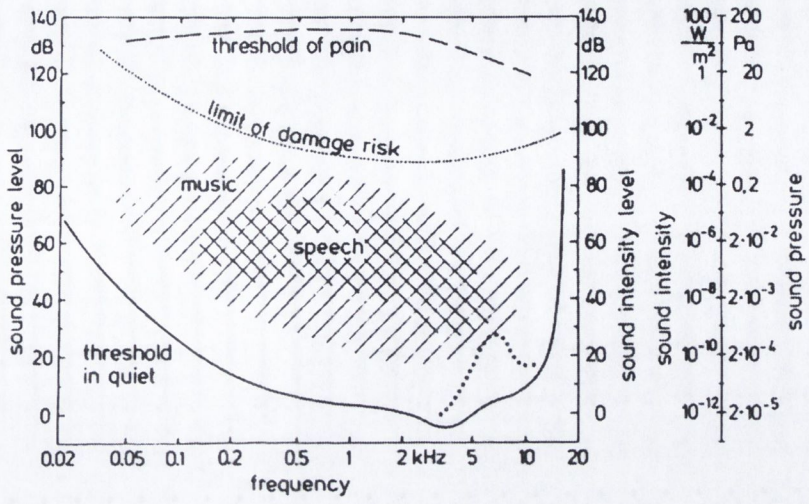


Figure 4.18: Threshold of audibility (threshold in quiet) curve [194].

of the outer and middle ear may thus be replicated by filtering the sound and passing the result to later auditory model stages. In [143], Pfleuger and Hoeldrich present a model of the auditory system which incorporates such a outer-middle ear filtering stage. The overall filter, $H_{OM}(z)$, is designed to simulate monoaural listening under free field conditions and is based on the inverted absolute threshold curve. $H_{OM}(z)$ consists of the cascade of an IIR, lowpass filter of order 8, $H_{LP}(z)$, with an IIR high pass filter of order 2, $H_{HP}(z)$. The system transfer functions for these filters are given below :

$$H_{LP}(z) = \frac{0.190z^7(z+1)}{z^8-2.5359z^7+3.9295z^6-4.7532z^5+4.7251z^4-3.5548z^3+2.1396z^2-0.9879z+0.2836} ,$$

$$H_{HP}(z) = \frac{z^2-2z+1}{z^2-2Rz+R^2} ,$$

$$H_{OM}(z) = H_{LP}(z)H_{HP}(z) .$$

The highpass filter $H_{HP}(z)$ is has one free parameter, R, to allow the model to adapt to recent deviations from the ISO 226 equal loudness contours found by some authors. Pfleuger and Hoeldrich suggest a value of 0.989 as providing a best match to the current data; this figure was therefore used for the EWD model. Figure 4.19 portrays the frequency response of the outer-middle ear model with $R = 0.95$, 0.96, 0.97, 0.98 and 0.99. The overall response provides the high and low frequency attenuation and also the increase in sensitivity of about 10 dB at around 3-4 kHz discussed earlier.

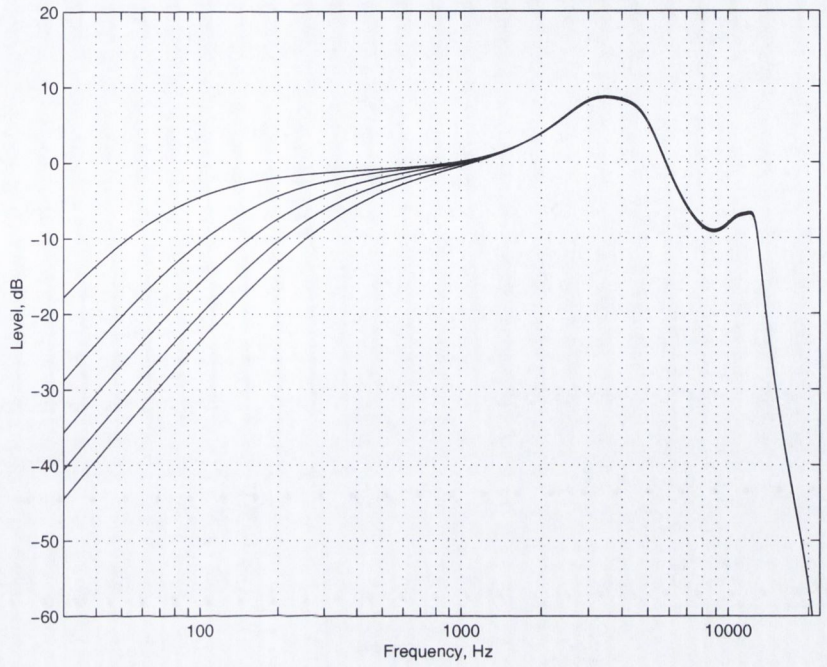


Figure 4.19: Outer-middle ear filter response. The values of R, from bottom to top are : 0.95, 0.96, 0.97, 0.98 and 0.99.

The effect of the outer-middle ear was incorporated indirectly into the design of the EWD by pre-filtering signals to be analysed by the above cascade filter, $H_{OM}(z)$. However, it was not always appropriate to include the effect of the outer-middle ear. Whether or not this pre-filtering was performed depended on the type of signal analysis being carried out using the EWD. In some circumstances, such pre-filtering was undesirable. For example, in Chapter 5, we compare the excitation patterns generated by the EWD to those generated by Glasberg and Moore's level dependent and level independent method of excitation pattern generation. Glasberg and Moore's method employs a slightly different outer-middle ear model. Therefore, to achieve a more direct comparison, outer-middle ear filtering was omitted for both methods. In Chapter 6, the EWD is employed to analyse and smooth HRTF (head related transfer function) data. Such data already contains the effect of the resonances of the auditory canal (outer-ear). Inclusion of the outer-middle ear filtering stage would result in these resonances being applied twice and because of this, the outer-middle ear filtering was not performed.

4.4 Computational Cost of the EWD

Determining the computational cost of calculating the EWD of a signal is complicated as the total number of multiplications and additions required depend on a number of parameters, namely the sampling frequency, the range, distribution and total number of frequency channels, the number of temporal windows used and the time interval at which the EWD is calculated. The calculation of one particular EWD frequency slice can be divided into three stages : *TCF calculation*, *TCF time smoothing* by the temporal window and *TCF frequency smoothing* by the frequency window. Note that the frequency smoothing window is pre-modulated so that this multiplication also transforms from the TCF to WDF domain. The following lists the various operations that must be performed to calculate a slice in frequency of the WDF.

- *TCF Calculation :* For each slice in frequency at fixed time, n, of the EWD we must calculate a TCF block of size $N_{Tmax} \times N_{Fmax}$ samples, where N_{Tmax} and N_{Fmax} are the lengths in samples of the longest time smoothing window, $h_{TA}[n, C_{LT}]$, and longest frequency smoothing window, $h_{FA}[l, C_{LF}]$ respectively. N_{Fmax} depends on the sampling frequency and the minimum channel frequency required. N_{Tmax} also depends on the sampling frequency and the minimum channel frequency, but also depends on the dynamic range D dB of the temporal window. This work found that there was rarely a need to use a temporal window of total dynamic range exceeding 50 dB—a dynamic range of 30 dB was found to be more than adequate for most analysis tasks. However, occasionally a higher dynamic range was useful, when, for example examining the temporal masking of impulsive stimuli. Therefore, in order to better appreciate the effect of these parameters on TCF size, Figure 4.20 presents a plot of TCF block size for different minimum channel frequencies. Here, $N_{Tmax} \times N_{Fmax}$ is divided by f_s^2 (the sampling frequency squared), therefore, in order to calculate the actual TCF size in samples, the ordinate can simply be multiplied by f_s^2. Each curve on the graph represents the TCF size for a different value of temporal window dynamic range D. The lowest curve shown is for a dynamic range of 10 dB. This range increases in steps of 10 dB up to a value of 100 dB for the top graph. The abscissa denotes the minimum frequency analysed; from the graph it is very clear that TCF size increases rapidly for low values of this parameter. This is due to a rapid increase in the size of the largest frequency smoothing window at such frequencies. It is also clear that for dynamic ranges greater than 60 dB, the TCF size grows dramatically. TCF block size also falls approximately exponentially with increasing minimum channel frequency.

Obviously, the total TCF size is also very dependent on the sampling frequency, f_s, being proportional to f_s^2. Since the TCF is Hermitian symmetric, the total number of multiplications required to calculate one TCF block of a real signal is

$$\frac{N_{Tmax} \times N_{Fmax}}{2} . \tag{4.18}$$

A given EWD is composed of slices in frequency spaced at regular intervals in time. If we denote the number of samples between adjacent frequency slices as T_{step}, and the number of frequency slices computed per second as f_{slice}, then

$$T_{step} = \frac{f_s}{f_{slice}} . \tag{4.19}$$

Typically, $T_{step} \ll N_{Tmax}$; there is therefore much data overlap between adjacent TCF blocks. This can be exploited to reduce the computational cost of each TCF block, apart from the first one. Therefore, for all but the first block, the number of multiplications per TCF block reduces to

$$\frac{T_{step} \times N_{Fmax}}{2} . \tag{4.20}$$

- *TCF Time Smoothing :* The number of multiplications needed to achieve a time smoothed TCF strip at each temporal window stage is the product of the *current* temporal window length times half the length of the longest frequency smoothing window in the temporal window frequency band. Estimating the total number of multiplications required for TCF time smoothing is therefore complicated by the fact that more than one temporal window is used, each being of different length. To simplify matters, we say that the total number of multiplications (for a real signal) is approximately equal to

$$\frac{\text{Number of Temporal Windows} \times N_{Tav} \times N_{Fav}}{2} , \tag{4.21}$$

where N_{Tav} and N_{Fav} are the average lengths of the time and frequency smoothing windows, respectively.

- *TCF Frequency Smoothing and Transformation to the WDF Domain :* The total number of multiplications required for TCF frequency smoothing is equal to the total length of all the frequency smoothing windows. Again, to simplify matters, we say

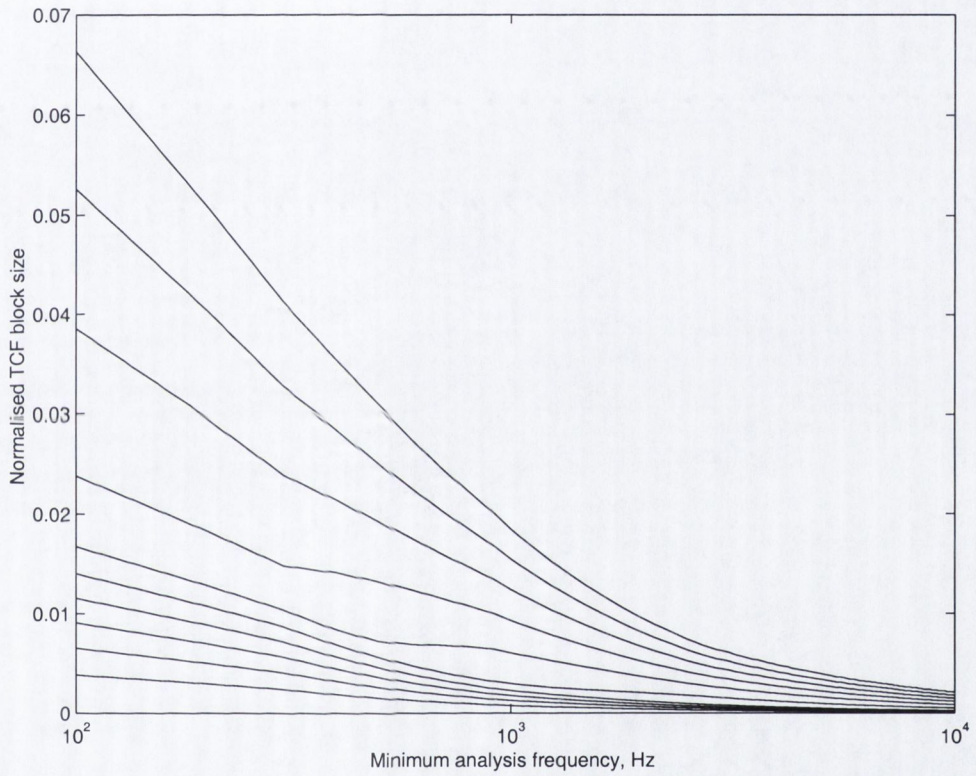


Figure 4.20: Normalised TCF block size for temporal window dynamic ranges of 10 to 100 dB (increasing in steps of 10 dB from bottom to top). The abscissa denotes the minimum frequency analysed.

that the total number of (complex) multiplications to perform frequency smoothing and transform into the WDF domain is

$$\frac{\text{Number of Channels} \times N_{Fav}}{2} , \tag{4.22}$$

where N_{Fav} is the average length of the frequency smoothing window. Assuming that it takes 4 real multiplications per complex multiplication, the total number of multiplications is

$$\text{Number of Channels} \times N_{Fav} \times 2 . \tag{4.23}$$

The total number of multiplications required to compute an EWD per second of signal is equal to the total number of multiplications above for TCF, time smoothing and frequency smoothing, times the number of EWD frequency slices computed per second, f_{slice}. This figure is

$$f_{slice} \times \left\{ \left(\frac{T_{step} \times N_{Fmax}}{2} \right) + \left(\frac{\text{Num. Temp. Wind.} \times N_{Tav} \times N_{Fav}}{2} \right) + \left(\text{Num. Ch.} \times N_{Fav} \times 2 \right) \right\} .$$

It is noteworthy that if the analytic version of the real signal is substituted in EWD calculation, the signal sampling frequency, f_s, can be halved [33], [28]. Hence, N_{Tmax}, N_{Fmax}, N_{Tav} and N_{Fav} are all half that required for the real signal. This would suggest a quartering of the multiplications required for TCF calculation and TCF time smoothing and a halving of the number of multiplications required for TCF slice frequency smoothing. However, if we assume that four real multiplications are required for each complex multiplication, and also that the computational cost of TCF calculation and time smoothing far outweigh the cost of TCF slice frequency smoothing, then the computational advantage of using the analytic signal is unfortunately lost.

4.4.1 Memory Requirements for EWD Calculation

The memory required for EWD computation is almost completely dominated by the memory required to store a single TCF block. According the the IEEE standard for floating point arithmetic, 4 bytes are needed to represent a floating point number; it therefore takes 8 bytes to store a complex floating point number. For double precision, which is used in

MATLAB, 8 bytes per float are used. Therefore a total of 8 bytes per real TCF sample and 16 bytes per complex TCF sample are used. Therefore, assuming N_{Tmax} and N_{Fmax} are halved for analytic signals and assuming that, due to TCF Hermitian symmetry, only half of each TCF block is stored, the following number of bytes are needed to store each TCF block in MATLAB :

$$N_{Tmax} \times N_{Fmax} \times 4 \qquad \text{for real signals samples at } 2f_s \text{ ,}$$
$$N_{Tmax} \times N_{Fmax} \times 2 \qquad \text{for complex analytic signals samples at } f_s \text{ .}$$

Therefore, although using the analytic signal does not result in computational savings, it does result in memory savings of a factor of 2. Using the analytic signal also removes cross term interference between positive and negative frequency components [18].

4.4.2 Real Time EWD Analysis

As an example, the number of multiplications required to compute a real time EWD of a typical audio signal is determined here. The cutoff of the audio signal is 11025 Hz, therefore, to satisfy the minimum sampling frequency for the WD, $f_s = 44100$ as the real signal was used.

The EWD slices in frequency are computed every 2.5 ms and 150 channels, spaced according to the ERB scale from 25 Hz to 11025 Hz are used. The dynamic range of the temporal window is set to 50 dB; seven temporal windows in all are used. For these values, the resulting parameter values are presented in Table 4.3 (the meaning of N_{Tmaxa} is explained shortly).

N_{Tmax}	N_{Fmax}	N_{Tav}	N_{Fav}	N_{Tmaxa}	T_{step}
4369	4401	3829	1144	1588	111

Table 4.3: Values of parameters N_{Tmax}, N_{Fmax}, N_{Tav}, N_{Fav}, N_{Tmaxa} and T_{step} for a particular case of real time EWD analysis.

The number of real multiplications required per second for each stage of EWD processing is presented in Table 4.4. The total number of multiplications required per second is in excess of 6 Gigaflops, which is high. For, example, four TigerSHARC DSP processors operating in parallel would be required to achieve this number of operations. Using 16-bit fixed point accuracy, however, the EWD could be computed on 1 TigerSHARC, albeit with a reduction in accuracy. By far the most computationally heavy operation is TCF

time smoothing, accounting for 98.4 % of the total number of multiplications in the given example. The cost of this could readily be reduced by lowering the dynamic range of the temporal windows and also lowering the number of temporal window stages. This would not result in a serious compromise of temporal resolution because the temporal window shape is relatively invariant across frequency.

TCF Calc.	TCF T. Sm.	TCF F. Sm	Total Mults/sec
$97.724.10^6$	$6.133.10^9$	$85.8.10^3$	$6.231.10^9$

Table 4.4: Total number of multiplications per second for TCF calculation, time smoothing and frequency smoothing/transformation for the particular example given.

In [18], Boashash and Black present a real time implementation of the discrete SPWD. As they note, there is an unavoidable latency in computing the DSPWD as TCF calculation requires signal samples both both positive and negative time. This latency is equal to $max(N_{Fmax}/2, N_{Tmaxa})$, where N_{Tmaxa} is the length of the positive half of the longest temporal window. For the example above, this latency is equal to N_{Fmax} and is 2201 samples, *i.e.* 50 ms. As the minimum channel frequency is reduced, $N_{Fmax}/2$ increases whereas N_{Tmaxa} increases as the dynamic range of the temporal window is increased. The greater of these determines the latency.

The memory requirement of the above example is dominated by the size of the TCF blocks. For MATLAB, 36.7 Mbytes for the real signal TCF block are required, each sample being stored using double floating point accuracy (8 bytes per sample). Floating point DSPs generally use 32-bit accuracy, therefore this figure may be halved for these devices, giving 18.35 Mbytes. If the analytic signal is used, then this figure may be halved again to give 9.17 Mbytes. If the analytic signal is used, and 16-bit fixed point arithmetic is used, the TCF memory requirements are halved again, giving 4.59 Mbytes. Smoothing window storage covers the majority of remaining memory requirements, typically adding another 10 % to the TCF figure.

4.4.3 Comparison of EWD and Filterbank-Based Auditory Models

The Wigner distribution algorithm is computationally expensive, especially for long signals. However, this thesis shows that the computational cost and memory requirements of the Wigner distribution may be very considerably reduced if the techniques outlined in this chapter are employed. In fact, surprisingly, the EWD algorithm has a computational complexity which is comparable to that of typical linear filterbank-based auditory models, as is

demonstrated here.

Filterbank models are generally based on an array of auditory filters which process the stimulus signal, followed by a nonlinear device (often square-law) and finally by a temporal integrator, or temporal window. One such linear model is the Spectro Temporal Excitation Pattern (STEP), proposed by Plack [150]. In this model, the input signal is first filtered by a bank of gammatone filters. The filterbank output is passed through a square law device and then smoothed using the temporal window model. Thus, the STEP is designed to model both auditory spectral and temporal resolution. If the STEP is computed at the same points in TF as the EWD, the number of multiplications per second is

$$\text{Number of Channels} \times f_{slice} \times \left\{ \left(\frac{Nt_{av}}{2}\frac{Nf_{av}}{4}\right) + Nt_{av} \right\} \tag{4.24}$$

Here, it is assumed that the temporal window shapes and gammatone filters used in STEP processing are equal in length to those used to design the EWD smoothing kernel. The filterbank model does not require the double signal sampling rate required by the WD. Also, autocorrelation of the frequency smoothing windows, which doubles their length, is unnecessary. Therefore the STEP temporal windows and gammatone filters are half and quarter the length of their EWD counterparts, respectively.

Equation 4.24 may be broken down as follows. At each TF analysis point, the signal must be convolved by the gammatone filter for a duration equal to the length of the temporal window. This corresponds to a total of

$$\text{Number of Channels} \times f_{slice} \times \left\{ \left(\frac{Nt_{av}}{2}\frac{Nf_{av}}{4}\right) \right\} \tag{4.25}$$

multiplications. The result is then smoothed by the temporal window and squared. This is an additional

$$\text{Number of Channels} \times f_{slice} \times \left\{ \left(\frac{Nt_{av}}{2} + \frac{Nt_{av}}{2}\right) \right\} \tag{4.26}$$

multiplications.

If the real time signal analysis example in Section 4.4.2 is used here, the total number of multiplications per second required by STEP analysis is $3.1555.10^9$. This is approximately half that required for EWD analysis. However, the computational load of both algorithms depend on a number of differing factors, including sampling frequency, the dynamic range of the temporal windows used and the intervals in time and frequency at which the analysis is to be performed. For example, the main processing load for STEP calculation can be attributed to frequency smoothing, whereas for the EWD it is temporal smoothing which

necessitates the bulk of the processing. The number of multiplications required for STEP is proportional to the number of frequency channels. A doubling of the number of frequency channels would cause a doubling of the number of STEP multiplications required, whereas the processing load for EWD analysis would be almost unchanged. In other circumstances, for example if the dynamic range of the temporal window is increased or the minimum analysis frequency is very low, the TCF block size for EWD calculation may become very large, resulting in a significant increase in computational load for the EWD. Therefore, it is not unreasonable to assume that, in general, the performance of the EWD in terms of computational load is comparable to that of linear filterbank models.

4.5 EWD Computation Platform

In this work, the EWD algorithm was coded in the MATLAB environment using the Signal Processing Toolbox on a PC equipped with an Intel Pentium-2 processor running at 266 MHz. The MATLAB environment was ideally suited to the task as many useful operations such as window and filter design, FFTs, Hilbert transforms, convolution, correlation and vector and matrix manipulation routines are provided via simple function calls. Furthermore, MATLAB provides excellent 2-D and 3-D plotting tools. As MATLAB is an interpretive language, it is far less efficient than C, however using its profiler, MATLAB code 'hot-spots' may be identified and efficiently coded in C as dynamically linked subroutines known as MEX-files. In this work, due to time constraints, the EWD was coded entirely in MATLAB. However, it is suggested that coding the TCF generation, TCF time smoothing and TCF frequency smoothing operations as MEX-files would result in a very considerable increase in EWD computation speed.

Chapter 5

Model Validation

This chapter seeks to determine whether or not we have met the design goals set out in Chapters 3 and 4, namely, whether or not the spectral and temporal resolutions of the EWD accurately match those of the gammatone filter and temporal window model of masking respectively. It is important to validate the EWD resolutions as the EWD represents a novel approach to modelling auditory masking, employing as it does the new kernel design techniques described in Chapter 4. Results showing that the EWD does indeed achieve the correct resolutions are presented in this chapter. The presence of EWD cross-term interference is also investigated as such interference can often confuse the interpretation of bilinear time-frequency distributions. It is found that for the EWD, cross-term interference is negligible as it is almost always suppressed below auto-term excitation level. For the particular case of a signal combined with a delayed reflection of itself, a direct and useful relationship between cross-term interference oscillating parallel to the frequency axis and auditory comb filtering is demonstrated. Cross-term interference and windowing artifacts have in the past limited the effective dynamic range of time-frequency distributions. The dynamic range of the EWD is therefore evaluated here and is shown to be effectively limited only by the dynamic range of the signal itself and not artifacts of the time-frequency signal analysis technique. Finally, some psychoacoustic limitations of the EWD are discussed.

5.1 EWD Properties

In Chapter 2, Table 2.1 presented the time-frequency properties of Cohen's class, and their corresponding kernel constraints. Table 5.1 now presents the case for the EWD, indicating the same properties and kernel constraints and whether or not they are satisfied. From the table it is clear that although the EWD satisfies time shift covariance, is real and even valued, all other properties are not satisfied. However this fact should not be seen as a

limitation of EWD analysis, but rather required features of the EWD's smoothing kernel. Firstly, the time and frequency marginals (P_4 and P_5) of the EWD do not equal the signal marginals, they equal the excitation pattern and averaged temporal excitation pattern (see Section 5.2) of the signal respectively. Since the marginals are not satisfied, the group delay and instantaneous frequency properties (P_{10}, P_{11}) are not satisfied. Furthermore, although the EWD is time shift covariant, it is not frequency shift covariant, its kernel being a function of frequency (P_7). The EWD kernel is not designed to satisfy the time and frequency support properties either as the kernel is designed to smear distribution energy in both time and frequency (P_8 and P_9). Finally, total EWD energy does not equal total signal energy (P_3) and the modulation and covariant properties are not satisfied (P_{12}, P_{13}).

Property	Kernel Constraint	EWD
P_1 : Reality	$\Psi^*(-\tau,-\theta) = \Psi(\tau,\theta)$	$\surd$
P_2 : Evenness	$\Psi(\tau,\theta)$ is real	$\surd$
P_3 : Energy Conservation	$\Psi(0,0) = 1$	
P_4 : Time Marginal	$\Psi(0,\theta) = 1$	
P_5 : Frequency Marginal	$\Psi(\tau,0) = 1$	
P_6 : Time Shift Covariance	$\Psi(\tau,\theta)$ is not dependent on time, t	$\surd$
P_7 : Frequency Shift Covariance	$\Psi(\tau,\theta)$ is not dependent on frequency, f	
P_8 : Finite Time Support	$\phi(t,\tau) = 0, \quad \lvert t/\tau \rvert > 1/2$	
P_9 : Finite Frequency Support	$\psi(\omega,\theta) = 0, \quad \lvert \omega/\theta \rvert > 1/2$	
P_{10} : Group Delay	$\Psi(\tau,0) = 1$ and $\frac{\delta}{\delta\theta}\,\Psi(\tau,\theta)\big\rvert_{\theta=0} = 0$	
P_{11} : Instantaneous Frequency	$\Psi(0,\theta) = 1$ and $\frac{\delta}{\delta\tau}\,\Psi(\tau,\theta)\big\rvert_{\tau=0} = 0$	
P_{12} : Convolution Covariance	$\Psi(\tau_1+\tau_2,\theta) = \Psi(\tau_1,\theta)\Psi(\tau_2,\theta)$	
P_{13} : Modulation Covariance	$\Psi(\tau,\theta_1+\theta_2) = \Psi(\tau,\theta_1)\Psi(\tau,\theta_2)$	

Table 5.1: Cohen's class properties and corresponding kernel constraints. A tick indicates that a property is satisfied by the EWD.

5.2 Spectral and Temporal Masking : Model Accuracy

As described in Section 3.2.5, the excitation pattern of a sound is an important indicator of the internal spectral representation of the sound. Typically the excitation pattern of a signal is composed of the individual contributions of a number of continuous sinusoids which are assumed to make up the signal. For example, Figure 3.12 in Section 3.2.5 demonstrated the

calculation of the excitation pattern of a complex tone consisting of the first 10 harmonics of a 300 Hz fundamental. The excitation pattern can be considered a smeared version of a signal's energy spectral density (for time limited signals) or power spectral density (for periodic signals). As such, temporal information is not represented, therefore excitation patterns cannot represent temporal masking. The term *temporal excitation pattern* was introduced by Plack [150] to describe a plot showing the output (in dB) of a temporal window function (of particular frequency) versus time. This plot is therefore a temporal equivalent of the (spectral) excitation pattern, representing auditory temporal resolution and temporal masking. Since a particular signal's energy is not, in general, evenly distributed across frequency, a plot of the outputs of several temporal windows of differing centre frequency versus time is necessary to adequately represent temporal resolution. It is convenient to therefore represent this information on a three dimensional plot with dimensions time, frequency and level.

As discussed in Chapter 2, one of the properties of the Wigner distribution is that the frequency marginal equals the energy spectral density of the signal. For the EWD to achieve accurate spectral resolution and thus model spectral masking, its frequency marginal should yield the excitation pattern of the signal. This is shown to be true in Section 5.2.1. Furthermore, for accurate modelling of temporal resolution and temporal masking, the time resolution of the EWD should equal that of the temporal model of masking. This may be tested by calculating the EWD of very short, impulsive sounds and comparing the EWD output of each channel along time with the temporal window shape for that particular channel. Both should agree closely for accurate temporal resolution modelling. This test is outlined further in Section 5.2.2, and results showing a very close match between EWD time resolution and temporal window shape are presented.

5.2.1 Spectral Resolution Accuracy

In order to test EWD spectral resolution, the EWD frequency marginals and excitation patterns can be compared for various signals. A method of generating excitation patterns is therefore needed. Section 3.2.5, summarised such a method proposed by Moore and Glasberg in 1983 based on the symmetric, level independent roex(p) filter shape [110]. The equivalence of this method to EWD excitation pattern generation is also demonstrated. Later work [111] by the same authors in 1987 introduces level dependency into the slopes of the upper and lower skirts of the filter and includes a Fortran program for the calculation of the resulting excitation patterns. The slopes of the filter shape depend on input level; for increasing level the lower skirt slope becomes more shallow, while the upper frequency

skirt becomes steeper. Further work [63] in 1990 more accurately measures auditory filter shape by incorporating a number of improvements into the measurement procedure. In this work it is found that changes in the upper frequency filter slope are much smaller and less consistent than changes in the lower frequency slope. These minor changes are therefore ignored; only the slope of the lower skirt of the auditory filter shape is assumed level dependent. Fortran code listings for the derivation of filter shapes from notched noise data and the calculation of excitation patterns are also included. The level dependency of the lower skirt is briefly summarised as follows. The shape of the level independent, symmetric reox(p) shape is

$$roex(p) = (1 + pg)\exp(-pg) \quad , \tag{5.1}$$

where g is normalised frequency ($g = |f - f_c|/f_c$), and p controls both the filter bandwidth and filter skirt slope. The value of p may be determined by from the filter centre frequency, f_c kHz, and equivalent rectangular bandwidth (ERB) from the following equations :

$$p = \frac{4f_c}{ERB} , \tag{5.2}$$

$$ERB = 24.7(4.37f_c + 1) . \tag{5.3}$$

In [63], Glasberg and Moore note that the auditory filter centred at 1 kHz is roughly symmetric for an input level of 51 dB/ERB. They also note that filter shapes at other centre frequencies are also symmetric for an input level equivalent to the level of 51 dB/ERB at 1 kHz. They summarise the level dependency of the low frequency skirt of the auditory filter by varying the level dependent low frequency skirt parameter, $p_{l(X)}$ according to the following equation :

$$p_{l(X)} = p_{l(51)} - 0.38(p_{l(51)}/p_{l(51,1k)})(X - 51), \tag{5.4}$$

Here, $p_{l(51)}$ is the value of p at f_c for an equivalent noise level of 51dB/ERB and $p_{l(51,1k)}$ is the values of p_l at 1 kHz for a noise level of 51 dB/ERB. The parameter X represents the actual input level at the filter centre frequency in dB/ERB. Both C and Fortran code implementing the generation of excitation patterns based on the above filter shape are available from the homepage of the Auditory Perception Group, University of Cambridge [8].

The magnitude response of the gammatone filter of order 4 is very close to that of the symmetric, level independent roex(p) filter shape of equal ERB. Excitation patterns generated using the EWD should therefore compare well with those generated from this shape. In order to generate such excitation patterns, a simple modification of the code

above was made, effectively disabling filter level dependency. In the following sections, excitation patterns generated using the EWD and both level dependent and independent auditory models are compared for a variety of signals. The signals used in the comparisons are sinusoids at 150, 1200 and 6000 Hz, a complex tone of 200 Hz fundamental, a notched noise masker plus tone and a synthetic vowel.

In order to eliminate slight differences between the outer-middle ear models used, the signals were not prefiltered by the outer-middle ear filter described in Section 4.3 for the EWD. Likewise, the option to apply equal loudness contour (ELC) 'correction' to account for the transmission characteristics of the outer and middle ear was not taken in the Cambridge program. For simplicity, from this point on we will refer to the level independent, symmetric roex(p) filter shape as the *symmetric roex(p) shape*, and the level dependent, asymmetric roex(p) shape as the *asymmetric roex(p) shape.*

In addition to a graphical comparison, the mean squared error (MSE) and percentage error difference (PED) were computed to obtain a measure of the difference between excitation patterns generated by the EWD and those generated by the symmetric and asymmetric shape. The MSE and PED were computed from the normalised linear magnitude of the EWD excitation patterns ($\hat{x}[k]$) and symmetric/asymmetric excitation patterns ($x[k]$) according to the following equations, where C was the number of frequency channels. For the cases of the single sinusoids (Figures 5.1 to 5.3), normalisation was carried out so that the level of the sinusoid in question was 1 (0dBs). For the case of the complex tone signal (Figure 5.4 and Figure 5.5), signal tone plus masker (Figure 5.6) and vowel sound (Figure 5.8) normalisation was performed so that the spectral level 200 Hz, 2 kHz, and 375 Hz was 0 dBs, respectively.

$$\text{PED} = \left\{ \sum_{k=1}^{C} (x[k] - \hat{x}[k])^2 \right\} / \left\{ \sum_{k=1}^{C} x^2[k] \right\}$$

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^{C} (x[k] - \hat{x}[k])^2$$

Single Sinusoids

Comparisons of the excitation patterns generated using the EWD and symmetric roex(p) filter shape were made for signals consisting of a single sinusoid. The sinusoid frequencies were 150, 1200, and 6000 Hz. As described earlier in Chapter 4, the effect of the outer-middle ear was not modelled. For the purposes of EWD analysis, the duration of each sinusoid was set to 100 ms. In order to prevent abrupt signal onsets and decays from causing spectral

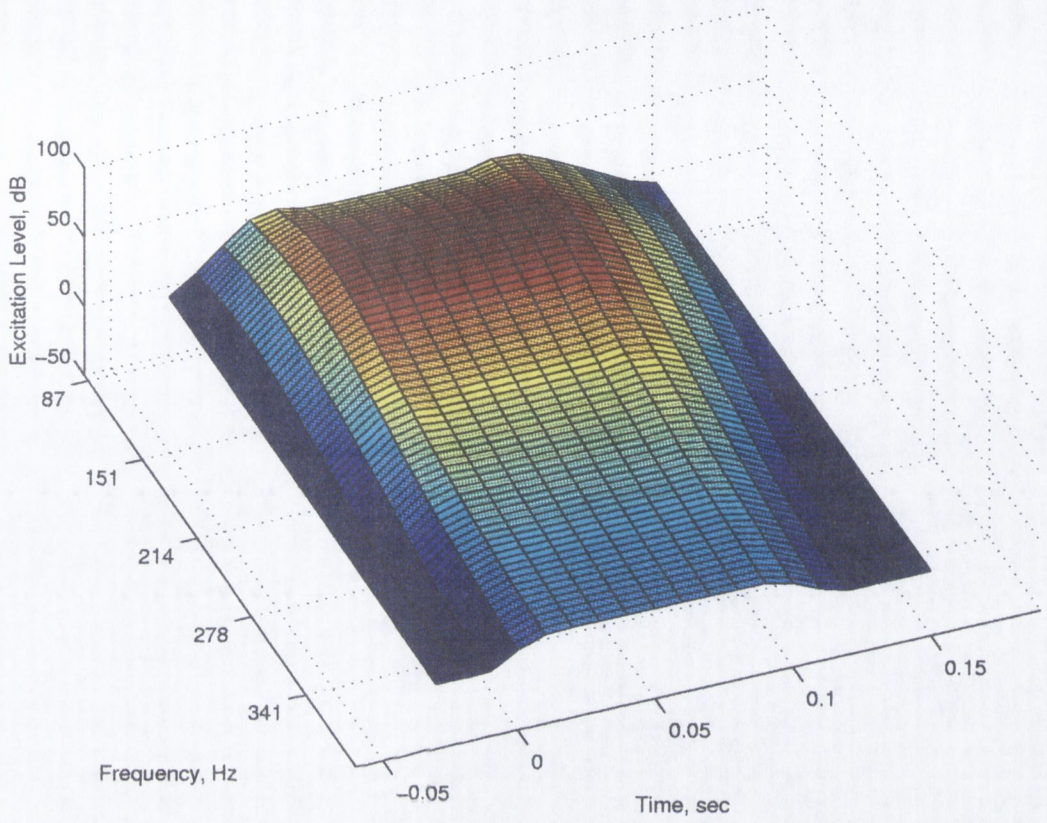
splatter, the first and last 5 ms of each sinusoid was ramped and damped by the left and right halves of a 10 ms hanning window. The distribution was calculated every 10 ms, from $t = -50$ ms to $t = 150$ ms, i.e. from 50 ms before signal onset to 50 ms after signal decay. The temporal window dynamic range was set to 50 dB for all three signals.

Figure 5.1 (a) shows the EWD for the 150 Hz sinusoid signal. The EWD is computed over a linear frequency range from 87 to 400 Hz, 87 Hz being equal to 3 on the ERB-rate scale, the minimum input frequency allowed by the Cambridge excitation pattern code. The 150 Hz signal produces excitation over a wide range of frequencies. Note that although the gammatone response used in EWD analysis is symmetric, the resulting TF excitation pattern is markedly asymmetric, showing a greater spread of excitation for frequencies above the signal frequency than for those below it. Also, the signal produces excitation outside of its duration, showing both forward and backward masking, forward masking clearly being the more dominant effect. The EWD thus displays both the spectral and temporal masking effects of this signal on one three-dimensional plot. Note that the EWD is plotted in units of dB, not dB SPL; the EWD units are not intended to reflect signal level in dB SPL. In fact, since the EWD is level independent, it will, in this work, generally be normalised to a level suitable for comparison with the symmetric and asymmetric roex(p) excitation patterns

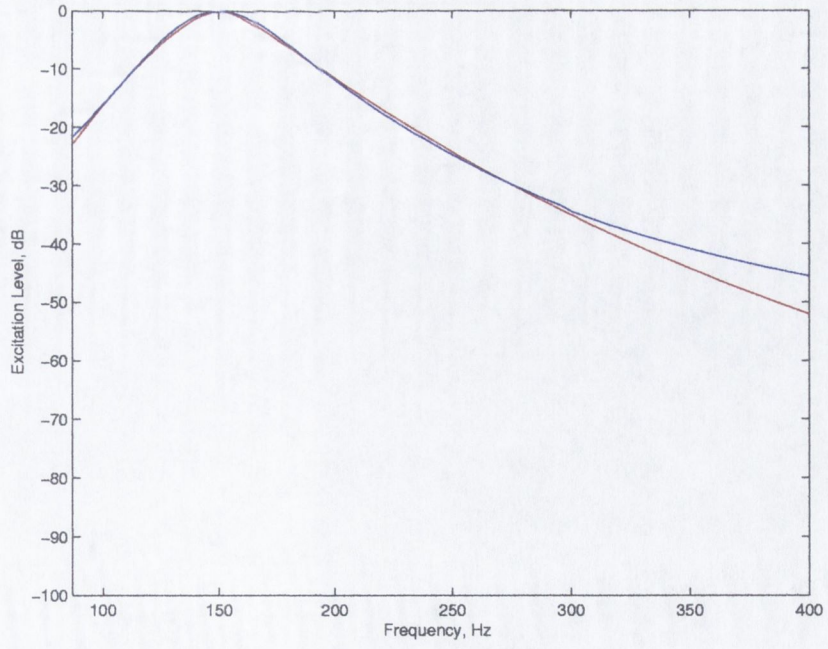
In order to compute the excitation pattern for the 150 Hz sinusoid, the EWD is simply summed along the time axis and plotted versus frequency. Figure 5.1 (b) presents the EWD excitation pattern (blue curve) with the symmetric roex(p) excitation pattern (red curve), for comparison purposes. The two excitation patterns agree very closely for the first 40 dB, after which the slope of the EWD excitation pattern begins to fall in comparison to the slope of the roex(p) excitation pattern. However, the gammatone filter and roex(p) filter shape are not identical so we can only expect a close match for band limited signals such as sinusoids over a limited dynamic range, beyond which the filter shapes begin to differ. The MSE and PED are both very low for this signal, being $3.71.10^{-4}$ and $2.40.10^{-3}$ respectively.

Figure 5.2 (a) and (b) and Figure 5.3 (a) and (b) present exactly the same scenario, now for frequencies of 1200 and 6000 Hz respectively. For the 1200 Hz signal, the analysis frequency range extends from 800 to 3000 Hz, and for the 6000 Hz signal the corresponding range is 4000 to 10000 Hz. The results are similar to those presented for the 150 Hz sinusoid, EWD excitation agreeing very closely to roex(p) excitation over a dynamic range of about 40 dB. Again, the MSE and PED are very low, being $2.89.10^{-4}$ and $3.11.10^{-3}$ respectively for the 1200 Hz signal and $4.04.10^{-4}$ and $3.32.10^{-3}$ respectively for the 6000 Hz signal.

These are encouraging results, as the excitation level well away from signal centre frequency for such simple signals depends on the low level filter response far from its passband.

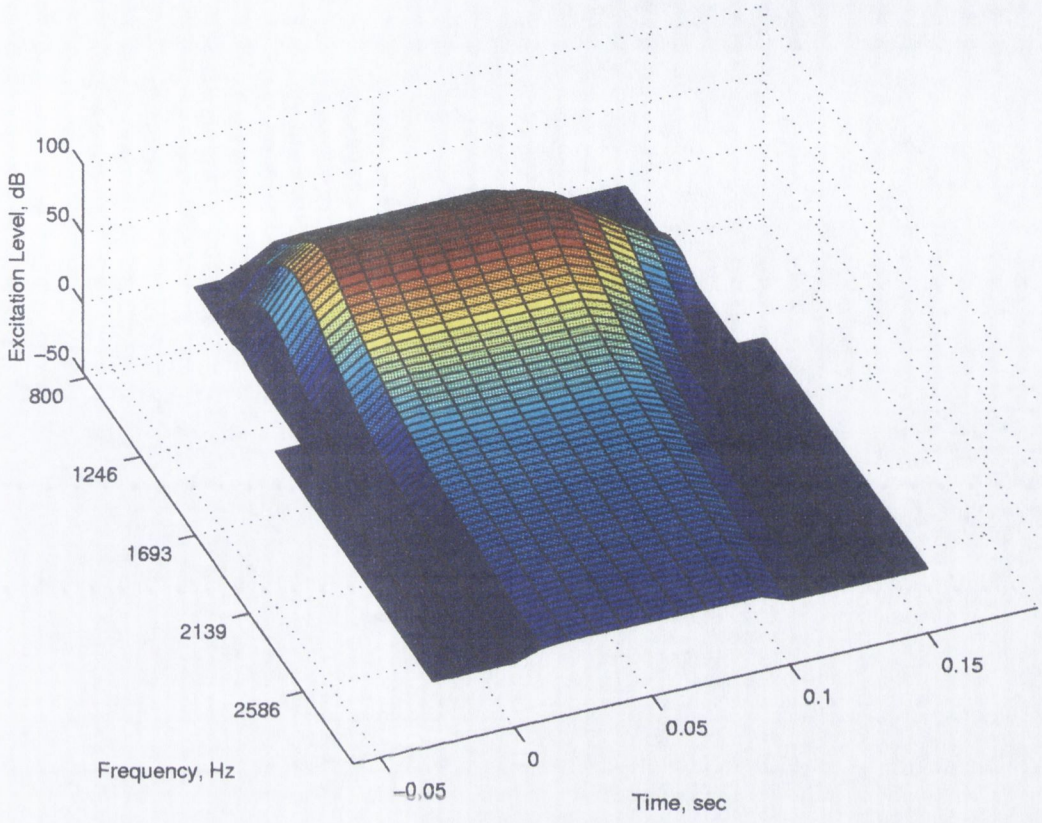


(a)

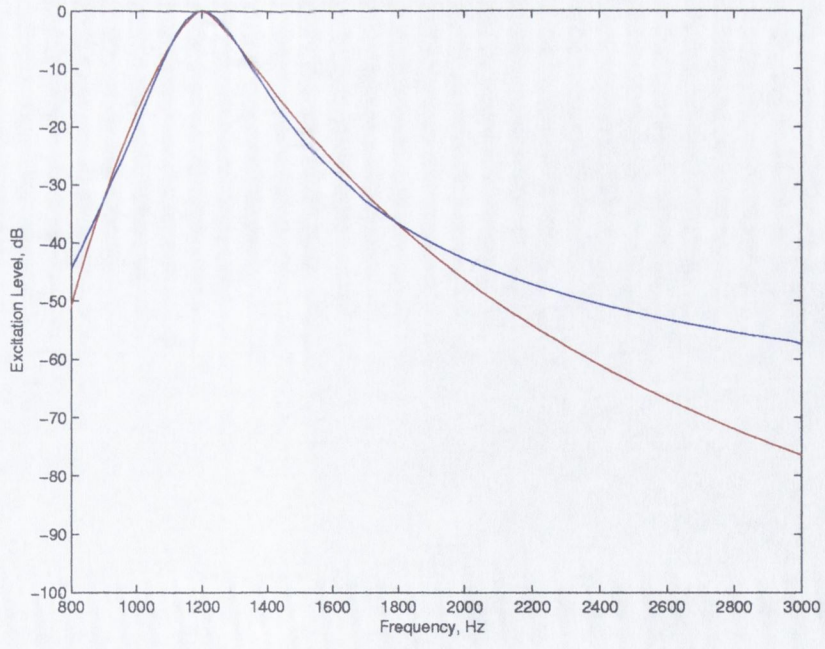


(b)

Figure 5.1: (a) EWD of a 150 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the symmetric roex(p) shape (red curve) and the EWD (blue curve). $MSE = 3.71.10^{-4}, PED = 2.40.10^{-3}$.

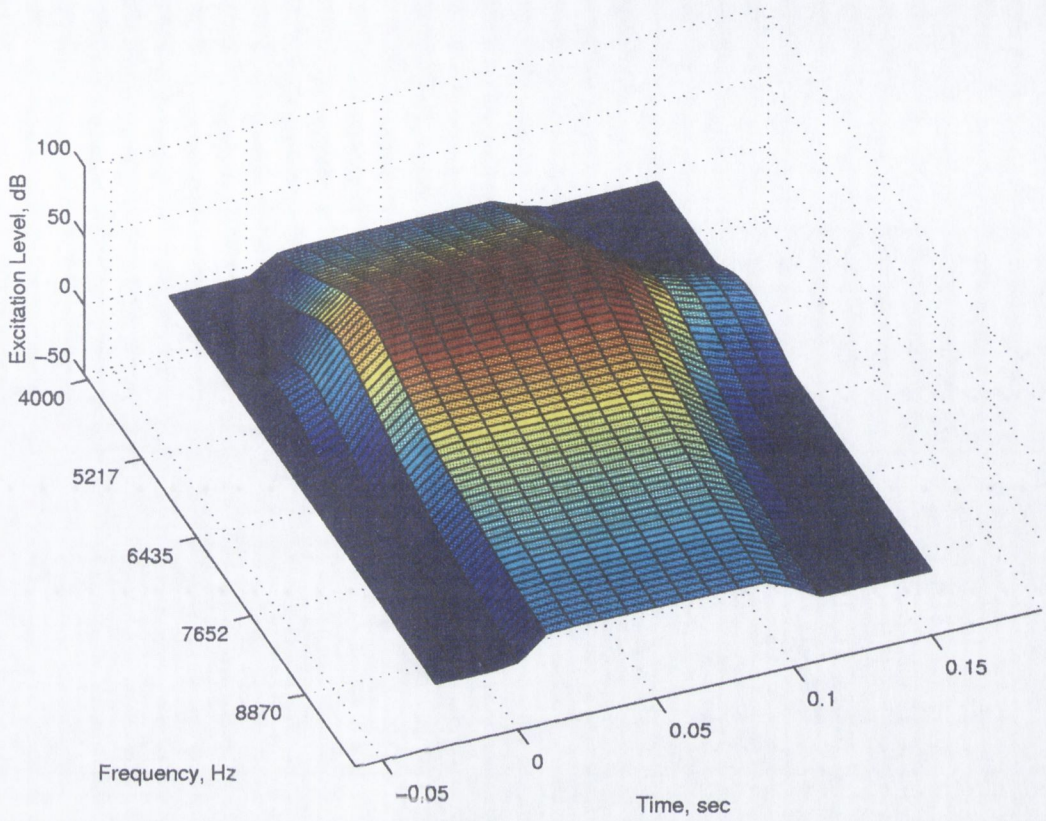


(a)

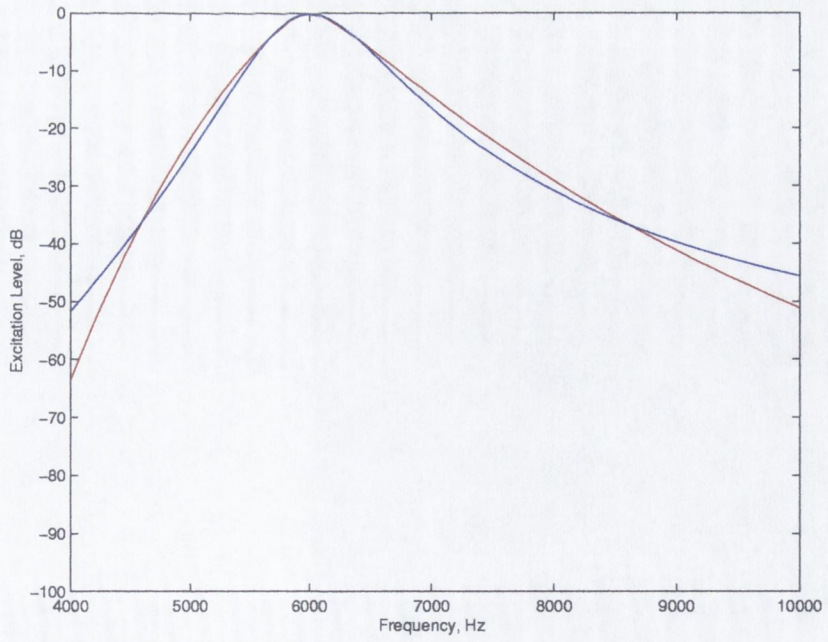


(b)

Figure 5.2: (a) EWD of a 1200 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the roex(p) shape (red curve) and the EWD (blue curve). $MSE = 2.89.10^{-4}, PED = 3.11.10^{-3}$.



(a)



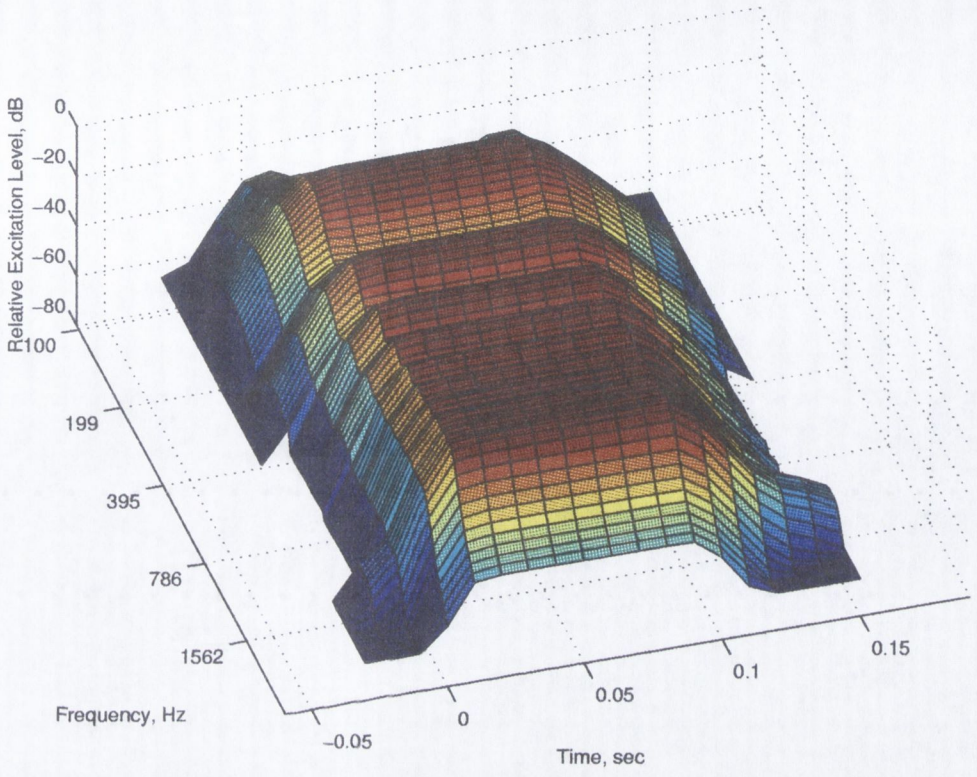
(b)

Figure 5.3: (a) EWD of a 6000 Hz sinusoid signal. (b) Comparison of the excitation patterns generated by the roex(p) shape (red curve) and the EWD (blue curve). $MSE = 4.04.10^{-4}, PED = 3.32.10^{-3}$.

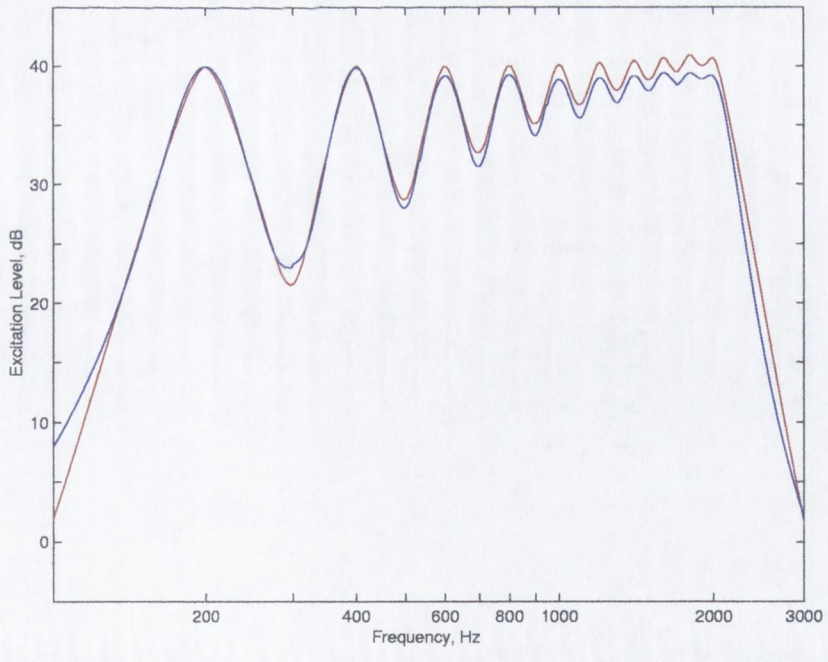
For the excitation patterns of such signals to match across all frequencies, the gammatone and roex(p) filter shape would have to match over a wide dynamic range. However, for more broadband (and realistic) signals, such as music or speech, excitation is primarily determined by filter behaviour nearer the peak of the filter shape. It is shown in the subsequent section that for such signals, the EWD and symmetric roex(p) excitation patterns agree very closely across the full signal bandwidth, and that these excitation patterns also agree well to those of the asymmetric filter shape for moderate signal levels.

Complex Tone

Figure 5.4 (a) presents the EWD of a signal consisting of the first 10 harmonics of a 200 Hz fundamental. The levels of the harmonics were set to be equal. The signal was ramped and damped using the hanning window in exactly the same manner as for the single sinusoid signal. The signal duration was 100 ms and again was not prefiltered by the outer-middle ear filter. The EWD was calculated over the frequency range 100 to 3000 Hz, with channel frequencies now spaced logarithmically within this range. The EWD shows no presence of cross-term interference whatsoever between any of the sinusoidal components. This is a very satisfying result which is discussed further in Section 5.3. Figure 5.4 (b) now compares the EWD (blue curve) and roex(p) (red curve) excitation pattern. The EWD excitation pattern is normalised to give a level of 40 dB at 200 Hz. There is a very good match between the two excitation patterns across all frequencies and levels. Note that the tiny discontinuity at exactly 300 Hz in the EWD excitation pattern is caused by cross-term interference between the 200 and 400 Hz harmonics and is an indication of the almost negligible effect of such interference on the calculation of excitation patterns using the EWD. The roex(p) curve shows a slight increase in excitation level for increasing frequency; this is due to the fact that the ERB of the auditory filter shape spans an increasing number of harmonics as frequency increases resulting in an increase in filter input level. However this is not seen in the excitation pattern of the EWD, which in fact shows a very slight drop in excitation level as frequency increases. This can be explained by the fact that the temporal window shape narrows for increasing frequency. For this particular signal this has the effect of just counteracting the increased input level of the auditory filters. However this should not necessarily be seen as a limitation of the EWD, as it reflects the combined effects of both spectral and temporal masking, whereas the roex(p) excitation patterns ignores temporal masking effects. Looking at the figures for MSE and PED we see that they are low but are still higher than those for the single sinusoids, being $7.51.10^{-3}$ and $1.52.10^{-2}$ respectively. This might seem surprising at first since the EWD and roex(p) curves match more closely



(a)



(b)

Figure 5.4: (a) EWD of a complex tone consisting of the first 10 harmonics of a 200 Hz fundamental. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) filter shape (red curve) and the EWD (blue curve). $MSE = 7.51.10^{-3}, PED = 1.52.10^{-2}$.

than for the single sinusoide. However the inceaase in MSE and PED can be explained by the fact that the spectral level for the complex tone is far higher than for the single sinusoids for most of the frequency range.

It is often stated that the gammatone filter is a reasonable approximation of auditory filter shape for moderate signal levels [143], [150], [131]. To demonstrate this, excitation patterns for the complex tone described above were generated for four component levels of 30, 50, 70 and 90 dB SPL using the asymmetric roex(p) filter shape. The resulting level-dependent excitation patterns are plotted as the black curves in Figure 5.5 and the corresponding MSE and PED values are listed in Table 5.2. Note that since outer-middle ear correction was not performed, the excitation levels of the components appear at their signal levels in dB SPL. The corresponding excitation patterns generated by the symmetric roex(p) filter shape (red curves) and EWD (blue curves) curves are also plotted at each level.

To facilitate visual comparison, the levels at 200 Hz of the EWD, symmetric and asymmetric excitation patterns are set equal for all four sets of curves. It can be seen that for the level of 50 dB SPL, the asymmetric and symmetric roex(p) excitation patterns agree almost exactly. For the increased levels of 70 and 90 dB SPL, the lower slope of the asymmetric roex(p) filter shape becomes more shallow, causing an upward spread of excitation. For the low level of 30 dB SPL, the lower skirt becomes steeper, resulting in the opposite effect—a reduction in excitation above the signal component frequencies. Looking at Table 5.2 we see that the MSE and PED were lowest for the symmetric filter shape. As the signal level increases from 30 to 90 dBSPL, the MSE and PED values for the asymmetric filter shape increase, indicating a worsening of fit. Overall, however, the EWD excitation patterns provide a very good match to the asymmetric roex(p) excitation patterns for moderate signals levels, and differences between the two only become significant for very high signal levels.

Models	Level/dBSPL	MSE	PED
EWD and symmetric shape	ALL	$7.51.10^{-3}$	$1.55.10^{-3}$
EWD and asymmetric shape	30	$9.01.10^{-3}$	$3.81.10^{-3}$
EWD and asymmetric shape	50	$1.07.10^{-2}$	$5.03.10^{-3}$
EWD and asymmetric shape	70	$3.53.10^{-2}$	$1.96.10^{-2}$
EWD and asymmetric shape	90	$10.47.10^{-2}$	$7.66.10^{-2}$

Table 5.2: MSE and PED values for a comparison of the EWD with the symmetric and asymmetric filter models for stimuli levels of 30, 50, 70 and 90 dBSPL.

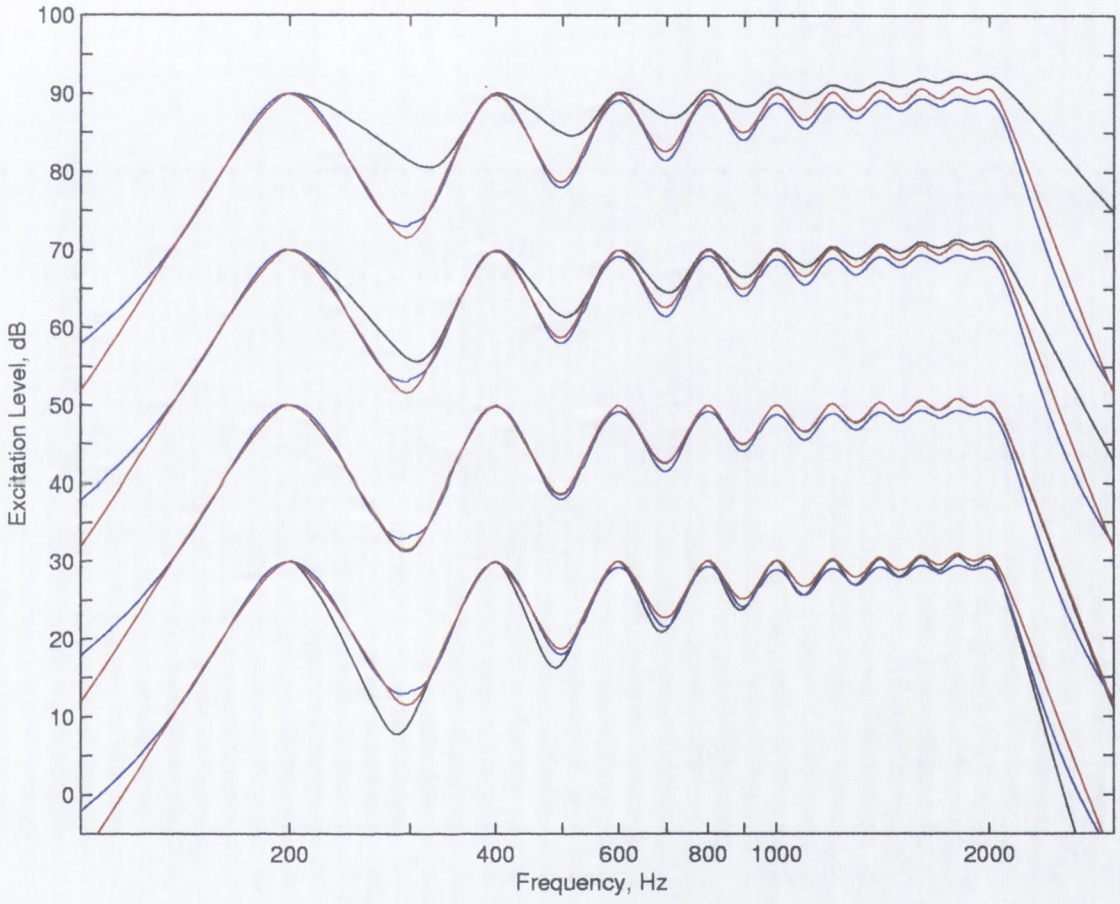


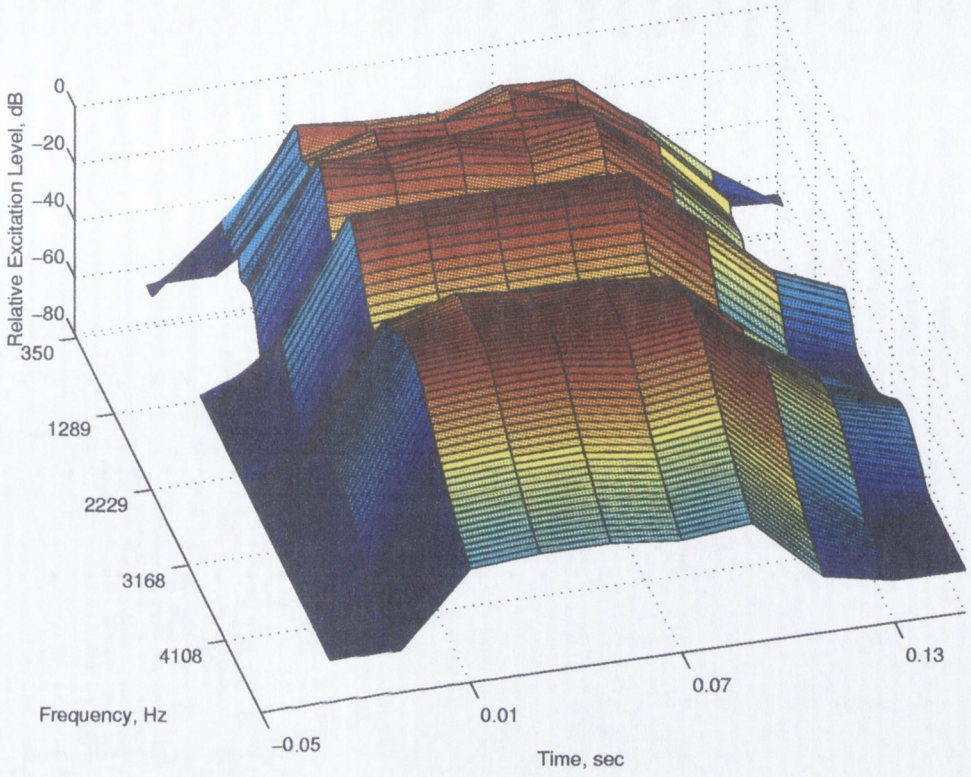
Figure 5.5: Excitation patterns generated by the asymmetric roex(p) filter shape (black curves), symmetric roex(p) filter shape (red curves) and EWD (blue curves) for the first 10 harmonics of a 200 Hz fundamental for four component levels of 30, 50, 70 and 90 dB SPL.

Notched Noise plus Tone

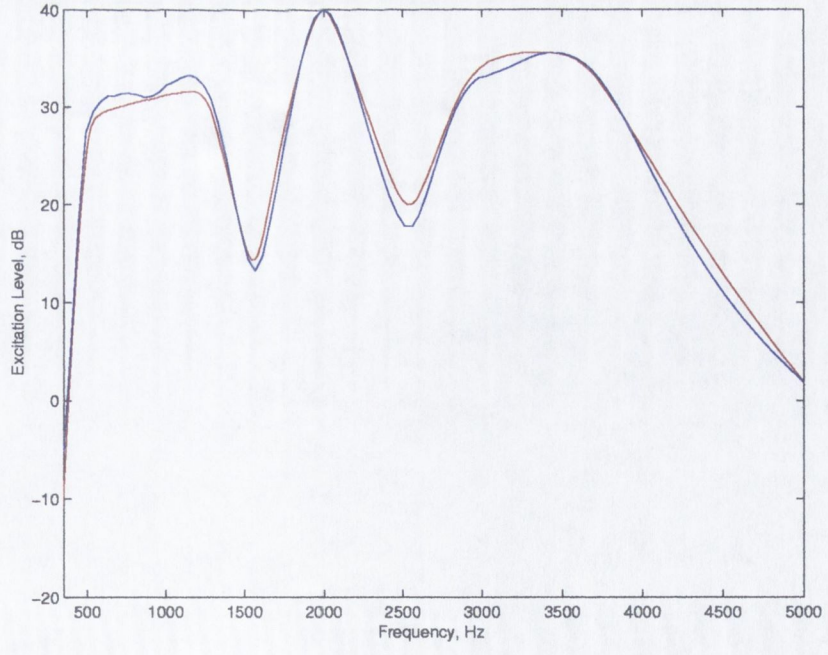
The auditory filter shape has often been estimated using a notched-noise masker with a signal tone placed between the notches. Such a signal was therefore analysed using the EWD and excitation patterns generated by the EWD and symmetric roex(p) filter shape were compared. For the roex(p) excitation pattern generation code, it is assumed that the signal consists of a finite number of sinusoidal components. Therefore, noise bands are composed of finely spaced sinusoids, the level of the sinusoids being multiplied by their spacing to ensure the correct noise power [63]. This method of noise generation was also used for the purposes of EWD analysis, except that the phases of the sinusoidal components were also randomised. The total duration of the signal and noise maskers was again 100 ms, and the same hanning ramping and damping was applied. The signal frequency was set to be 2 kHz. The upper and lower edges of the lower noise band were 500 and 1300 Hz, giving a bandwidth of 800 Hz. The upper noise band also had a bandwidth of 800 Hz, its upper and lower edges being at 2900 and 3700 Hz respectively. The noise bands were therefore placed asymmetrically about the signal, the signal tone being 700 Hz from the edge of the lower noise band and 900 Hz from the edge of the upper noise band. The signal tone was set to a level of 40 dB SPL and the noise spectrum level was 10 dB SPL. Figure 5.6 (a) shows the EWD of the overall signal. The analysis time interval is 10 ms. It is clear from the EWD plot that the signal is well above the masking threshold of the noise. Also note that the upper noise band shows less fluctuations than the lower noise band because the higher auditory filter ERB causes an increased amount of spectral smoothing. Figure 5.6 (b) plots the excitation patterns generated by the symmetric roex(p) filter shape (red curve) and the EWD (blue curve) described above. The random fluctuations of the EWD are seen to average to give an excellent match to the symmetric roex(p) excitation pattern. The MSE and PED are very low, being $1.32.10^{-3}$ and $1.04.10^{-2}$ respectively. Results not presented here show that, not unexpectedly, the EWD excitation patterns become even closer to the roex(p) excitation patterns as the signal length is increased.

Synthetic Vowel

As a final test of spectral resolution of the EWD, a comparison was made between the EWD and roex(p) excitation patterns for the synthetic vowel /I/ described in [111]. Again this sound was composed of a finite number of harmonics, spaced every 125 Hz from a fundamental frequency of 125 Hz to 4 kHz. The levels of each of the harmonics are plotted in Figure 5.7. Again, for the purposes of EWD analysis, the signal length was set to 100 ms and was ramped and damped using the 10 ms hanning window. Also, no outer-middle



(a)



(b)

Figure 5.6: (a) EWD 2 kHz signal tone plus noise masker. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) shape (red curve) and the EWD (blue curve). $MSE = 1.32.10^{-3}, PED = 1.04.10^{-2}$

ear filtering was performed. The analysis was performed from 125 Hz to 4 kHz, or from approximately 4 to 27 ERBs. The phases of the harmonics were randomised. Figure 5.8 (a) shows the resulting EWD, computed at time intervals of 10 ms. In order to obtain a smooth frequency marginal (Figure 5.8 (b)), 400 channels were used in the analysis, however for clarity, only every fourth channel is displayed in the EWD surface plot (Figure 5.8 (a)). No presence of cross-term interference is detectable. Figure 5.8 (b) shows the excitation patterns of the vowel sound generated using the EWD and both asymmetric and symmetric filter shapes.

Table 5.3 lists the MSE and PED values for the comparsion of the EWD and symmetric and asymmetric filter shapes. For the asymmetric filtershape, the vowel sound was analysed at two different levels of 50 and 80 dB SPL, as carried out in [111]. The symmetric (level independent) and EWD excitation patterns agree very well across all frequencies and have a very low MSE and PED of $2.74.10^{-4}$ and $4.10.10^{-3}$, respectively. The MSE and PED obtained for the asymmetric filter for a signal level of 50 dBSPL are actually slightly less than for the symmetric filter shape, indicating a slightly better fit. However, the fit worsens for the higher signal level of 80 dBSPL, with MSE and PED being $3.9.10^{-3}$ and $4.69.10^{-2}$, respectively. Overall, the EWD and asymmetric (level dependent) roex(p) excitation patterns agree very closely for the lower signal level, and are still in good agreement for the higher signal level.

Models	Level/dBSPL	MSE	PED
EWD and symmetric shape	ALL	$2.74.10^{-4}$	$4.10.10^{-3}$
EWD and asymmetric shape	50	$2.05.10^{-4}$	$3.20.10^{-3}$
EWD and asymmetric shape	80	$3.9.10^{-3}$	$4.69.10^{-2}$

Table 5.3: MSE and PED values for a comparison of the EWD with the symmetric and asymmetric filter models for stimuli levels of 30, 50, 70 and 90 dBSPL.

5.2.2 Temporal Resolution Accuracy

In order to test the temporal resolution of the EWD, we can calculate the EWD of a very short, impulsive signal. The output excitation level against time for frequencies covering all of the temporal window frequency bands can then be compared with the shape of each of the 7 temporal windows used by the EWD. There should be close agreement between the EWD output and the (time-reversed) temporal window shape for accurate EWD temporal resolution to be attained. In some sense this test is akin to estimating the impulse response

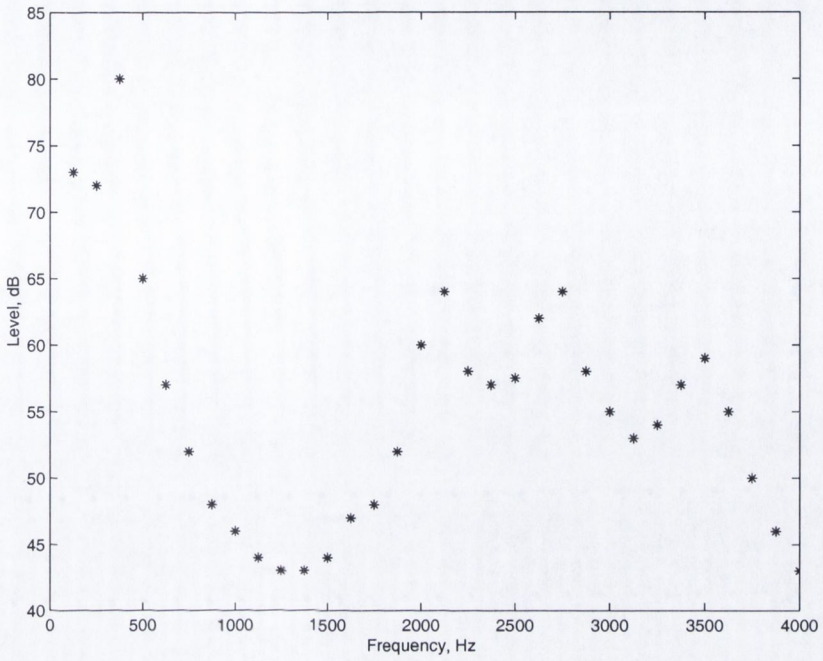
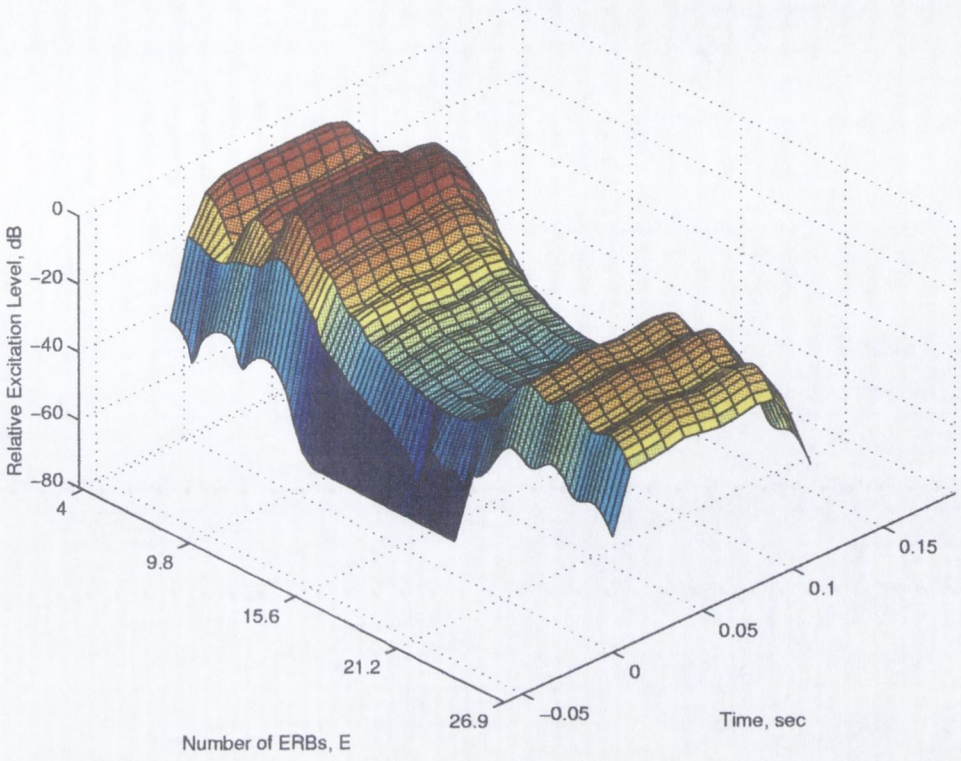


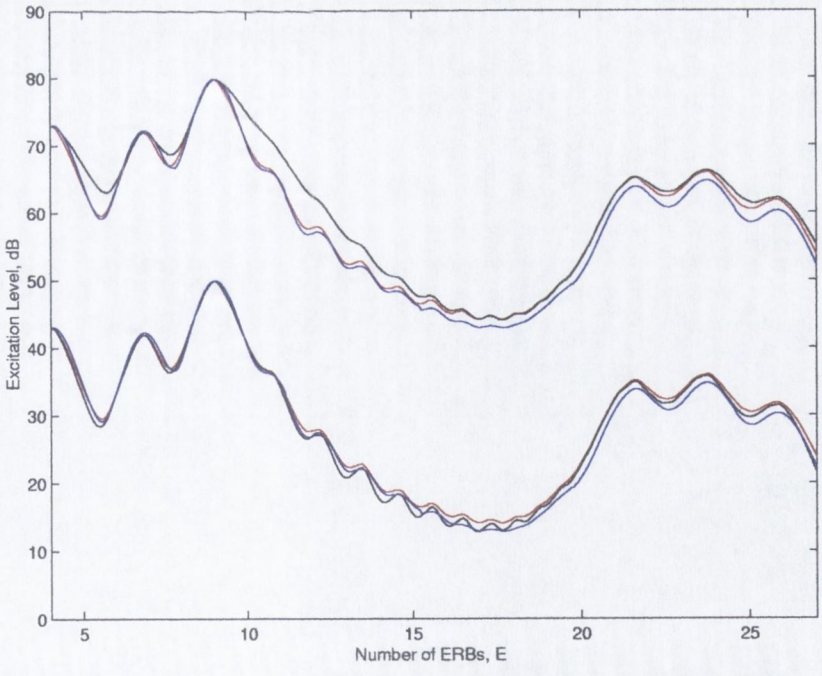
Figure 5.7: Harmonic levels of /l/ vowel sound, from fundamental at 125 Hz to 32nd harmonic at 4 kHz. (data taken from [111]).

of a system by exciting the system with an impulse and measuring the output, except here we 'measure' a time-frequency response, rather than simply an impulse response.

Figure 5.9 (a) and 5.9 (b) plot the waveform and magnitude response of the signal used to perform this resolution test. The signal is a 20th order lowpass filter of total duration equal to just 1 ms so as to allow an accurate estimation of the EWD temporal resolution. The cutoff of the filter is set at 9 kHz. In order to estimate the temporal resolution of the EWD, the EWD of the lowpass filter was calculated at 7 frequencies, each of which was inside one of the 7 temporal window frequency bands. Therefore, all 7 of the temporal windows are used in the resolution test. In the analysis, the dynamic range of the temporal window was set to 80 dB and the analysis was performed at time intervals of 10 ms from $t = -100\ ms$ to $t = 150\ ms$. The results of this EWD analysis are plotted in Figure 5.10 (a). The response at each of the frequencies analysed bears remarkable similarity to the window shapes plotted in Chapter 4, Figure 4.16 (b), once time-reversed. As frequency increases, the peak level of the EWD rises, reflecting the widening ERB of the auditory filters. Figure 5.10 (b) compares the output with time of each of the the EWD frequency channels with the 7 temporal window shapes. There is an almost perfect match between the two for all 7 (time-reversed) windows, except at the peaks and 'knees' of the temporal window shapes where the 10 ms time interval of the EWD results in an approximation to

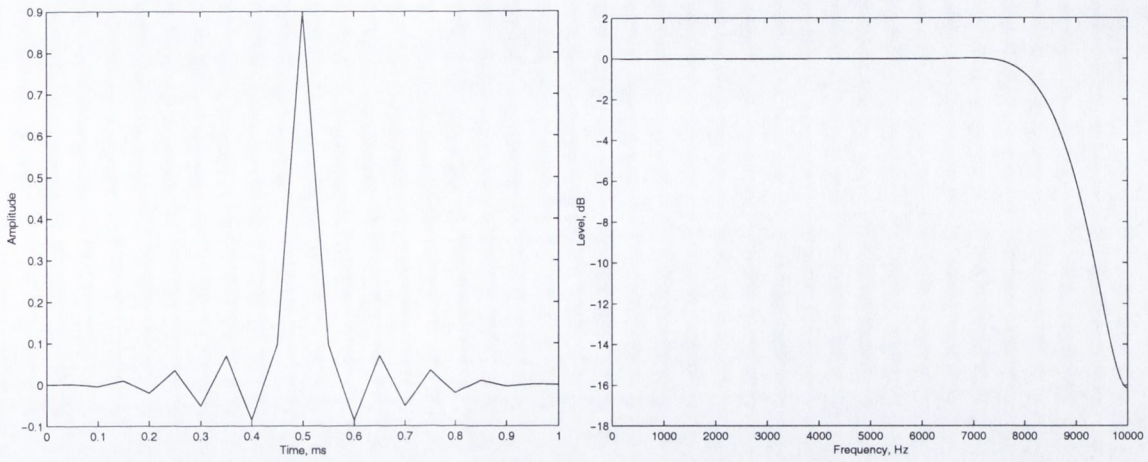


(a)



(b)

Figure 5.8: (a) EWD of a /I/ vowel sound, fundamental 125 Hz. (b) Comparison of the excitation patterns generated for this signal by the symmetric roex(p) shape (red curve), symmetric roex(p) shape (black curve) and the EWD (blue curve) for stimulus levels of 50 and 80 dB.



(a) (b)

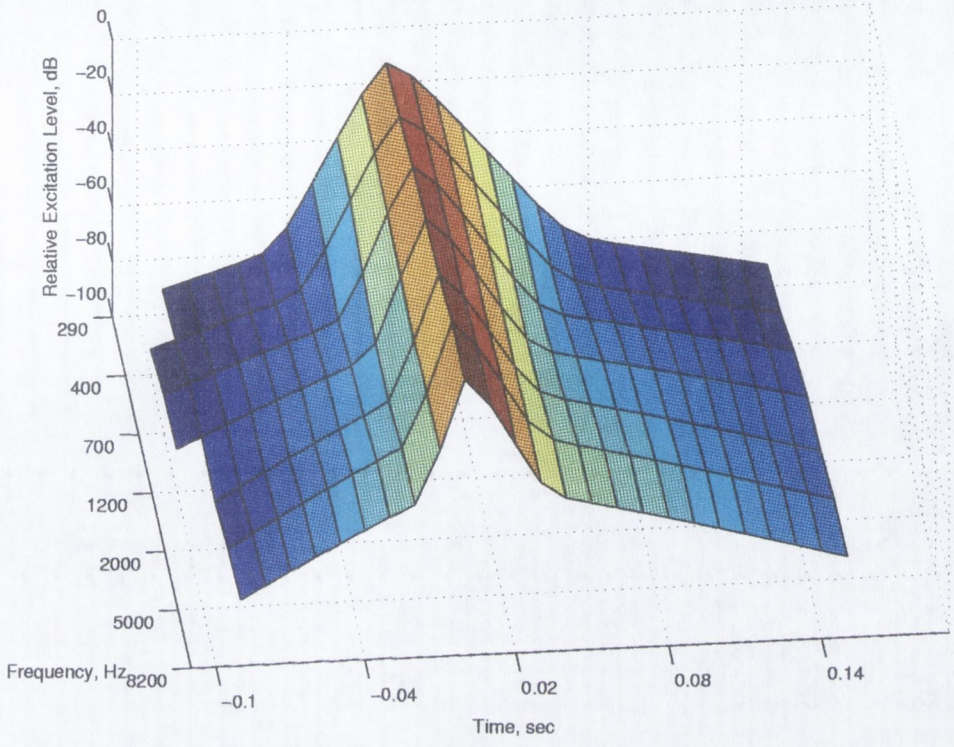
Figure 5.9: (a) Waveform and (b) frequency response of the 20th order FIR lowpass filter used to test EWD temporal resolution.

the much smoother window shapes. Overall, this resolution test reveals that the temporal resolution of the EWD almost exactly equals that of the temporal window model of auditory temporal resolution described in Chapters 3 and 4, and hence accurately incorporates this model of temporal masking.

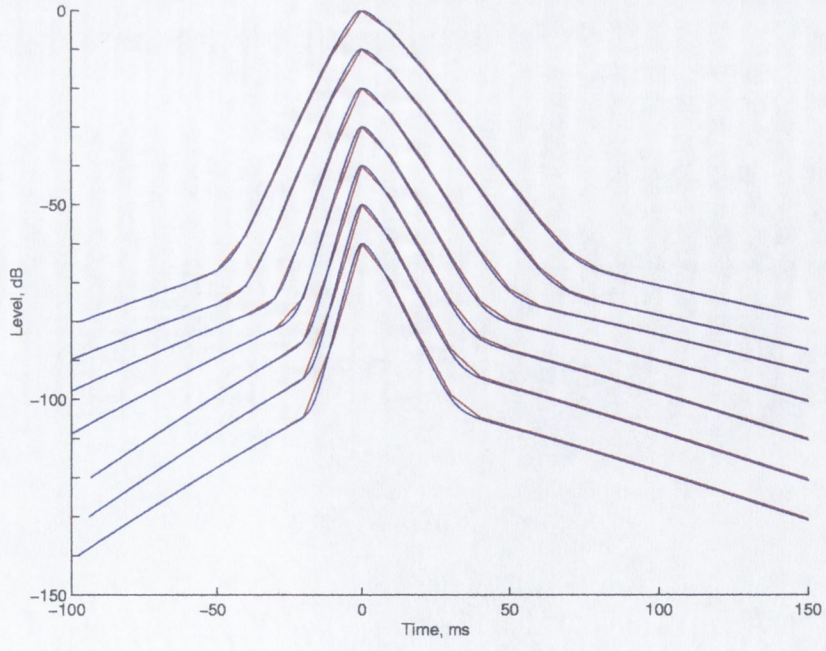
5.2.3 Comparison of EWD and Spectrogram

In Chapter 2 the main disadvantage of the spectrogram compared with the Wigner distribution is highlighted, namely that the time and frequency resolutions of the Wigner distribution may be independently set whereas those of the spectrogram may not. The consequence of this is that temporal and spectral masking cannot be simultaneously modelled using the spectrogram. In order to highlight this fact Figure 5.11 presents the EWD of the same complex tone presented in Figure 5.4 now plotted on a linear rather than logarithmic frequency axis. The tone consists of the first 10 harmonics of a 200 Hz fundamental. Again, both forward and backward masking can clearly be seen, forward masking being the more dominant of the two. Also, the transition from one temporal window to the next can be seen at frequencies 300, 533 and 1687 Hz, illustrating the frequency dependance of the temporal resolution of the EWD. The decreasing inter-ripple height of the harmonics with frequency also clearly demonstrates that spectral masking is being modelled. This is possible because of the variable frequency resolution of the EWD.

In order to use the spectrogram for auditory modelling, or indeed for any type of signal analysis, we must first choose an analysis window. So, for example, if temporal masking is



(a)



(b)

Figure 5.10: (a) EWD of the 20th order FIR filter for 7 analysis frequencies, using all 7 temporal windows. (b) A comparison of the EWD output and the actual temporal window shape. The curves plotted represent temporal window 1 to 7 (top to bottom), with the graphs normalised and offset by 10 dB for clarity.

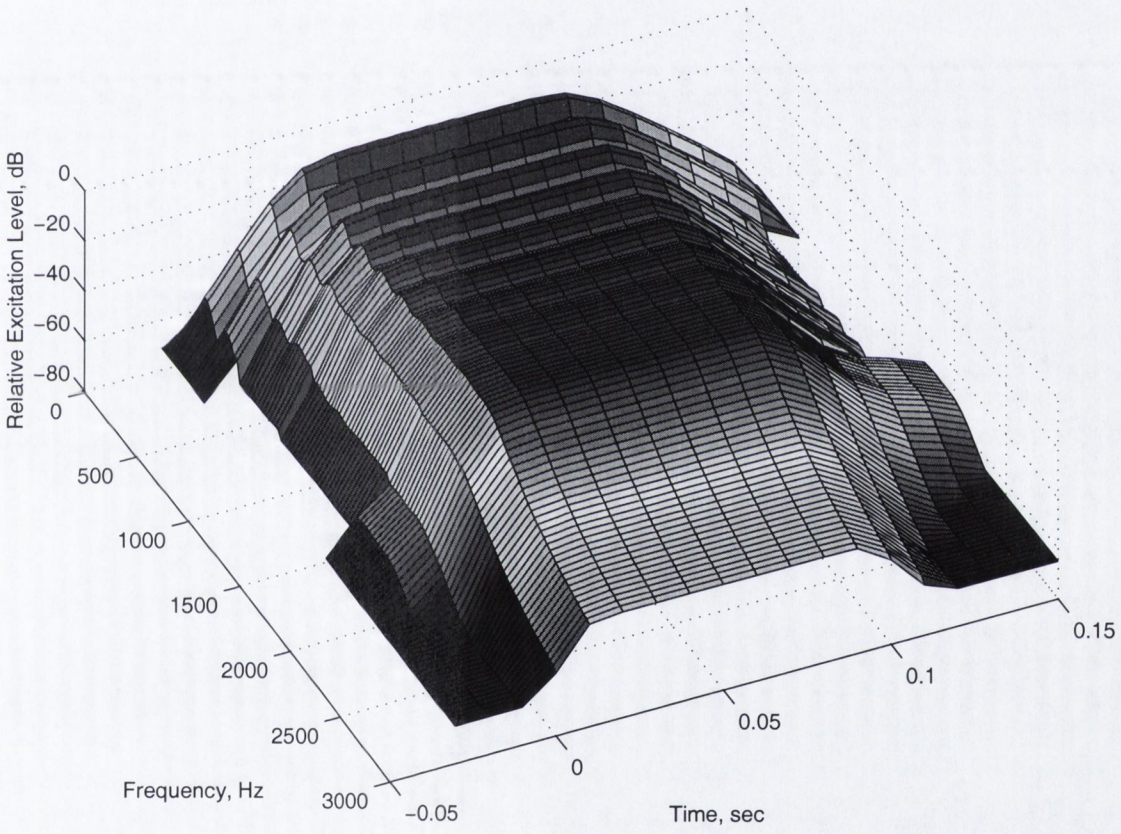


Figure 5.11: EWD of the first 10 harmonics of a 200 Hz fundamental.

to be modelled, we can choose an analysis window based on the temporal window model[1]. Figure 5.12 presents the resulting spectrogram where the analysis window chosen is based on the temporal window of centre frequency of 1000 Hz. Comparing the distribution with the EWD of Figure 5.11, we see that the temporal resolution matches that of the EWD at that centre frequency. However, since only one analysis window may be chosen with the spectrogram, the temporal resolution stays constant with frequency. Furthermore, spectral masking is not modelled, as the spectral resolution is constant and equal to the Fourier transform of the temporal window at 1000 Hz (see Equation 2.7).

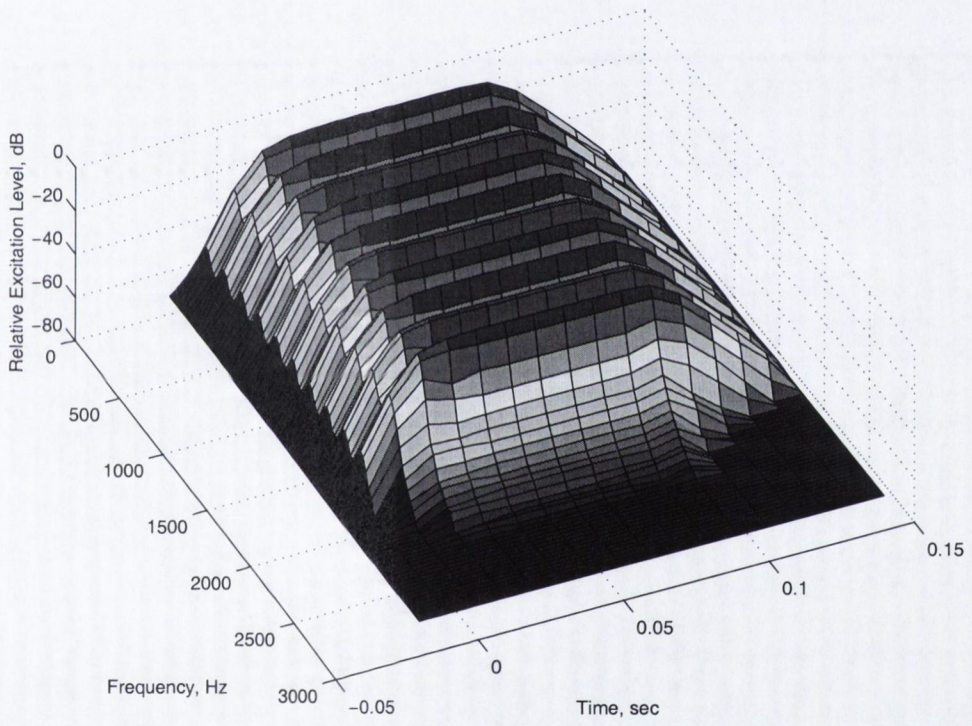


Figure 5.12: Spectrogram of the first 10 harmonics of a 200 Hz fundamental, using the 1000 Hz temporal window as the analysis window.

It is not possible to model spectral masking using the spectrogram, since its frequency resolution cannot be made frequency dependent. Figure 5.13 illustrates this point, where we have used a gamma window of centre frequency 1000 Hz as our analysis window. Here the spectral resolution matches that of the EWD of Figure 5.11 at 1000 Hz only. Spectral masking is therefore not modelled since frequency resolution is fixed. Also, the temporal

[1]Note that the square root of the temporal window must be used here, because the temporal window model is given in units of power. After the frequency slices of the STFT are computed using this window, the magnitude is taken and is squared to give each frequency slice of the spectrogram, thus ensuring the correct temporal resolution.

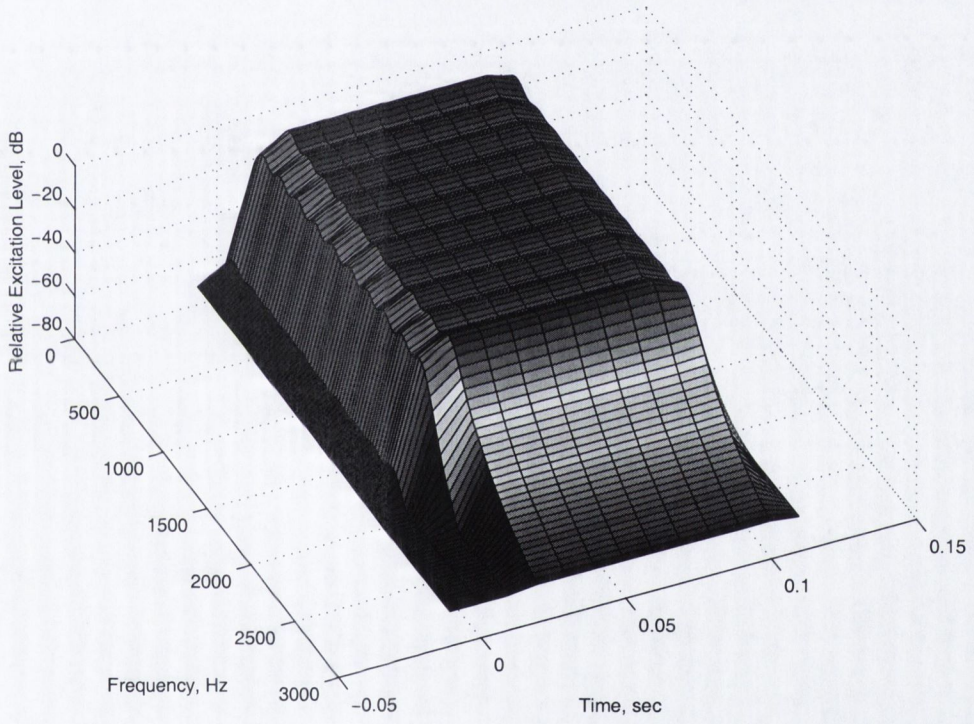


Figure 5.13: Spectrogram of the first 10 harmonics of a 200 Hz fundamental, using the 1000 Hz gammatone window as the analysis window.

resolution of the distribution is equal to the shape of the gamma window. Consequently, temporal masking is not modelled here. Overall, is is readily apparent that the spectrogram is incapable of providing an adequate model of temporal and spectral masking because the spectrogram cannot achieve independent control of time and frequency resolution and the spectrogram cannot have a frequency dependent kernel.

5.3 Cross-Term Interference

In Chapter 2, the cross-term interference inherent to the Wigner distribution was discussed. It was shown that any two auto-terms separated in the time-frequency plane interfere to create a cross-term located at their geometrical midpoint. The cross-term oscillates at a frequency proportional to the auto-term separation. It was also demonstrated that cross-term interference could be suppressed by smoothing. In the previous section we presented a series of results demonstrating the accuracy of the spectral and temporal resolutions of the EWD. Furthermore, it was noted that the EWD showed very little evidence of any cross-term interference whatsoever. This section explains why this is so.

Figure 5.14 portrays the EWD of a signal consisting of a sinusoid of constant frequency 250 Hz and duration 500 ms, plus a linear chirp signal of same duration sweeping from 250 Hz to 4 kHz. Both signals were ramped and damped using the left and right halves of a 10 ms hanning window, as in the previous sections. The signal was analysed from $t = -50$ ms to $t = 580$ ms; 100 channels were spaced linearly between 150 and 5 kHz. The temporal window had a dynamic range of 50 dB. Cross-term interference is visible on the plot between the sinusoid and chirp signal components (auto-terms) along a line running from 250 Hz at $t = 0$ ms to 2125 Hz at $t = 500$ ms. It is, however, well attenuated, only just protruding above auto-term excitation level. Since the temporal and spectral smoothing of the EWD kernel is lowpass in nature, and since cross-term oscillation frequency is proportional to auto-term separation, we would expect that cross-term interference be serious for small auto-term separation. However, examination of Figure 5.7 shows that although the cross-term level does indeed rise as auto-term separation falls, it does not rise sufficiently to protrude above auto-term excitation level for small auto-term separation. For wide auto-term separation, such as at $t = .5$, where the sinusoid and chirp are separated by 3.75 kHz, the combination of auto-term spectral masking and temporal smoothing ensure that cross-term interference only barely protrudes above auto-term excitation level.

It should be emphasised that the example above tests the worst-case scenario for the presence of cross-term interference in the EWD. This is because the signal used consists of

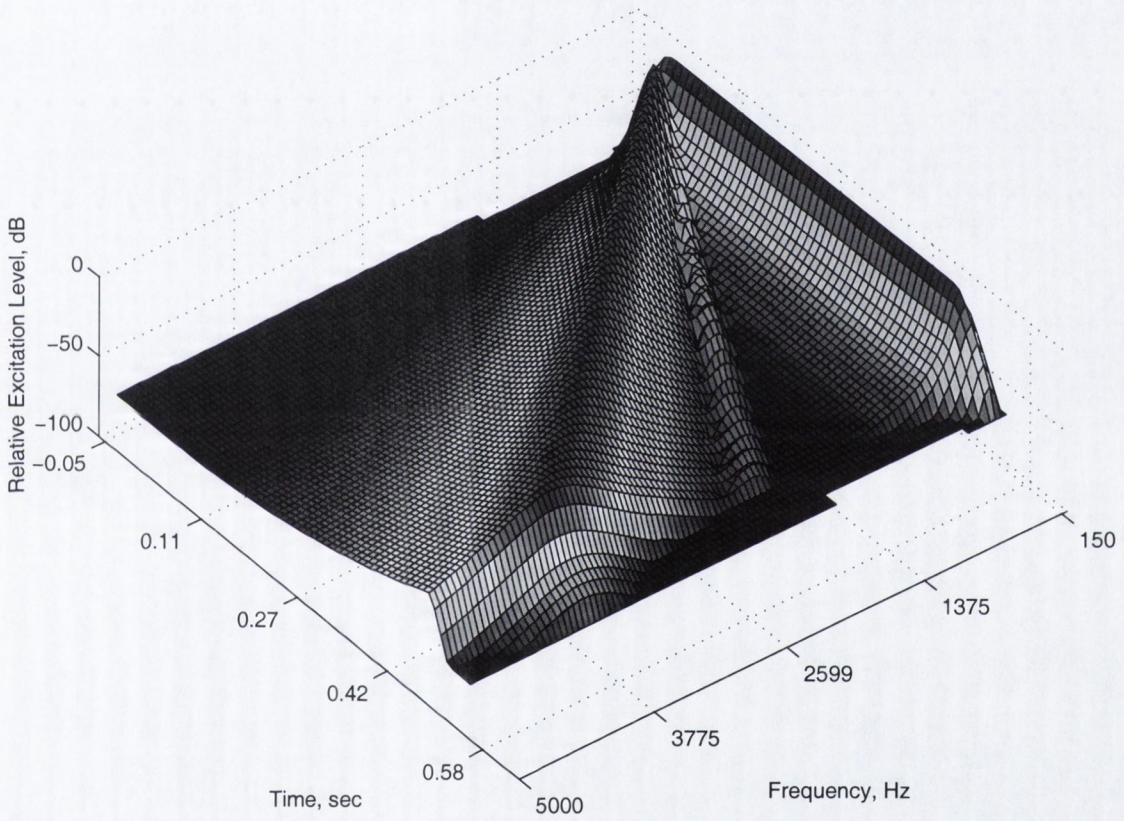


Figure 5.14: EWD mesh plot illustrating the cross-term interference between a sinusoid of 250 Hz and a linear chirp signal, sweeping from 250 Hz to 4 kHz.

two narrowband signal components well separated in the time-frequency plane. The results of the previous section demonstrate that for more broadband signals such as the complex tone or synthetic vowel signal, cross-term interference is completely masked under auto-term excitation level. Therefore, EWD analysis of real acoustic signals such as music or speech will to all intents and purposes never be affected by unwanted cross-term interference, since such interference will be well suppressed below auto-term excitation level.

5.3.1 Frequency Response of the Temporal Window

In Chapter 2 it was demonstrated that cross-term interference oscillating parallel to the time axis could be suppressed by temporal smoothing. The amount of cross-term suppression may be determined by the frequency response of the temporal component of the separable smoothing kernel. Pielemeier's work on the time-frequency analysis of musical signals using the *modal distribution* [147], [146], [148] relies on this fact to deliver excellent suppression of cross-term interference between musical signal partials. Prior to modal distribution analysis, the inter-partial spacing of a particular musical note is evaluated, giving an estimation of the oscillation frequency of the cross-term interference parallel to the time axis. This frequency is then used in the design of a lowpass temporal smoothing window providing sufficient attenuation at that oscillation frequency. Hence the temporal resolution of the modal distribution depends on the signal being analysed. If the signal's inter-partial spacing is narrow, then a long temporal smoothing window is required, resulting in a poor temporal resolution. If the signal has a wide inter-partial spacing, then a short temporal window is sufficient, resulting in good temporal resolution.

As Pielemeier notes, the temporal smoothing window effectively determines the temporal resolution of the modal distribution. Similarly, Plack notes that the frequency response of the temporal window shape determines auditory temporal resolution, i.e. the ability of the ear to track changes in the envelope of a sound with time [150]. Although the temporal window model of temporal resolution is not designed with the suppression of cross-term interference in mind, it has been shown that, in conjunction with the frequency smoothing provided by the gammatone model of spectral masking, sufficient suppression of cross-term interference is provided. In order to evaluate the attenuation of cross-terms provided by the temporal window, Figure 5.15 plots the frequency response of the 7 temporal windows used by the EWD. It is clear the temporal windows are lowpass in nature. The responses are all initially flat below 10 Hz, then begin to fall off, achieving a slope of about 35 dB/decade after 100 Hz. The responses of each of the temporal windows are similar, with the lowest (*i.e.* longest) frequency window achieving the highest attenuation.

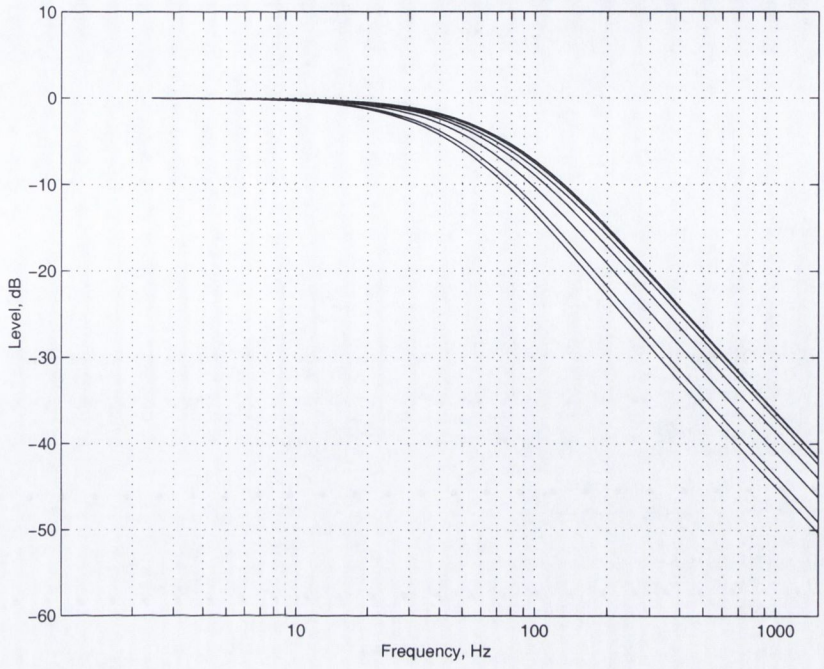


Figure 5.15: Frequency response of all 7 temporal windows.

5.4 Effective Dynamic Range

The design of DSPWD smoothing kernels is typically concerned with achieving a compromise between a high distribution resolution and a high suppression of unwanted cross-term 'clutter'. However the kernel smoothing action can introduce distribution clutter of its own. For example, as Pielemeier reports [148], if an autocorrelated hanning window is used as frequency smoothing window, the peak sidelobe level is about 41 dB down from the mainlobe. This sidelobe structure limits the useful dynamic range and frequency resolution of a DSPWD using such a frequency smoothing window because the lobing structure can be confused with signal properties. The overall useful dynamic range is therefore in general limited by the presence of both window artifacts and cross-term interference. In Section 5.3, the EWD was shown not to be adversely affected by cross-term interference. Also, in Chapter 4, Figures 4.11 and 4.8 demonstrate that the gamma window does not suffer from a lobing behaviour. Therefore, the EWD is cross-term and clutter-free over a very wide dynamic range, as demonstrated in this section.

The dynamic range of the ear, given by the difference between the threshold of audibility and the threshold of pain, is in excess of 100 dB from 100 Hz to 10000 Hz. In an article discussing the links between the responses and subjective quality of audio equalisers, Gerzon [61] points out that : "There is an urgent need to refine existing methods of simul-

taneous time/frequency analysis to maximise the amount of fine low-level detail that can be seen. In computer-based analysis packages, this means carefully devising the filtering and 'windowing' used to minimize all discontinuity, resonance and aliasing artifacts, and using very high quality graphics to display the results on a very fine time/frequency grid."

Unfortunately, the majority of DSPWD research to date has generally been plotted on a linear amplitude scale, which is not suitable for audio. Pielemeier's modal and constant-Q modal distributions go a considerable way towards rectifying the situation, being time-frequency distributions with very high time and frequency resolutions and good suppression of clutter. The distributions are designed for the analysis of musical signals which are modelled according to the following equation :

$$\hat{s}(t) = \sum_{l=1}^{N} \hat{B}_l(t) exp(j\phi_l(t)) \tag{5.5}$$

where $\hat{s}(t)$ is one instrument, consisting of finite number of signal partials represented by N complex exponentials covering both positive and negative frequency with amplitudes and phases $\hat{B}_l(t)$ and $\phi_l(t)$. The instantaneous frequency of each of the partials is given by the derivative of the phase $\phi_l(t)$ [33]. The amount of time smoothing is chosen prior to analysis to sufficiently attenuate cross-terms interference between the most closely spaced signal partials. The modal distribution is designed to give a constant frequency resolution, suited to the particular signal analysis task at hand, whereas the frequency resolution of the constant-Q modal distribution dilates with increasing frequency and is usually computed at 1/3 semitone intervals. Pielemeier's constant-Q and modal distributions are also computationally efficient and are suitable for the analysis of signals of long duration and/or high sampling frequency. However as he reports [147], due to cross-term interference and window artifacts, the effective dynamic range is limited to about 40 - 60 dB for these distributions.

The results of this chapter have so far demonstrated that the EWD does not suffer from unwanted cross-term interference, and that the frequency and time smoothing produced by its kernel is clutter free. Since the constant-Q modal distribution is the most related member of Cohen's class to the EWD, this section compares these two distributions for samples of a flute and vibraphone attack taken from the McGill University Master Samples (MUMS) CDs [122]. These CDs contain digital audio recordings of a large number of musical instruments, played chromatically throughout each instrument's register using a number of performance techniques. The constant-Q modal distribution was implemented in MATLAB by the author and there was an excellent correspondence between results for this implementation of the distribution and those presented by Pielemeier in [148].

Vibraphone Attack

The constant-Q distribution and EWD of the vibraphone hard mallet attack are shown in Figures 5.16 and 5.17 respectively. The constant-Q distribution is designed for musical signal analysis and its frequency bins are spaced according to the musical scale at 1/3 semitone intervals. According to the musical scale, an octave represents a doubling of frequency, with each octave containing 12 semitones. The note A4 has a standardised fundamental frequency of 440 Hz; this note is assigned to semitone number 69. Therefore, the frequency, F, of a particular semitone may be calculated from

$$F = 440 \times 2^{\frac{S-69}{12}} , \tag{5.6}$$

where S is the semitone number. The vibraphone note value is F#4 (semitone 66), giving a fundamental frequency of f_0=370 Hz. The signal sample is taken from MUMS Volume 3, Track 6, Index 14. The signal was downsampled by a factor of 2 for the purposes of EWD calculation to give a sampling frequency of 22050 Hz. Both the EWD and constant-Q distribution are displayed at 1/3 semitone intervals from semitone 64 (329.6 Hz) to 106 (3801.8 Hz); this is exactly the same range used by Pielemeier in [148]. Also, the time and frequency smoothing windows used by Pielemeier were also used here. The constant-Q modal distribution and EWD are computed for the first 130 ms and 120 ms of the signal, respectively. However, the EWD is computed starting at t=-10 ms in order to examine backward masking effects. For the purpose of comparison, the signal was not pre-filtered by the outer-middle ear filer for EWD analysis.

Following the example of Pielemeier, the constant-Q distribution is computed and displayed at intervals of 1.1 ms, however the EWD is only computed and displayed at an interval of 2 ms since the EWD is a smoother distribution of both frequency and time. Both distributions show an strong impulsive signal onset, followed by the emergence of a partial at f_0 (370 Hz), $4f_0$ (1480 Hz) and an inharmonic partial at 9.5 f_0 (3515 Hz). Very brief partials at $2f_0$, $3f_0$ and $5f_0$ at signal onset are also visible on close examination; these decay rapidly after about 3-5 ms.

The overall envelope and form of the impulsive onset and partial amplitudes agree closely for both distributions. The constant-Q distribution is designed with maximum temporal and frequency resolution in mind, therefore the signal partials are represented by sharp fins for this distribution. In contrast however, the EWD is designed to reflect auditory temporal and spectral masking and indicates that the partials cause excitation over a much broader range of frequencies than shown by the constant-Q distribution. Also, clutter is not fully

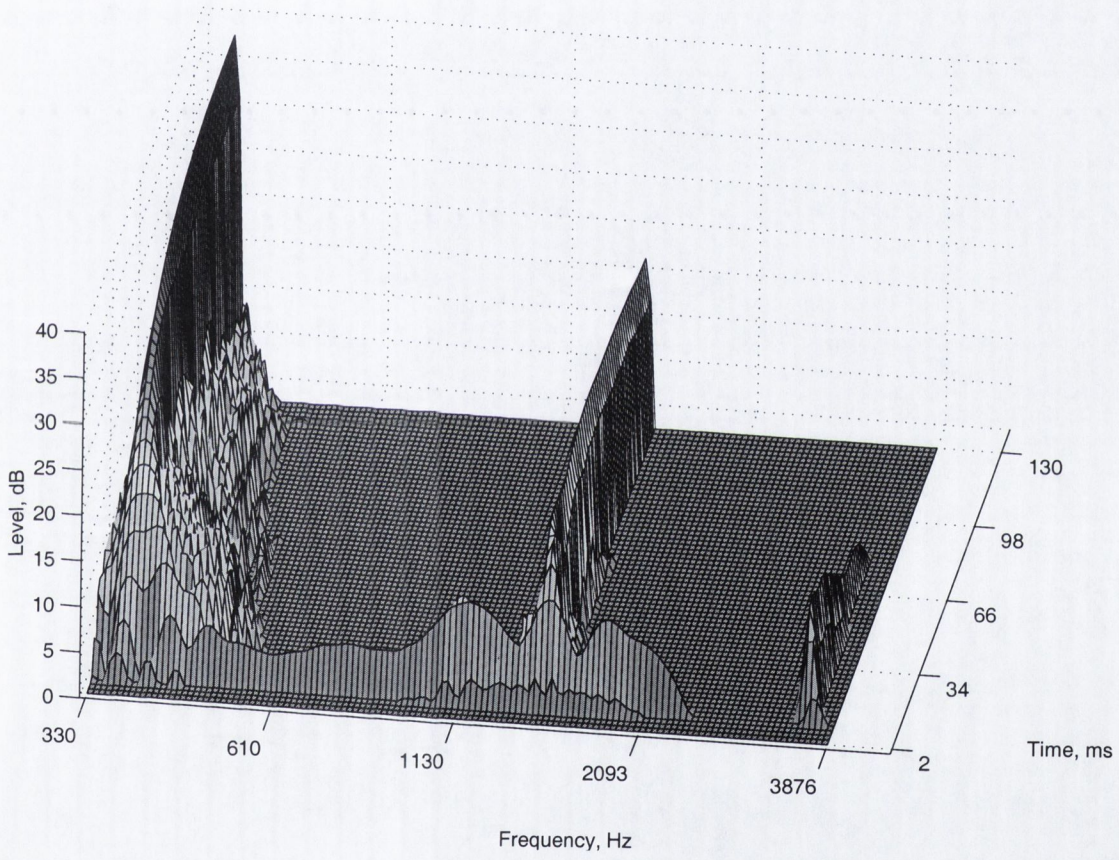


Figure 5.16: Constant-Q distribution of a vibraphone note F#4 (hard mallet attack), showing the first 40 dB of the distribution.

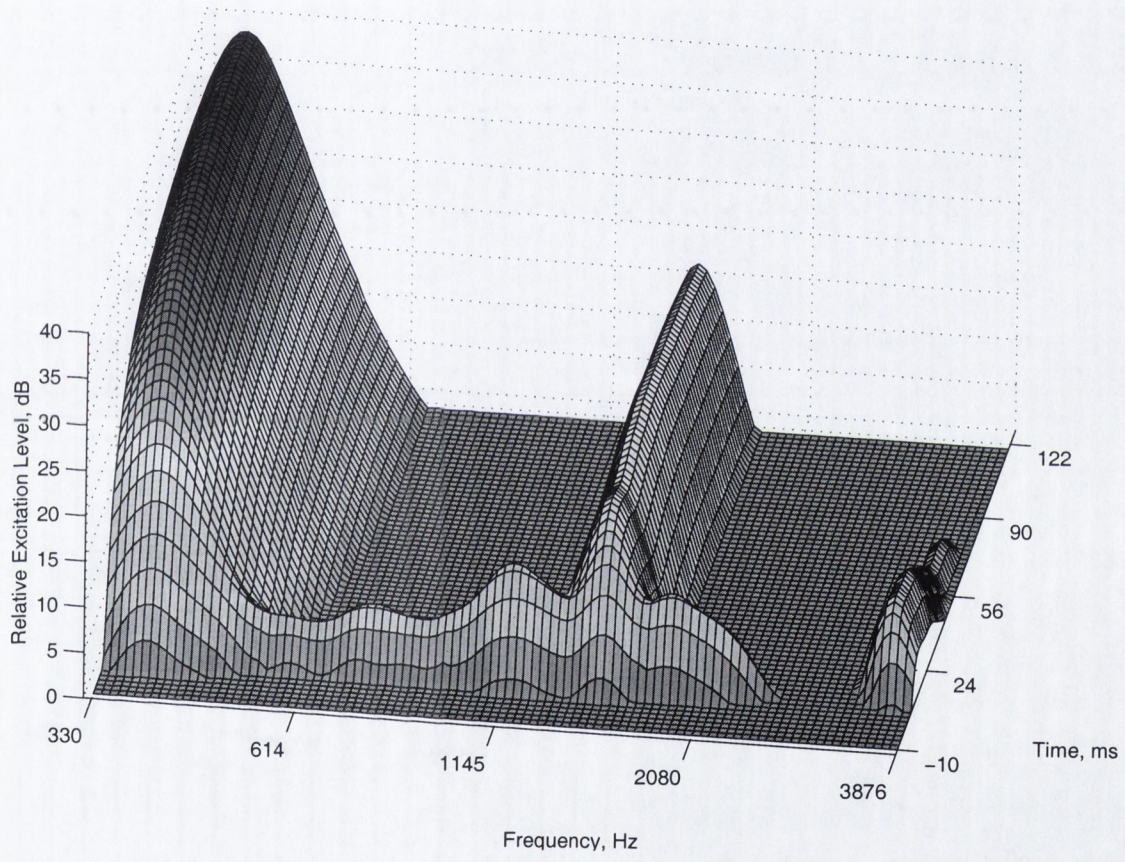


Figure 5.17: EWD of a vibraphone note F#4 (hard mallet attack), showing the first 40 dB of the distribution.

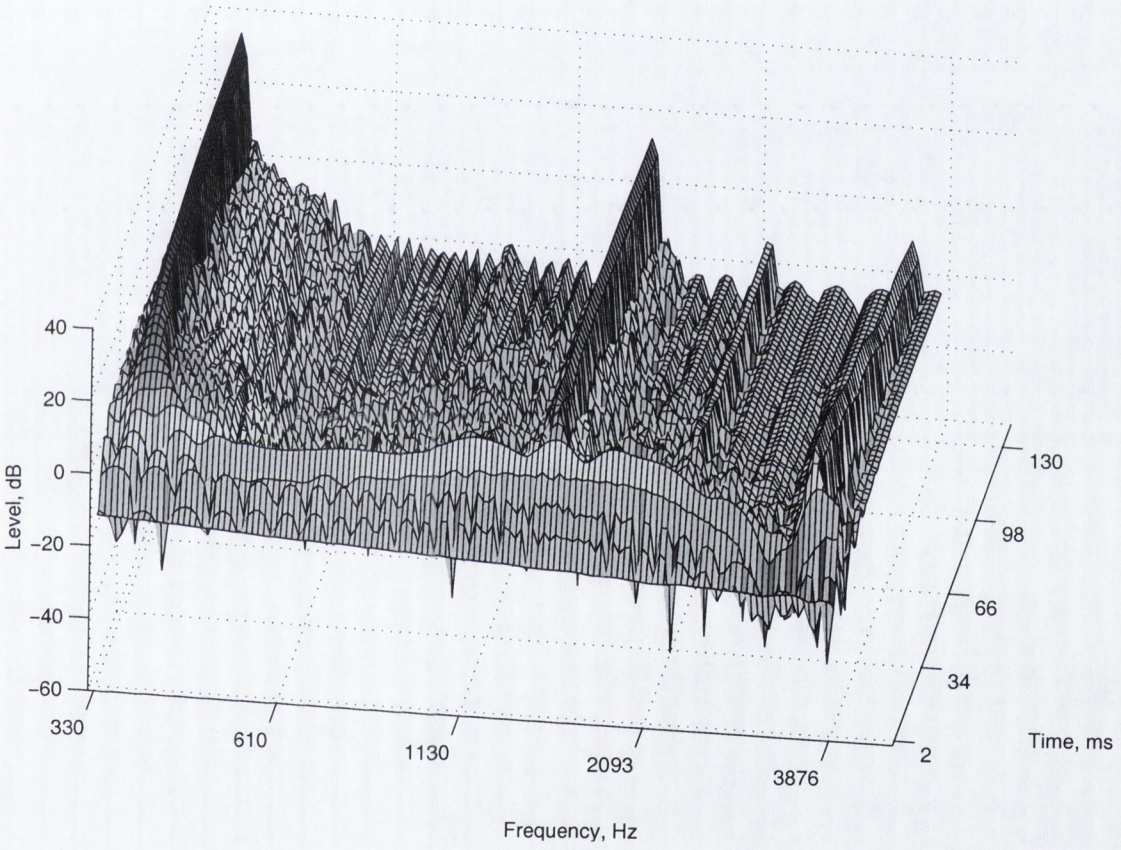


Figure 5.18: Constant-Q distribution of a vibraphone note F#4 (hard mallet attack). The full dynamic range of the distribution is shown.

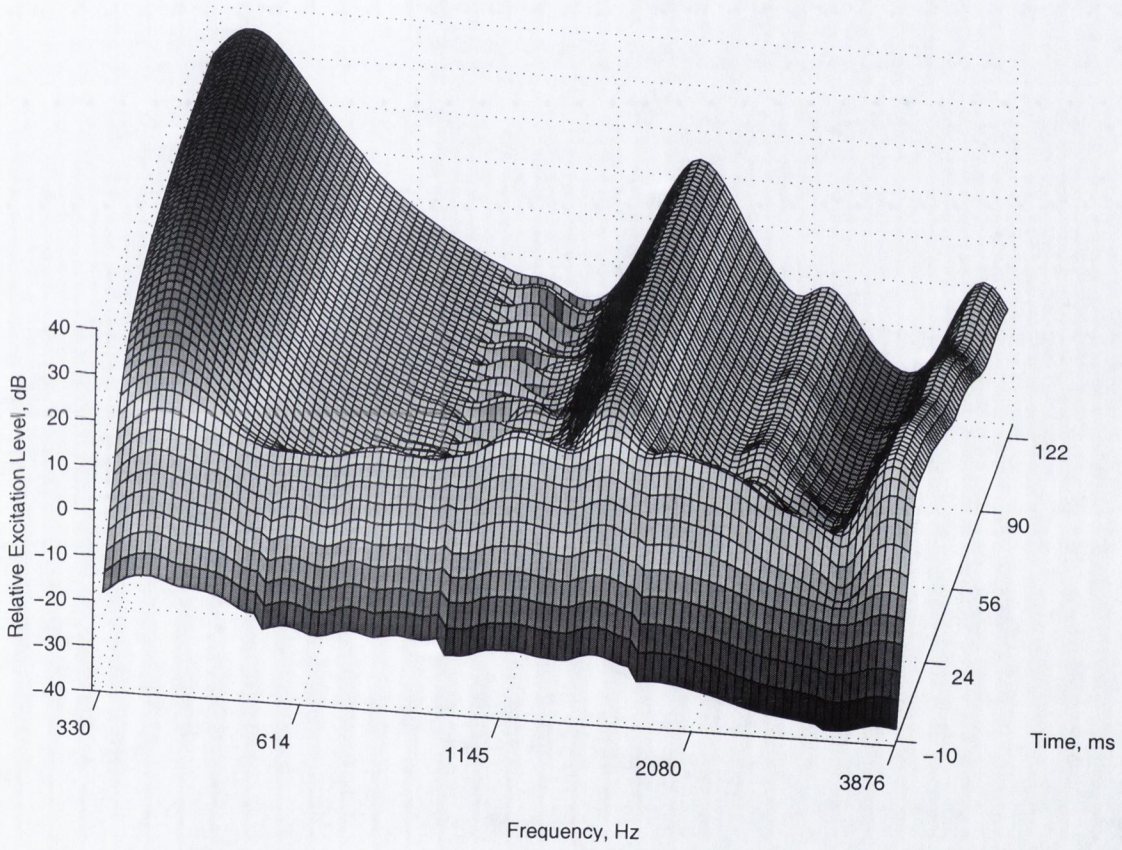


Figure 5.19: EWD of a vibraphone note F#4 (hard mallet attack). The full dynamic range of the distribution is shown.

suppressed for the constant-Q distribution, even though the distribution is plotted only over a 40 dB dynamic range. This clutter can be seen above the fundamental, from about 400-500 Hz. In contrast, the EWD shows no evidence of clutter whatsoever in this range.

Figures 5.18 and 5.19 present exactly the same distributions, now displaying the full range of the data. The constant-Q-distribution dynamic range is compromised by cross-term interference and windowing artifacts (lobing) resulting in a noise floor about 50-60 dB below its peak. In contrast, the EWD distribution is effectively free from all clutter over a dynamic range of about 80 dB. Only slight evidence of cross-term interference can be seen between the partial at f_0 and $4f_0$. However, the vibraphone signal is unusual in that the spacing between the first 2 signal partials is quite wide, because partials 2 and 3 are effectively missing. This results in a relatively large gap between the partial at f_0 and $4f_0$, leaving the EWD somewhat exposed to cross-term interference at $2.5f_0$. As shown in Figure 5.19, this interference is nevertheless very well attenuated, being almost 80 dB below the level of the fundamental. Thus the cross-term suppression and lack of clutter evident for the EWD is far superior to that exhibited by the constant-Q distribution. Finally, it is noted that the backward masking produced by the impulsive signal attack is clearly visible for the EWD.

Flutter-Tongued Flute Attack

Figure 5.20 and 5.21 compare the constant-Q modal distribution and EWD of the first 160 ms of a flutter-tongued flute signal. The signal is taken from Volume 2, Track 2, Index 13 of the MUMS samples and again is downsampled by a factor of two prior to EWD computation. The note value is C5 (semitone 72), giving a fundamental of f_0=523.5 Hz. The distributions are both computed and displayed from semitone 71 (493.9 Hz) to 113 $\frac{1}{3}$ (5696.3 Hz). The constant-Q modal distribution is computed and displayed at a time interval of 1 ms, following the work of Pielemeier [148], whereas the EWD is computed and displayed every 2 ms. Also, the time and frequency smoothing windows used by Pielemeier in [148] were used here. Again, this signal was not pre-filtered by the outer-middle ear filter for EWD analysis. Also, the EWD is computed before signal onset in order to evaluate backward masking effects.

From Figures 5.20 and 5.21 we see that the first four harmonics of the signal are clearly visible, all showing amplitude modulation characteristic of flutter tonguing. The modulation depth of the fundamental is far less than that of harmonics 2, 3 and 4. Also, as noted by Pielemeier, the analysis shows an inharmonic partial present for the first 50 ms after signal onset rising from 3729 to 5587 Hz. This partial can clearly be heard as a brief rising whistle

during the note attack.

Comparing the EWD and constant-Q modal distribution, we see that the effective dynamic range of the constant-Q modal distribution is limited to about 60 dB, whereas the EWD is clutter free over a much larger range in excess of 120 dB. Furthermore, although ostensibly the constant-Q modal distribution displays more signal detail, it is generally difficult to separate important signal information from clutter for much of the dynamic range shown. Displaying an even greater dynamic range exacerbates this problem further. In contrast, the EWD displays all of the important signal information shown by the constant-Q modal distribution but without any of the clutter. In fact, the overriding feature of this signal, namely the partial amplitude modulation caused by the flutter tonguing, is in fact more clearly brought out by the EWD than by the constant-Q modal distribution. This example demonstrates an important general feature of the EWD, namely that perceptually important signal features are preserved by the distribution, whereas undesired, perceptually masked signal detail, and/or distribution clutter is hidden.

5.5 Psychoacoustic Limitations of the EarWig Distribution

This chapter has been primarily concerned with evaluating the validity of the EWD as a simultaneous model of both spectral and temporal masking. This is carried out by comparing both its spectral and temporal resolutions with those of the psychoacoustic models employed. Issues particular to bilinear time-frequency distribution design such as cross-term interference, windowing artifacts and distribution dynamic range have also been addressed.

In Section 5.2.1, the accuracy of the spectral resolution of the EWD was evaluated by testing its effectiveness in generating excitation patterns. Excitation patterns generated by the symmetric and also asymmetric roex(p) filter shape were compared with those generated by the EWD for a variety of test stimuli. It was found that the EWD was in very good agreement with the level independent roex(p) shape for all of the stimuli. Furthermore, differences between the EWD and the level dependent, asymmetric roex(p) shape were found to be significant for high stimuli levels only. In Section 5.2.2, the temporal resolution of the EWD was also tested and found to agree exactly with the temporal window model of masking for all 7 of the temporal window shapes employed. Overall, the EWD is in very good agreement with the above models of spectral and temporal masking. Therefore, the psychoacoustic limitations of the EWD depend primarily on the limitations of the gammatone filter and temporal window model, discussed below.

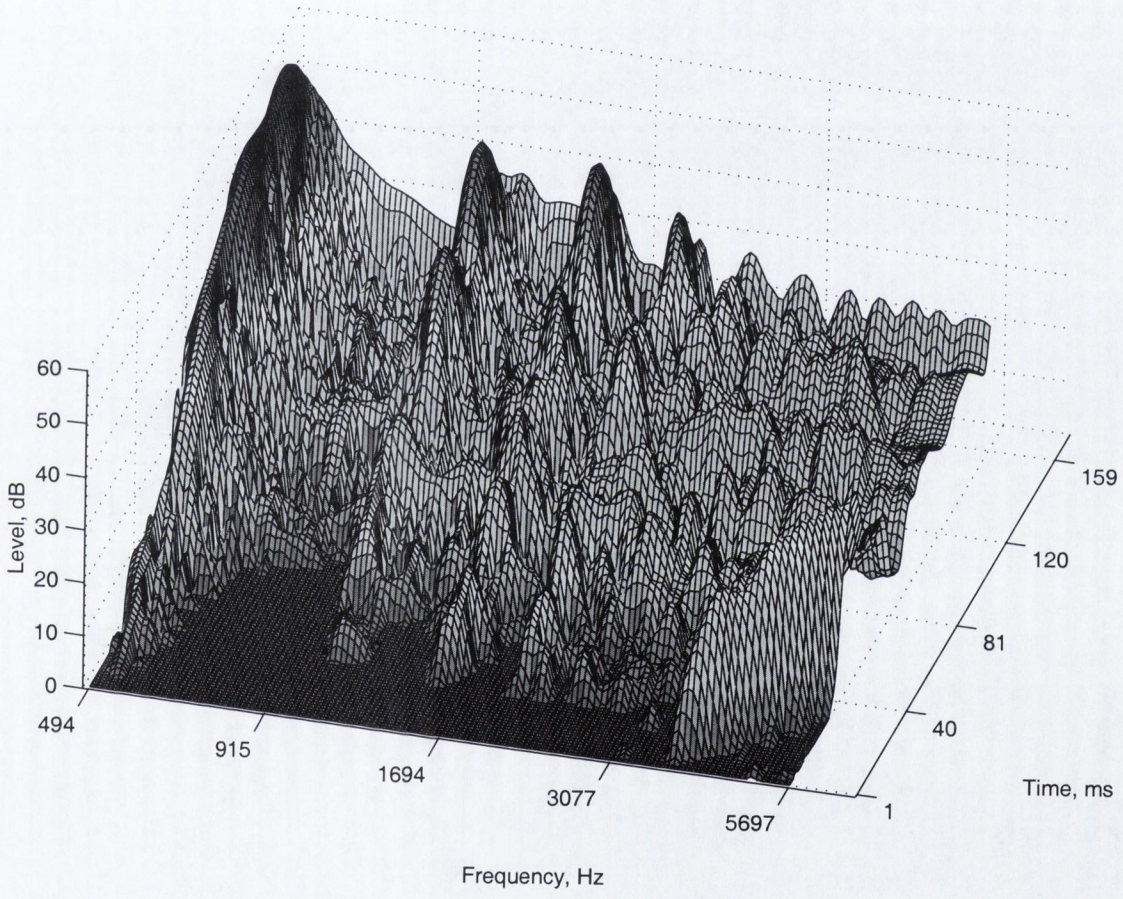


Figure 5.20: Constant-Q distribution of a flutter-tongued flute attack for the note C5. The first 60 dB of the distribution is shown.

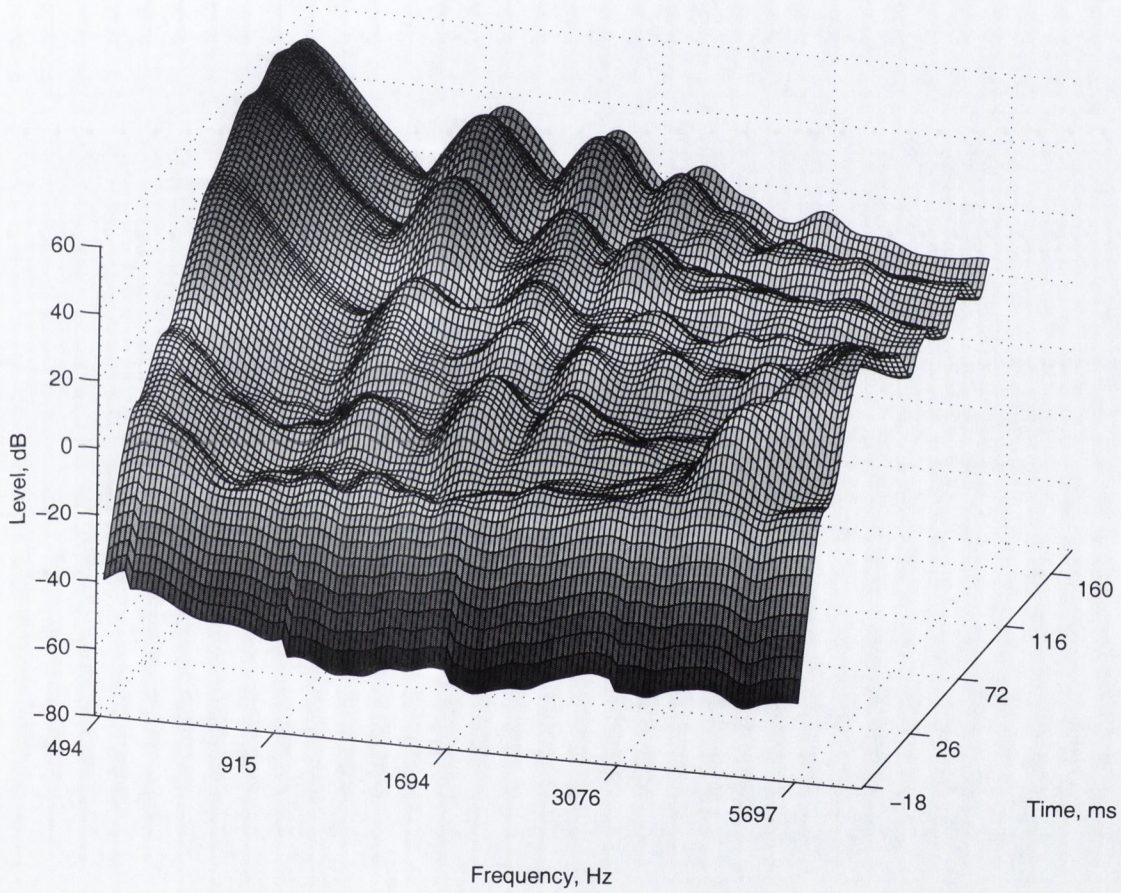


Figure 5.21: EWD of a flutter-tongued flute attack for the note C5. The full dynamic range (140 dB) of the distribution is shown.

5.5.1 Spectral Resolution

The validity of the EWD as a model of auditory spectral resolution depends on the usefulness of the gammatone filter model. In particular, Moore and Glasberg [111] make some salient points regarding possible caveats of the asymmetric roex(p) shape and excitation pattern models which also apply to the gammatone filter shape and EWD. They note that the accuracy of auditory filter shapes derived from notched noise data is limited by the assumptions of the power spectrum model of masking. They also point out that there is still some disagreement over the value of ERB for low centre frequencies. Furthermore they note that any generic auditory filter model is naturally compromised by the fact that frequency selectivity differs slightly from individual to individual; these differences being more pronounced for the hearing impaired.

Moore and Glasberg also note that the effect of *suppression* is not modelled by either asymmetric, symmetric or gammatone filter models, Suppression refers to a reduction in neural activity at one centre frequency caused by the presence of activity at a neighbouring centre frequency [108]. However, Moore and Glasberg note that not modelling suppression is not a serious limitation, since suppression has only a minor effect on simultaneous masking and is therefore not important for excitation pattern calculation.

A further limitation relevant to the EWD may be added to those given by Moore and Glasberg above, that being that the EWD filter shape is not level dependent. However, such level dependency may be incorporated if the kernel is made level dependent following the design approach outlined in Chapter 4. As noted earlier however, level dependency has been shown in Section 5.2.1 to be important for excitation pattern calculation for high stimuli levels only.

5.5.2 Temporal Resolution

It has been pointed out above that using a level independent kernel for the EWD is not a serious limitation for modelling spectral resolution. However, level dependency may not be so easily disregarded if temporal masking is to be accurately modelled. This is all the more true for signals exhibiting abrupt changes in level, where temporal masking effects become more important. The success of the EWD at modelling temporal masking is essentially governed by the limitations of the temporal window model itself. However, the ability of the EWD to accurately reflect auditory temporal masking may be further constrained by the fact that it does not incorporate the level dependency of the temporal window model itself.

In [150], Plack notes that the temporal window shape is linear for a given stimulus level.

Therefore nonlinear effects such as the *masker overshoot effect* [194], the *nonlinear growth of forward masking* [152] and the *nonlinear additivity of forward and backward masking* [49] are not incorporated. Masker overshoot refers to the fact that masking may be greater at signal onset than during the steady state portion of the masker. The nonlinear growth of forward masking refers to the fact that for forward masking a given increase in masker level does not requires a corresponding increase in signal level for the signal to remain at threshold. This effect is dependent on signal level. For low signal levels, an increase in masker level only requires a much smaller increase in signal level to keep the signal at threshold. As the signal level increases, however, the masking function steepens, and becomes close to unity for high signal levels. Recent work by Plack and Oxenham [152] addresses the nonlinear effects of non-simultaneous masking. For simultaneous masking, the ratio of signal-to-masker at threshold is constant for a wide range of levels. The nonlinear additivity of forward and backward masking refers to the fact that the masking effects of two or more maskers do not obey the normal rules of addition of intensity [194]. The additivity of masking is explored by Oxenham and Moore in [125].

Plack also adds to the above limitations the fact that the temporal window model does not take into account the phenomena of *adaption* and *fatigue*. Adaption describes the rapid reduction in neuron firing rates after the onset of a stimulus; Plack notes that adaption is probably not responsible for non-simultaneous masking. Fatigue refers to the temporary loss of sensitivity (a temporary threshold shift) caused by exposure to intense stimuli of long duration. However the effect of fatigue is minor for stimuli levels below about 80 dB SPL, and only becomes noticeable after durations of several minutes.

A further limitation of the model discussed by Plack, which also applies to the EWD is that information present in the fine structure of a stimuli is ignored because only changes in signal envelope are registered. Therefore, the well known *phase locking* ability of auditory neurons is not incorporated in the temporal window model and EWD. Phase locking refers to the tendency for auditory neurons to fire on one particular phase of a signal waveform only. Information from phase locking is probably important for pitch detection [108] and localisation tasks [16]. However, Plack points out that this is not necessarily a limitation of a model of auditory resolution, and makes several salient points which are also relevant to the EWD : "The model might appear at first to be incompatible with a representation based on information about the fine structure of signals. However, this is not necessarily the case. The model is intended to describe to what extent the auditory system can resolve changes in a stimulus over time ... A system using phase locking might be able to determine very accurately the frequency of a continuous sinewave, for example, but this does not imply that

changes in the pattern of phase locking could be accurately resolved over time. There is a distinction therefore between the resolution itself and the quantity being resolved. Rather than say that the auditory system's ability to use phase locking information implies high temporal resolution (to the level of neural spikes), it might be more appropriate to regard the frequency information derived from phase locking as a quantity whose changes are not perfectly resolved. The model is intended primarily as a model of resolution and not a complete description of auditory processing."

We also add to the limitations of the temporal window model listed above the fact that the EWD does not incorporate the level dependency of the temporal window model itself. The temporal resolution of the EWD is dependent on frequency only. However, level dependency may be incorporated using the existing EWD framework by making the smoothing kernel signal dependent. The temporal resolution of the EWD may be made level dependent without affecting the spectral resolution in any way, since the kernel is separable. The scope of this work was limited to level independent modelling due to time constraints. However, incorporating temporal resolution level dependency should be the next step in improving the EWD model. This improvment is readily achievable, since the incorporation of new or revisions to existing auditory models into the EWD's kernel design is greatly facilitated by the independence of EWD spectral and temporal resolutions.

The freedom to independenly choose time and frequency resolution is a key advantage of using a Cohen's class time-frequency distribution to model auditory processing. In contrast, independent control of the time and frequency resolution of linear time frequency distributions, such as the wavelet transform or the spectrogram, is not possible. Controling the resolutions of filterbank based auditory models such as the STEP is also complicated by the fact that the temporal resolution of such models depends on both the temporal window smoother and also on the filterbank time domain responses. Substituting a new filterbank to change the frequency resolution of such models will have an unavoidable effect on their temporal resolution. Such an effect may not always be desirable.

As a joint time-frequency distribution incorporating the temporal and spectral resolutions of the auditory system, the EWD is strategically situated between standard TF signal analysis techniques, designed to extract signal properties, and auditory models, designed to investigate auditory signal perception. It is thus a very useful, perceptually motivated joint time-frequency signal analysis tool. Its motivation is psychophysical rather than physiological, providing a link between a signal's physical properties and its auditory perception, rather than being a detailed model of the internal auditory system. Adding further complexity to the EWD model by incorporating phase locking, for example, would

not necessarily prove advantageous, as the EWD would gravitate towards physiological auditory modelling and move away from the elegant mathematical signal analysis framework provided by Cohen's class. The following chapter provides clear examples of the utility of the signal property-perception link provided by the EWD. Here, the EWD is applied to two important areas of audio signal processing, namely that of head related transfer function smoothing for 3-D surround sound system design and the analysis of perceptible degradation in synthetic reverberation algorithms. These two examples highlight the spectral and temporal resolutions of the EWD respectively. In Chapter 6 the frequency marginal of the EWD is used to remove redundant spectral detail from HRTF data. In Chapter 7, temporal excitation patterns of signal partials are extracted from EWDs using the temporal marginal property. The smoothness, lack of cross-term interference and lack of clutter exhibited by the EWD is crucial to the success of both applications. Together, these chapters strongly illustrate the utility and general applicability of the EWD and highlight the advantages of incorporating both spectral and temporal resolutions into a single TF distribution.

Chapter 6

Head Related Transfer Function Smoothing

When one considers the vast changes that have taken place over the last few decades in the realms of computing, digital multimedia, telecommunications and the like, it is remarkable to note that one of the fundamental principles behind the reproduction of audio, namely stereo sound (or *intensity stereo*) [17], has remained unchanged for over 50 years. Of course, the introduction of digital audio, the audio CD, and minidisk technologies have resulted in improvements in the form of less audible noise and greater audio clarity. However, despite these improvements, almost all audio reproduction today is still compromised by one of the fundamental limitations of the ubiquitous stereo specification, namely that stereo imaging is limited to a periphery bounded by the two loudspeakers [93]. Put more simply, audio images generated by stereo sound can only be panned along an arc somewhere between the two loudspeakers, and not outside this limit. Also, stereo panning is limited to the horizontal plane only [182]. It is obvious that in real listening conditions sound can approach us from any direction and also that each sound we hear is accompanied by a sense of this direction. Thus, the ear can simultaneously *localise* many sound sources. For example, if we cycle down a busy street, sound approaches us from all directions providing us with valuable information about our environment. We may hear the conversation of pedestrians to our left, an aircraft passing overhead, or even perhaps the sound of a motorist not overly sympathetic to cyclists passing dangerously close on the right! In a more relaxed environment, for example sitting near the front row at a concert hall performance, we can readily identify the locations of the various instruments in front of us (and also possibly the people coughing all round during quiet movements !). The aim of a '3-D surround-sound system' or 'virtual auditory display' is to re-create this aural experience by 'fooling' the

ears with synthetically generated audio localisation cues. Because of the limitations of the standard stereo reproduction system, the localisation cues stereo provides can only generate a crude approximation to the real aural experience [15]. Furthermore, as stated earlier, the images generated by stereo are limited to the region between the two loudspeakers.

In spite of the fact that virtual auditory display technology has seen recent application in such areas as air traffic control and in computer game audio, surround-sound has seen very little application in general consumer audio, for example in the music industry. This is perhaps due to an unwillingness on the industry's part to invest heavily in a new technology, especially when stereo is perceived as being 'good enough' by most people. It may also be due to an ignorance on the part of consumers of the potential benefits of more realistic stereo imaging. Indeed, the technology and expertise now exist for the design of surround-sound recording studios capable of producing immersive recordings lush with virtual images. Furthermore, such recordings do not necessarily require any additional expenditure on audio equipment by the end user as realistic virtual audio images may actually be generated using only two speakers, *i.e.* the standard stereo setup [182].

As we will see in the following sections, the auditory cues important for localisation for a particular person may be summarised by that person's set of head related transfer functions (HRTFs) or head related impulse responses (HRIRs). A HRTF set encapsulates the frequency dependent gain and phase changes sound undergoes as it interacts with a listener's body before it reaches the left and right eardrums. A head related impulse response is the time domain description of a given HRTF. These effects are primarily caused by the reflection and diffraction effects of a person's torso, shoulders, head and ears. However, because no two people are physically identical, there is considerable inter-subject variability in HRTF measurements. Therefore, a major difficulty in auditory display design is in achieving a generic HRTF database which works well for everyone, or at least for the vast majority of the population [185] [186]. Furthermore, the timbre of the resulting virtual images is of great importance. In addition, because of spectral masking, there is a considerable amount of spectral detail in HRTF measurements which is not relevant to localisation. However, discriminating between the relevant and irrelevant detail is a difficult task. This chapter therefore explores the use of the EWD for HRTF analysis. The EWD is demonstrated to be ideally suited to the removal of redundant high frequency spectral detail not relevant to localisation. Such information is shown to be spectrally masked and is removed by the critical band smoothing inherent to the EWD. Furthermore, as the EWD removes only spectrally masked detail, a significant reduction in inter-subject variation is achieved without removing HRTF features important to localisation. Finally, the HRTFs generated

by EWD smoothing are compared both visually and subjectively to another form of data smoothing, using principal component analysis (PCA)[89]. The EWD-smoothed HRTFs are shown to be superior in terms of both virtual imaging naturalness and imaging accuracy. One particular currently favoured method of generating a generic HRTF set is to deduce features important to localisation from HRTF measurements taken from a large number of listeners. Such measurements generally contain unwanted, spectrally unresolvable high frequency detail such as notching caused by the complicated fine structure of the pinna. Sometimes the particular measurement technique used is also culpable. It is desirable therefore to remove such spectral artifacts without compromising other, more important HRTF details. Again, the EWD is highly suited to such a task as it removes only spectrally masked detail.

6.1 Auditory Localisation and Localisation Cues

When the perception of an acoustic event occurs, it is accompanied by a sense of the location of the sound source. This is referred to as *spatial hearing* [16]. The ability of the auditory system to determine the location of a sound source is termed *localisation.* Certain sounds are more easily localised than others. As an example, the impulsive sound of a hand clap is easier to localise than a sound comprised of continuous tones, such as the ringing sound of a mobile phone. The listening environment also has an influence on localisation. A reverberant listening environment is usually characterised in terms of *direct, early* and *reverberant fields* [15]. In such an environment, for example a concert hall, sound first reaches the ears via a direct path between sound source and ears. This initial sound is known as the direct field. The early field refers to a period up to 100 ms after the direct field, during which sound reaches the ears via reflections off nearby surfaces, such as the walls or floor. The reverberant field occurs after the early field and may last for several seconds, during which a very large number of reflections reach the ears from all directions. One might expect therefore, the auditory system's ability to localise in such an environment to be very poor. However, the well known *precedence effect* [16] causes the auditory system to give much more weight to the acoustic cues present in early incident sound waves (up to a delay of about 2 ms) than to cues present in subsequent reflections. Therefore only very early reflections (less than 2 ms) significantly hinder localisation. Generally, reflections caused in reverberant environments modify the perception of the timbre and loudness of a sound and give a sense of spaciousness and envelopment. Also, the ratio of direct to reverberant signal energy is an important cue to the distance between listener and sound source [15].

The greater the ratio of direct to reverberant energy, the closer the sound source appears. Reverberation is discussed further in Chapter 7, where natural and synthetic reverberation are analysed using the EWD.

The primary influences on localisation are the reflection and diffraction effects that sound undergoes as it interacts with the listener's body. For a given angle of incidence, these effects lead to unique modifications of sound before it impinges upon the eardrums. Furthermore, localisation is severely impeded under monoaural conditions and therefore depends heavily on binaural cues, namely the cues of *interaural level difference* (ILD) and *interaural time difference* (ITD), both of which are frequency dependent. The most significant physical contributors to these cues are outlined below and Table 6.1 lists in order of importance the most dominant cues for horizontal and vertical localisation.

- *The Torso and Shoulders* : The effect of a listener's torso and shoulders has been shown to be important, primarily for vertical localisation. The 'shoulder bounce cue' refers to the reflection of elevated sound from the top of a listener's shoulders. The shoulders and torso also have some influence on localisation in the horizontal plane, causing a measurable directional influence of up to $\pm$ 5 dB in certain frequency bands [15].

- *The Head* : The listener's head is perhaps the single most important provider of auditory localisation cues. The difference in the path length for a sound travelling to the left and right ears causes the interaural time difference cue, the most important cue for horizontal localisation. The *acoustic shadowing* effect of the head causes a reduction in the signal level at the far ear. ILDs derived from the listener's head are the second most important horizontal localisation cue [182].

- *The Outer Ear* : The outer ear consists of the pinna and the ear canal, the pinna being the external part of the ear made of a single complex shaped cartilage, and the ear canal being the narrow tunnel about 2-3 cm in length leading from the pinna to the ear drum. The complex shaping of the pinna causes frequency dependent constructive and destructive interference. This spectral modification depends on the angle of incidence of the sound source producing spectrally varying ILD cues which are very important for horizontal and vertical localisation. Both the outer and inner pinna structure are important for localisation, the outer pinna structure being important for horizontal localisation and the finer inner pinna structure being more important for vertical localisation. Lastly, the resonant frequency of the cavum conchae, the largest resonant cavity at the opening of the ear canal, has been shown to increase with

increasing source elevation and is therefore an important localisation cue [15].

The seven most important horizontal localisation cues	The seven most important vertical localisation cues
ITDs due to head	ILD's due to cavum conchae
ILDs due to head	ILD's due to inner pinnae structure
ILDs due to outer pinnae structure	ILD's due to outer pinnae structure
ILDs due to inner pinnae structure	ITD's due to head
ILDs due to cavum conchae	ILD's due to listener's head
ILDs due to shoulders	ILD's due to 'shoulder bounce cue'
ILDs due to torso	ILD's due to torso

Table 6.1: Horizontal and vertical localisation cues listed in order of importance.

6.1.1 Head Related Transfer Functions

The reflection and diffraction effects of the combined torso, shoulders, head and pinna can be modelled as linear, time invariant transfer functions known as head related transfer functions (HRTFs). For any given sound location, a HRTF measurement for both left and right ears is sufficient to fully describe all the ILD and ITD localisation cues. HRTFs can be generated by measuring the impulse response between source and listener for a given sound location and transforming the data into the frequency domain. The response is usually measured at or close to the eardrum in order to capture the effect of the ear canal, as well as the effect of the outer and inner pinna, head, shoulders and torso. Figures 6.1, 6.2 and 6.3 illustrate the HRTF gain functions for the left and right ears of three subjects for a source localised in the *horizontal* plane (*i.e.* elevation is 0^o). The data is based on measurements taken at the Waisman Centre at the University of Wisconsin [187], discussed further in Section 6.2.1. The data is plotted up to a frequency of 15 kHz for an azimuth of 0^o to 180^o, with zero degrees corresponding to a source directly in front of the listener and azimuth angle increasing clockwise around the listener's head. The azimuth angle is plotted in reverse for the right ear so that the left and right ear data can be better compared for each subject. It is clear from the illustrations that there is a great deal of inter subject variability between subjects, particularly for frequencies above 5 kHz. These high frequency spectral details are mainly due to differences in the complex inner pinnae shapes of each listener and may also be due to artifacts of the measurement process. Spectral detail below 5 kHz is mainly caused by the shadowing effect of the head and also the outer pinna structure,

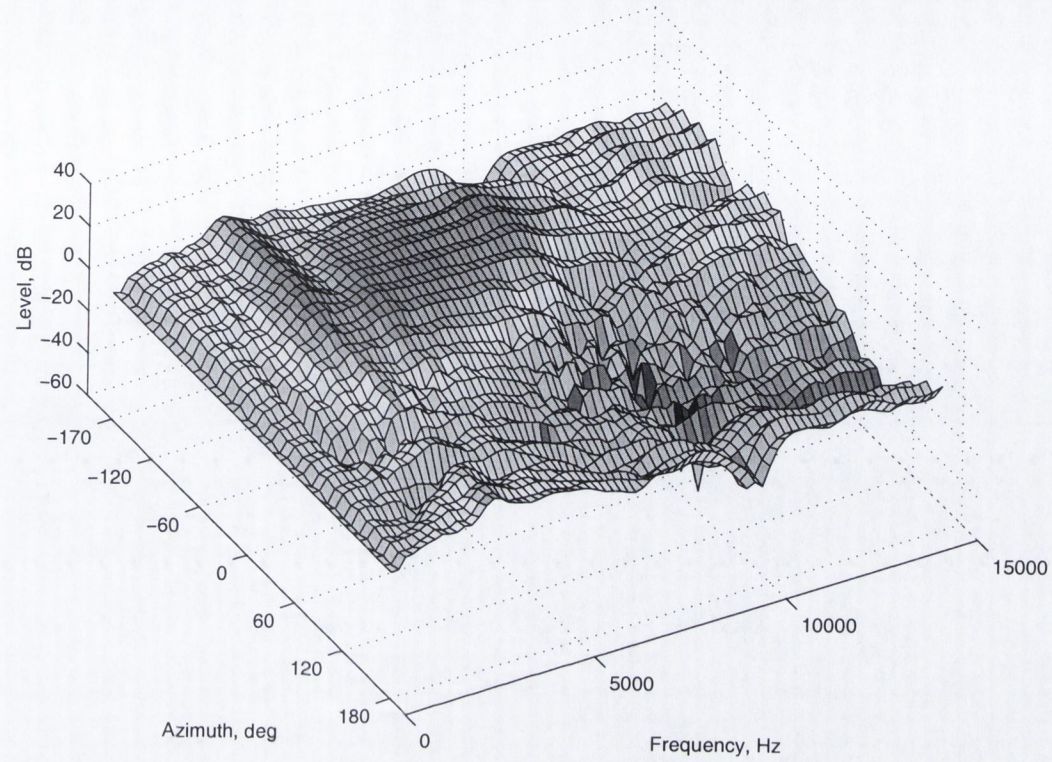
which exhibits much less listener specificity. It is also interesting to note that there are measurable differences between HRTFs of the left and right ears of each subject.

During the 1980's, developments in the field of digital signal processing led to a much greater flexibility in filter design and in particular made possible the design of IIR and FIR filter structures based on complex HRTF specifications. A given individual's HRTF set could be measured and then used to filter sound for binaural presentation over headphones [185], [186]. In this manner, the localisation cues generated by an individual's physical features could be replicated to a very accurate degree, resulting in very convincing virtual imaging. However, such a system can only work well for the individual in question. In order to design an auditory display suitable for many individuals, a more generic HRTF database, incorporating the commonalities of HRTF features for many listeners, is required. Some possible methods of achieving such a database are described below.

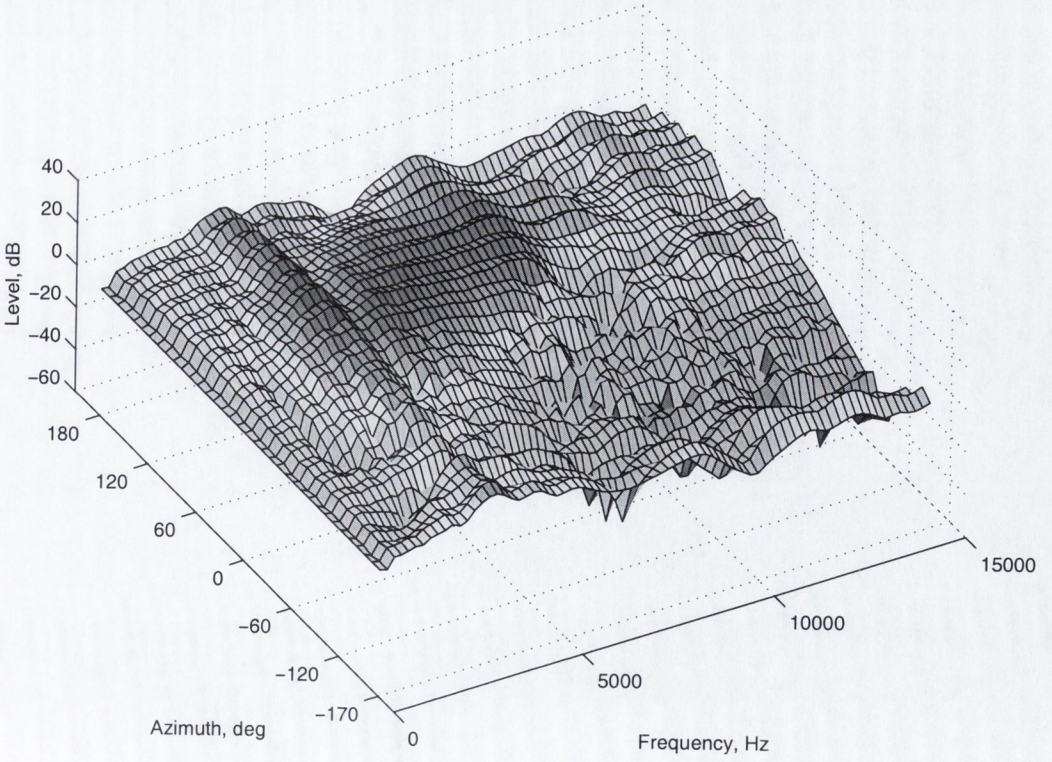
6.2 Obtaining a Generic HRTF Set

The success of any virtual auditory display system depends on its ability to produce convincing virtual images for a high proportion of its listeners. As demonstrated in the previous section, HRTF data shows much listener specific detail which tends to overshadow the commonalities or shared features of listener HRTFs for any given azimuth. Therefore, much work has been carried out in attempting to identify those HRTF features which contribute to localisation. The goal of such work is the specification of a generic or 'average' HRTF set capable of synthesising convincing virtual images for as many listeners as possible. Approaches to the design of such a generic HRTF set can generally be classified as either *analytic* or *empirical*. An example of the *analytic* approach is Blumlein's original intensity stereo scheme [17], which still remains by far the most widespread stereophonic-format in use today, despite the fact that it was conceived almost 70 years ago. Intensity stereo seeks to provide ITD based localisation cues by changing the ratio between the left and right loudspeaker levels. ITD cues are provided because these level changes are actually equivalent to a phase difference between the signals reaching the listener's left and right ears. Intensity stereo is designed so that these phase differences equate to appropriate ITD cues. Hence, an impression of panning is achieved.

Intensity stereo yields natural sounding images, but is limited in that it generates correct ITD cues only. Furthermore, intensity stereo is unable to produce virtual images outside the loudspeaker arc. An extension of this simple approach, known as spectral stereo, is taken by Cooper and Bauck [38]. This approach seeks to incorporate the diffraction effects

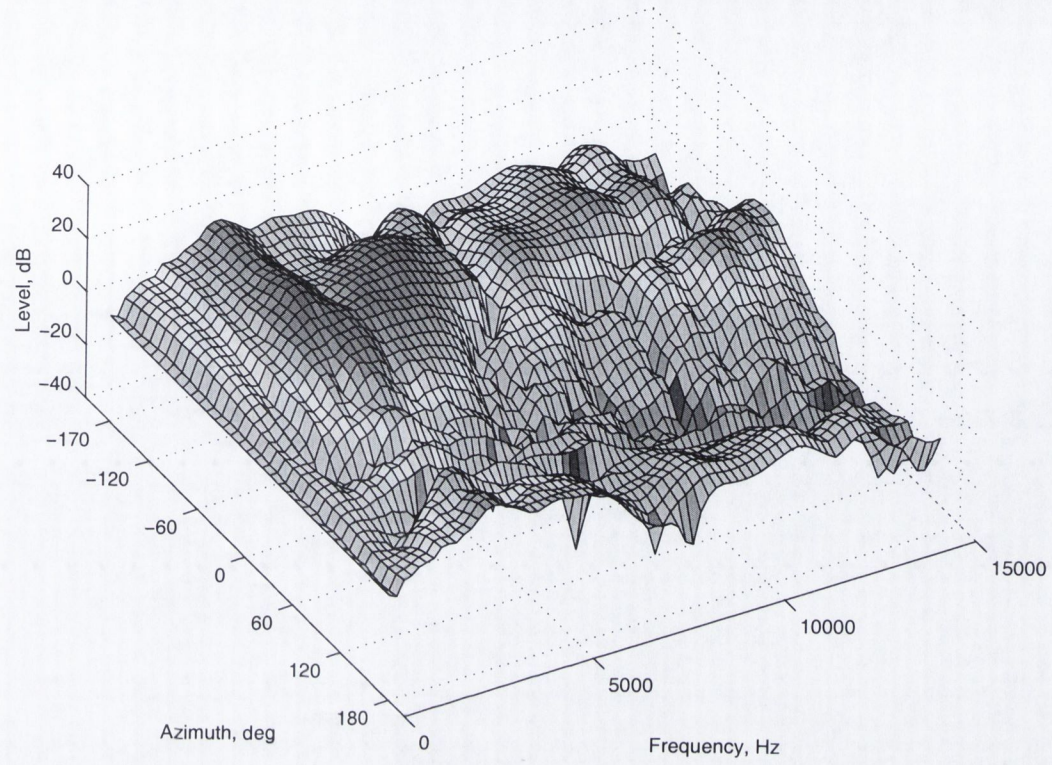


(a)

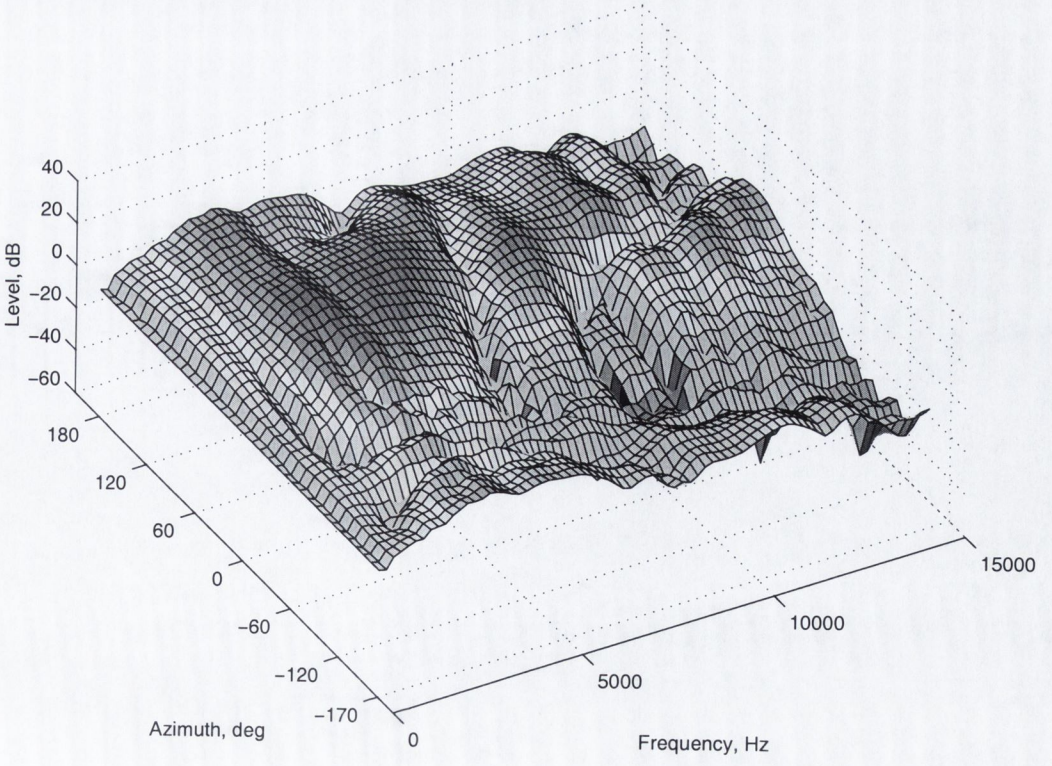


(b)

Figure 6.1: HRTF magnitude responses for subject AFW : (a) left ear and (b) right ear

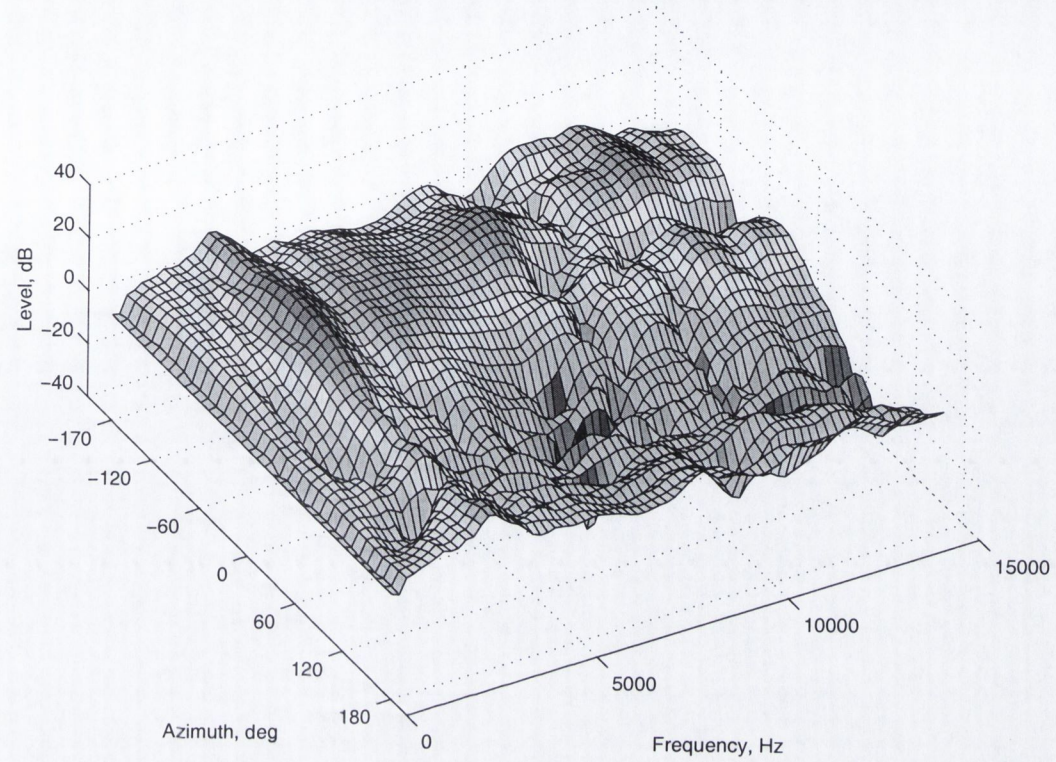


(a)

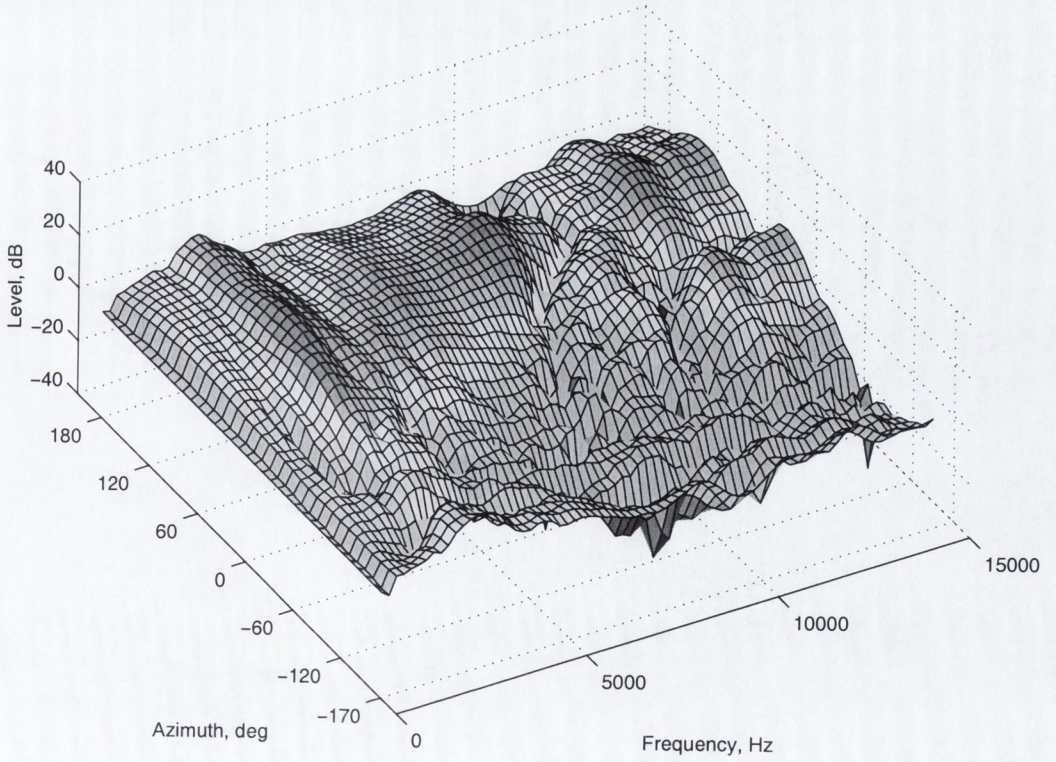


(b)

Figure 6.2: HRTF magnitude responses for subject SOW : (a) left ear and (b) right ear



(a)



(b)

Figure 6.3: HRTF magnitude responses for subject SOU : (a) left ear and (b) right ear

of the head by modelling it as a spherical surface, resulting in simplified, synthetic HRTFs knows as diffraction coefficients. More recently, parametric mathematical models of the head and pinna have been employed in an attempt to investigate their influence on the final HRTF function [48], [47] and to characterise an average head. One popular *empirical* method of generating an average HRTF database is to take measurements from a real model of an averaged head, called a 'dummy head'. Such dummy heads usually consist of a smooth approximation of the human head and torso and feature simplified pinnae shapes. Unfortunately, they cannot therefore include the complex diffraction and reflection effects of the outer and inner pinna structure.

A second empirical method of obtaining an average HRTF set is to use some form of mathematical averaging of a large database of HRTF measurements of real subjects. As illustrated in Figures 6.1, 6.2 and 6.3, HRTF measurements exhibit much high frequency spectral detail showing considerable variability between subjects. However, much of this detail is not resolvable by the auditory system. For frequencies below about 5 kHz, much less inter-subject variability is exhibited. A successful mathematical averaging approach would therefore produce a HRTF set containing the spectral features relevant for localisation for a high proportion of listeners while discarding irrelevant or spurious spectral features.

Principal component analysis, a method of statistical data reduction, has been investigated at Trinity College Dublin by Walsh for the purposes of HRTF smoothing [182]. This work is reviewed in Section 6.2.1. However, as discussed below, PCA smoothing employs no psychoacoustic criterion in smoothing HRTF data; it is a statistical technique which does not discriminate between the type of data to be analysed. The use of the EWD is therefore proposed in this thesis as an alternative method of HRTF smoothing since it directly incorporates the critical band masking (smoothing) of the auditory system and therefore introduces an appropriate degree of smoothing at all frequencies. This approach is outlined in Section 6.3 and is a continuation of the work carried out by Walsh. It is demonstrated in the following sections that EWD analysis is much more suited to the task of smoothing HRTF data than PCA as it reduces inter-subject variability and spurious spectral detail not resolvable (masked) by the auditory system while preserving the shared features exhibited by HRTF data.

6.2.1 PCA HRTF Smoothing

Principal component analysis (PCA) was introduced in 1933 by Hotelling, who applied the technique to the scoring of intelligence tests. The mathematical details of PCA are not covered here in detail. It suffices to say that PCA re-express data as a linear combination of

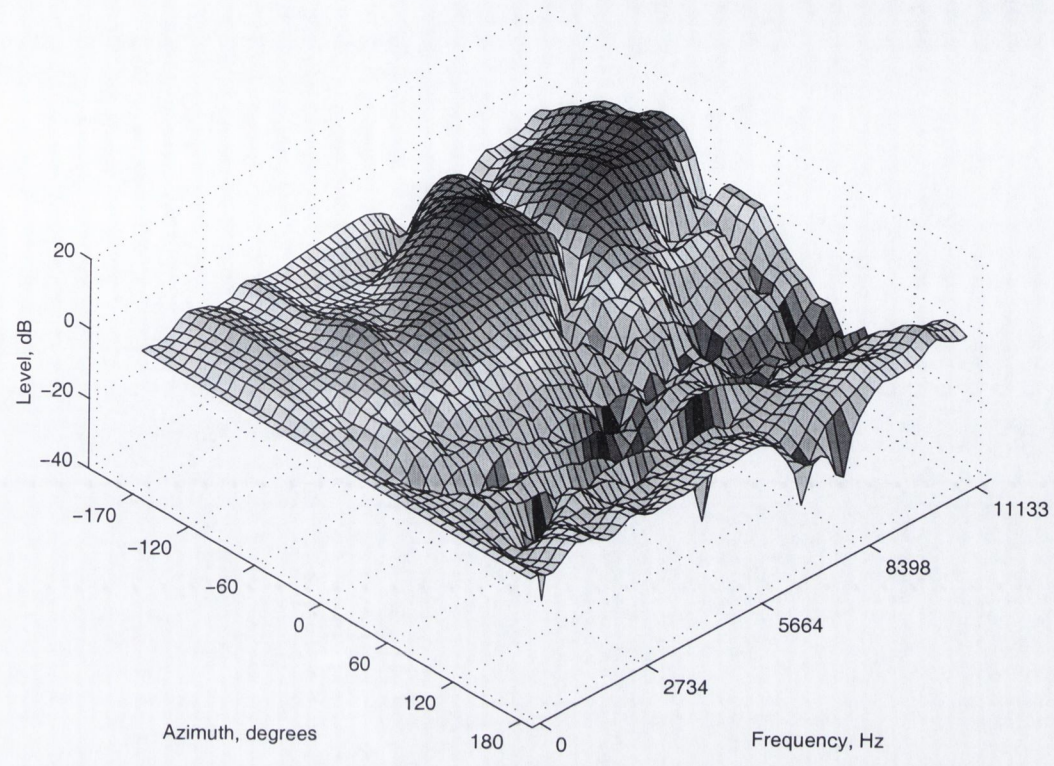
a set of orthonormal basis functions, or principle components (PCs). These basis functions are derived by calculating the eigenvectors of a matrix of correlation coefficients derived from the original data set. The number of PCs used to represent the data is known as the PCA order. The reconstructed data becomes more like the original data as the PCA order is increased. PCA is effectively therefore a useful method of data reduction.

In [89], Kistler and Wightman performed the PCA of HRTF data. The HRTFs were measured from both ears of 10 subjects at 256 source positions. They found that fewer than 5 PCs were necessary to account for over 90 % of the variation in the data. Listening tests revealed that subjective judgements of the apparent directions of headphone-presented stimuli were almost identical for both the original HRTF measurements and the PCA modelled HRTFs. In other words, listeners found the PCA data to be a realistic reconstruction of their HRTFs. Principal component analysis (PCA) [84] was therefore proposed and investigated by Walsh [182] as a possible method of HRTF averaging or smoothing. By varying the number of PCs, a greater or lesser degree of smoothing could be achieved.

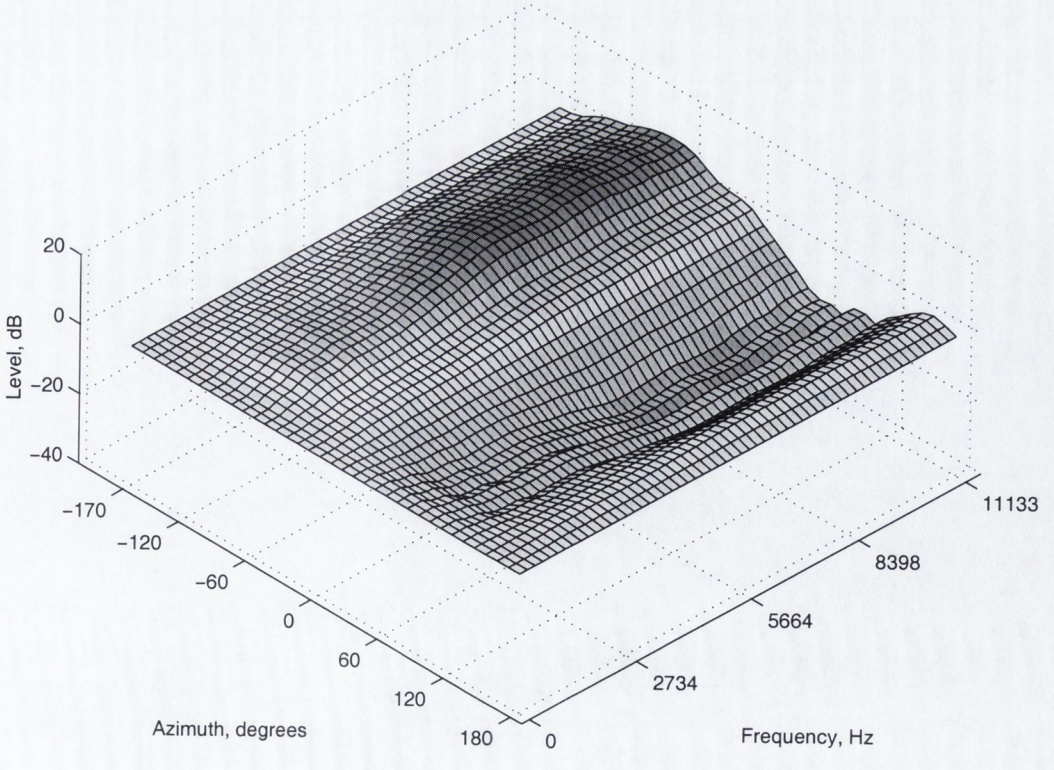
The HRTF data used by Walsh was a subset of measurements made at the Waisman Center by Fredrick Wightman and Doris Kistler [187]. This database consists of HRTF measurements of five subjects for azimuths from 0^o to 350^o and elevations from 80^o to -50^o taken at 10^o intervals. Subjects are referred to by the initials: AFW, SOW, SOU, SOS and SJX. HRTFs taken from this database in the horizontal plane only are considered in this study, and for the purposes of PCA analysis the frequency range is limited to 58 bins between 400-11025Hz. Therefore, a total of 360 HRTFs are analysed using PCA—36 HRTFs for both left and right ears of all five subjects. PCA of the data for orders 1 to 7 was carried out by Walsh; Table 6.2 presents the amount of variance preserved by this analysis.

PCA operates on real-valued data only, therefore PCA was performed on the log magnitude of the directional transfer functions (DTFs) of each subject. A DTF set is the HRTF set of a particular subject minus the across azimuth mean (AAM) for both ears. Thus, a DTF only contains the direction dependent information of a HRTF as direction independent information such as the ear canal resonance is factored out [90].

Since PCA used the log magnitude of the DTFs the HRTF frequency dependent phase response was lost. Filter design requires the specification of both a magnitude and phase response. However, Kistler and Wightman [90] demonstrate that the frequency dependence of the interaural time delay is not perceptually relevant and that only the overall ITD between the left and right ear HRTFs for a given source azimuth is important. They therefore represent the ITD at each source position by a constant value and furthermore

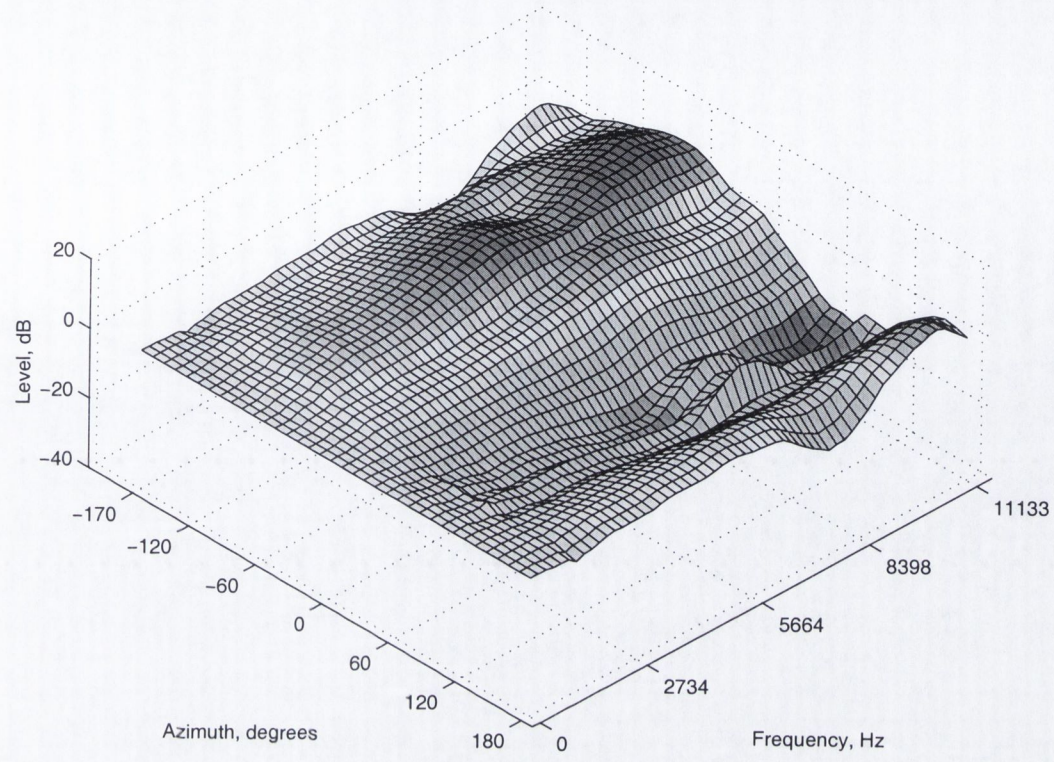


(a)

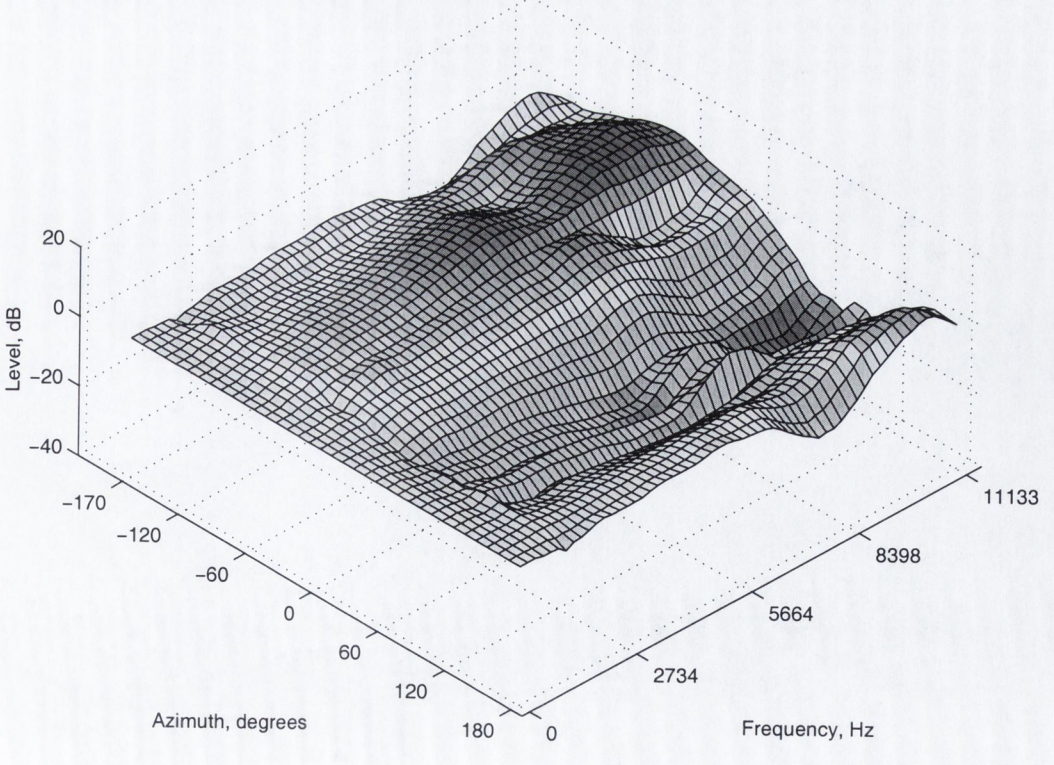


(b)

Figure 6.4: (a) Raw DTF data and (b) PCA-smoothed data (1 PC) for the left ear of subject SOW.

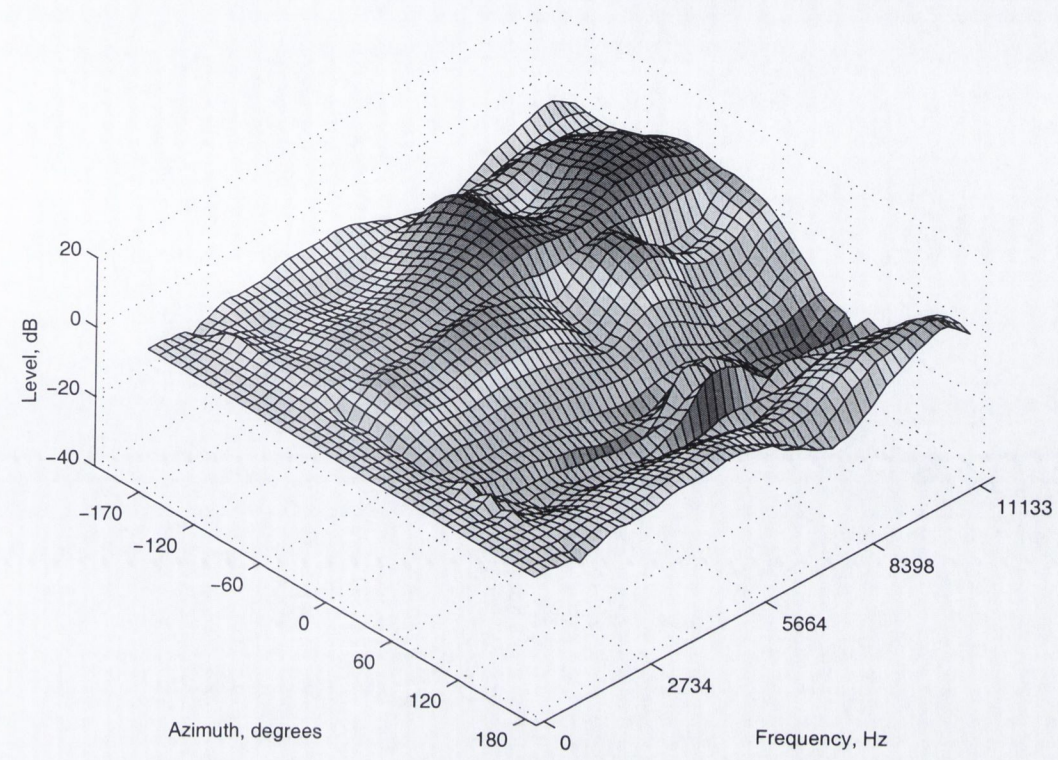


(a)

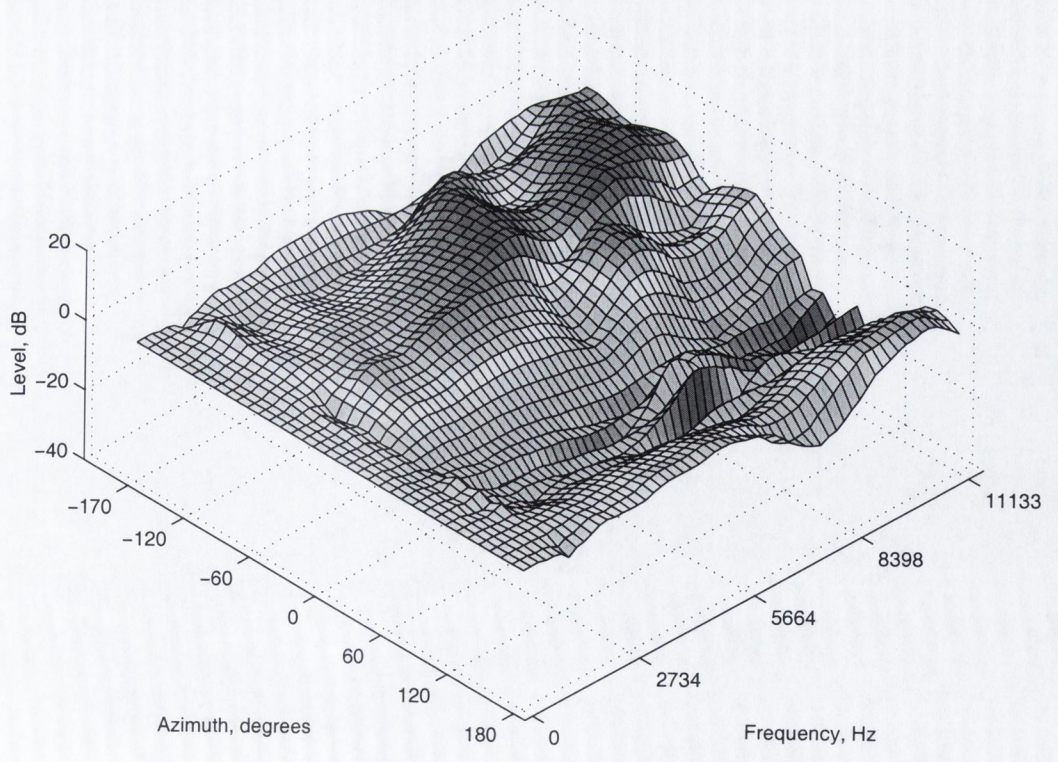


(b)

Figure 6.5: PCA-smoothed data for the left ear of subject SOW. (a) 2 PCs (b) 3 PCs.

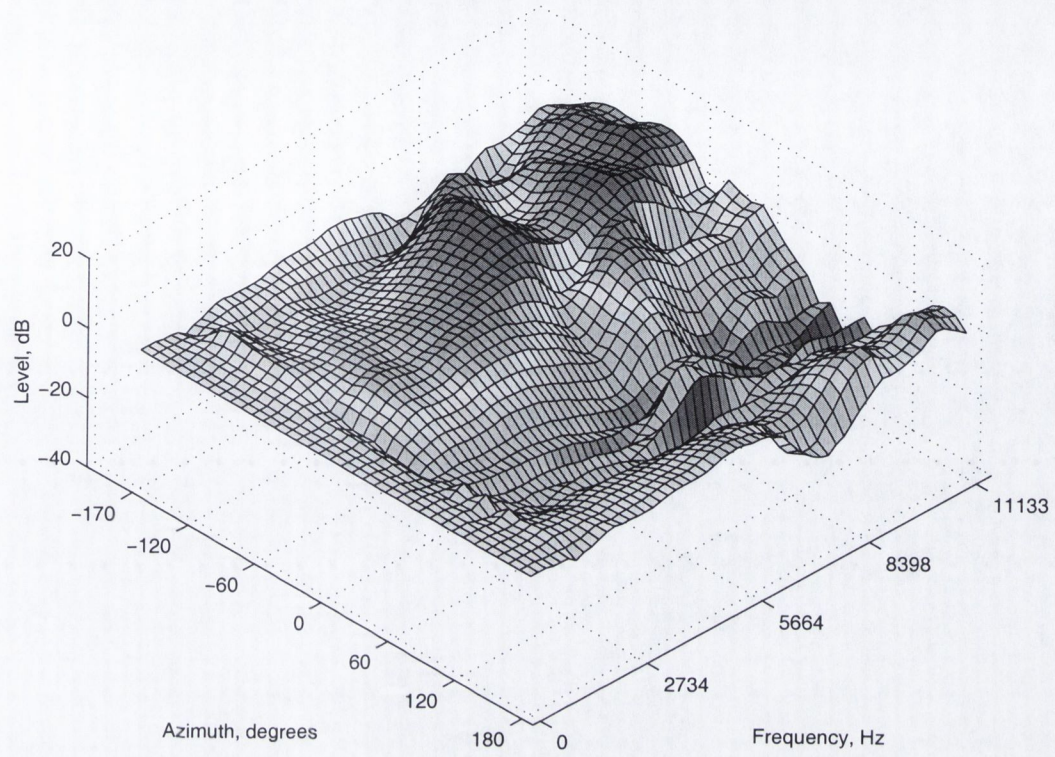


(a)

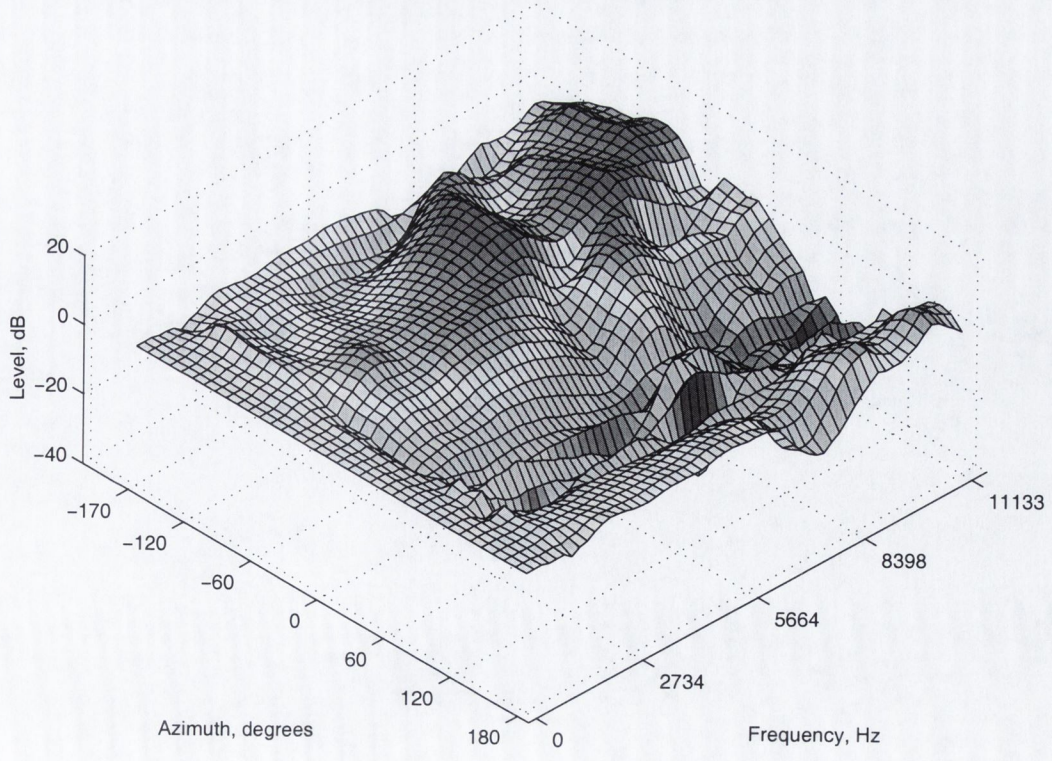


(b)

Figure 6.6: PCA-smoothed data for the left ear of subject SOW. (a) 4 PCs (b) 5 PCs.



(a)



(b)

Figure 6.7: PCA-smoothed data for the left ear of subject SOW. (a) 6 PCs (b) 7 PCs.

Number of PCs Used in Reconstruction	Percentage of Variance Preserved
1	80.51
2	84.94
3	88.14
4	90.72
5	92.62
6	93.75
7	94.58
8	95.20
9	95.78
10	96.30

Table 6.2: Percentage of variance preserved in DTF data after PCA.

assume that HRTFs can be modelled as minimum phase functions [123]. They demonstrate that a given left-right pair of HRTFs can be modelled by computing the minimum phase functions of the PCA plus across azimuth mean for both ears, and adding an appropriate interaural time delay. This may be expressed as

$$\frac{HRTF_L(\theta_s)}{HRTF_R(\theta_s)} \approx \frac{HRTF_L(\theta_s)|_{MIN}}{HRTF_R(\theta_s)|_{MIN}} H_{AP}(ITD(\theta_s)) \ , \tag{6.1}$$

where $HRTF_L(\theta_s)$ and $HRTF_R(\theta_s)$ represent the HRTF functions at a source angle of θ_s for the left and right ears respectively, $HRTF_L(\theta_s)|_{MIN}$ and $HRTF_R(\theta_s)|_{MIN}$ represent their minimum phase equivalents and $H_{AP}(ITD(\theta_s))$ represents an allpass filter with linear phase providing the appropriate pure interaural time delay for a sound source at θ_s.

Figures 6.4, 6.5, 6.6 and 6.7 present the results of Walsh's PCA analysis for the left ear of subject SOW. Figure 6.4 (a) illustrates the DTF of the original HRTF data. Comparing this with the plot of the corresponding HRTF data, shown in Figure 6.2 (a), we see that the primary effect of removing the across average mean is to eliminate the broad resonance around 3 kHz due to the ear canal[1]. The DTF of the original, raw data is quite 'rough' and exhibits deep notches at mid to high frequencies, many of which are unlikely to be resolvable by the auditory system.

Figures 6.4 (b) and Figures 6.5, 6.6 and 6.7 plot the PCA derived DTFs for PCA

[1]Note that the HRTF data plotted in Figure 6.2 (a) is plotted over the range 0-15 kHz, whereas the data in Figure 6.4 (a) is only plotted from 0-11025 Hz.

orders 1, 2-3, 4-5 and 6-7 respectively. It is clear that PCA analysis of order 1 produces a DTF containing very little spectral detail. Such a DTF is only capable of reproducing the gross ILD localisation cues of the head, but not of the more detailed spectral cues of the outer and inner pinna structure. PCA of orders 2, 3 and 4 yields DTFs which exhibit an increasing amount of spectral detail, whereas DTFs produced by PCA of orders 5, 6 and 7 are almost identical, and are quite close to the original data. These trends were observed for all subjects.

For all orders, principal component analysis is seen to smooth out the roughness and deep notches present in the original data. However, despite the fact that PCA for orders 5, 6 and 7 contains over 92% of the variance of the original data, such analysis does not always faithfully preserve the features observed in this data. A close comparison of the mesh plot for the PCA analysis of order 7 (Figure 6.7 (b)) and the original data (Figure 6.4 (a)) reveals that features of the original data can sometimes be absent or somewhat displaced in the reconstructed PCA data. This point is explored further in Section 6.3.1. Furthermore as PCA smoothing does not distinguish between high frequency data and low frequency data; the same degree of smoothing is introduced at all frequencies. However, as pointed out in Section 6.2, it is desirable to introduce more smoothing at high frequencies than at low. A more psychoacoustically motivated method of smoothing than PCA is therefore required to improve matters. The following section describes the use of the EWD for this purpose, and presents convincing evidence of its applicability to the task of HRTF smoothing.

6.3 EWD Derived HRTF Smoothing

It was demonstrated in Chapters 4 and 5 that the EWD marginals could be used to generate excitation patterns of stimuli. Such excitation patterns reflect only unmasked spectral detail. Therefore, EWD derived HRTF smoothing was performed by calculating the EWDs of individual HRIRs and summing the resulting time-frequency distribution along the time axis. Thus, the excitation pattern of each HRIR, as generated by the EWD, was effectively used as a smoothed HRTF. Since excitation patterns simulate the spectral masking of the auditory system, it was expected that such EWD-smoothed HRTFs would exhibit far less inter subject variability because spectral detail not resolvable by the ear would be removed. This is possible because, although there is a great deal of variability in our physical makeup, and especially in our pinna structure, there is a high degree of agreement between the auditory filter shapes observed in subjects with normal, unimpaired hearing. In other words, spectral masking shows relatively little inter-subject variability, except for

the hearing impaired. In contrast to PCA smoothing, EWD smoothing is based on psychoacoustics and therefore reflects the varying amount of spectral masking with frequency. It was expected that such processing would be able to retain perceptually relevant detail and reject or smooth irrelevant (masked) features such as the severe high frequency notches observed in many raw HRTF measurements.

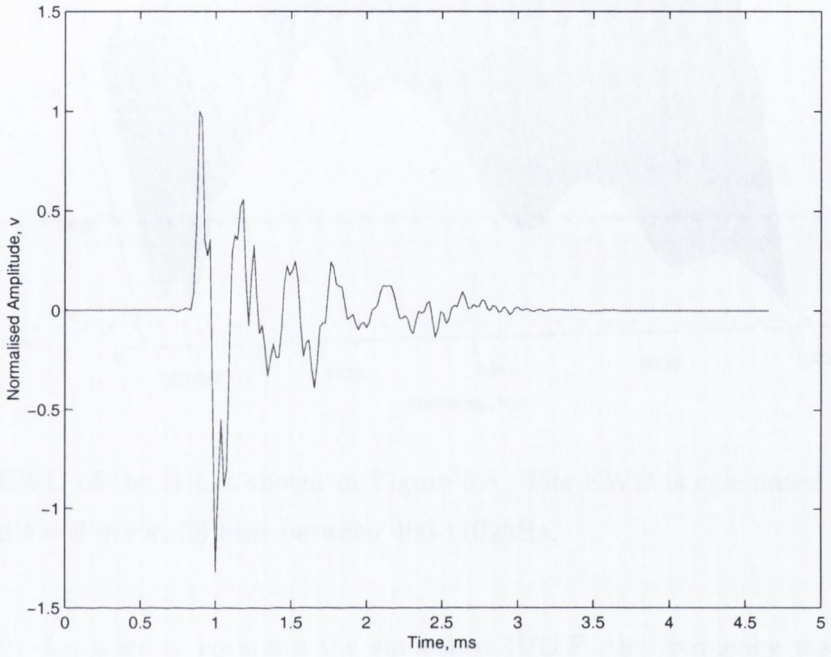


Figure 6.8: A typical HRIR measurement (subject SOU).

In order that a fair comparison be made between PCA and EWD smoothing, exactly the same data used by Walsh to investigate PCA smoothing for orders 1 to 7 was used to investigate EWD smoothing. Therefore a total of 360 HRIRs corresponding to the left and right ears of the 5 subjects AFW, SOW, SOU, SOS and SJX were analysed using the EWD. Figure 6.8 plots one of these HRIR measurements. The duration of the HRIR is small (about 2-3 ms) compared with the duration of the temporal window model of temporal masking; this the case for all HRIRs in the database[2]. Therefore, apart from gross changes in level, the resulting EWDs change very little with time and need only be calculated at a few points in time in order to generate an accurate HRIR excitation pattern.

Figure 6.9 illustrates the EWD for this HRIR, calculated at times 0, 3 and 6 ms, and at the frequencies computed for PCA analysis by Walsh described above (58 bins between

[2]Note that the delay before the onset of the HRIR is important for determining ITD information, which is typically computed by measuring the time lag to the peak of the cross correlation of the left and right HRIRs for a given azimuth.

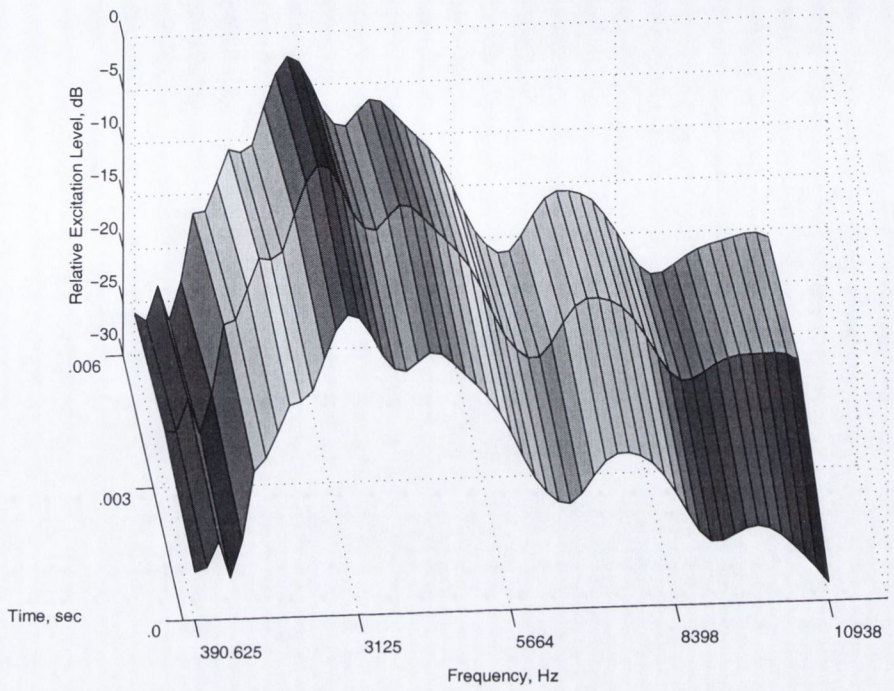
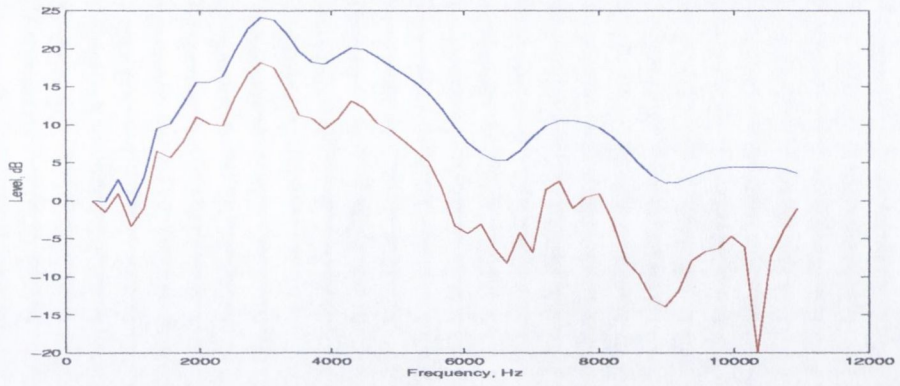


Figure 6.9: EWD of the HRIR shown in Figure 6.8. The EWD is calculated at $t = 0$ ms, $t = 3$ ms and $t = 6$ ms at 58 bins between 400-11025Hz.

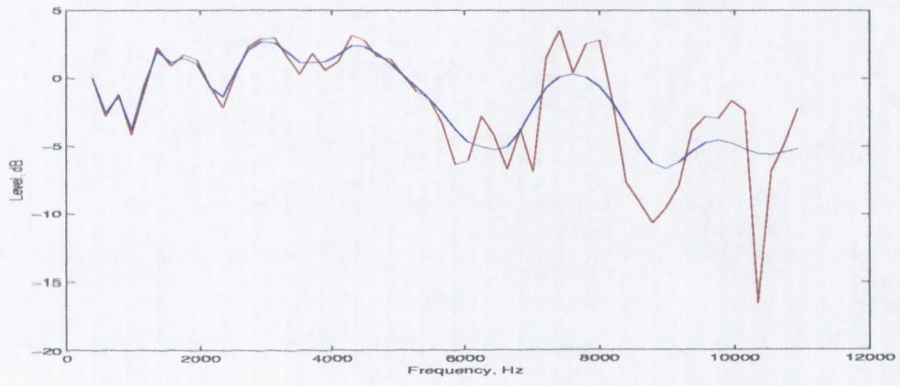
400-11025Hz). In order to compute the smoothed HRTF, the frequency marginal of the EWD was computed. Figure 6.10 (a) compares this smoothed HRTF (blue curve) with the HRTF of the original data (red curve) over the range of frequencies used in the PCA analysis of Walsh. Note that, as for PCA, EWD smoothing discards the phase information contained in the HRIR. It is clear from the figure that the EWD introduces more smoothing at high frequencies. However the widening bandwidth of the auditory filters also introduces a high frequency bias. Although this bias is an accurate reflection of auditory spectral resolution, it is undesirable for the purposes of HRTF smoothing. In order to remedy this, the across average means of the smoothed data were removed to yield smoothed DTFs. This removed all direction independent spectral detail, including the effects of the ear canal resonance and also the high frequency bias introduced by the EWD smoothing process. To reconstruct smoothed HRTFs, the across average means of the *original* data may be added to the smoothed DTFs[3].

Figure 6.10 (b) illustrates the DTFs of both original and EWD-smoothed data. The

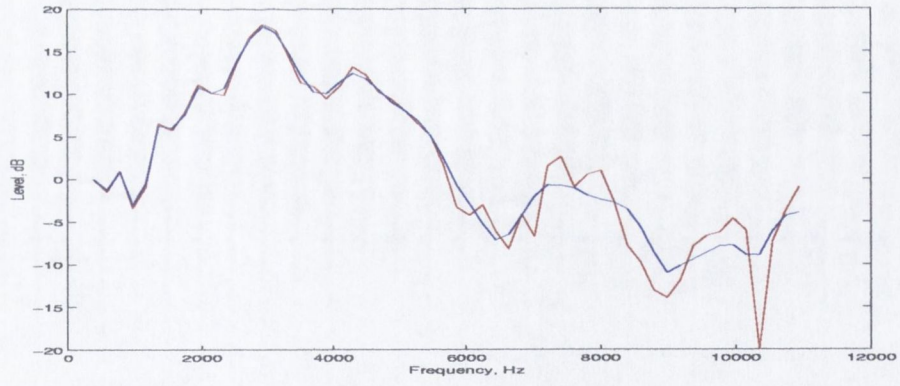
[3]Only loudspeaker presentation was considered in this study. For such presentation, the across average means do not have to be re-inserted since the loudspeaker filter functions consist of rational functions expressed in terms of HRTFs, with an equal number of HRTF terms in both numerator and denominator. Therefore, the across azimuth means cancel in such expressions.



(a)



(b)



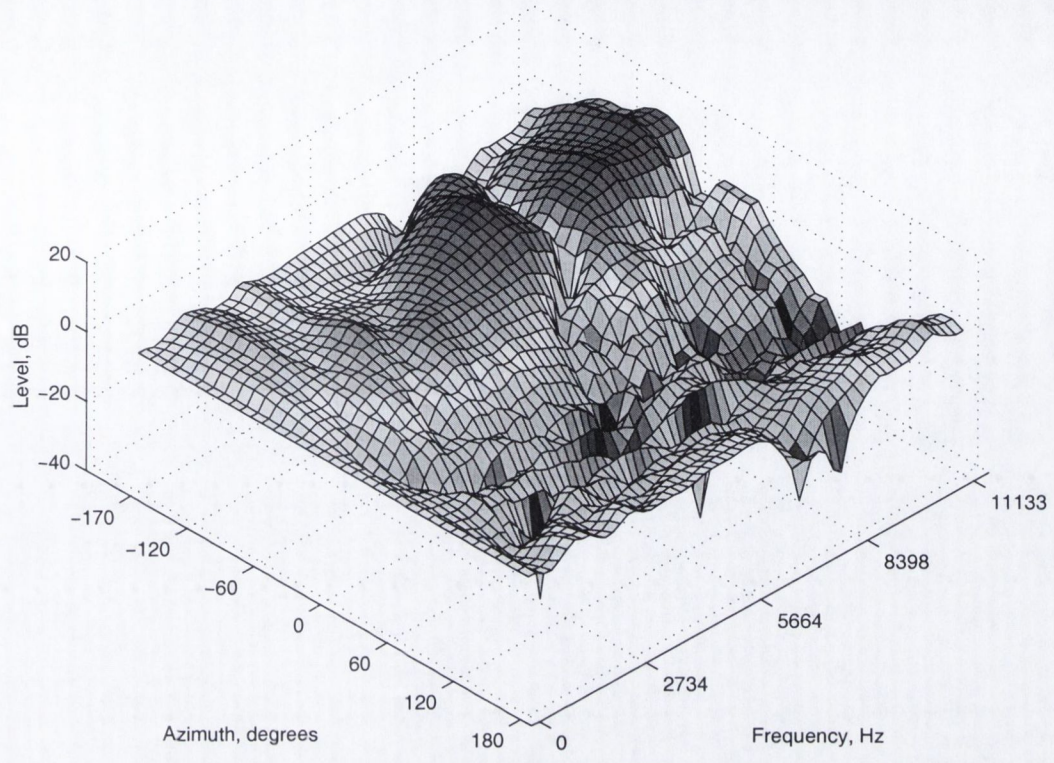
(c)

Figure 6.10: (a) A comparison of the original HRTF response (red curve) and the EWD-smoothed HRTF response (blue curve) illustrating the high frequency bias introduced by EWD smoothing at high frequencies. (b) A comparison of the corresponding DTFs, with the original DTF (red) and the smoothed DTF (blue). (c) A comparison of the unsmoothed HRTF (red) and the smoothed DTF, with the AAM information re-inserted (blue).

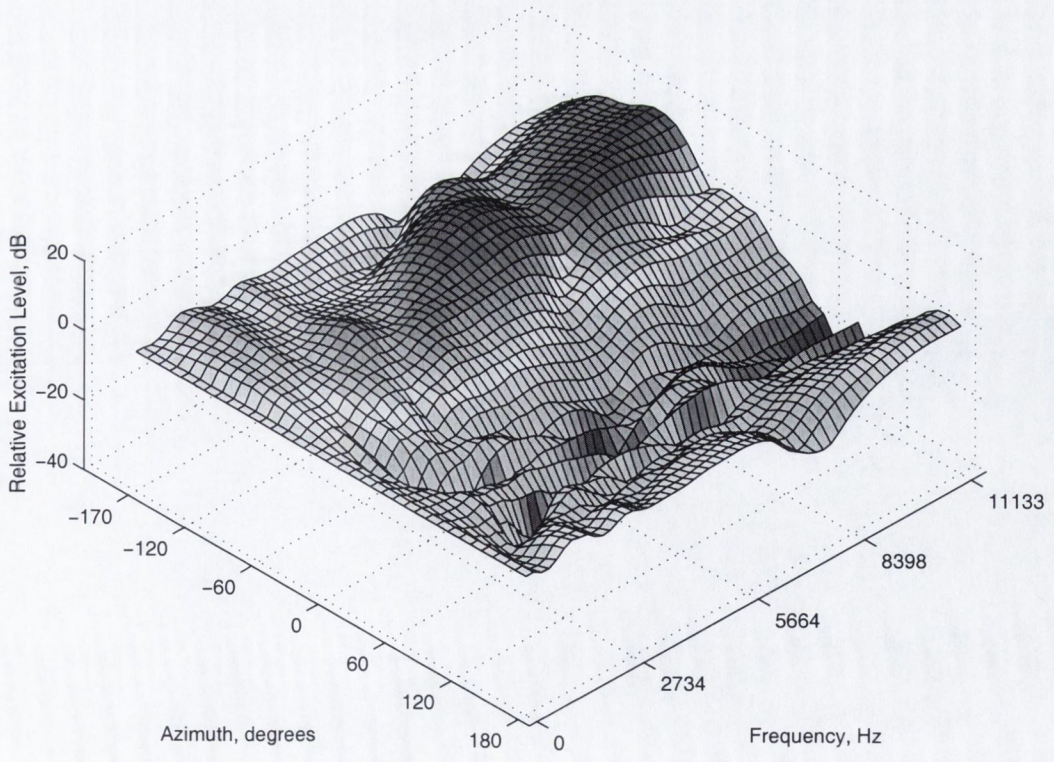
smoothing performed by the EWD is more easily seen here. It is clear that for frequencies below about 5 kHz, the smoothed data follows the original data very closely, only smoothing out 'roughness' in the data. Above this frequency the original data becomes a lot more irregular, displaying some sharp resonances and notches. Such irregularity at high frequencies was noted for much of the Waisman Center HRTF data. For frequencies between 5 kHz and 9 kHz, EWD smoothing is seen to reduce the severity of the perturbations in the data, while still following its overall trend, giving a smooth dip at around 6.5 kHz and 9 kHz and a smooth peak at 7.5 kHz. For frequencies above about 9 kHz, increased smoothing ensures that the very severe (and very unresolvable !) notch at 10.5 kHz is effectively eliminated.

For completeness, Figure 6.10 (c) compares the original HRTF data and the EWD-smoothed DTF with the AAM re-inserted. It is clear that the bias introduced by EWD analysis has effectively been eliminated, as has the high frequency roughness and notching. However, the re-insertion of the AAM information is not necessary for loudspeaker presentation, as shall be explained in Section 6.3.2. Finally, Figure 6.11 presents a comparison of the unsmoothed DTFs and EWD-smoothed DTFs for the same ear as illustrated for PCA analysis of orders 1 to 7 in Figures 6.4 to 6.7. It is very clear that the EWD-smoothed data follows the major spectral trends in the original DTF data while still removing roughness and notching at high frequencies.

The HRTF critical-band smoothing method described here depends entirely on the frequency marginal property of the EWD. One might therefore ask whether or not there was any advantage in using the EWD over a more conventional method of generating excitation patterns, such as Moore and Glasberg's method, described in Chapter 3. In this instance EWD analysis showed that the HRIR data did not change significantly with time. Therefore, the EWD smoothed HRTFs were accurately generated by averaging slices in frequency at only 3 points in time. For the particular case of EWD HRIR analysis presented here, the argument in favour of a more conventional method of critical band smoothing such as Glasberg and Moore's method is a valid one. However, as discussed in Chapter 8, the audible effects of filter functions such as audio equalisation filters can rarely be estimated from their spectra alone. Transient effects of equalisation impulse responses are known to have a surprisingly audible effect [61]. The audibility of such effects are strongly influenced by both spectral and temporal masking. Therefore, an investigation of the HRIR data using the EWD was very worthwhile since both temporal and spectral masking are incorporated. The fact that the HRIR EWDs did not change significantly with time adds support the assertion of Wightman and Kistler noted in Section 6.2.1 that the frequency dependance of the interaural time delay of HRIR data is not perceptually relevant. This assertion is



(a)



(b)

Figure 6.11: (a) Raw DTF data and (b) EWD-smoothed data for the left ear of subject SOW.

important because it allows a given left-right HRTF pair to be modelled by their minimum phase equivalents plus an interaural pure time delay (Equation 6.1). This fact is the basis of their reconstruction of HRTF data using PCA and is the basis for the EWD and PCA smoothing described here and in [182].

A detailed examination of virtual auditory display systems later in this chapter highlights the audibility of even minor roughness in HRTF data. This reinforces the importance of EWD smoothness, dynamic range and lack of clutter. Overall, the chapter highlights the utility of the spectral masking property of the EWD. Chapter 8 correspondingly demonstrates the usefulness of the EWD's temporal masking property. Together, the two application areas provide ample evidence of the general applicablilty of the EWD to audio TF analysis and provide a background for several further EWD application areas suggested in Chapter 8.

6.3.1 Empirical Comparison of PCA and EWD Smoothing

It was demonstrated in the previous section that of the seven PCA orders tested, only orders 2 to 4 produced a degree of HRTF smoothing appropriate for the design of a more generic head model than the raw HRTF data. Order 1 PCA had a pleasant timbre but did not contain enough high frequency detail for accurate localisation. In contrast, orders 5 to 7 exhibited reasonable localisation but did not smooth high frequency detail sufficiently to achieve a pleasant timbre.

Results of subjective listening tests carried out by Walsh in Trinity College Dublin and also those presented here suggest that the performance of PCA-smoothed HRTFs derived from each of the 5 subjects was similar for orders 1 and 2. Also, informal listening tests revealed that as the number of PCs was increased beyond this point, listeners tended to prefer HRTFs derived from a single subject, with the best performance being achieved for either 3 or 4 PCs. So, for example, one listener might prefer the HRTFs derived from the first 4 principal components of SOU, while another might prefer the HRTFs derived from the first 3 principal components of AFW. Overall, order 3 PCA was found to apply the most appropriate degree of smoothing.

In order, then, to compare PCA smoothing with EWD smoothing, Figures 6.12 to 6.17 compare the DTFs of the original, EWD-smoothed and PCA-smoothed (order 3) data. In each figure, the upper panels plot the original DTFs, the middle panels EWD-smoothed DTFs and the lower panels plot the order-3 PCA-smoothed DTFs. Each figure plots the data at two particular azimuths for the right ears of all 5 subjects. The azimuth starts at 0^o for Figure 6.12 and increases in steps of 30^o, reaching 330^o for Figure 6.17. The

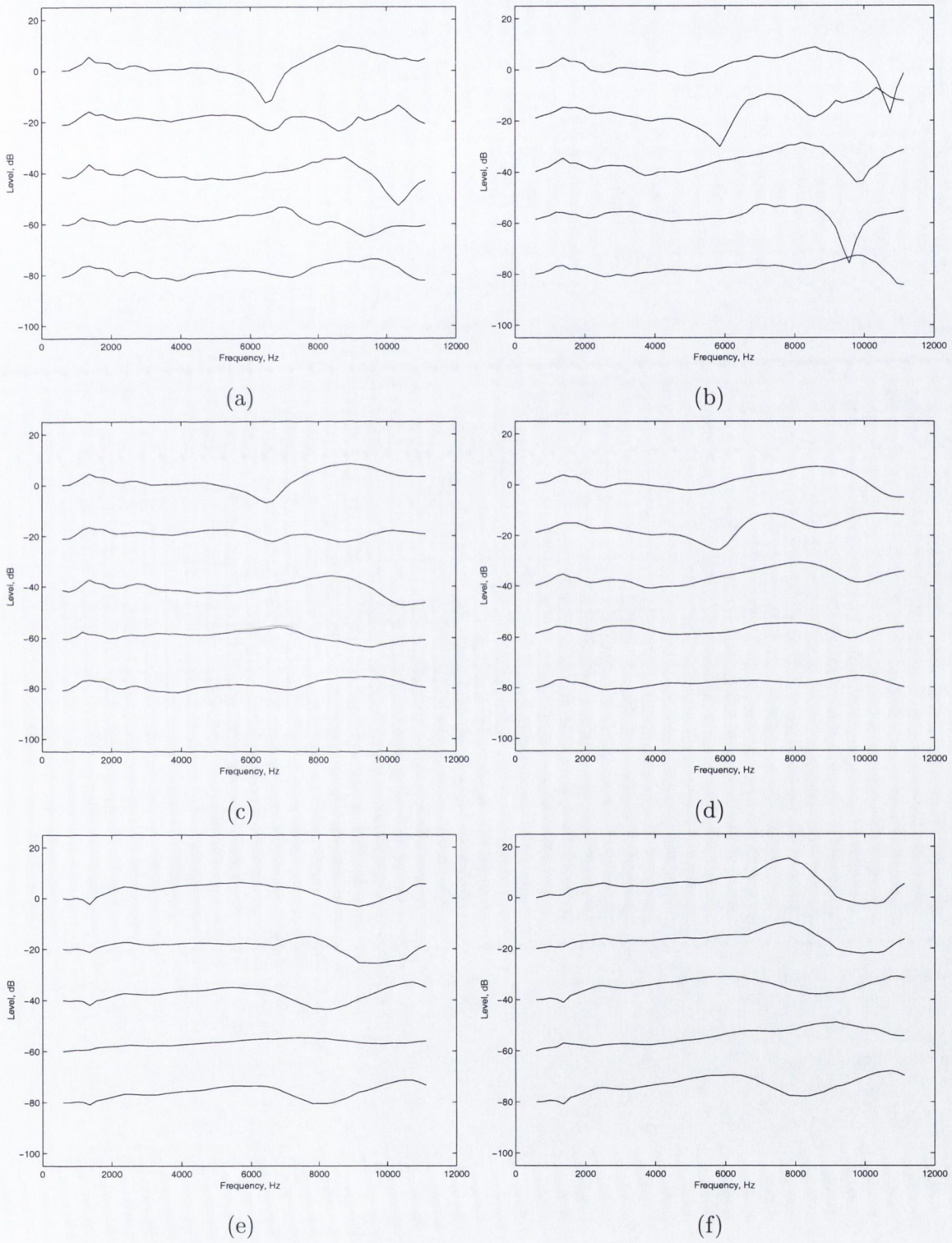


Figure 6.12: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 0^o and 30^o respectively.

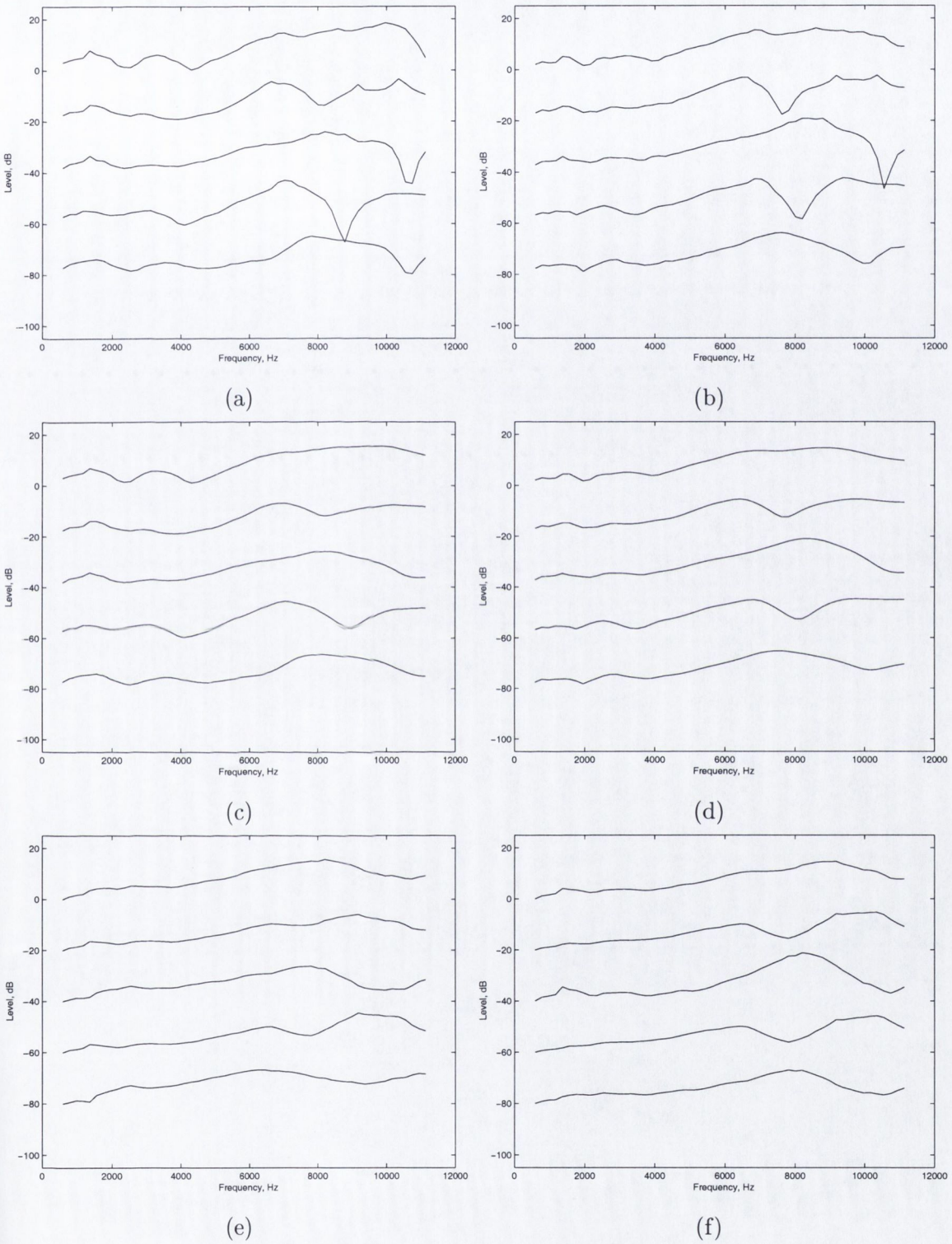


Figure 6.13: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 60^o and 90^o respectively.

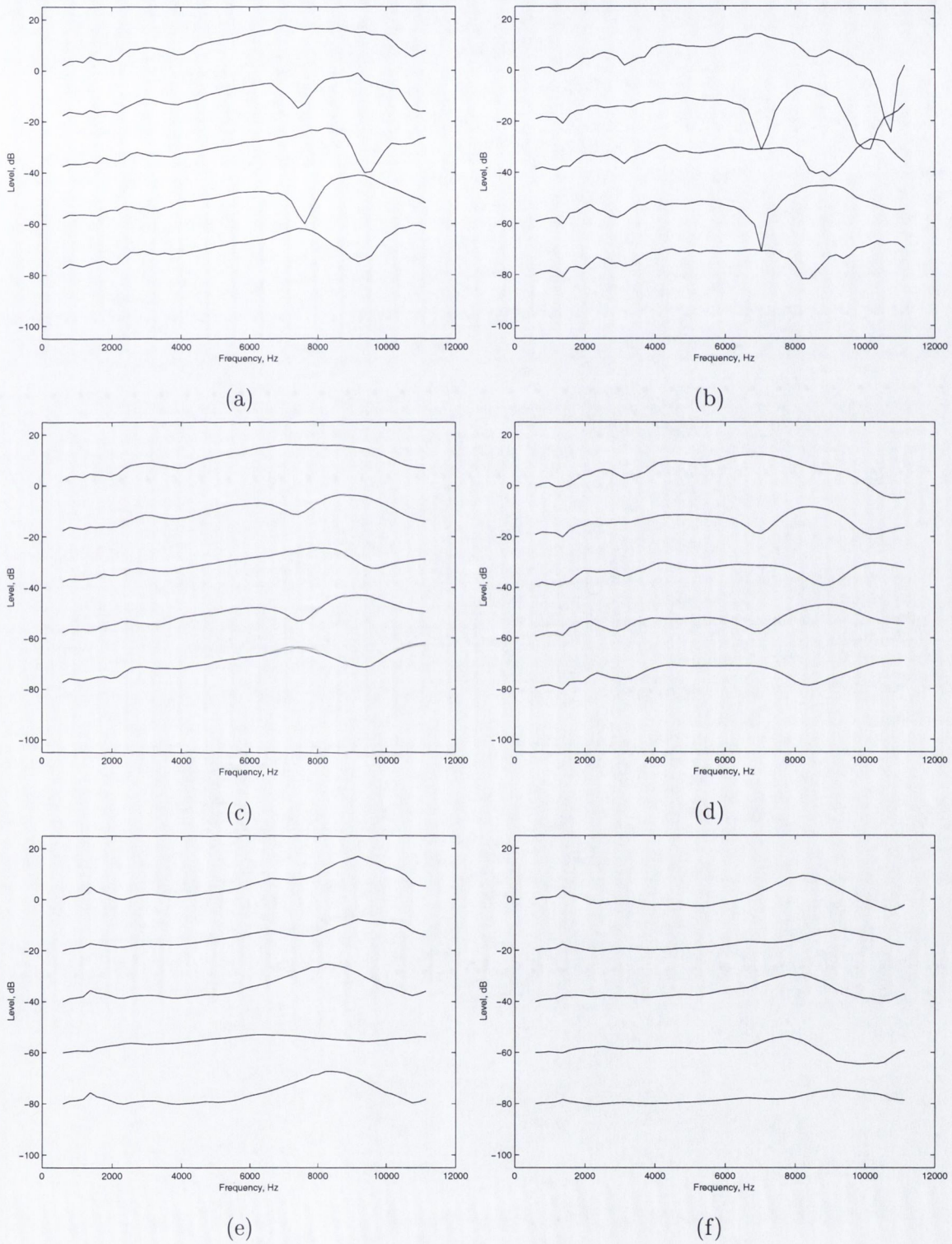


Figure 6.14: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 120^o and 150^o respectively.

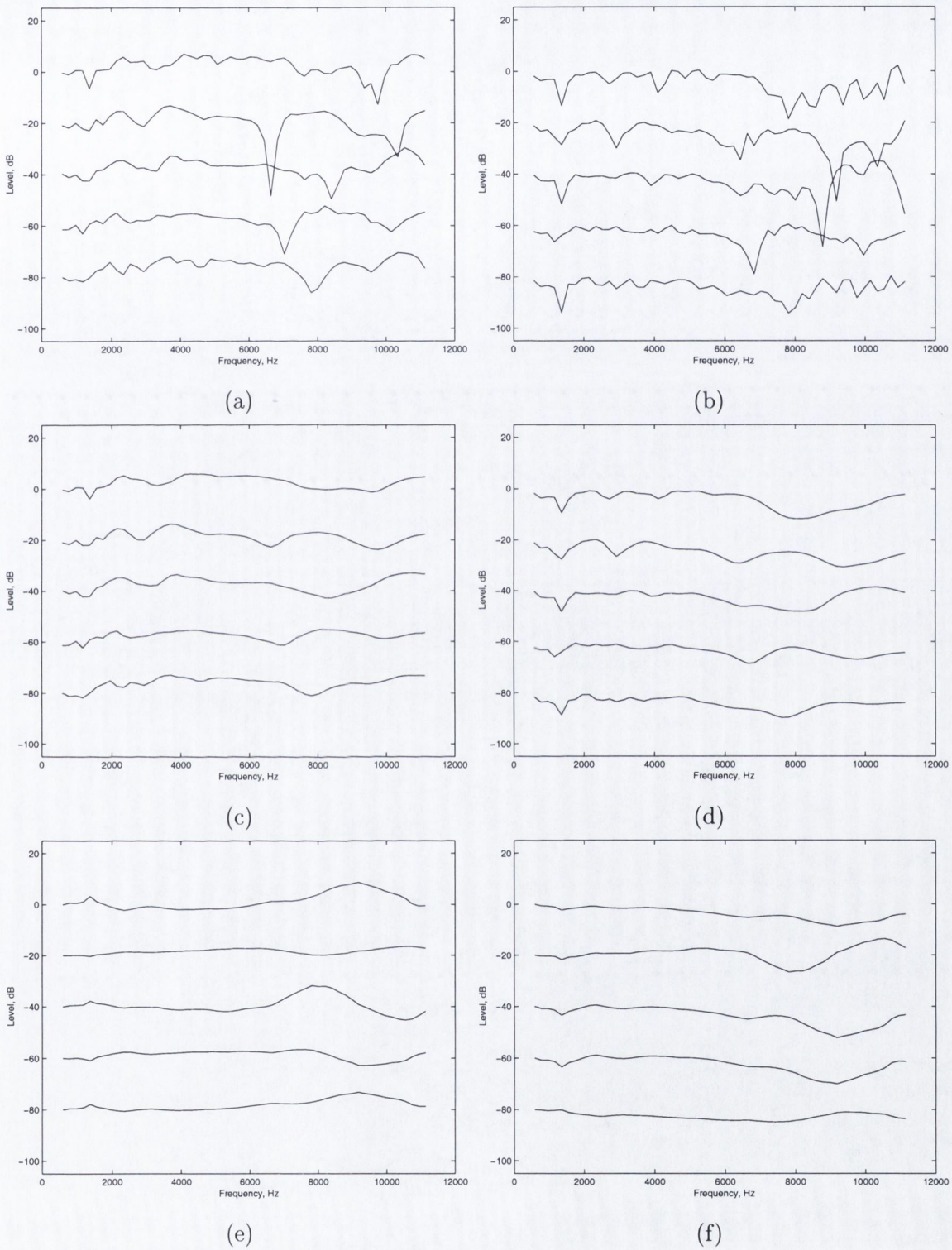


Figure 6.15: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 180^o and 210^o respectively.

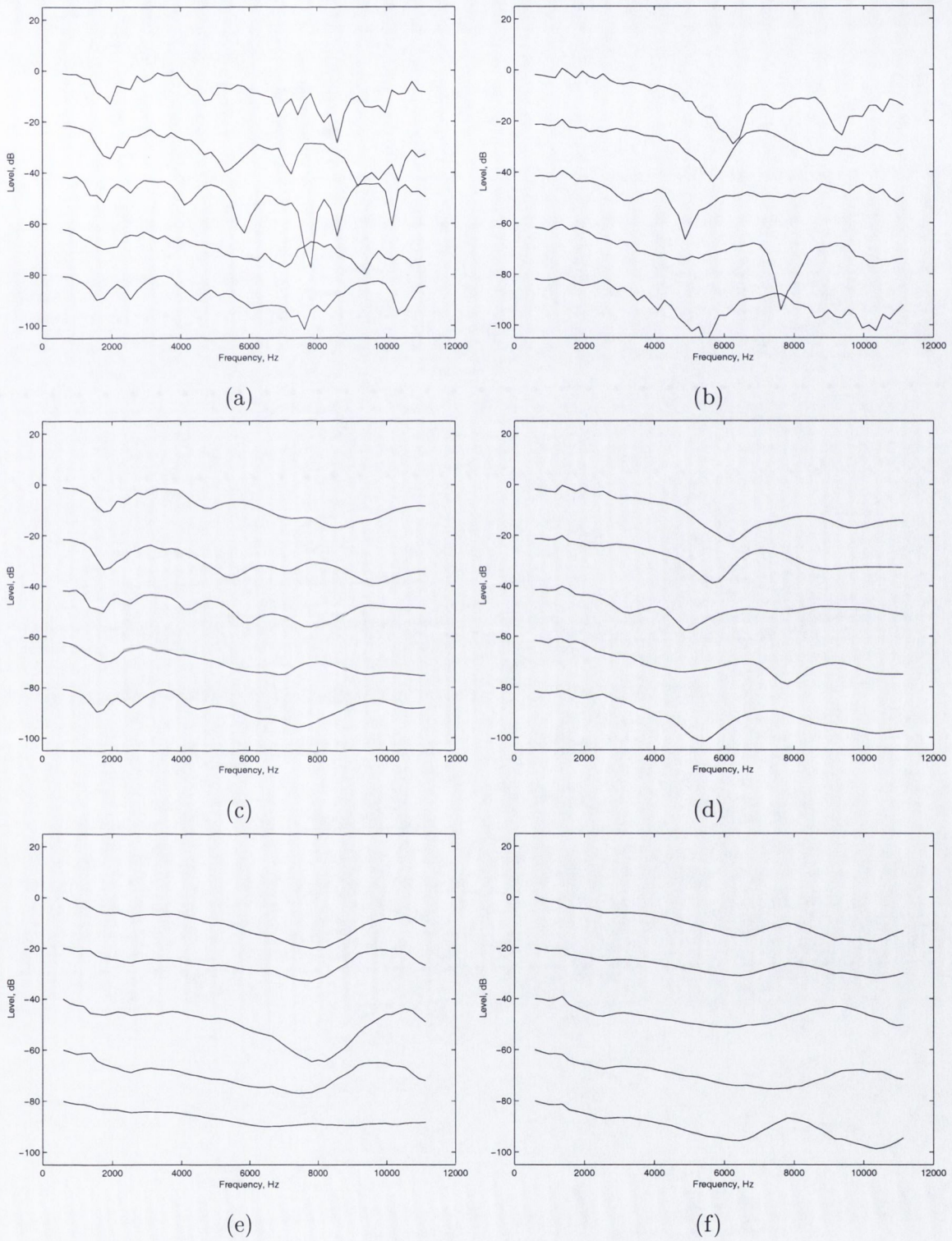


Figure 6.16: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 240^o and 270^o respectively.

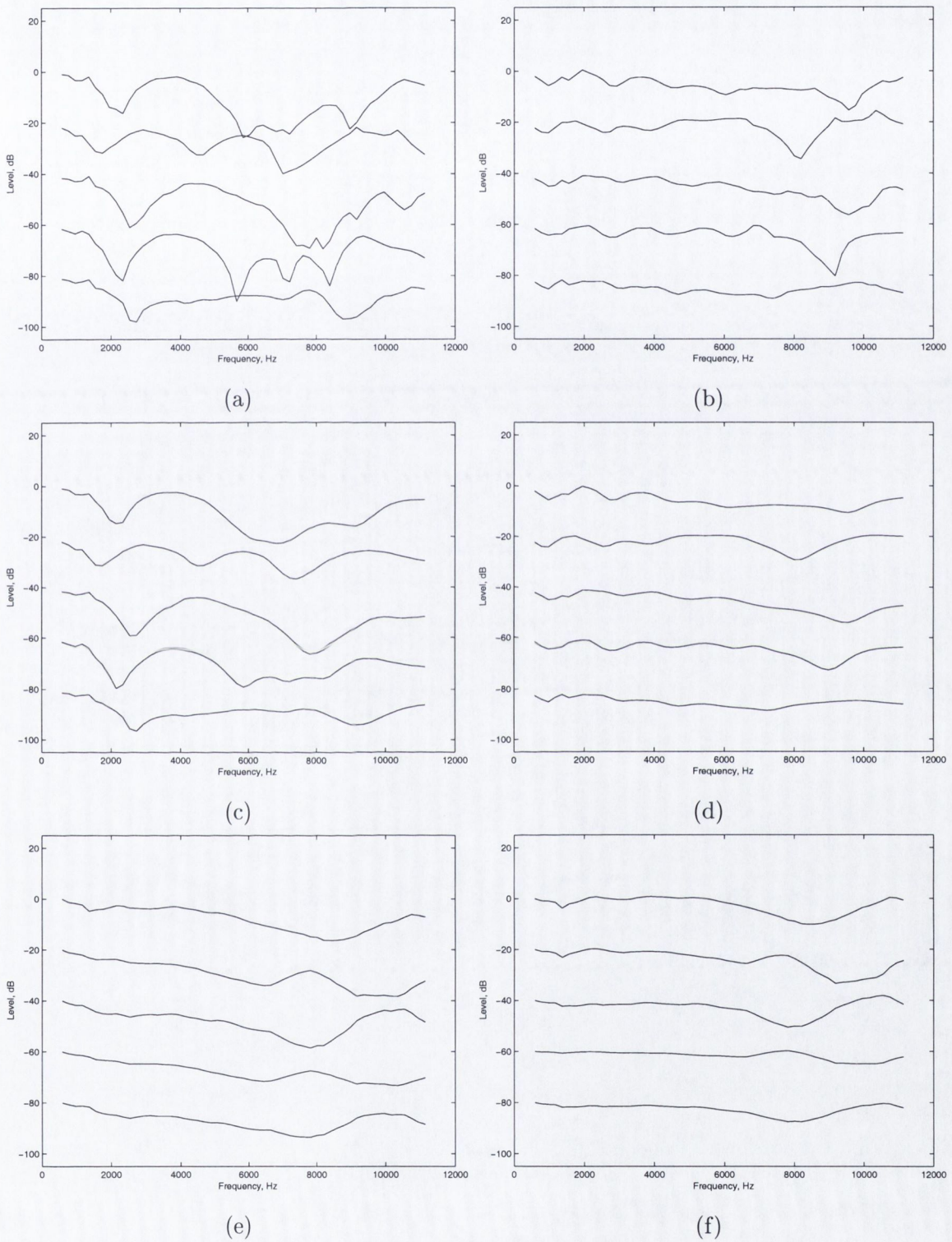


Figure 6.17: A comparison of the unsmoothed (top panels), EWD-smoothed (middle panels) and (order-3) PCA-smoothed (bottom panels) DTF magnitudes for the right ears of all 5 subjects. The left and right panels present the data for azimuths of 300^o and 330^o respectively.

HRTF data is offset by 20 dB for clarity, the order of the subjects from top to bottom being AFW, SOW, SOU, SOS and SJX. The superiority of EWD smoothing over PCA smoothing is very evident from these plots. EWD analysis introduces sufficient smoothing at high frequencies to remove perturbations and notches in the data above about 5 kHz in all cases. This also has the desirable effect of reducing the inter-subject specificity of the data. For lower frequencies, for which the original data is less listener specific, only enough smoothing is applied to remove roughness. In contrast, the PCA data shows too little low frequency detail. Furthermore, PCA introduces some features not present in the original data, especially for higher frequencies. PCA is seen to only really reflect overall trends in the level changes of the original HRTF data with azimuth, and does not really preserve more detailed spectral cues. Overall, EWD smoothing can therefore be deemed to be a much more effective method of reducing inter-subject variability in HRTF measurements then PCA smoothing as EWD smoothing removes only perceptually redundant spectral detail without sacrificing larger scale HRTF features.

6.3.2 Subjective Comparison of PCA and EWD Smoothing

In order to carry out a subjective comparison of PCA and EWD smoothing, the same experimental setup used by Walsh for the subjective evaluation of PCA using a two-channel, two-loudspeaker model was reused in this work. Figure 6.18 presents a schematic of the two-channel, two-loudspeaker model for a virtual source located to the right side of the user at an azimuth of θ_s. The HRTFs associated with a real source subtending an angle θ_s to the listener are given by $HRTF_L(\theta_s)$ and $HRTF_R(\theta_s)$ for the left and right ears respectively. Those HRTFs associated with the left and right loudspeaker positions are denoted $HRTF_L(LF)$ and $HRTF_L(RF)$ for the left ear and $HRTF_R(LF)$ and $HRTF_R(RF)$ for the right ear. The purpose of the β_{2spk} and $[\alpha/\beta]_{2spk}$ filters is to process the input signal so that the resulting speaker outputs produce the illusion of a sound source at θ_s, i.e. a source filtered by $HRTF_L(\theta_s)$ for the left ear and by $HRTF_R(\theta_s)$ for the right. Therefore, an analysis of the configuration for the left and right ears, respectively, gives

$$\beta_{2spk}\frac{\alpha_{2spk}}{\beta_{2spk}}HRTF_L(LF) + \beta_{2spk}HRTF_L(RF) = HRTF_L(\theta_s) \tag{6.2}$$

$$\beta_{2spk}\frac{\alpha_{2spk}}{\beta_{2spk}}HRTF_R(LF) + \beta_{2spk}HRTF_R(RF) = HRTF_R(\theta_s)\ . \tag{6.3}$$

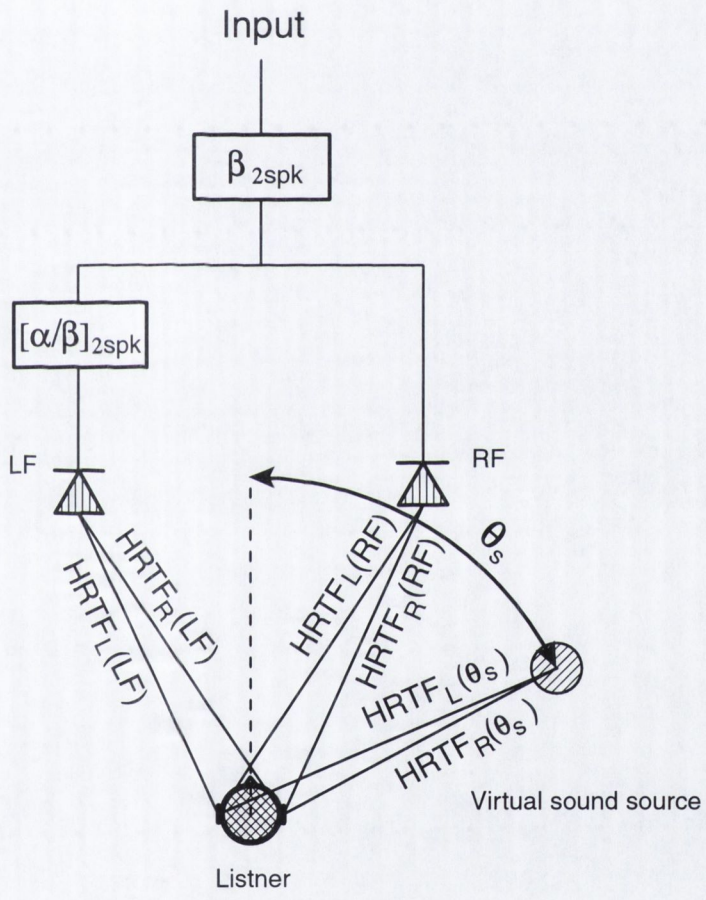


Figure 6.18: Virtual auditory display for a two loudspeaker setup.

Solving the above for $[\alpha/\beta]_{2spk}$ and β_{2spk} yields

$$[\alpha/\beta]_{2spk} = \frac{HRTF_L(\theta_s)HRTF_R(RF) - HRTF_R(\theta_s)HRTF_L(RF)}{HRTF_R(\theta_s)HRTF_L(LF) - HRTF_L(\theta_s)HRTF_R(LF)} \tag{6.4}$$

$$\beta_{2spk} = \frac{HRTF_R(\theta_s)HRTF_L(LF) - HRTF_L(\theta_s)HRTF_R(LF)}{HRTF_R(RF)HRTF_L(LF) - HRTF_L(RF)HRTF_R(LF)} . \tag{6.5}$$

In this configuration, the β_{2spk} filter is shared by both speakers; it is responsible for equalising the virtual sound source. The $[\alpha/\beta]_{2spk}$ filter provides the temporal localisation cues. For images located at azimuths from 0^o to 180^o, the $[\alpha/\beta]_{2spk}$ is causal and therefore realisable because the left signal must lag the right signal. However, outside this range the situation is reversed, requiring a non causal $[\alpha/\beta]_{2spk}$ filter. This can be remedied if we assume that the head is symmetrical enough to allow the approximations :

$$HRTF_L(\theta_s) = HRTF_R(-\theta_s) \quad \text{and} \tag{6.6}$$

$$HRTF_R(\theta_s) = HRTF_L(-\theta_s) . \tag{6.7}$$

These assumptions are indeed valid under most circumstances; we can therefore produce images on the listener's left hand side by simply swapping the loudspeaker feeds [182]. As outlined in Section 6.2.1, both PCA and EWD analysis discard the ITD information contained in the HRIRs. However, using the minimum-phase approximations of Equations 6.1, this information can be reinserted resulting in β_{2spk} and $[\alpha/\beta]_{2spk}$ equations for 2-speaker panning. Since these equations are lengthy they are not included here; the interested reader is instead referred to [182]. These exact filter specifications, used by Walsh for PCA evaluation, were used to generate the EWD β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications. These were then used to design 32nd order IIR filters using a filter fitting technique based on Prony's method [97]. This technique achieved very good fits for all the EWD β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications.

Figures 6.19 (a) and (b) present a typical example of such a fit to the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications respectively, where the desired magnitude and phase responses are plotted as red curves, the IIR fits as blue curves and the difference as green curves. The fit is seen to be very accurate, a difference of less than 1 dB for filter magnitude and 0.25 radians for filter phase being typically observed for all but the extremes of the frequency range.

Figure 6.20 portrays the experimental setup used for comparison and evaluation of PCA and EWD imaging. The IIR filters generated above were implemented on a DSP32C DSP board, linked to a PC. One channel of the analog output of a Marantz CD player acted as the signal source. This signal was digitized using a Burr Brown ZPD 1001 Dual Channel 16

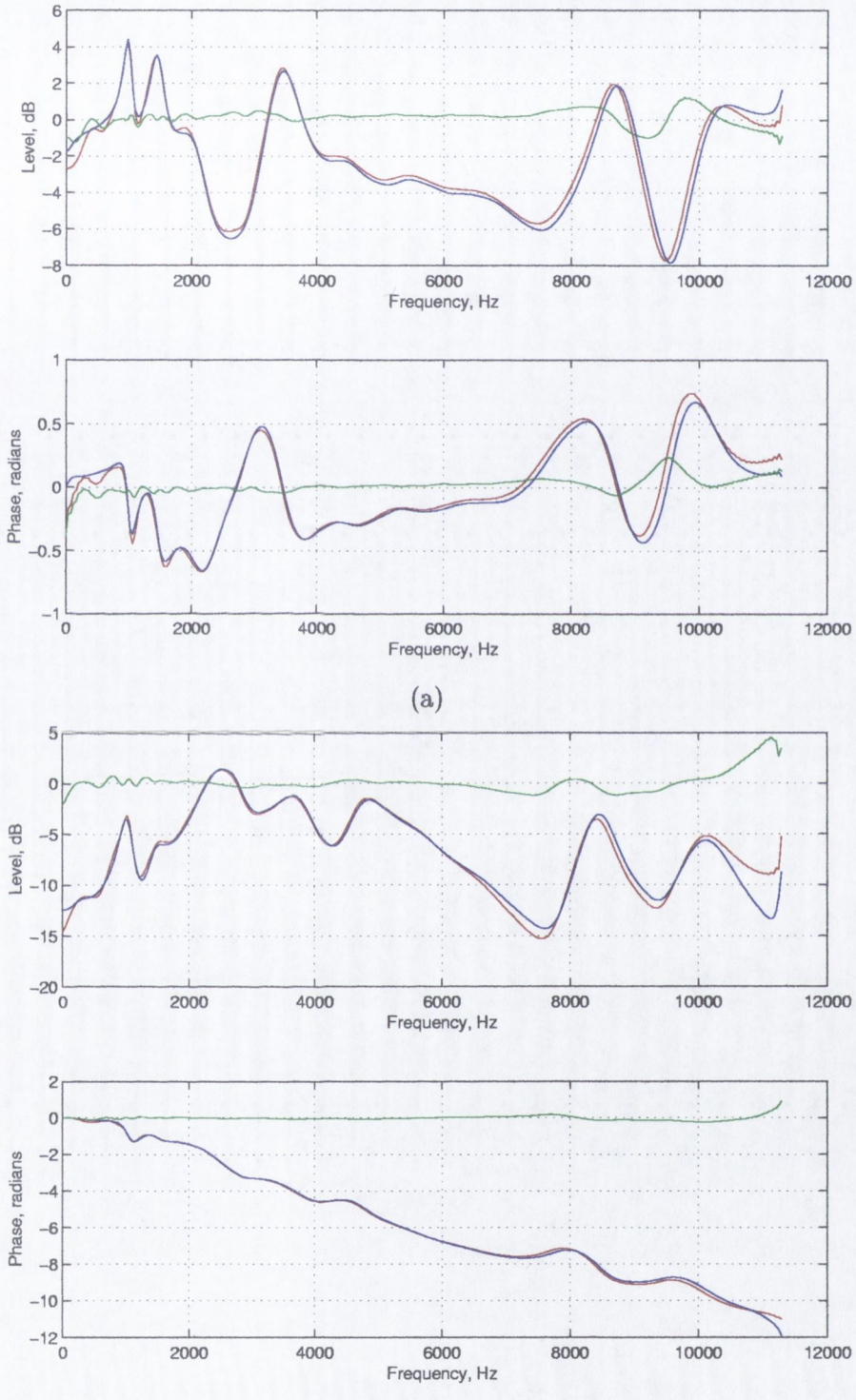


Figure 6.19: (a) A comparison of the desired β_{2spk} filter magnitude and phase response (red curves) and the IIR fits achieved by the filter fitting process (blue curves). The green curves plot the difference. (b) The same scenario, now presented for the $[\alpha/\beta]_{2spk}$ filter.

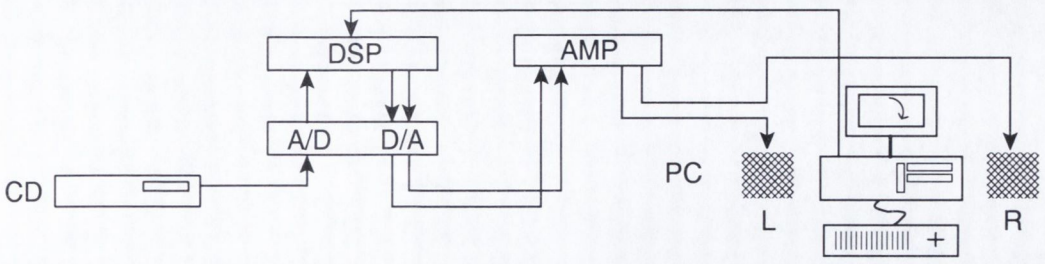


Figure 6.20: Experimental setup used for comparison of PCA and EWD-smoothed HRTFs.

bit D/A and A/D, with the sampling frequency set to 22050 Hz. The output of the A/D was used as input to the DSP board, which performed the β_{2spk} and $[\alpha/\beta]_{2spk}$ filtering and output the results using both channels. These were then converted back to analog format using the afore-mentioned Burr Brown D/A and fed to an Arcam Delta 60 stereo amplifier. The outputs of this amplifier were used as speaker feeds to two JPW Mini Monitors, which were placed to subtend an angle of 30^o at the users ears. The filtering performed by the DSP was controlled by the user via a graphical user interface and a keyboard; the user could adjust the panning angle in steps of 10^o between 0^o and 350^o. The overall experimental setup thus used for subjective listening tests was very close to that used by Walsh, the only difference being the choice of loudspeakers and CD player.

In the subjective assessment of PCA carried out by Walsh, the performance of all five of the HRTF sets for all seven PCA orders was judged in terms of imaging naturalness and imaging accuracy. This was carried out informally by a small number of listeners. There were a of 35 PCA models in all. For good imaging naturalness, a balanced and pleasant timbre was required for each azimuth position. Imaging accuracy referred to the ability of the PCA model to 'trick the ears' into localising the virtual sound source at the required azimuth. As stated earlier, optimum PCA performance was usually experienced by listeners for one or two principle component 'heads', presumably because these heads were similar to the listener's own HRTFs. Particular attention was therefore given by Walsh to the seven PCA models derived from subject SOS, as he found this subject to be judged the most agreeable.

Walsh tested the virtual imaging accuracy of the model derived from SOS's HRTFs using both lowpass and high pass noise of varying bandwidths, with the noise panned to a virtual angle of 90^o. For the lowpass noise stimuli, no spectral smearing[4] was noted for

[4]Spectral smearing occurs when the high frequency components of a virtual image begin to move away

the model derived from one principal component up to a noise bandwidth of 3 kHz. This bandwidth increased with increasing number of principal components up to a limit of 4.5 kHz for the models derived from four to seven principal components. The results of lowpass analysis demonstrated that low order PCA yields models which only pan low frequencies well, higher order PCA being required to pan higher frequencies. For the highpass noise stimuli, no virtual imaging was achieved beyond the loudspeaker arc with a highpass cutoff greater than 2 kHz for the model derived from one PC. This increased to 2.5 kHz for two to four PCs and 3 kHz for models derived from five to seven PCs. The results for the highpass noise signal indicate that the presence of low frequency information was very important for effective virtual imaging, since these frequencies were the most convincingly panned. Also, high frequency imaging improved with increasing PCA order.

An assessment of imaging naturalness for noise, music and speech stimuli carried out by Walsh showed that the model derived from one PC generated images that sounded very natural for all azimuths. As the number of PCs was increased however, imaging naturalness decreased. The author adds that for high order PCA, abrupt changes in timbre were often experienced between adjacent panning positions. Walsh notes that the imaging naturalness of models based on a *particular* subject for five to seven PCs was almost identical. However, the author notes that there was much variability in imaging naturalness *across* subjects for the same number of PCs. Therefore, the inter-subject variability inherent in HRTF measurements was not necessarily reduced by PCA analysis.

The performance of the PCA and EWD models were estimated for all orders. Due to time constraints this was carried out only by the author and was not conducted blind. In this comparison the EWD and PCA models were simply judged in terms of image naturalness and image localisation on a scale from 1 to 10. A variety of styles of music were used as stimuli. Since these results reflect the findings of the author only, they are liable to bias and should only be considered a pilot study. Table 6.3 presents these results.

From Table 6.3 we can see that imaging naturalness was found to be excellent for the EWD model; EWD smoothing was deemed to sound at least as natural as all PCA orders for each of the subjects. This was an encouraging and somewhat surprising result as it was not expected that EWD smoothing would sound as natural as PCA analysis for low orders. Furthermore, there was less variation in imaging naturalness across subjects for EWD smoothing than for PCA smoothing. The plots of Figure 6.12 to 6.17 support

from the desired panning azimuth and towards the closest loudspeaker. This means that some of the signal's frequency components appear to emanate from the loudspeaker and the remainder from the desired azimuth resulting in a split or smeared virtual image.

	PCA 1		PCA 2		PCA 3		PCA 4		PCA 5		PCA 6		PCA 7		EWD	
	L	N	L	N	L	N	L	N	L	N	L	N	L	N	L	N
AFW	5	7	6	5	6	5	6	5	6	5	7	5	7	6	7	8
SOW	4	8	5	5	5	6	4	6	5	6	7	6	7	6	7	8
SOU	4	7	6	5	5	6	6	7	6	6	6	5	6	5	8	8
SOS	5	8	5	6	5	4	5	6	6	6	7	4	7	5	9	9
SJX	4	7	5	5	5	6	5	7	5	5	6	5	6	5	8	8

Table 6.3: Subjective comparison of PCA and EWD models for all orders. 'L' denotes localisation and 'N' denotes naturalness.

the theory that EWD smoothing was successful at reducing inter-subject variability in HRTF measurements. Furthermore, the abrupt changes in timbre between adjacent azimuth positions noticed for the PCA models was much alleviated in the EWD models. A detailed comparison of the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications for PCA and EWD models revealed that the EWD filter specifications contained comparable detail to those for the mid to upper PCA orders. However, they were in general more smooth and did not suffer from peaking and notching to the extent that the PCA specifications did. A more detailed examination of the underlying causes of such artifacts and their consequences on virtual imaging filter specification is carried out in Section 6.3.3.

Again, due to time constraints, a detailed investigation of the imaging accuracy performance of the EWD model using the lowpass and highpass noise stimuli described above for the PCA analysis conducted by Walsh could not be performed. However, a subjective comparison of PCA and EWD imaging using musical stimuli was performed. Results of this study, presented in Table 6.3 show that EWD imaging accuracy was better at all azimuths than PCA imaging accuracy for one to three PCs. EWD imaging was also marginally better than PCA imaging for four or more PCs up to an azimuth of about $\pm 130^o$. Beyond this azimuth, EWD and PCA imaging accuracy were approximately equal. Again, these results are open to bias since the comparison was conducted by the author only.

For PCA models, spectral smearing often occurred for musical instruments with high frequency content such as snare drum hits or high piano notes. High frequency content tended to veer towards the loudspeakers for rearward azimuths. This was usually accompanied with the perception of a somewhat 'hissy' or 'screeching' musical timbre. This spectral smearing effect was much less noticeable for EWD models, which did not suffer to the same extent from such degradation in timbre. Thus, a possible connection between poor image

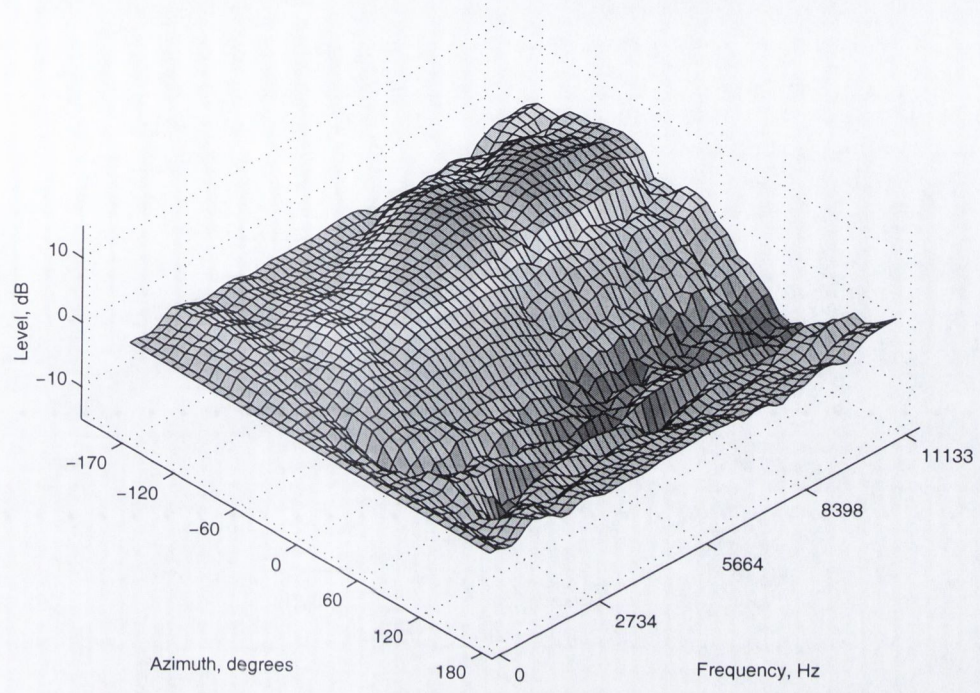
localisation accuracy and image timbre for high frequency signal components was identified. Panning movement was also perceived to be smoother for the EWD models as there were less abrupt changes in timbre from one azimuth position to the next. Although no noticeable improvement in image accuracy over PCA models with 4 or more PCs was observed beyond $\pm 130^o$, the timbre at such azimuths was superior for the EWD models. Finally, there was less inter-subject variability in imaging accuracy for the EWD models than for the PCA models, for a fixed number of PCs.

6.3.3 Comparison of 'Raw' and EWD Averages

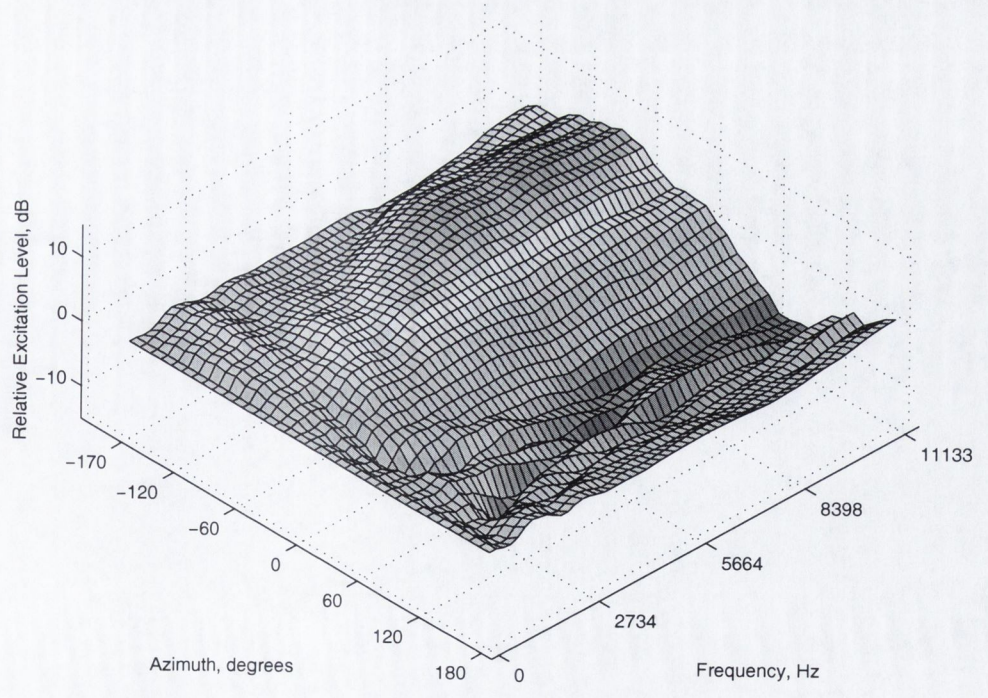
In order to further investigate the process of HRTF smoothing, two 'overall average' HRTF sets were generated, one being the average of the original raw data, and the other being the average of the EWD-smoothed data. These were both calculated by averaging the log magnitude of both raw and EWD-smoothed HRTFs for all of the five subjects. This data was then used to calculate β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications in exactly the same manner as described above for PCA and EWD analysis, with the same ITD information being re-inserted for both.

Figure 6.21 (a) and (b) compare the DTFs of the left ear of both raw and EWD-smoothed averages. Both sets of data are in excellent agreement; this indicates that EWD smoothing did not introduce bias over any frequency range. The PED and MSE are low, being $8.34.10^{-3}$ and $8.27.10^{-4}$, respectively. The dynamic range of both data sets is only about 15 dB. This is very low when compared to the 30 dB range observed for first order PCA, the 35 dB dynamic range observed for EWD analysis and PCA of order two to seven, and the 50 dB dynamic range observed in the raw, unsmoothed data. It was therefore expected that both raw and EWD-smoothed averages would yield very natural sounding images. There are some minor differences between the raw and EWD-smoothed averages. The raw average data is somewhat 'rougher', showing minor perturbations, mainly above 6 kHz. Calculation of the difference between raw and EWD-smoothed averages showed that the difference was generally about 1 - 2 dB only. It was therefore expected that these differences would not be noticeable, as β_{2spk} and $[\alpha/\beta]_{2spk}$ filter calculation would 'average out' these minor perturbations. However, this was not the case, as is explained below.

An informal subjective comparison of the two models based on the raw and EWD-smoothed averages was carried out using a variety of musical styles as stimuli. This comparison was similar to that carried out above for the PCA and EWD models. Due to time constraints, the comparison was carried out by the author and one other person only and was not carried out blind. Imaging accuracy was equally poor for both models, being com-

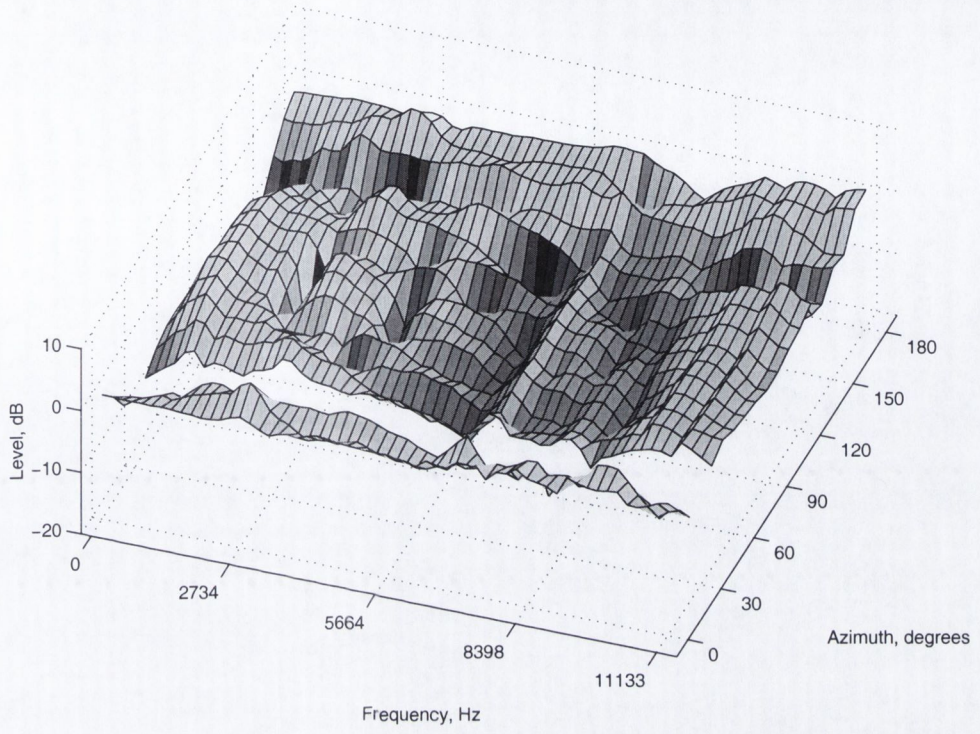


(a)

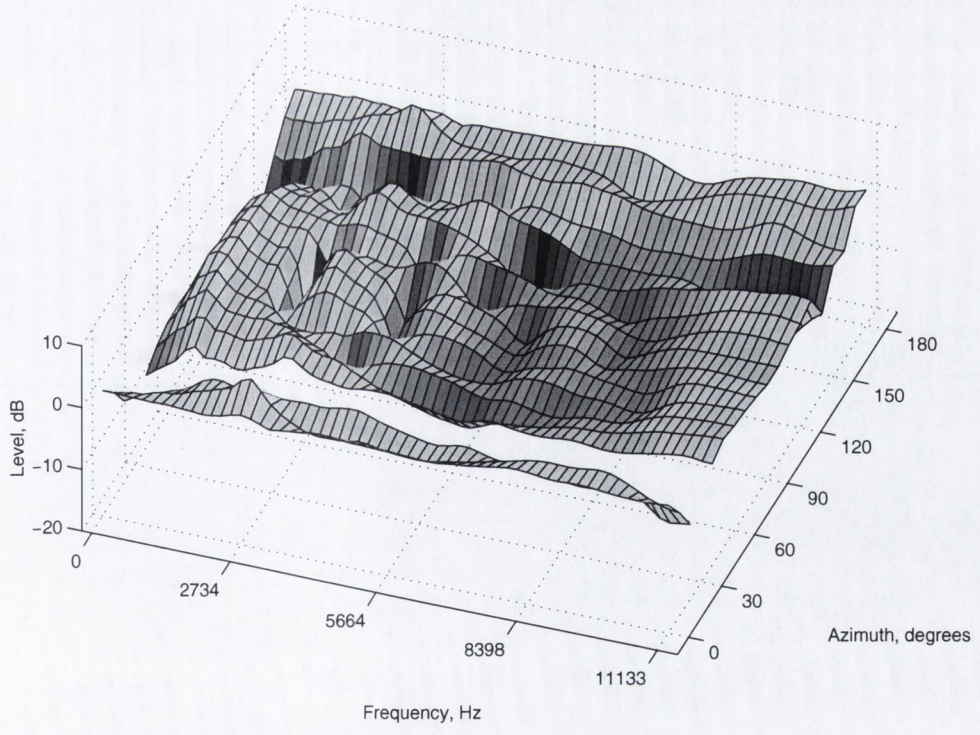


(b)

Figure 6.21: (a) Average 'raw' DTFs for the left ear of subject SOW. (b) Average EWD-smoothed DTFs for the same ear. $PED = 8.34.10^{-3}$, $MSE = 8.27.10^{-4}$



(a)

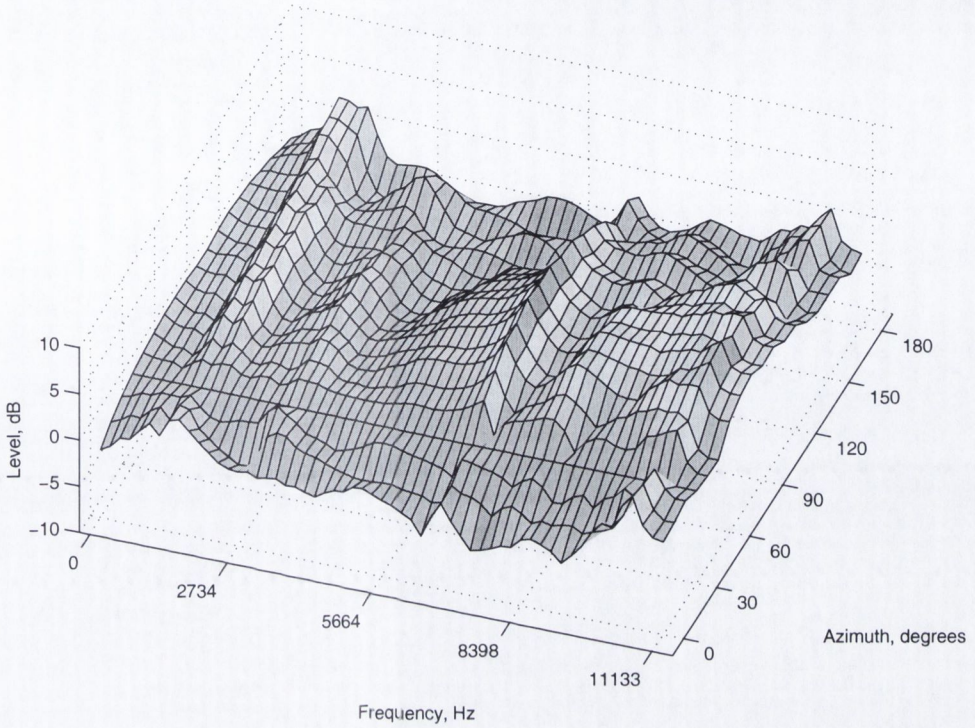


(b)

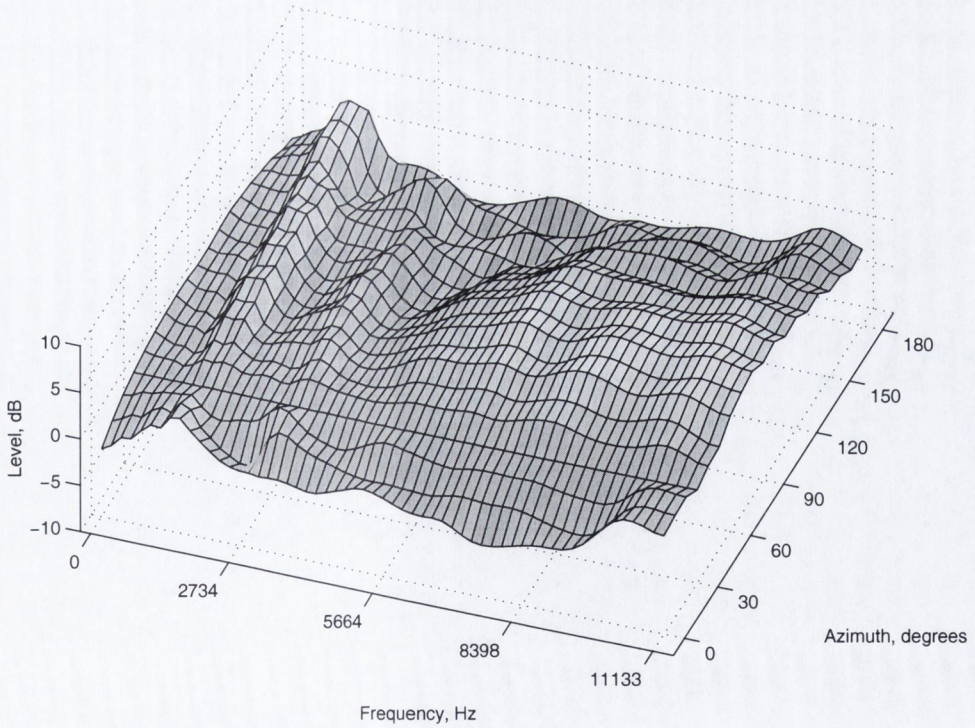
Figure 6.22: (a) Magnitude response of the $[\alpha/\beta]_{2spk}$ filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data.

parable to that achieved by first order PCA analysis. In general, very little imaging was achieved beyond $\pm 90^o$. This can be attributed to the lack of high frequency detail in the DTF specifications, the diffraction and reflection effects of the pinna having been 'averaged out'. In contrast to the poor imaging accuracy, imaging naturalness was excellent for both models, and was superior to all the previous PCA models and EWD models of individual heads described above. As stated earlier, it was expected that the minor perturbations in the average of the raw data would not be noticeable. However, although imaging naturalness was excellent for both raw and EWD averages, the timbre of the EWD average was somewhat better. Figure 6.22 presents a comparison of the magnitude of the $[\alpha/\beta]_{2spk}$ filter specification across frequency for all 18 azimuth positions from 0^o to 180^o (for images beyond this range, the speaker feeds were swapped, as described in Section 6.3.2). The filter specification for the averaged 'raw' data is presented in the top plot, (a), whereas the bottom plot, (b), presents the specification for the EWD-smoothed data. The gap at 30^o is due to the fact that for this azimuth, the left speaker is 'off' and the signal is fed directly to the right speaker only. It is clear that below about 7 kHz, both specifications are very smooth and agree closely, with the specification for the EWD-smoothed data being the smoother of the two. However, above this frequency clear resonances emerge for the raw average data which are not observed in the EWD-smoothed data. These resonances are at fixed frequencies which do not vary with azimuth. Close examination of a plot of their difference reveals that the resonances are at the frequencies 6836 Hz (bin 36), 8398 Hz (bin 44), 9961 Hz (bin 53) and 10547 Hz (bin 55). A dip is also observed at 8789 Hz (bin 46). The amplitudes of the dip and resonances are all about 2 dB, which is minor in comparison to the severe notching observed in the raw HRTF responses. Figures 6.23 (a) and (b) present a comparison of the β_{2spk} filter magnitude. Again, the specification for the raw average is corrupted at exactly the same frequencies, now by a difference of about 3 dB.

Finally, we examined the specification for the cascade of the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter functions. This specification is used to feed the left speaker for images panned to the right side of the listener and the right speaker for images panned to the left of the user. Therefore, any distortion common to both the β_{2spk} or $[\alpha/\beta]_{2spk}$ filter function will be doubled on loudspeaker presentation. To demonstrate this, Figure 6.24 (a) and (b) present the product of the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications. It is evident that the distortion is now serious and would result in a noticeable degradation in imaging naturalness. In this case, this degradation was worst for azimuths from 40^o to 130^o and -40^o to 130^o, or over half of the panning positions.

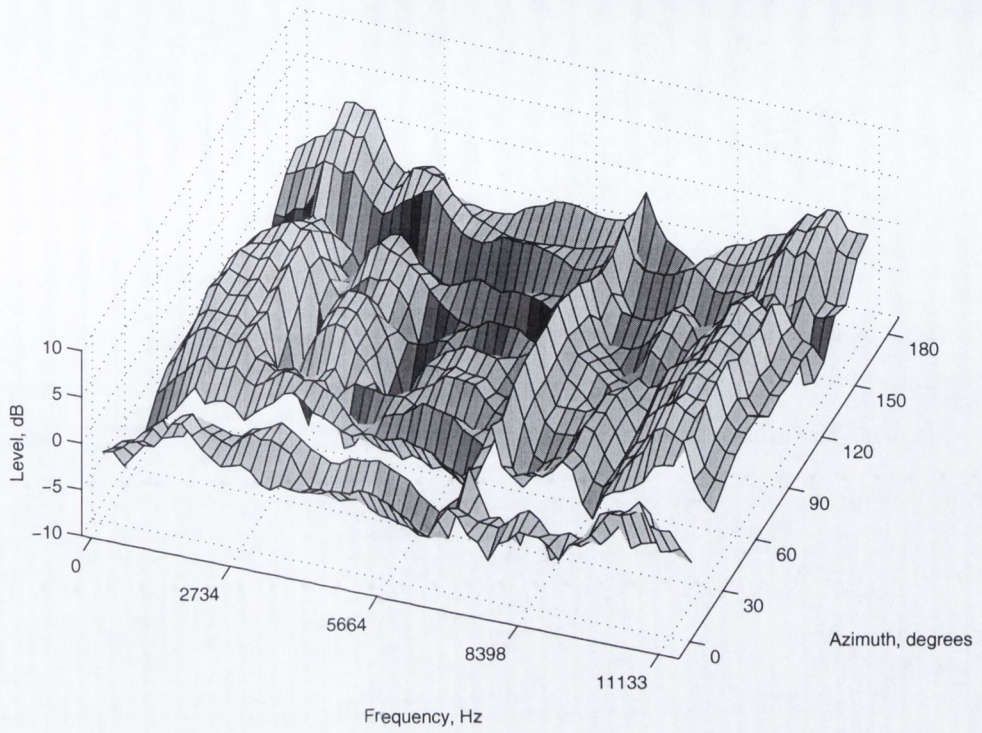


(a)

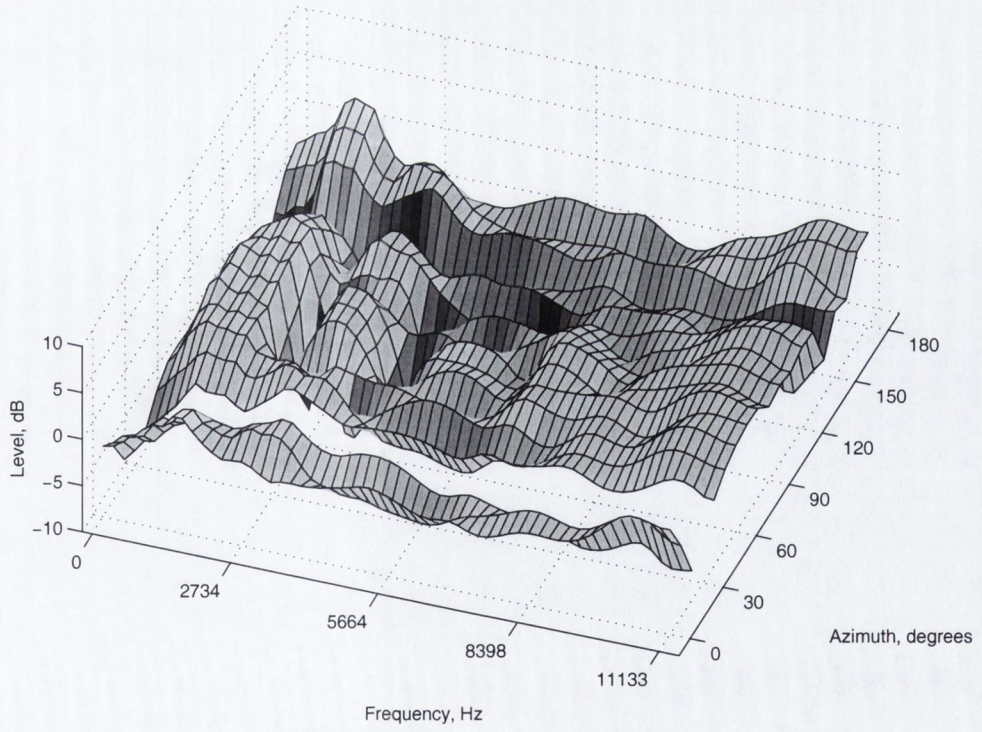


(b)

Figure 6.23: (a) Magnitude response of the β_{2spk} filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data.



(a)



(b)

Figure 6.24: (a) Magnitude response of the $[\alpha/\beta]_{2spk} \times \beta_{2spk}$ filter specification for the average of the 'raw' data. (b) Magnitude response of the same filter specification for the average EWD-smoothed data.

The fact that the degradation described above remained relatively fixed with azimuth suggests that it is caused by terms in the β_{2spk} and $[\alpha/\beta]_{2spk}$ equations that are present at all azimuths. An examination of Equations 6.4 and 6.5 shows that it is the HRTFs associated with the left and right loudspeakers which are present at all azimuths. These are denoted $HRTF_L(LF)$ and $HRTF_L(RF)$ for the left ear and $HRTF_R(LF)$ and $HRTF_R(RF)$ for the right ear. Since the left and right loudspeakers were placed at -30^o and 30^o respectively, $HRTF_L(LF)$ and $HRTF_L(RF)$ become $HRTF_L(-30)$ and $HRTF_L(30)$, and $HRTF_R(LF)$ and $HRTF_R(RF)$ become $HRTF_R(-30)$ and $HRTF_R(30)$ in the configuration. Figure 6.25 (a) to (d) presents a comparison of $HRTF_L(-30)$, $HRTF_R(30)$, $HRTF_L(30)$ and $HRTF_R(-30)$ for the raw average (dashed curves) and EWD-smoothed average (solid curves). Examination of these graphs reveals that whereas the HRTFs derived from the average of the EWD are smooth for all frequencies, the HRTFs derived from the raw average are quite jagged above about 6 kHz. It is in this region that the differences between the EWD and raw averages become most serious. Close examination of these graphs reveals that the resonances present in the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications at 6836 and 10547 Hz are due to the sharp notches at these exact frequencies in the $HRTF_R(30)$ filter specification for the raw average model. These notches are completely removed for the EWD-smoothed average. Note that, similar to this study, Walsh used the approximations of Equations 6.6 and 6.7 in his implementation. Therefore, the filter $HRTF_R(30)$ was also used for $HRTF_L(-30)$, which appears in every denominator for the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter specifications for PCA smoothing. Also, the peak at 9961 Hz and dip at 8789 Hz can be attributed to the corresponding peak and dip seen in Figure 6.25 (c) for $HRTF_L(30)$ for the raw average. Both $HRTF_L(30)$ and $HRTF_R(-30)$ appear several times in the β_{2spk} and $[\alpha/\beta]_{2spk}$ filter equations. Finally, we note that the assumptions of Equations 6.6 and 6.7 (*i.e.* that $HRTF_L(\theta_s) = HRTF_R(-\theta_s)$ *etc.*) hold well across all frequencies for the EWD-smoothed average, whereas they are only true for the raw average below about 6 kHz.

It can be readily appreciated from this investigation that removing notches from HRTF data is vital for attaining a good timbre. This is especially true for the particular HRTFs associated with the left and right loudspeakers, as demonstrated above, since these HRTFs affect the β_{2spk} and $[\alpha/\beta]_{2spk}$ filters at all azimuth positions. Furthermore, the differences between the average of the raw and EWD-smoothed data were all relatively minor, being 3 dB or less in magnitude, yet their effect was perceptible and resulted in a degradation in image naturalness. Therefore, any roughness at all in DTF data can have a detrimental effect on timbre. The removal of notches from raw, unsmoothed HRTF data can be appre-

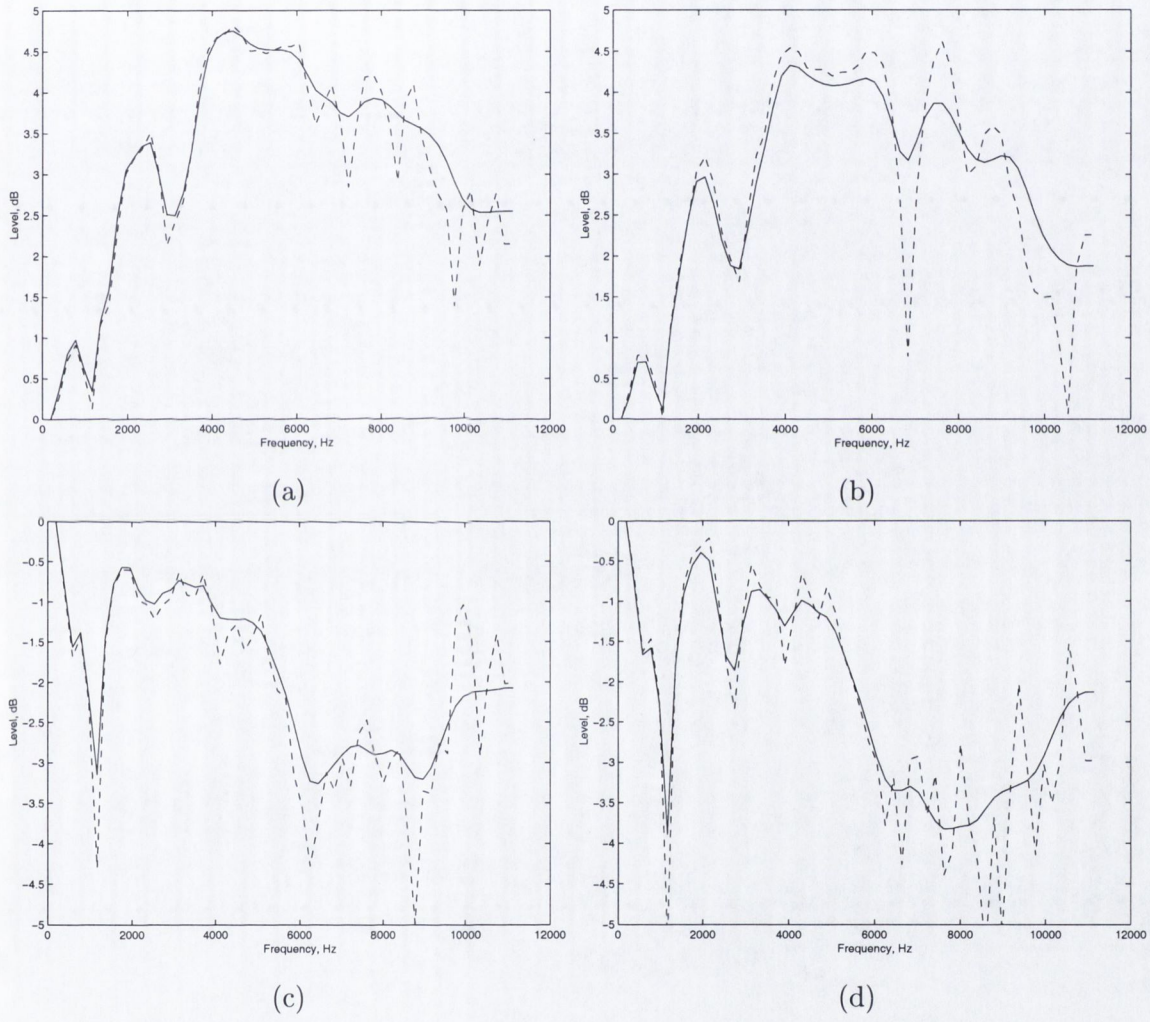


Figure 6.25: A comparison of the average 'raw' (dashed curves) and EWD-smoothed (solid curves) HRTF filter specifications common to all azimuth positions : (a) $HRTF_L(-30)$, (b) $HRTF_R(30)$, (c) $HRTF_L(30)$ and (d) $HRTF_R(-30)$.

ciated as being of paramount importance, since the magnitude of such notches can actually exceed 15 dB. The smoothing provided by the EWD has been shown to be ideally suited to the removal of such artifacts, especially at high frequencies where they are most prevalent. Notches are removed by EWD smoothing while still preserving the overall trend of the high frequency data. For lower frequencies, less smoothing is applied, or needed, since raw HRTF data is not typically degraded by notches for this frequency range, and also suffers from less inter-subject variability. In a comparison with PCA smoothing, the EWD outperformed all orders for image naturalness, and gave imaging accuracy equal or better to the best accuracy achieved by PCA smoothing. PCA of order three was deemed to achieve the best compromise between imaging accuracy and naturalness. However, Figures 6.12 to 6.17 clearly demonstrate that third order PCA over-smooths low frequency information while under-smoothing high frequency information. Also, the curves of the high frequency detail can sometimes appear 'cornered' or 'jagged'. In contrast, the curves of the EWD-smoothed data are very smooth and rounded. These details are only noticeable on very close examination of the individual filter functions and are not always apparent from the mesh plots of the DTFs of the PCA data shown in Figures 6.4, 6.5, 6.6 and 6.7. The previous comparison of the raw and EWD-smoothed averages demonstrates clearly a direct link between such minor perturbations and a perceptible degradation in timbre.

Overall, the spectral masking properties of the EWD have been shown to be very suited to the task of removing only undesired spectral detail and thus reducing inter-subject variability. The techniques of PCA and simple averaging of the log magnitude of the HRTFs have been shown to be less successful as they are not psychoacoustically motivated. It is the frequency marginal property of the EWD that makes this HRTF smoothing possible, as integrating along the time axis produces a smoothed frequency marginal equal to the the excitation pattern of the signal. It is also worth noting that achieving the degree of smoothing described above is possible using the EWD only because great care has been taken to ensure that the EWD itself is cross-term and clutter free over a very large dynamic range. Without this important property, HRTF smoothing would be compromised by the unwanted artifacts of bilinear time-frequency analysis, namely windowing and interference artifacts.

Again, it must be emphasised that the subjective assessment of imaging was performed by the author only and was not conducted blind. A more detailed psychophysical experiment is clearly required to fully assess the benefits of HRTF critical band smoothing via the EWD. Such an experiment was not possible because of time constraints and because of the scope of the work presented in this thesis. However, even without a thorough psychophysical

evaluation, the results presented do point to the usefullness of critical band smoothing as an effective and valid method of reducing inter-subjective variability in HRTF data. The EWD analysis of HRIR data shows that HRIR filtering is relatively invarient with time. This supports the PCA HRTF reconstruction methodology of Wightman and Kistler and their assertion that the frequency dependence of the ITD is not perceptually important. It is also worth noting that, in their present form, the EWD-smoothed HRTFs do not constitute generic HRTF sets. An analysis of many HRTF sets is obviously required to construct an overall, generic HRTF set which would contain the features important for localisation for as large a proportion of the population as possible. However, EWD smoothing is well suited as a preprocessing stage to such analysis, as it reduces inter-subject variability and undesirable artifacts by removing only spectrally masked features.

Chapter 7

Reverberation Analysis

Even though we may not always be aware of it, practically all of the sound we hear is 'acoustically processed'. When we listen to sound in a reverberant environment, as for example in a concert hall, the sound reaches our ears not only via a direct path from the source, but also via many indirect paths. These reflections modify the perception of the sound, affecting its timbre and loudness. They also contribute to the sound's overall spatial impression. Whether reverberation is desirable or not depends on the type of sound being listened to. For example, excessive reverberation can lead to a reduction in the intelligibility of speech. Consequently university lecture theatres are generally designed to avoid such effects. Music, on the other hand sounds very lifeless without reverberation. The appropriate amount of reverberation, however, depends on the style of music being played; excessive reverberation can cause some music to sound 'muddy' or 'washed out'. A great deal of care is therefore invested in the architectural design of new concert halls to ensure that the resulting reverberation enhances the type of music for which the hall is intended.

Almost all music recorded nowadays is processed by artificial reverberation. This reverberation is typically implemented using digital filtering techniques on DSPs. Reverberation algorithms typically strive to achieve as realistic and natural an effect as possible. In some cases artificial reverberation algorithms are designed to replicate the 'sound' of real rooms or performance spaces. Here, the aim is to design a reverberator which is perceptually indistinguishable from a real room. This involves identifying the perceptually important characteristics of the room acoustic response and generating a digital filter structure which replicates these characteristics. Unfortunately, distinguishing between the perceptually relevant and irrelevant characteristics of an acoustic response is still an issue of contention. Furthermore, the 'tuning' of the filter parameters of artificial reverberators in general often involves a good deal of 'trial and error' analysis, an approach which can be quite time con-

suming. With these difficulties in mind, the EWD is proposed as a tool for reverberation analysis. As the EWD is perceptually based, the analysis and comparison of the EWDs of both real and synthetic reverberation might provide clues as to the relationship between the physical and perceptual characteristics of performance spaces. Also, an examination of the EWDs of artificial reverberation containing unpleasant or unnatural audible artifacts might offer a better understanding of the flaws in artificial reverberator design. This chapter describes such an analysis.

7.1 Natural Reverberation

When a vibration occurs in a reverberant environment, the sound that reaches our ears first travels via a direct path from sound source to listener. This is known as the *direct sound* [15]. The earliest reflections after the direct sound reach our ears via indirect paths involving a small number of bounces, perhaps off the floor or nearby walls. Reflections involving just one bounce are termed *first order reflections*, those involving two reflections are termed *second order reflections* and so on. Such early reflections are not usually heard as separate from the direct sound. However, they modify the timbre and loudness of the sound and also its spatial characteristics. The *early sound* [15] refers to a period of up to 100 ms after the direct field, during which these early reflections occur. For increasing delays, the path from the source to the ear becomes increasingly complicated and the number of reflections reaching the ear increases to such an extent that individual reflections cannot be individually resolved. Furthermore, the absorption of the reflective surfaces at each bounce and the increasing path length results in a progressive reduction in the signal energy of each reflection. The overall effect of these multiple reflections is known as *late reverberation* (or the *reverberant field*) and can be characterised by a dense, noise like collection of echoes whose envelope decays exponentially with time [86], [115]. The reverberant field contributes to the perception of the background ambience of a performance space and may last for several seconds, during which a very large number of reflections reach the ears from all directions.

In order to better characterise reverberation, we can consider the reverberant environment as a linear time invariant system (LTI), where the system input is the signal source (for example a soloist in an orchestra) and the system outputs are the two signals reaching the eardrums of a listener. For a given source-receiver pair the input-output behaviour of a performance space can be characterised by its *binaural impulse response*. Typically, this is measured by exciting the room using an impulsive sound source, such as an electrical

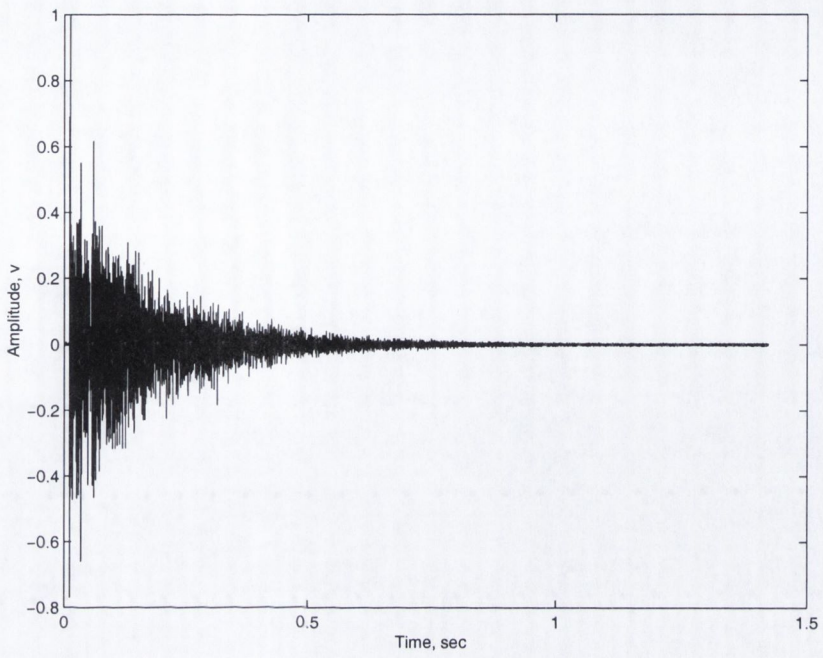


Figure 7.1: An echogram of a typical room.

spark discharge or air pistol and recording the acoustic response [51]. Steady-state sound sources such as noise or maximum length sequences can also be used [88], [60]. Figure 7.1 presents the monoaural impulse response (or echogram) of the *Teatro Comunale di Ferrara* concert hall in Italy. The initial direct sound can clearly be seen, followed by a few early reflections. The onset of late reverberation occurs after about 80 ms. For late reverberation, the *echo density*[1] becomes very high and individual reflections cannot be resolved; the response becomes quite like a noise waveform [86]. The reflections come from many different directions simultaneously, hence the intensity of late reverberation is independent of position within the reverberant environment. The reverberation time of a reverberant environment (or 'room') is given by the time it takes for the sound intensity to drop by 60 dB. A typical lecture theatre or cinema might have a reverberation time of about half a second, whereas a concert hall might have a reverberation time in excess of 2 seconds. Although the figure of reverberation time is useful as a single figure measure of the characteristics of a listening environment, the rate of decay of sound actually changes with frequency, with most environments exhibiting a faster decay at higher frequency. Therefore, reverberation time is often measured at different frequencies. This is usually done by exciting a room with wideband noise and then interrupting the noise signal. By filtering the resultant reverberant decay using octave filters at a range of centre frequencies and plotting the output (the

[1]The echo density is the average number of reflections per second.

impulse decay), the reverberation time at different centre frequencies may be estimated [51]. The frequency dependent reverberation time may also be calculated from the energy decay curve (EDC) given by the inverse integral (Schroeder integral) of the impulse response of the room [51], [86] :

$$EDC(t) = \int_t^\infty h^2(\tau)d\tau \ . \tag{7.1}$$

This integral computes the ene$EDC(t) = \int_t^\infty h^2(\tau)d\tau$ ulse response after time t. The EDC may be calculated at different frequencies by narrowband filtering the impulse response, $h(t)$, prior to EDC calculation. Hence the EDC at different centre frequencies may be displayed as a 3-D surface with dimensions of time, frequency and level. The *energy decay relief*, $EDR(t, \omega)$, is a more rigorous extension of this concept proposed by Jot [86]. The EDR is, in practice, calculated in a very similar manner to the EDC above except that the integration is performed on a time-frequency representation of reverberation such as the spectrogram, rather than the reverberation impulse response. In such as manner, the EDR is a very useful summary of important monoaural reverberation indexes, such as the reverberation time, early decay time and direct to reverberant ratio. In particular, the value of the EDR at $t = 0$ gives the overall frequency dependent gain of the room, which is clearly audible, whereas the frequency dependent reverberation decay time may be intuitively represented across frequency.

7.2 Reverberation Algorithms

Most music nowadays is listened to in highly non-reverberant environments such as living-rooms, automobiles or over headphones. To create ambience, reverberation is therefore needed. Reverberation has been added in the past by either recording the music in 'live' reverberant rooms, or by adding artificial reverberation to 'dry' studio recordings. Past approaches to artificial reverberation generation have used mechanical devices such as spring or plate reverberators to approximate the behaviour of real performance spaces. However, nowadays, artificial reverberation is typically implemented using digital filtering algorithms on DSP processors, allowing a much greater degree of control and accuracy over more traditional methods. These algorithms typically aim to reproduce the aural impression of real or imagined performance spaces. Therefore, identifying and understanding the perceptually relevant features of reverberation is an important issue. The processing demands typically placed on reverberation units are generally high. For example, studio reverberation units must be able to generate reverberation in real time, often for several signal feeds. Immer-

sive 3-D audio simulations must typically be able to handle both HRTF and reverberation processing for multiple source-listener positions. Efficiency of implementation is a crucial issue.

There are two basic approaches to reverberation algorithm design : the *physical approach* and the *perceptual approach* [60]. The physical approach aims to replicate exactly the input-output behaviour of a given real or imagined room. For real rooms, this simply involves measuring the binaural impulse response, as described earlier. For imaginary rooms, the behaviour of the room can be modelled by taking into account the exact geometry of the space, and the absorption of all the surfaces therein. The propagation of sound from source to receiver may be predicted using the laws of acoustic wave propagation and interaction with surfaces. Once estimated, the room response is used as a FIR filter and may be convolved with any given sound source to simulate the effect of listening to the sound in such a room. The overall procedure is referred to as *auralisation* and has been used to predict the acoustical quality of concert hall designs prior to construction. While the physical approach to artificial reverberation is useful for the evaluation of concert hall designs, it is too computationally expensive for more general use. This is because the processing required involves convolving the source signal with reverberation responses of several seconds in duration. Furthermore, achieving real-time parametric control is impractical since each parameter change necessitates the calculation of a new (and very long) FIR reverberation response [86].

In contrast to the physical approach outlined above, the perceptual approach to reverberation algorithm design seeks to reproduce the perceptually relevant reverberation features only. For the case of real rooms, these features are extracted from the room response. These are then used to parametrically control a reverberation algorithm based on highly efficient IIR filters. The advantage of this approach is that not only is it efficient, but it also allows real-time control of the perceptually relevant reverberation characteristics, since the IIR filter coefficients may be dynamically updated without incurring a heavy computational burden. One general reverberation algorithm may be sufficient to model all types of reverberation. A major problem with the perceptual approach is that there is much contention over exactly which properties of reverberation are perceptually relevant and which are not. It has also been difficult to replicate those effects deemed perceptually relevant using digital filtering techniques [60], [86]. Therefore, much current reverberation algorithm design is carried out with the aim of designing a reverberator which is perceptually indistinguishable from a given real room, without necessarily identifying or providing parametric control of all perceptually relevant features. The following section presents a brief review of important

reverberation algorithms based on the perceptual approach. Particular attention is paid to models of late reverberation and especially to Schroeder's reverberator as a study of the audible degradations of this reverberator using the EWD is carried out later in this chapter.

7.2.1 Simulating Early Reverberation

As outlined earlier, early reverberation consists of a relatively small number of discrete reflections of low order. As such, the pattern of these reflections may be calculated by considering the geometry of the room to be modelled. As long as the dimensions of the reflective surfaces of the room are large in comparison to the wavelength of the sound to be reflected, the interaction of the sound with a room's surfaces can be modelled using techniques similar to ray tracing in computer graphics [20]. A single sound source with multiple reflections is usually modelled as a series of multiple sound sources with no reflections. This method is termed the *source image method* [60]. If N is the number of wall surfaces, then the number of reflections of order k can be approximated as N^k. Therefore, the source image method is only practical for modelling early reverberation, where the reflection order is low.

The overall early reverberation impulse response may be generated by summing the contributions of each reflection. The delay time of each reflection may be calculated by measuring the total path travelled between source and listener for that reflection and dividing by the speed of sound in air. If the frequency dependency of the losses due to air propagation and surface absorption are ignored then the amplitude of each reflection is inversely proportional the total distance travelled and is proportional to the product of the absorption coefficients of all of the surfaces encountered. A more accurate model [95] takes into account the frequency dependent nature of the losses due to air propagation and absorption at each surface encountered using standard linear systems analysis. Thus, the transfer function, $A(\omega)$, of each reflection may be expressed as

$$A(\omega) = G(\omega) \prod_{j \in S} \Gamma_j(\omega) \quad , \tag{7.2}$$

where S is the set of walls encountered, $\Gamma_j(\omega)$ is the transfer function representing the frequency dependent absorption at each wall, and $G(\omega)$ models the frequency dependent losses due to air propagation and also the time delay from source to receiver. Naturally, reflections reach our ears from different directions. Therefore, the HRTFs of both left and right ears for the given direction may be used to provide the listener with this directional information. If only headphone presentation is required, for example, the complicated α and

α/β filter specifications described in Chapter 6 are unnecessary; we may simply convolve the signal directly with the HRTF functions associated with the direction of the impinging reflection. Therefore, the final transfer functions from source to eardrums, $Y_L(\omega)$ and $Y_R(\omega)$, taking into account the effects of absorptive losses due to the air and walls, the ITD and ILD localisation cues and the time delay due to the total path length are given by

$$Y_L(\omega) = HRTF_L(\omega)G(\omega)\prod_{j\in S}\Gamma_j(\omega) \tag{7.3}$$

$$Y_R(\omega) = HRTF_R(\omega)G(\omega)\prod_{j\in S}\Gamma_j(\omega)\ , \tag{7.4}$$

for the left and right ears respectively.

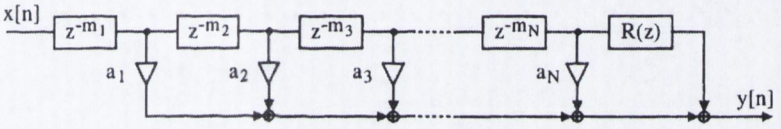


Figure 7.2: Schroeder's filter structure implementing early echoes and late reverberation [60]

The filter structures associated with early reverberation are quite simple if we model each reflection as a unit impulse in the early reverberation response. The reverb response can then be realised as a simple tapped delay line structure, with delays m_i and gains a_i associated with each reflection. Figure 7.2 depicts this structure, initially proposed by Schroeder [169]. The filter $R(z)$ is included to simulate late reverberation; such filters are considered in Section 7.2.2. Since each reflection is represented simply by a unit impulse the reverberation can have a discrete sounding quality [115]. This can be easily overcome by replacing the gain sections a_i with lowpass filter functions [60]. Furthermore, if the absorptive losses of the air and room surfaces and localisation cues are to be reproduced, then each tap in the filter structure may be followed by a filter $A(z)$ and filters $HRTF_L(z)$ and $HRTF_R(z)$, for the left and right ears, respectively, as described in Equation 7.3 and 7.4. This is depicted in Figure 7.3 for two taps. Calculating the absorptive losses and localisation cues for each reflection is a rather exhaustive approach as early reflections are generally not perceived individually [15]. A simpler and more efficient scheme, such as that of Jot [87], associates with each echo an individual gain and interaural time and intensity difference. Echoes with similar absorption and direction are then grouped in sets and processed by an average directional filter to provide the remaining spectral cues. This therefore reduces the amount of processing required while still preserving the most important reflection localisation cues.

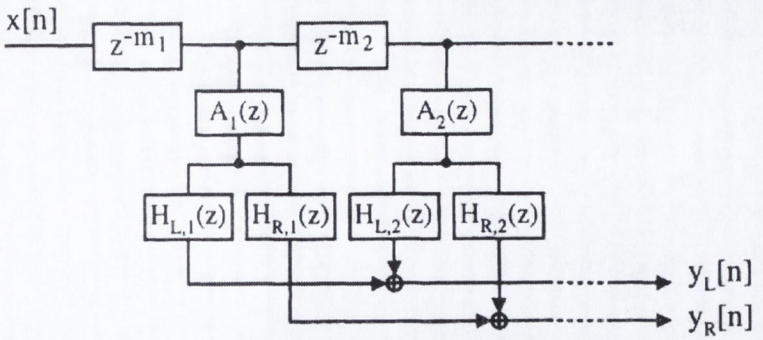


Figure 7.3: Processing of early reflections by absorptive and localisation filters [60].

7.2.2 Simulating Late Reverberation

As described earlier, late reverberation is characterised by a dense, noise-like ensemble of reflections arriving at the ear from all directions. The level of late reverberation is not dependent on the positions of source or receiver. In contrast, the direct sound level decreases with increasing distance between source and receiver. Therefore, late reverberation does not provide directional information. However, late reverberation does provide a very important distance cue as the ratio of direct sound to (late) reverberation decreases as the source-receiver distance increases. Because of its noise-like nature, statistical measures are often useful for describing the behaviour of late reverberation. One such measure is the *modal density*, D_m, which refers to the average number of modes per Hz of the frequency response of the room. These modes arise due to resonances caused by the geometry of the reflective surfaces of a room. For example, two reflective, parallel walls resonate in a manner similar to that of a closed pipe, causing resonances at frequencies $c/2L$, $2c/2L$, $3c/2L$ and so on, where c is the speed of sound in air and L is the distance between the walls. In a real room with complex shaped walls, floor and ceiling, many modes exist involving reflections off a variable number of surfaces. It can be shown [60] for rectangular rooms that the modal density is approximately proportional to the square of frequency, f :

$$D_m(f) = 4\pi V \frac{f^2}{c^3} \,. \tag{7.5}$$

For frequencies above the Schroeder frequency [86], the average spacing between modes is less than the bandwidth of the modes. This results in an overlap between adjacent modes. Having a high modal overlap ensures that the reverberation does not sound *coloured*, *i.e.* no prominent resonances are audible. This is an important feature of pleasant sounding

reverberation because the rapidly varying, random peaks of the frequency response no longer correspond to the room modes and are too fine to be resolvable by the ear.

Another important measure associated with late reverberation is the *echo density*, D_e. This figure gives the average number of reflections per second and is approximately proportional to the square of time :

$$D_e(t) = 4\pi c^3 \frac{t^2}{V} . \tag{7.6}$$

Thus, for real reverberation, the echo density increases rapidly with time. In a similar manner to the case for modal density, the random, noise-like fluctuations of the late reverberation of a real room do not correspond to individual reflections. In fact, Moorer has noted that Gaussian noise multiplied by an exponentially decaying envelope can sound remarkably like natural late reverberation [115]. Achieving a high echo density is important in artificial reverberation design in order to attain a reverberant decay free from audible artifacts such as fluttering or grainyness. There is some uncertainty concerning the number of echoes per second necessary to achieve a realistic sounding reverberant decay. Schroeder [168] has suggested that an echo density of about 1000 is sufficient to produce natural sounding late reverberation, while Greisinger [66] has suggested that 10000 may be a better figure. The following section outlines some of the most important work on artificial reverberators producing late reverberation, with particular attention being paid to the work of Schroeder.

7.2.3 Late Reverberation Algorithms

The first model of late reverberation employing digital filtering techniques was suggested by Schroeder in 1962 [168]. The model is based on the combination of several comb filters of mutually prime delays connected in parallel, followed by one or two allpass filters. The purpose of the comb filters is to generate a pseudo random echo pattern which the allpass filter then multiplies to generate a sufficiently high echo density. Figure 7.4 shows a single comb filter, with its associated impulse response, magnitude response and pole zero diagram.

The z transform of the comb filter is :

$$H(z) = \frac{z^{-m}}{1 - gz^{-m}} . \tag{7.7}$$

Here, g is the feedback gain and m is the delay in samples. Looking at Figure 7.4, we see that the time response of this filter is an exponentially decaying pulse train, spaced evenly every m samples. The frequency response displays peaks spaced at intervals of $1/mT$ Hz, where T is the sampling frequency. In order to achieve a flat frequency response, and thus

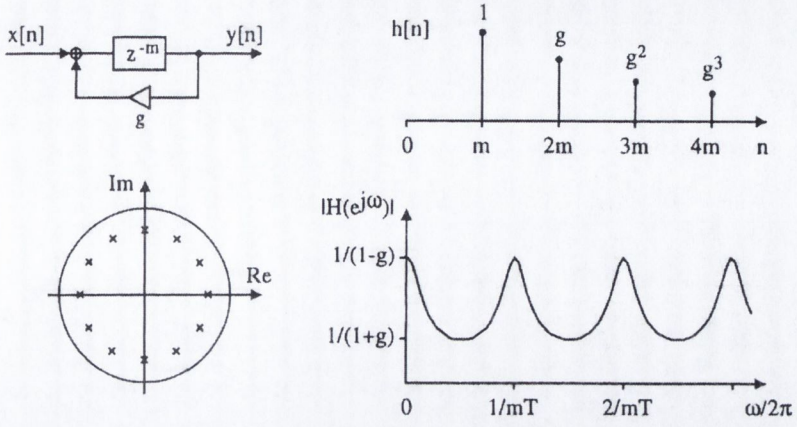


Figure 7.4: Filter structure, impulse response, pole-zero diagram and spectrum magnitude of a single comb filter [60].

avoid possible colouration, Schroeder suggested the use of allpass filters instead of comb filters. Figure 7.5 portrays such a filter, with its associated impulse response, pole-zero diagram and magnitude response. For this filter, the effect of the zeros outside the unit circle balances that of the poles inside the unit circle resulting in a flat frequency response. The impulse response decays at the same rate as that of the comb filter, namely by a factor of g every mT seconds.

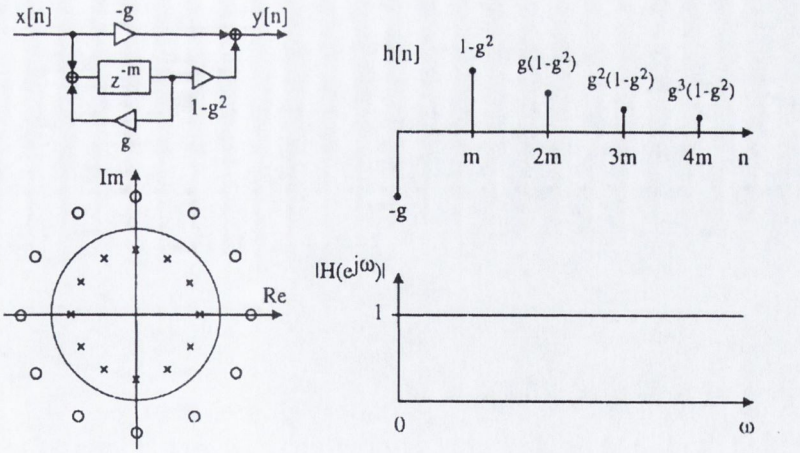


Figure 7.5: Filter structure, impulse response, pole-zero diagram and spectrum magnitude of a single allpass filter [60], [168].

In actual fact, the impulse responses of allpass and comb filters sound quite similar. This is due to the fact that the Fourier transform is defined over an infinite extent, whereas

the ear analyses sound on a short-time basis. Furthermore, real rooms do not have flat frequency responses, but consist of a large number of overlapping resonances. However, by connecting several comb filters in series and ensuring that the comb filter delays are mutually prime or incommensurate, the density of the combined comb filter peaks may be made sufficiently high to approximate the modal density of real reverberation. Furthermore, one or more allpass filters may be placed after the parallel comb filter section in order to greatly increase the echo density. This is because every echo produced by the comb filter section results in a series of echoes being produced by the allpass filter section.

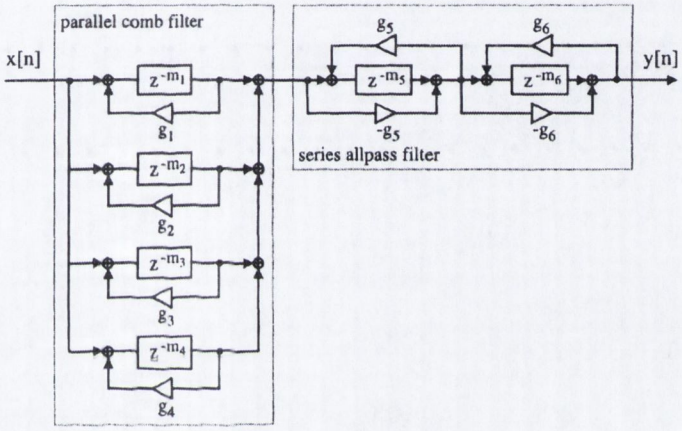


Figure 7.6: Schroeder's reverberator [60]

Figure 7.6 portrays the resulting structure, consisting of four parallel comb filters and two series allpass filters, known as Schroeder's reverberator. Schroeder suggested delays between about 30 to 35 ms for the comb filter section, and much shorter delays of 1.7 and 5 ms for the allpass filter section. The gains of the allpass filter sections are given a nominal value of 0.7. The gains of the comb filter section are set so that the comb filter outputs decay uniformly to produce the correct reverb time [60]. From Figure 7.4 we see that the amplitude of the echo falls by a factor of g_i every $m_i T$ seconds. For a reverb time of T_r seconds the echo amplitudes of each echo should drop by 60 dBs. We can use these facts to calculate the feedback gains, g_i, of each comb filter stage :

$$g_i = 10^{-3m_i T/T_r} \ . \tag{7.8}$$

Overall, the comb filter section produces the long reverberant decay and modal density, whereas the allpass section multiplies the resulting echoes to produce a sufficiently high echo density without affecting the overall magnitude response.

Schroeder's original reverberator sounds good as long as the reverberation to direct sound level is moderate and the reverb time is quite short. However, outside of these bounds the reverberation begins to suffer from unpleasant audible artifacts. These include an audible discrete quality or grainyness for the initial reverberation response and a fluttering or beating heard in the reverberant decay, especially for long reverb times. Also, the reverberation can be tonally coloured and sound overly harsh or metallic. In Section 7.3 an investigation of artificial reverberation degradation is carried out and reverberated signals are assessed for such degradation. A CD accompanying this thesis includes some of these signals in both unreverberated and reverberated form so that the reader may better appreciate this degradation.

The performance of the Schroeder reverberator is very sensitive to the choice of parameters, particularly the choice of delay lengths for the comb filter sections. In order to overcome some of these problems, Moorer proposed increasing the number of comb filter stages from four to six in order to increase the echo density (and therefore the modal density). He also placed a one-pole lowpass filter in the feedback loops of each comb filter in order to model the absorptive losses of air propagation. This had the effect of reducing the reverb time for higher frequencies and resulted in a more natural, less metallic sounding decay. It also reduced the discrete quality or grainyness of the reverb. Less fluttering was also observed. However, the artifacts were not completely eliminated.

Since the early work of Schroeder and later Moorer, there has been much research carried out into reverberation algorithm design. Gardener has investigated the use of nested allpass filters [59], initially suggested by Schroeder [168]. Such a filter structure produces an echo density that increases with time; this is observed for real reverberation. The use of feedback delay networks has been suggested by Gerzon [62] for the production of multichannel reverberation. More recently, Jot has proposed a reverberation algorithm based on the feedback delay network concept which allows specification of the frequency dependent reverb time, $T_f(\omega)$, and frequency response envelope, $G(\omega)$. A *reference filter*, whose impulse response is perceptually indistinguishable from exponentially decaying Gaussian noise is generated using a feedback delay network. The frequency dependent reverb time is tuned by associating an absorptive filter with each delay unit in the feedback loop [85]. The algorithm allows for fast dynamic control of the reverberation parameters since it is based on efficient IIR structures, rather than FIR convolution.

A more detailed description of these reverberation algorithms is not presented here as they are not directly relevant to this work. However, the interested reader's attention is drawn to the excellent summary of these algorithms presented by Gardener [60]. Earlier in

this chapter it was pointed out that there is still considerable uncertainty over the features of real reverberation which are perceptually important and those which may be ignored. This uncertainty results in a difficulty in designing artificial reverberators since this ultimately results in reverb design being part engineering, part trial-and-error experimentation. The presence of perceptible degradation in artificial reverberators is also unpredictable, being sensitive to the choice of reverberator parameters and the signal characteristics. Important features of reverberation, such as the fusing of early reflections and the resulting colouration are directly linked to temporal masking and comb filtering respectively. These two effects have been shown to be very well modelled by the EWD. Achieving a sufficiently high modal density, an issue of contention in reverberator design, is important for achieving a colourless response. This depends on the degree of frequency smoothing afforded by the ear, or spectral masking. Again, this is exactly incorporated in the design of the EWD.

It was noted in Section 7.1 that early reflections, although visually resolvable in the reverb response, are generally not heard as distinct from the direct sound, but fuse with its perception causing, instead, a change in signal loudness, colouration and spatial impression. The EWD has been shown to model the effect of temporal masking very well, and also to incorporate the effects of comb filter colouration. It was also pointed out that having a high modal density was important for achieving a colourless artificial reverberator. However, there is uncertainty over the modal density figure necessary to achieve this. Achieving a sufficiently high modal density depends on the frequency smoothing afforded by the ear, a feature which is again very well modelled by the EWD. The EWD is therefore proposed as an ideal tool for the analysis of reverberation responses and reverberated signals. Such an analysis is described in the following sections. In order to limit the scope of the work, the analysis focuses on the comparison of real reverberation with a single reverberation algorithm, namely the Schroeder reverberator. This particular algorithm is chosen because of its simplicity and ubiquitousness; its merits and flaws are well documented in the literature. Its basic structure is also the basis for more advanced reverberation algorithms. In this work, the influence of parameter changes on the resulting reverberation is investigated. Parameter choices resulting in both good and bad reverberation are made for comparison purposes. Furthermore, the reverberation produced is compared both visually (using the EWD) and aurally so that perceptible degradation could be identified aurally and characterised in the time-frequency domain. Thus a link between reverberation deficiencies and their effect on the reverberated signal can be formed.

7.3 EWD Analysis of Reverberation

Figure 7.7 (a) and (b) compare the responses of a real and Schroeder reverb (both monoaural) for a reverb time of just over one second. The real reverb is a monoaural response measured at the *Teatro dell'Aquila* in Italy [53]. Four comb filters and two allpass filters were used for the Schroeder reverb, with the mutually prime comb filter delays being 29, 37, 41 and 47 ms.

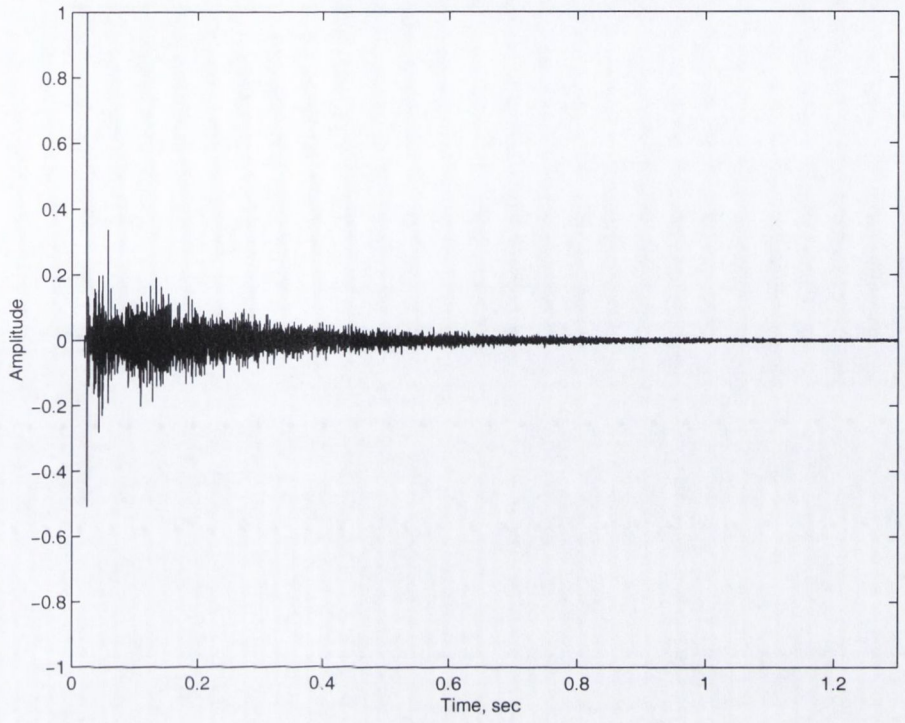
Examination of the two figures shows up several differences. The real reverb exhibits some early reflections up to about 30 ms after the direct sound, after which the echo density becomes very high. In contrast, the Schroeder reverberator produces a constant echo density. In addition, the Schroeder response has a discrete and almost periodic quality. Listening to these responses reveals that the real reverb sounds like a hand-clap in a relatively large room, whereas the Schroeder reverb sounds more like a metallic and discrete ringing.

Figures 7.8 (a) and (b) show the EWDs of the real and synthetic responses respectively up to an analysis frequency of 5 kHz. The real reverb has a longer reverb time for low frequencies, whereas the Schroeder reverb exhibits a fairly constant reverb time for all frequencies. An almost periodic fluctuation with time for much of the bandwidth shown is also evident for the Schroeder reverb's EWD. This is perhaps due to the somewhat periodic nature of its impulse response seen earlier in Figure 7.7 (b).

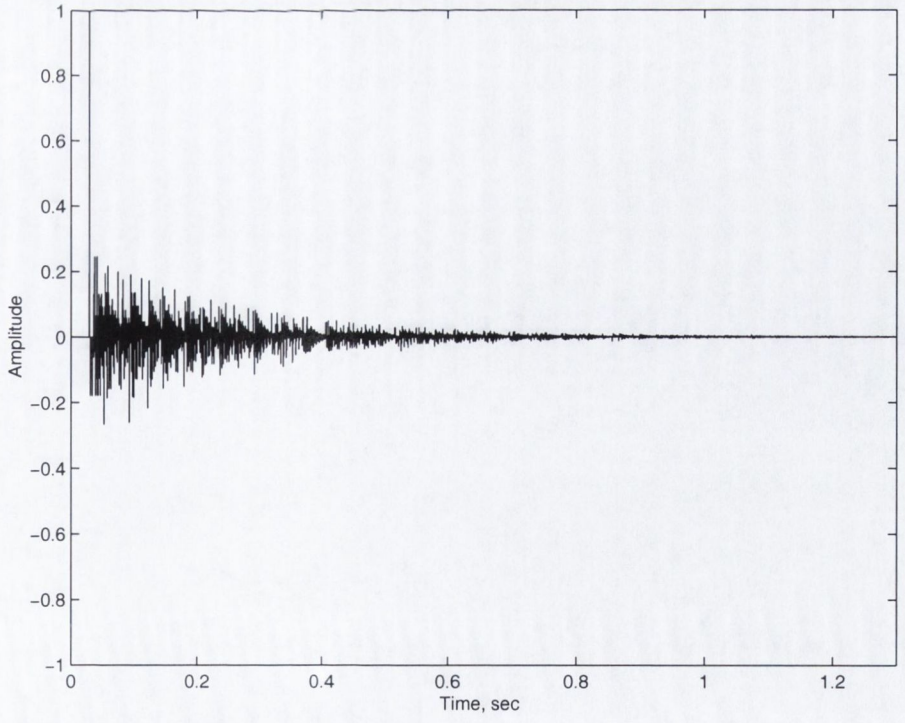
In order to test the subjective quality of the reverbs, they were convolved with an monophonic, anechoic sample of a female voice taken from track 4 of the Archimedes CD [12]. The recording was made in the anechoic room at the Acoustics Laboratory, Technical University of Denmark, Copenhagen using a B&K 4003 microphone pointed directly at the speaker. The speech section was taken from last two seconds of this track, the speech being "...changed in eighteen seventy five". Figure 7.9 (b) shows the EWD of this signal in its unreverberated form and Figure 7.9 (a) plots the actual signal. Vowel sounds are generally below 1 kHz, with the plosive and fricative sounds producing excitation at higher frequencies.

Figure 7.10 (a) and (b) present the EWDs for the speech signal reverberated with the real and synthetic reverbs, respectively. It is apparent from the plots that the fine details of the reverberated speech signal presented in Figure 7.9 (a) are now blurred, with the real reverb smoothing these details slightly more than the artificial reverb. Listening to the reverberated signals reveals a reduction in speech intelligibility which is slightly greater for the real reverb than the Schroeder reverb.

Since the overall signal length is now over 3 seconds in duration for both, it is quite

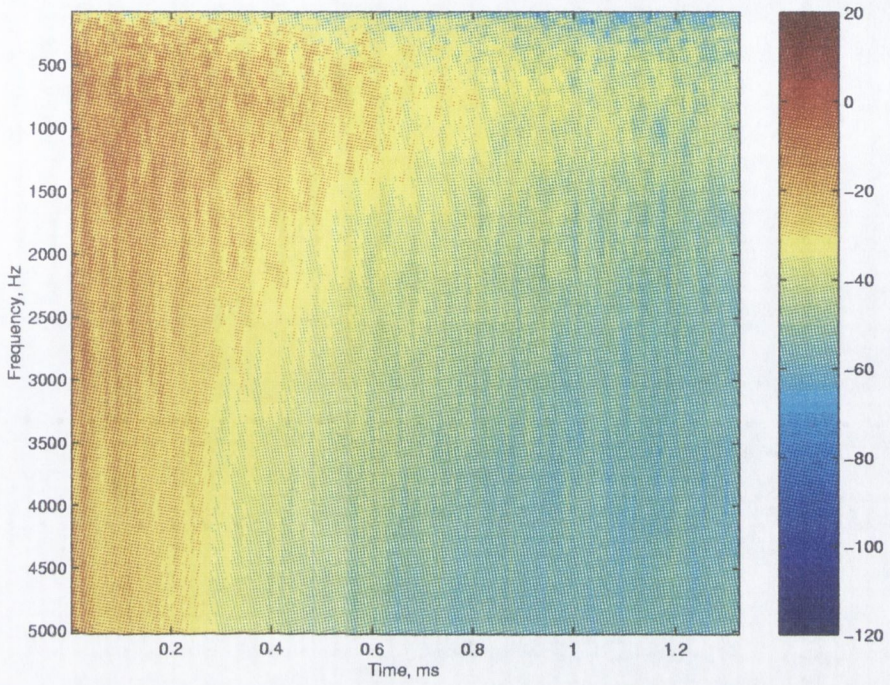


(a)

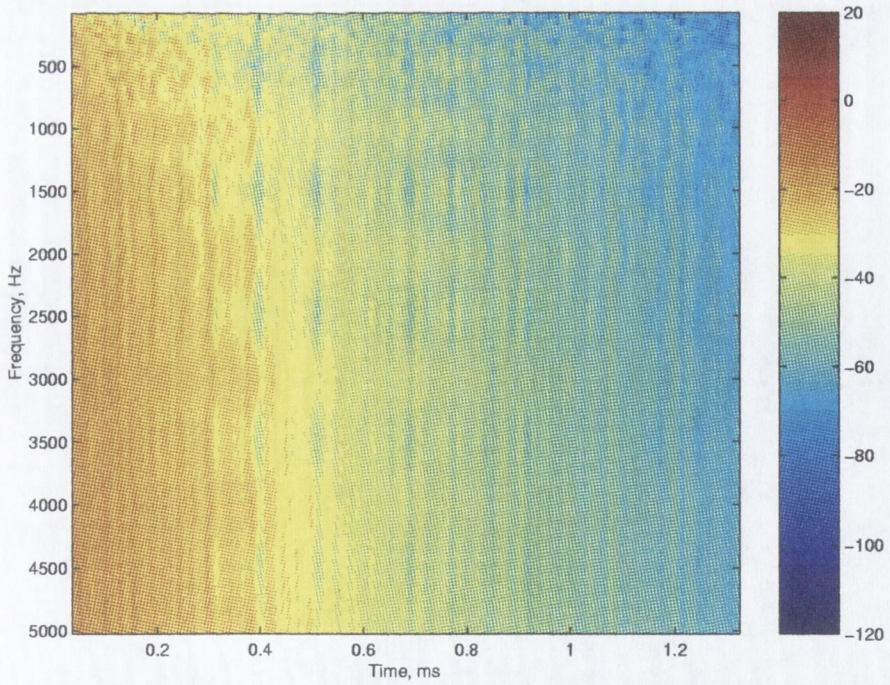


(b)

Figure 7.7: A comparison of (a) real room response and (b) Schroeder reveberator response.

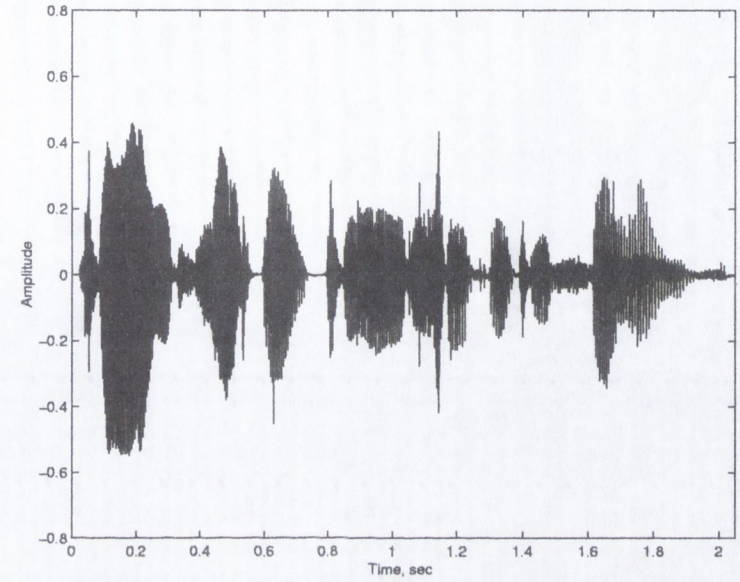


(a)

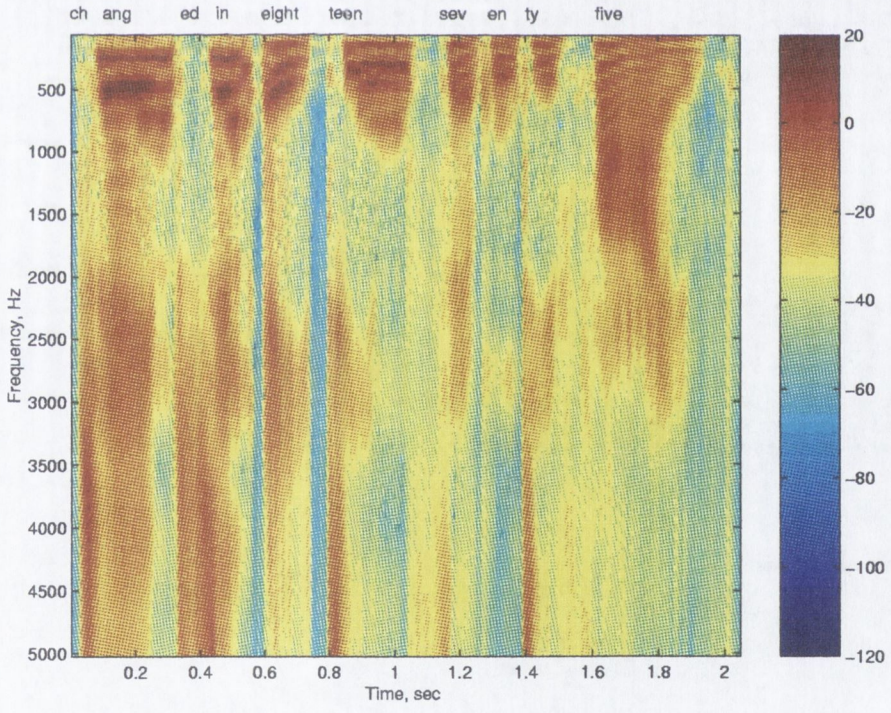


(b)

Figure 7.8: A comparison of EWDs of (a) real room response and (b) Schroeder reveberator response. Amplitude unit is relative excitation level (dBs.)

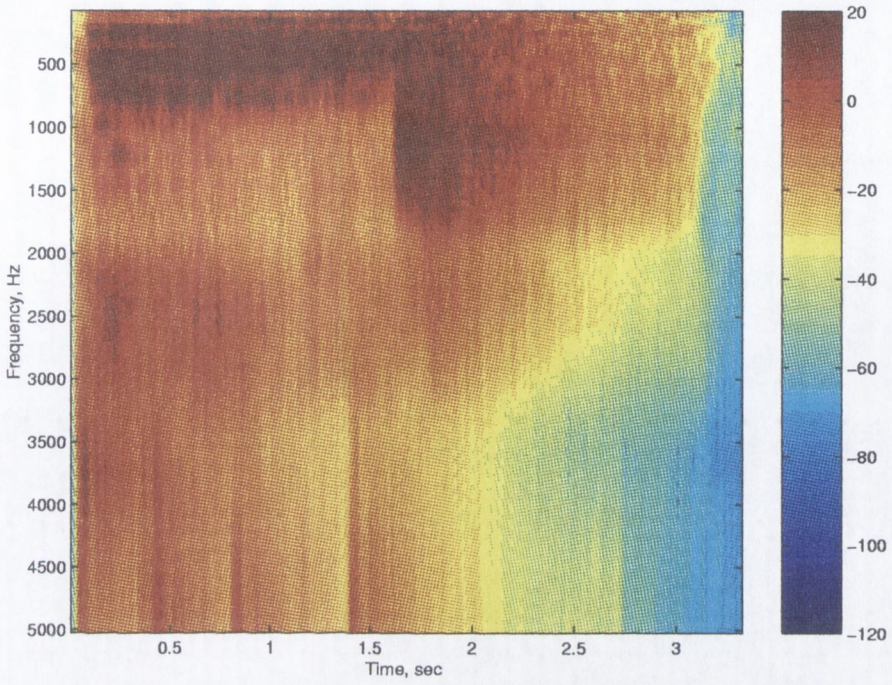


(a)

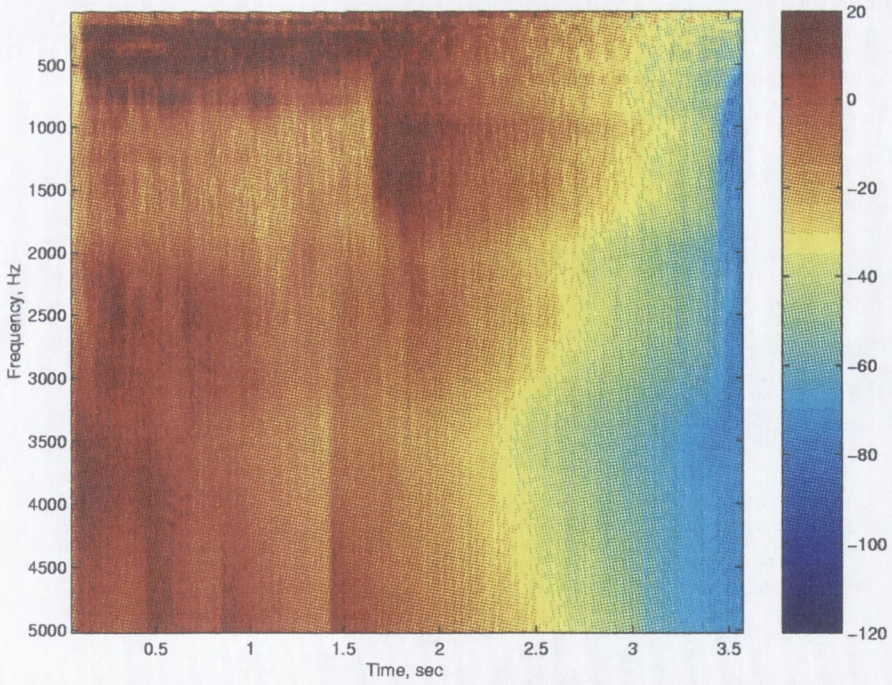


(b)

Figure 7.9: (a) Speech signal '*... changed in eighteen seventy five.*' (b) EWD of same signal. Amplitude unit is relative excitation level (dBs.)



(a)



(b)

Figure 7.10: (a) EWD of speech signal reverberated with (a) real room response and (b) Schroeder response. Amplitude unit is relative excitation level (dBs.)

difficult to compare the fine details of the plots at the scale shown. However, it is clear that the Schroeder reverb produces a more uniform decay across frequency, whereas the decay of the real room is slower for low frequencies. The greater amount of high frequency excitation for the Schroeder reverb results in a more intelligible, albeit slightly more grating sounding speech than the real reverb, which is quite dull sounding. The Schroeder reverb caused only a slightly metallic coloration of the speech signal. Overall, the Schroeder reverb compared quite well with the real reverberation for this particular reverb time and signal despite the fact that its impulse response sounded far more unnatural than the real reverb impulse response.

Reverb (time)	Synthetic	Clarinet	Vibraphone	Snare Drum	Flute	Violin
Real (1 sec)	-	-	G	-	-	-
Real (3 sec)	-	-	-	-	-	-
Prime (1 sec)	-	-	G	GM	-	-
Non-Prime(3 sec)	-	G	GGM	FGGM	-	-
Prime (3 sec)	M	FM	FFGG	GGMM	FM	-
Non-Prime(3 sec)	M	FM	FGGG	FFFGGMMM	M	-

Table 7.1: Audible degradation noted in a variety of musical signals. 'G' denotes grainyness, 'F' denotes fluttering and 'M' denotes a metallic timbre. The number of each letter denotes the presence of a weak, moderate or strong degradation.

Instrument	Note	f_0 /Hz	Volume	Track	Index	Style
Clarinet	$B\flat 4$	233.1	2	11	4	No vibrato
Vibraphone	$F\sharp 4$	370	3	6	4	Hard mallet
Snare Drum	-	-	3	11	7	Snare 'on'
Flute	$C\natural 5$	523.3	2	2	13	Flutter tongue
Violin	$G\natural 3$	196	1	1	1	Open string
Synthetic	$G\natural 3$	200	-	-	-	-

Table 7.2: Details of the samples taken from the MUMS database where f_0 is the fundamental frequency of the particular note. The synthetic signal was generated in MATLAB and is included for completeness. It has a note value of approximately $G\natural 3$.

As noted in [115], the presence of audible degradation in Schroeder-reverberated signals depends quite strongly on the signal itself. Impulsive sounds often seem to highlight the deficiencies of Schroeder-reverberated signals and are often the worst affected by the arti-

facts of grainyness and fluttering [60]. The speech signal presented above tended to hide or mask such artifacts. Therefore, in order to better investigate the degradation associated with the Schroeder reverberation, a wide variety of musical signals taken from the MUMS database [122][2] were convolved with Schroeder reverbs with reverb times of 1 and 3 seconds using both prime and non-prime comb filter delays. The prime delays used were again 29, 37, 41 and 47 ms, whereas the non-prime delays were 30, 40, 42 and 48 ms. It was expected that the non-prime delays would increase the presence of audible degradation. For three out of four cases, the difference between the prime and non-prime delays was only 1 ms. However, this was indeed sufficient to effect a considerable change in the presence of audible artifacts.

Table 7.1 presents results of subjective listening tests carried out by the author for some of the signals used, with the volume, track and index number of the MUMS samples being listed in Table 7.2. Due to time limitations, these listening tests were performed by the author only and were not conducted blind. However, the tests were conducted solely with the purpose of identifying candidate signals exhibiting audible degradation for further analysis and are not meant to be an objective measure of such degradation. The CD rom accompanying this thesis includes the reverberated vibraphone and snare drum signals so that the reader may gain a better appreciation of these artifacts.

The synthetic signal was composed of the first 30 harmonics of a 200 Hz fundamental, with harmonic amplitudes chosen to match those typically exhibited by a violin signal. The pitch of the signal was therefore a slightly sharp $G\natural$. The signal was 2 seconds in total duration, and had a rise time of 0.2 seconds and a fall time of 0.6 seconds. Figure 7.11 portrays the partial amplitudes of the signal, plotted against time and frequency. The onset and decay of the lower frequency partials precedes that of the upper partials. Overall, the signal had a timbre resembling that of a violin for its steady state, but sounded more like a reed instrument for the onset and decay.

Looking at the subjective results for the signals presented in Table 7.1, we see that the real reverbs did not suffer from any of the artifacts of grainyness or fluttering and had a natural, colour-free timbre for all signals (except for the 1 second real reverb and vibraphone signal which had some slight graineyness). The degree to which the four Schroeder reverbs performed depended very much on the signals chosen. For the (vibrato-free) clarinet signal and the (flutter-tongued) flute signal, there was a minor level of fluttering introduced for the

[2]Note that the MUMS database contains samples recorded in a variety of recording spaces, depending on the instrument. The particular samples chosen, while not anechoic, were recorded in a 'dry' studio environment and were therefore deemed to be sufficiently reverberation-free to be suitable for subjective evaluation of the Schroeder reverb.

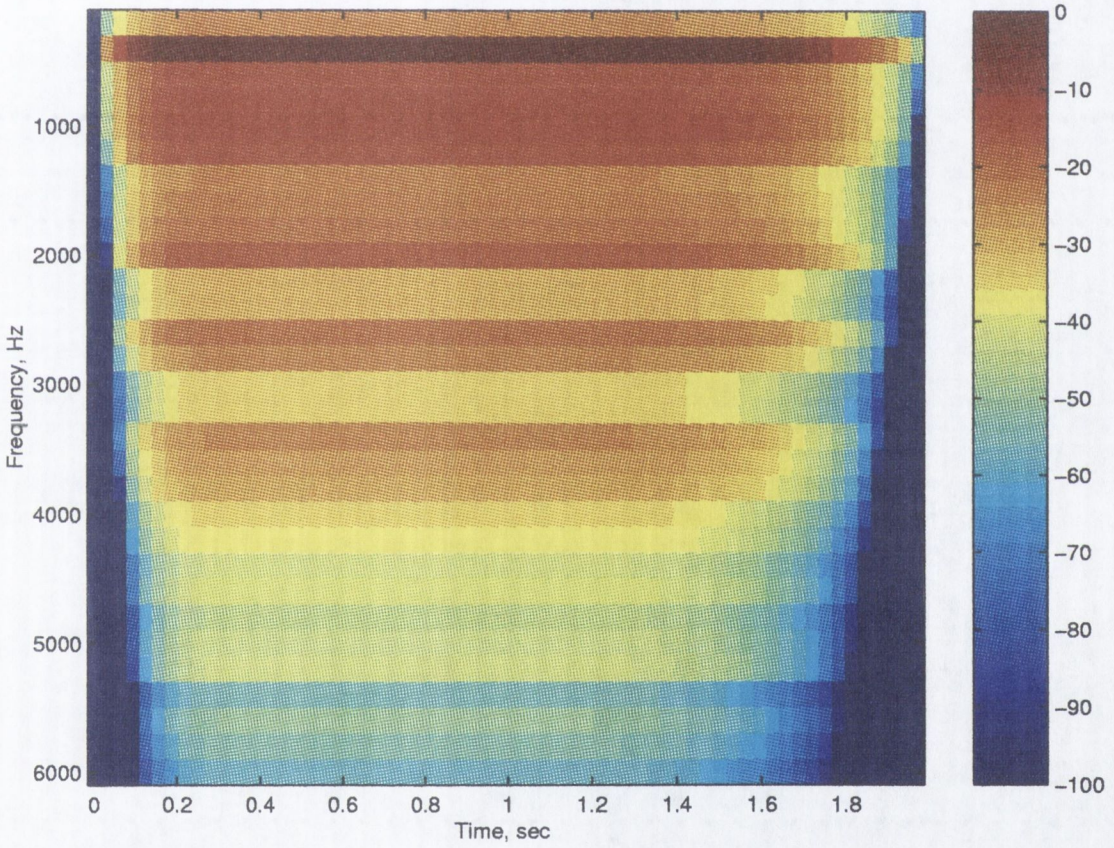


Figure 7.11: Partial amplitudes of the synthetic signal. Amplitude unit is relative excitation level (dBs.)

3-second, prime Schroeder reverb. However, this fluttering sounded quite natural for both signals—indeed for the flute signal it was difficult to separate it from the flutter-tonguing attack of the signal itself. It is interesting to note that this fluttering was not present for the non-prime, 3-second reverb. Finally, for both non-prime and prime 3-second Schroeder reverbs, there was a certain metallic quality to the reverberant decay. This was, however, not deemed serious.

In contrast to the wind instruments discussed above, the percussive instruments of snare drum and vibraphone (hard mallet attack) suffered very audible reverberation degradation which worsened with increasing reverb time. The snare drum suffered from a metallic timbre which was most noticeable for the 3-second, non-prime Schroeder reverb. The decay for this reverb also was seriously degraded by fluttering. Grainyness was also noticeable for the snare drum sample, more so for the 3-second reverbs, which were quite metallic sounding. The vibraphone signal did not suffer from a metallic timbre (except slightly for the 1 second non-prime reverb), but did suffer from strong grainyness, which was worst for the 3-second, non-prime reverb. In contrast, fluttering was moderate for the 3-second prime reverb, but faint for the 3-second non-prime reverb.

Overall, an increase in the reverb length of the Schroeder reverbs was noted to have the most prominent effect on the subjective quality of the reverberation. This was in keeping with the observations of [60] for the Schroeder reverb. Increasing the reverb length from 1 to 3 seconds resulted in a marked increase in grainyness, fluttering and timbre colouration for all signals exhibiting such artifacts. Furthermore, substituting non-prime delays for prime delays resulted in an increase in grainyness for all signals, which was expected, but a decrease in fluttering for some signals only, which was unexpected. The reasons for this decrease are not clear. It is also worth noting that the synthetic signal did not suffer from either grainyness or fluttering for any of the Schroeder reverberations. It did, however, suffer from strong resonances at the 5th partial (1 kHz) and 10th partial (2 kHz) which were more serious for the reverberation generated using prime delays than non-prime delays.

In order to investigate in more detail the suitability of the EWD for reverberation analysis, a closer examination of the EWD distributions was required. It was therefore decided to examine thoroughly the effect of changes in reverb length and comb filter delays on a single signal and to compare in detail the signal partial amplitudes derived from the resulting EWDs. Since the vibraphone signal used above exhibited both grainyness and fluttering, it was chosen as the candidate signal. The next section details an examination of the partial amplitudes of this signal for real reverbs and prime and non-prime Schroeder reverbs of reverb-times between 0.5 and 3 seconds.

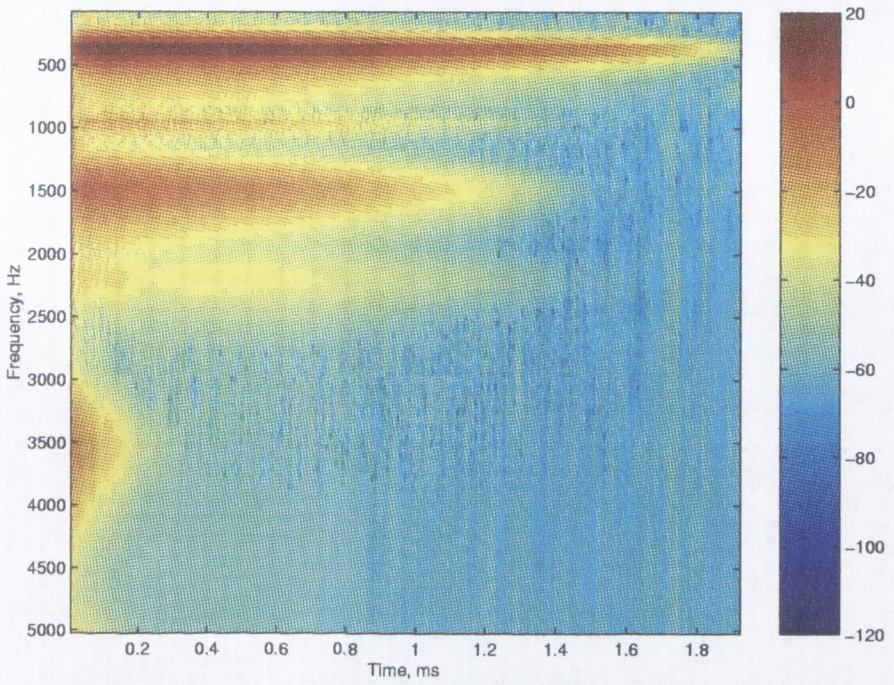
7.3.1 Vibraphone Analysis

The vibraphone signal was convolved with Schroeder reverberation having reverb times equal to 0.5, 1, 2 and 3 seconds. Prime and non-prime comb filter delays were used at each of the four reverb times, with the comb filter delays being equal to those chosen above in Section 7.3. For comparison purposes, real reverberation data was taken from a database of mainly Italian concert hall and outdoor auditorium responses made by Angelo Farina. This database is available from his homepage [53], along with several dummy head HRTF databases and a variety of anechoic recordings. The responses chosen were those of the *Teatro di Comunale di Firenze*, *Teatro Alcorcon* (in Madrid), *Teatro dell'Aquila* and the *Palasport MO*; for the sake of simplicity, these will be referred to as real halls 1, 2, 3 and 4 from here on. To compare the effects of the real and Schroeder reverbs, all Schroeder and real reverbs were convolved with the vibraphone signal. While the reverb times of the real halls did not agree exactly with those of the Schroeder reverbs at all frequencies, there was generally quite a close match, as is demonstrated later. For real halls 2 and 4 (the *Teatro Alcorcon* and *Palasport MO*), only binaural reverberation responses were available. Therefore, for these halls, monophonic reverberation signals were generated by summing the left and right channels.

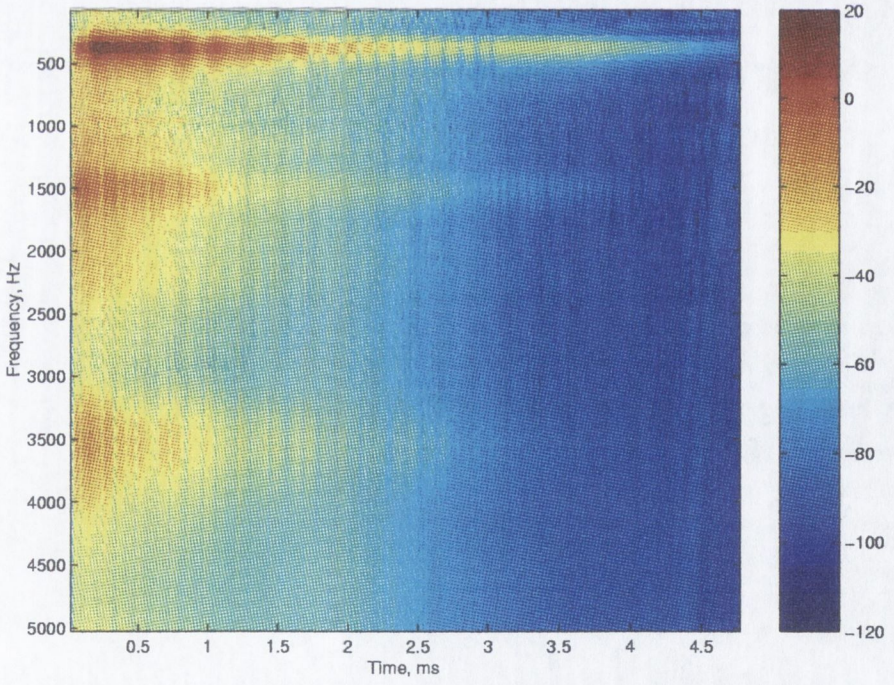
Figure 7.12 (a) plots the EWD of the vibraphone signal used[3]. The fundamental frequency of the vibraphone tone is $f_0 = 370$ Hz; the EWD reveals that after its impulsive onset (due to the hard mallet attack), signal partials emerge at f_0 (370 Hz), $4f_0$ (1480 Hz) and at $9.5\ f_0$ (3515 Hz). In order to give an example of the typical effects of grainyness and fluttering on the EWDs of the vibraphone signal, Figure 7.12 (b) presents the EWD of this signal reverberated with the 3-second, non-prime Schroeder reverb. From $t = 1.5$ to $t = 3$ seconds, the fundamental at 370 Hz shows a periodic amplitude modulation. This modulation is also clearly audible in the fundamental partial. The perception of the modulation also ceases at exactly 3 seconds. Furthermore, the re-emergence of the fundamental at $t = 3$ seconds is audible on careful listening. Comparison of Figure 7.12 (a) and Figure 7.12 (b) also reveals a rapid modulation of the partials at 1480 and 3515 Hz for the reverberated signal which is not present in the original signal. The exact nature of this modulation is not clear at the scale shown but the intervals over which the fluctuations occur correspond very closely to those marking the perception of grainyness for this signal.

To fully examine the differences between each of the 12 reverberations and investigate

[3]This is the same vibraphone signal used in Chapter 5 to compare the performance of the EWD and constant-Q modal distribution (Figures 5.16, 5.17, 5.18, 5.19). However, the time and frequency analysis points chosen here are different to those presented in Chapter 5.



(a)



(b)

Figure 7.12: A comparison of EWDs of (a) the unreverberated vibraphone signal and (b) the vibraphone signal reverberated with the 3-second, non-prime Schroeder reverb. Amplitude unit is relative excitation level (dBs.)

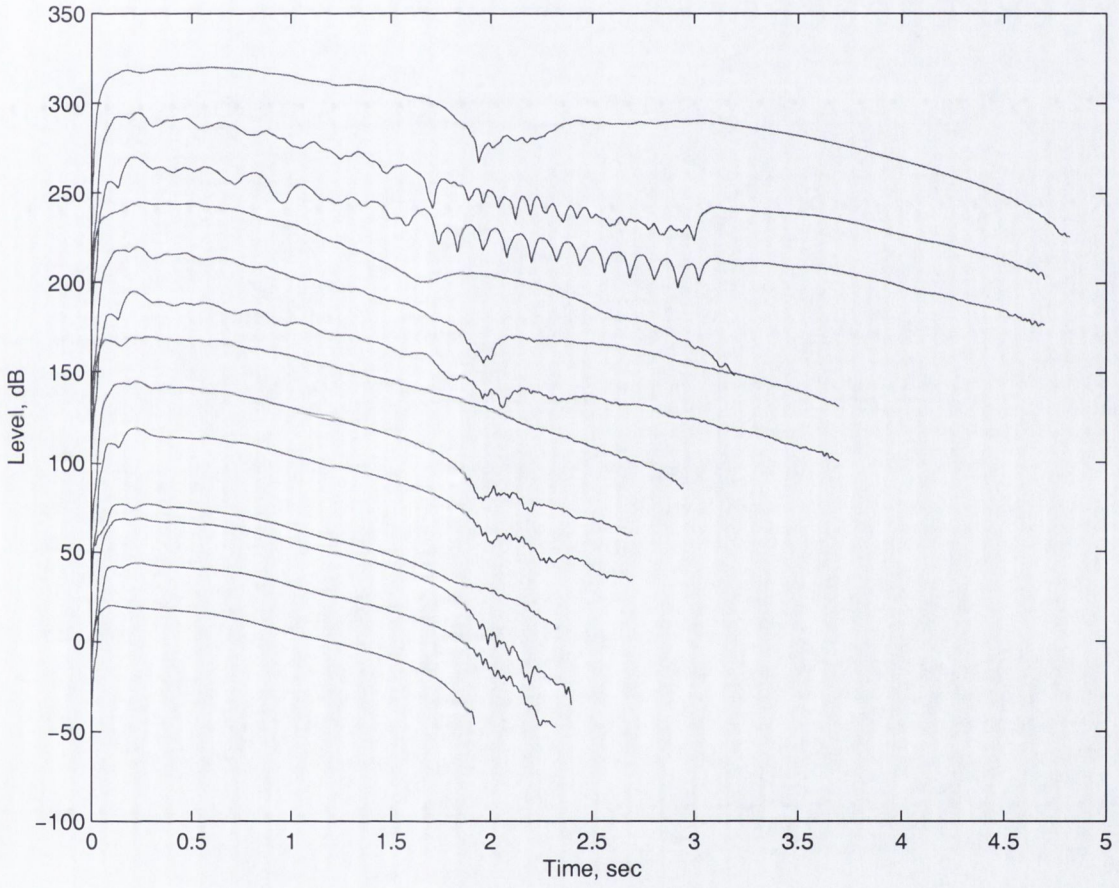


Figure 7.13: The temporal marginals for the vibraphone fundamental, f_0 (370 Hz).

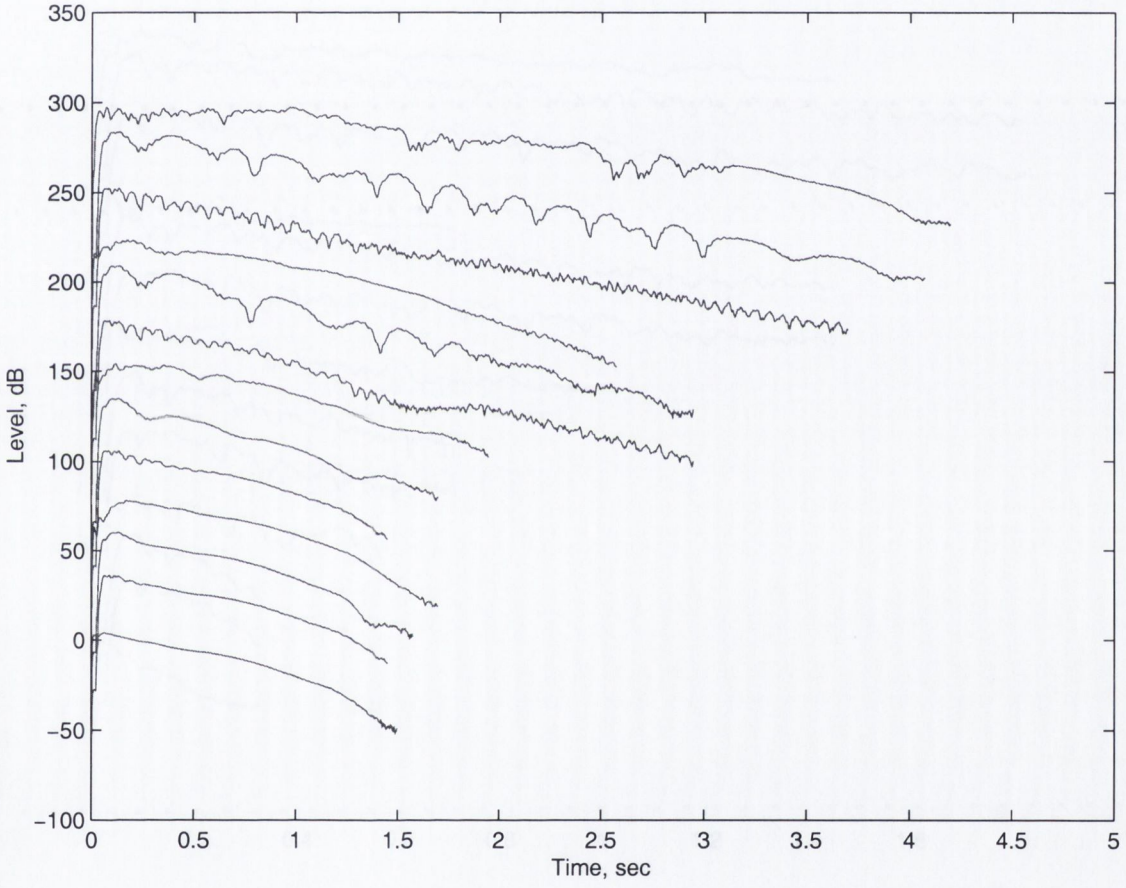


Figure 7.14: The temporal marginals for vibraphone partial, $4f_0$ (1480 Hz).

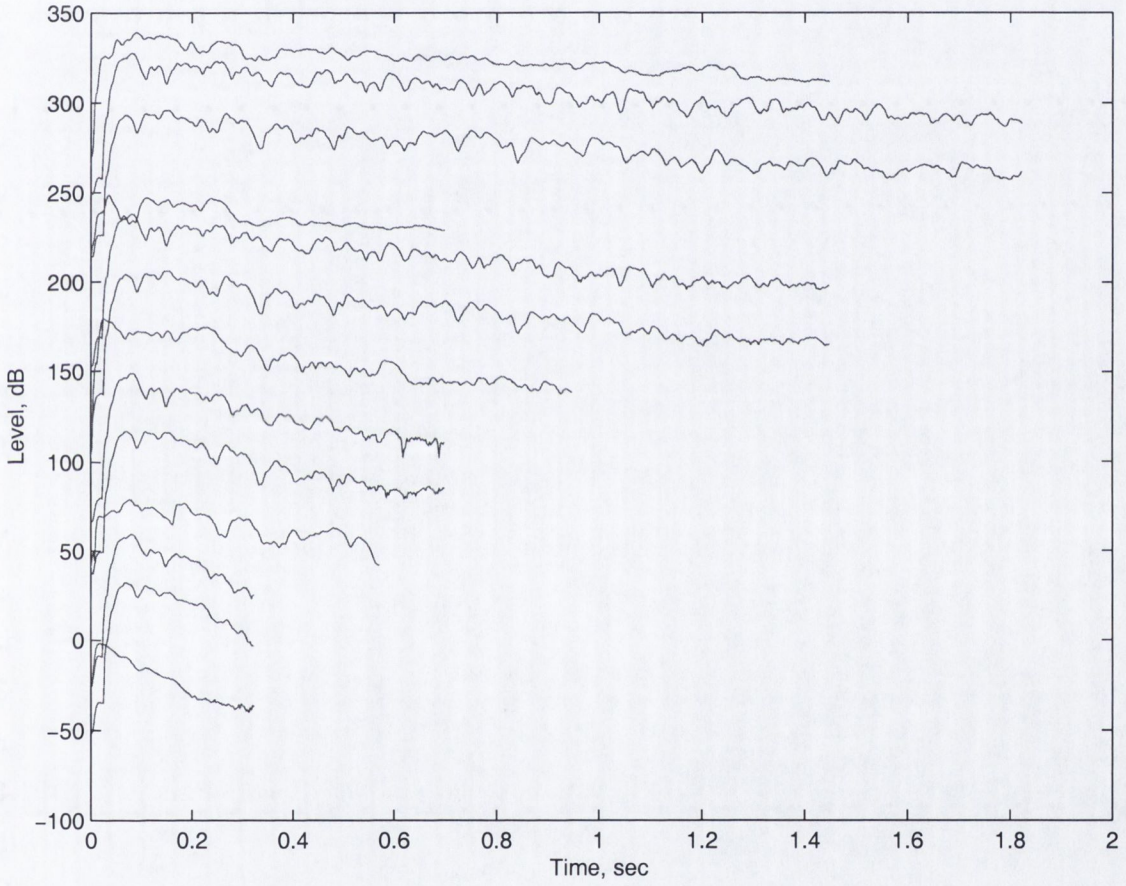


Figure 7.15: The temporal marginals for the vibraphone partial, $9.5f_0$ (3515 Hz).

more fully the modulations demonstrated above, the EWDs of the unreverberated and 12 reverberated vibraphone signals were computed, giving a total of 13 EWDs. The frequency range chosen for all EWDs was 100 Hz to 5 kHz, with 100 frequency channels in all being computed. The EWDs were computed every 3 ms over the entire duration of each signal. The temporal marginals of the partials at f_0, $4f_0$ and $9.5f_0$ were computed for all 13 EWDs. The power estimator methodology of Pielemeier was utilised for this calculation [147]. This gives a reliable estimate of partial strength with respect to time. This method is not detailed here; suffice to say that it involves choosing a frequency region around each partial of interest and performing a summation over this region[4].

Once these partial amplitudes were calculated, the signal was judged acoustically to determine whether or not the EWD partial amplitudes and auditory perception of these partial modulations corresponded. The following discussion argues that there was a strong correspondance between the EWD and auditory perception. However, this should be considered a pilot study since the tests were conducted by the author only. Furthermore, the experiment may be open to bias since both acoustic stimuli and partial modulation representation are judged subjectively. Further detailed experimentation is required in order to objectively determine the degree to which the EWD mirrored auditory perception.

Figures 7.13, 7.14 and 7.15 present the resulting temporal marginals for the partials at f_0 (370 Hz), $4f_0$ (1480 Hz) and $9.5f_0$ (3515 Hz) respectively. The lowest curve shown in each of the three figures is the temporal marginal of the unreverberated signal. Above this, the plotting order from bottom to top is : 0.5 second non-prime Schroeder, 0.5 second prime Schroeder and hall 1, then 1-second non-prime Schroeder, 1-second prime Schroeder and hall 2, then 2-second non-prime Schroeder, 2-second prime Schroeder and hall 3, and finally 3-second non-prime Schroeder, 3-sec prime Schroeder and hall 4. Each curve is offset by 25 dB, to allow easy comparison of the partial marginals for all 13 signals. Note that because the EWD incorporates spectral masking, there was a higher spread of excitation above each partial than below. Upper partials also produced excitation over a greater bandwidth than lower partials. The frequency regions chosen depended on the partial and are presented in Table 7.3. Since the gamma window does not exhibit a lobing structure integration was typically performed over a frequency region covering the first 20-30 dBs of each partial's

[4]The frequency region used in the work of Pielemeier is generally chosen to equal the width of the main lobe of the frequency smoothing window. This region should be free from cross-term interference and lobing artifacts. Pielemeier notes that excellent estimates of partial amplitude can be attained using this method and that the method is quite robust to the symmetry of the particular frequency region chosen. A frequency estimator methodology is also described by Pielemeier but is not relevant to the reverberation analysis presented here.

excitation. It is also worth noting that although the ordinate ranges for all three figures are the same, the relative level of the partials across figures is not an indication of relative partial strength as this level is influenced by the width of the integration regions chosen.

Partial Number	f_c	f_{min}	f_{max}
1	370	200	750
2	1480	1150	1950
3	3515	2950	4450

Table 7.3: Integration regions for the partials at f_0, $4f_0$ and $9.5f_0$ of the vibraphone signal. f_c is the partial frequency and the integration is performed over the region f_{min} to f_{max}.

An examination of the temporal marginals shown in Figures 7.13, 7.14 and 7.15 reveals some very interesting signal behaviour. After its rapid onset, the temporal marginal of the fundamental of the unreverberated signal, shown as the lowest curve in Figure 7.13, is very smooth and decays at a more or less exponential rate with time. The temporal marginals of the real reverberated signals (represented by the 4th, 7th, 10th and 13th curves from the bottom) are also smooth, apart from an audible, but natural-sounding broad dip in the curve for hall 4 at around $t = 2$ seconds (top curve). In contrast, the curves shown in Figure 7.13 for the Schroeder reverb (represented by the 2nd, 3rd, 5th, 6th, 8th, 9th, 11th and 12th curves from the bottom) betray a considerable amount of modulation of the fundamental. This modulation gets progressively more severe as the reverb time increases. For the 0.5 and 1 second prime and non-prime Schroeder reverbs, some very low level noise-like modulation is observed in the temporal marginals after 2 seconds. However, this modulation is not audible for the 0.5 second case and only barely noticeable for the 1 second case. The modulation for the 2 and 3 second Schroeder reverberation is more pronounced, and is most audible for the 3 second non-prime Schroeder-reverberated signal (3rd curve from the top). For the first 1.5 seconds, approximately, this modulation is non-periodic and 'smooth' and is heard as a random fluctuation in partial strength. However, after about $t = 2$ seconds, a point, incidently, corresponding to the end of the unreverberated signal, a distinctly periodic modulation emerges, being clearly audible as a regular 'beating' or fluttering. At about $t = 3$ seconds, this beating ceases and the fundamental grows in strength slightly and then decays evenly, an effect which is actually audible on careful listening. Overall, there is a very close correspondence between the perception of the modulation of this partial and its temporal marginal, as extracted from the EWD. This is also true for the case of the 3 second prime reverb, which is less degraded by fluttering and for the 2 second prime and non-prime

reverberated signals, which exhibit only a moderate degree of fluttering. Other interesting features of these plots include the difference in the partial onset between the prime and non-prime cases for reverb times of 1, 2 and 3 seconds. The onset of the fundamental for each the non-prime reverberated signals of 1, 2 and 3 seconds (5rd, 8th and 11th curves from the top) exhibits a dip at about $t = 12.5$ ms which is clearly audible, the fundamental 'kicking in' strongly after this dip. This effect is not noticeable for the prime reverbs and does not appear in their corresponding temporal marginals.

Figure 7.14 presents exactly the same set of curves, this time for the partial at $4f_0$ or 1480 Hz. The marginals computed for the unreverberated and reverberated signals for the reverb times of 0.5 and 1 seconds (i.e. the first 7 curves from the bottom) show little evidence of temporal modulation for this partial. This agrees well with the results of subjective listening tests, with only a very minor degree of fluttering being evident for the 1 second case, and no audible degradation being noticeable for the 0.5 second case. The top six graphs, however, tell a different story. The curves representing the temporal marginals for the non-prime and prime 2-second Schroeder reverberation (the 8th and 9th curves from the bottom), are noteworthy. It is clear that the non-prime reverberation (the 8th curve from the bottom) in this case produces a prolonged, rapid and noise-like modulation. This manifests itself, as expected, as an audible perception of grainyness for this partial. In contrast, the modulation caused by the prime reverberation (9th curve from bottom) is smoother and less rapid and is heard as a fluttering effect. The upper three curves present three contrasting partial amplitudes corresponding to the 3-second non-prime, prime and real reverberations. A prolonged, noise-like partial modulation is visually evident for the non-prime reverberation that is clearly audible as a strong, rasping or grainyness. For the prime reverberation, a much slower, almost periodic modulation is visually apparent which is also clearly audible as a fluttering of this partial. Finally, the top curve, representing the temporal marginal for the real reverb case of hall 4, shows three distinct regions of grainyness. All three intervals, from $t = 0$ to $t = 0.6$, $t = 1.5$ to $t = 2$ seconds and $t = 2.5$ to $t = 3.1$ seconds, are accompanied by a perception of moderate grainyness. Furthermore, the 3 dips in the last region ($t = 2.5$ to $t = 3.1$ s)are audible, creating a sense of brief fluttering. Lastly, Figure 7.15 presents the temporal marginals of the final partial at $9.5f_0$ (3515 Hz). Although most of the marginals shown exhibit much noise-like modulation, it is barely audible. This is because the level of this partial is very low in comparison with the previous two partials discussed.

Overall, the analysis demonstrates a strong degree of correspondence between the signal properties extracted from the EWD and perception of the audible degradations of grainyness

and fluttering. Grainyness seems best characterised by a rapid, noise-like modulation of signal partial amplitudes, whereas fluttering seems to be produced by a slower, almost periodic modulation. The artifacts become more noticeable as the depth of modulation increases. In hindsight, it is not surprising that the artifacts of fluttering and grainyness are caused by such modulations. However, the degree to which their manifestation in the EWD and their audibility corresponded is very encouraging. The EWD very accurately predicts the time intervals over which these artifacts are noticeable. EWD analysis demonstrates clearly the strong detrimental effect of increasing reverb time on the quality of the Schroeder reverb algorithm. This is a well known characteristic of this algorithm. EWD analysis also revealed that grainyness is worse if non-prime comb filter delays are used. Fluttering, however, behaves in a more unpredictable fashion, at least for the signals analysed. With this in mind, the Section 7.3.3 presents an investigation of the influence of comb filter delay choice on the presence of undesirable artifacts for the Schroeder reverberation algorithm. However, first an interesting analysis of the modulation spectra of the partial amplitudes is presented.

7.3.2 Spectral Analysis of Partial Amplitude Modulation

Due to time constraints, the subjective visual and auditory comparison of partial amplitude modulations were carried out only by the author and were not conducted blind. It is therefore reasonable to argue that such comparisons are open to bias. Therefore, in order to achieve a more objective measure of the temporal modulations plotted earlier, a preliminary investigation of their modulation spectra was conducted. It was hoped that such an investigation might ascertain whether or not perceptible degradations could be identified from these spectra. If so, such spectral analysis might prove to be a useful tool in automatically identifying the presence of reverberation degradation.

It was decided to investigate the modulation spectra of the partial at $4f_0$ of the vibraphone signal for non-prime, prime and real reverbs for reverb times of 2 and 3 seconds. These modulations are plotted as the top six curves in Figure 7.14 and were chosen because they exhibited a combination of graineyness and fluttering that was readily audible. The modulation spectra were generated by reconverting the amplitude partial modulations to linear amplitude, removing the DC component to eliminate bias, calculating the Fourier transform and displaying the result in dBs.

Figure 7.16 presents the modulation spectra for the non prime (green), prime (blue) and real (red) reverbs for a reverb times of 2 seconds, which corresponding the the 6th, 5th and 4th curves from the top in Figure 7.14. Figure 7.17 present the corresponding spectra for

the 3 second reverb. These are the spectra of the top three partial modulations plotted in Figure 7.14. All spectra are plotted up to a frequency of 80 Hz. It is uncertain how much of this frequency range is relevant to the perception of graineyness or fluttering. However, modulations beyond 80 Hz are unlikely to percieved as graineyness (and definitely not as fluttering), if they are perceptable at all.

It is interesting to note the striking correspondence between the two spectra for the non-prime reverbs and also the correspondence for the prime reverb spectra. Of course, changing the reverb time of a particular reverb does not change the echo pattern, merely the rate of decay of this echo pattern. However, it is very satisfying that the modulation spectral for different reverb times should match to such a high degree. This may be taken as a clear indication of the accuracy of the EWD and partial power estimation methodology adopted in this chapter.

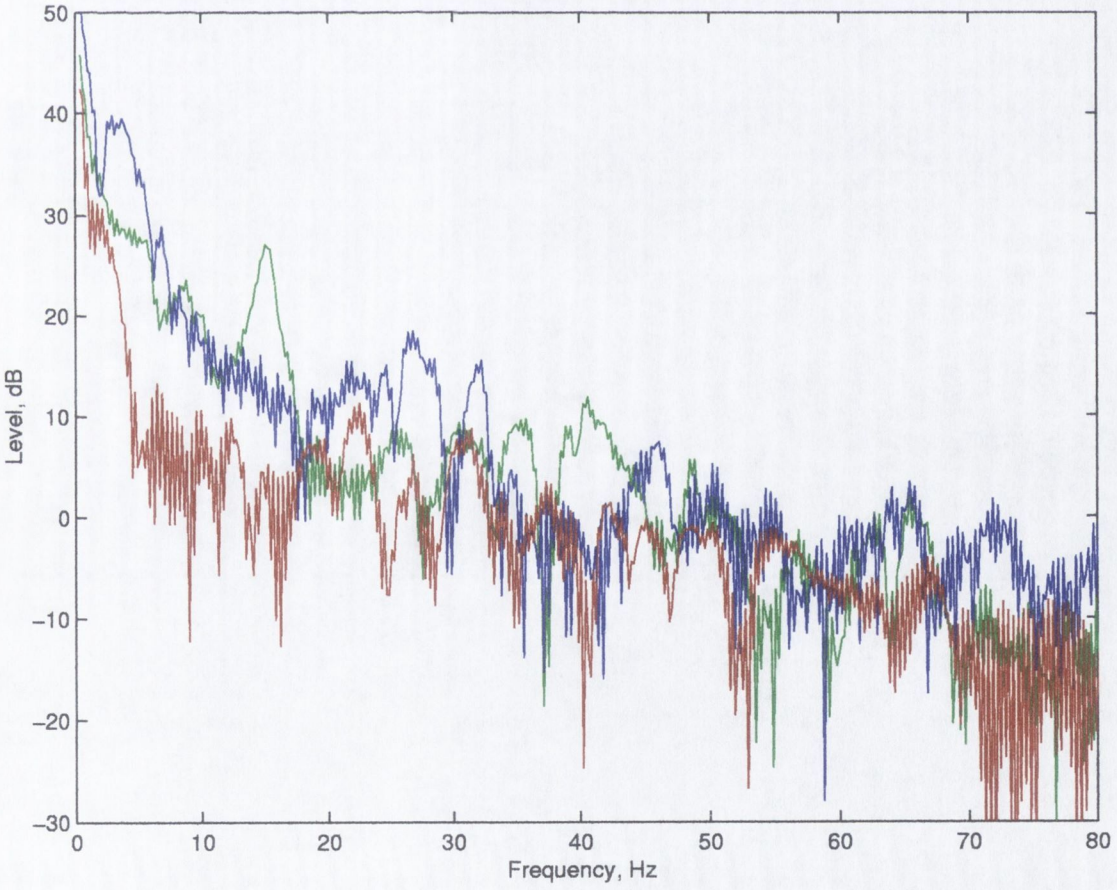


Figure 7.16: Modulation spectra of the vibraphone partial at $4f_0$ (1480 Hz) for the 2 second non-prime (green), prime (blue) and real (red) reverberations.

Looking at the first 20 Hz of the modulation spectra for the 2 second reverbs in Fig-

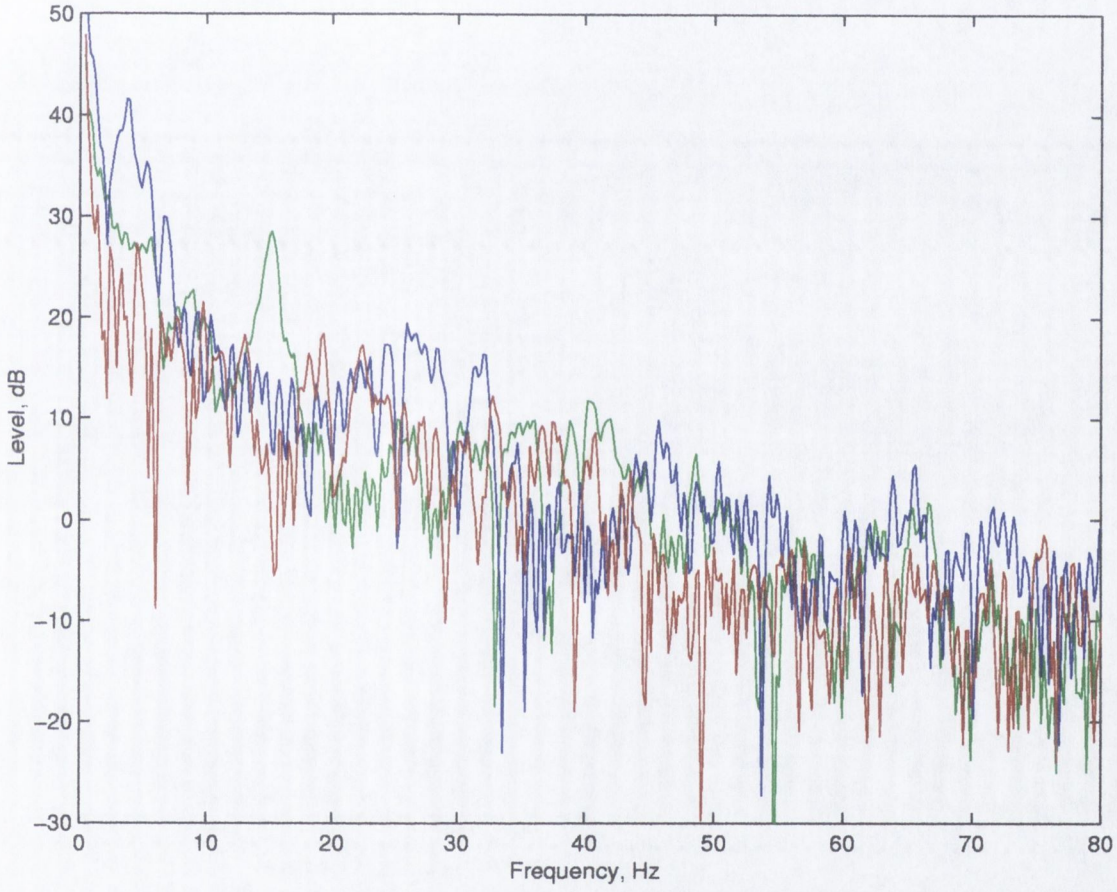


Figure 7.17: Modulation spectra of the vibraphone partial at $4f_0$ (1480 Hz) for the 3 second non-prime (green), prime (blue) and real (red) reverberations.

ure 7.16, we see that the real reverb's modulation spectrum (red curve) is concentrated below about 3 Hz. This is expected as the real reverb decay was very smooth for this partial. The spectral levels for the prime and non-prime reverb exceed that of the real reverb by about 10 dB for most of the frequency range 3-20 Hz. The modulation spectrum for the prime reverb (blue curve), which exhibited strong fluttering, has a broad peak at about 4 Hz. In contrast, the modulation spectrum for the non-prime reverberation, which exhibited strong graineyness, has a marked peak at about 15 Hz. These same peaks are also evident for the 3 second reverb. The modulation spectrum for the 3 second *real* reverberation has a higher spectral level than that of the 2 second real reverberation, especially in the 3-20 Hz region. As noted earlier, this may be attributed to the 3 distinct regions of mild graineyness exhibited by this partial. It is more difficult to comment on the modulation spectra beyond 20 Hz, as the differences between the real and non-prime and prime spectra lessen. Two lower level spectral peaks at about 27 and 32 Hz are noted for both the 2 and 3 second prime reverberations. Unlike the real reverb modulation spectra, the modultion spectra for the 2 and 3 second reverbs are remarkably similar across the whole frequency range displayed.

While the findings presented here are only preliminary, they suggest that the artifacts of graineyness and fluttering could perhaps be deduced from the partial modulation spectra of certain reverberated signals. An objective measure of graineyness and fluttering could perhaps be made based on the levels of the modulation spectra in certain frequency regions. The strong correspondence between the modulation spectra of the 2 and 3 second Schroeder reverbs do suggest that the echo pattern of a reverb has a consistent and identifiable effect on a signal's partial amplitudes. This correspondence also is a strong indication of the accuracy of the analysis approach take here. Therefore, the EWD is proposed an an excellent tool for pursuing this area further. However, further work is necessary before it can be ascertained whether or not an objective measure of the the perceived quality of a reverberant decay may be derived from the modulation spectra of reverberated signal partials. If so, appropriate criteria for reliable and objective measures of graineyness and fluttering may be arrived at.

7.3.3 Analysis of the Effects of Comb Filter Delays

In [115], Moorer notes the sensitivity of the Schroeder reverberator to the choice of filter parameters. As Schroeder points out in [168], "To avoid echo cancellation and superposition, it is advisable to use *incommensurate* delay ratios ..." For any two comb filters connected in parallel, echo superposition occurs periodically with period equal to the least common multiple (LCM) of the two delays. To ensure that the superposition is kept to a minimum, as high a LCM as possible is desirable. Thus, prime comb filter delays are generally pre-

ferred, the LCM of two prime numbers being equal to their product. The previous section demonstrates that very minor changes to the comb filter delays of a Schroeder reverberator can effect marked changes in reverberation quality. Row one and two of Table 7.4 present the comb filter delays used in Section 7.3.1 for the Schroeder reverbs with both prime and non-prime delays. The lowest LCM for each reverberator is also included. It is clear that the minor changes in the comb filter delays caused a large reduction in the lowest LCM from 1073 to 120. This change resulted in a noticeable increase in grainyness for all reverberated signals exhibiting this artifact. It also affected fluttering, albeit in a more unpredictable fashion, causing fluttering to increase for some signals and to decrease for others.

	D_1	D_2	D_3	D_4	D_5	D_6	Min. LCM
Prime A	29	37	41	47			1073
Non Prime A	30	40	42	48			120
Prime B	29	37	42	47			1073
Non Prime B					28	30	420
Prime + Non Prime B	29	37	42	47	28	30	84, 210

Table 7.4: Feedback delays for the prime, non-prime and variable-gain, non-prime comb-filter sections of the Schroeder reverberator. Delay time is in ms.

In order to investigate the influence of comb filter sections with low LCM values, a new Schroeder reverberator was designed, this time with 6 comb filter sections. The first four had delays which were almost identical to the prime delays presented in the first row of Table 7.4, with the exception of the substitution of the 42 ms delay for the 41 ms delay. This substitution was done to introduce the lower prime factors of 2, 3 and 7. However, because the other 3 delays of 29, 37 and 47 ms were still prime, the effect of this 42 ms delay on echo superposition for this 4-comb-filter section was negligible. The last 2 comb filters had non-prime delays of 28 and 30 ms. At 420, their LCM was still relatively high and thus was not responsible for excessively rapid echo superposition. However, the gain, G, at the output of the two-comb-filter section was variable, as shown in Figure 7.18. Thus, by increasing this gain, echo superposition could be markedly increased because the delays of 28 and 30 ms interacted with the 42 ms delay producing echo superposition every 84 and 210 ms respectively.

To examine the effect of gradually introducing more rapid echo superposition, the gain, G, of the two-comb-filter section was set to 5 values : 0, 0.25, 0.5, 0.75 and 1. Four different reverb times were investigated : 0.7, 1, 2 and 3 seconds. This gave a total of 20

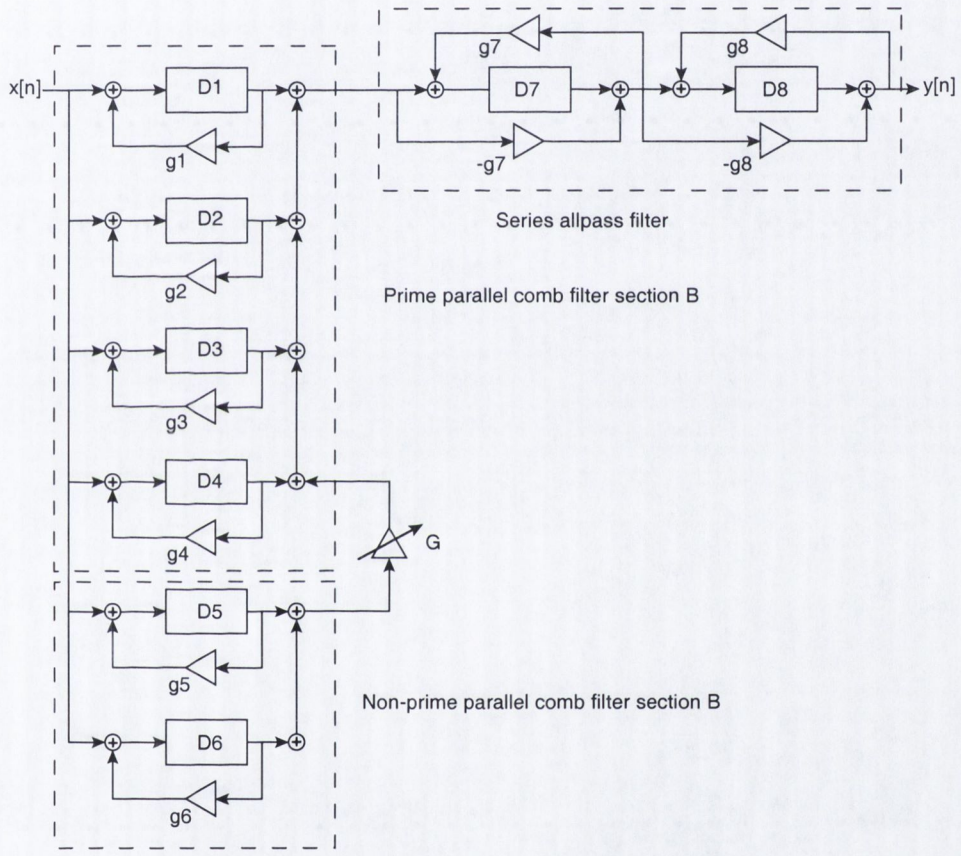


Figure 7.18: Schroeder's reverberator with additional variable-gain, 'non-prime' comb filter section.

reverberations which were convolved with the vibraphone signal to generate 20 reverberated vibraphone signals. The EWDs of these signals were computed and the temporal marginals of the signal partials were extracted in the exact manner outlined earlier in Section 7.3.1. It was hoped that such an analysis would produce a gradual change in the audibility of the reverberation artifacts of grainyness and fluttering that would be mirrored in the resulting EWDs and partial marginals. This indeed was the case, depending on the reverb length and partial examined. For the reverb times of 0.7 and 1 seconds, little change was noted as the gain, G, was increased apart from a slight increase in grainyness for the second partial at $4f_0$ (1480 Hz). This change was reflected by the EWD as a slight increase in rapid perturbations of this partial's temporal marginal. There was almost no change in the temporal marginals of the fundamental or partial at $9.5f_0$. The overall lack of change for these reverb times is not surprising as the Schroeder reverb algorithm is typically quite good sounding and free from undesirable audible artifacts, for short reverb lengths.

In contrast to the lack of change noted for the short reverb times of 0.7 and 1 seconds, the 2 and 3 second reverbs were much more influenced by the effect of increased echo superposition produced as G was increased. For the 2 second case, an increase in fluttering was observed for the fundamental as the gain was increased. Figure 7.19 portrays the temporal marginal of this partial. The plotting order, from bottom to top, corresponds to values of G of 0, 0.25, 0.5, 0.75 and 1, respectively. For the lowest 2 curves, a weak fluttering was apparent (and audible) for the interval from $t = 0.15$ to $t = 1.3$ seconds. For the upper 3 curves, this fluttering became less rapid, but more pronounced as G increased from 0.5 to 1. This behaviour was readily audible on attentive listening. The marginals of the partial at $4f_0$ for the 2 second reverb, shown in Figure 7.20, exhibited moderate fluttering, upon which some evidence of grainyness is superimposed. The grainyness worsens slightly as G increases. However, this effect was not audible on listening, perhaps because of its relatively low level in comparison to the simultaneous fluttering observed.

The temporal marginals in the fundamental for the case of the 3-second reverb, shown in Figure 7.21, are quite striking. The fluttering behaviour is very similar to that observed for the 2-second reverb, but it is more pronounced and audible. The more rapid fluctuations, occurring from $t = 1.8$ to $t = 3$ seconds were of too low a level to be audible, however, the re-emergence of the partial at $t = 3$ seconds was noticeable on careful listening. Strong fluttering was also observed for the partial at $4f_0$ for the 3-second reverberation. This is plotted in Figure 7.22. Furthermore, in contrast to the 2-second reverberation, the superposition of this grainyness on the fluttering was audible and worsened as the gain, G, increased. Finally, the temporal marginal of the partial at $9.5f_0$ exhibited considerable

grainyness for all reverb times. However this grainyness was not really audible since the level of the partial was very low. The grainyness was also relatively invariant to changes in G, but was more affected by changes in the reverb time, showing a greater amount of fluctuation for longer reverb times than shorter.

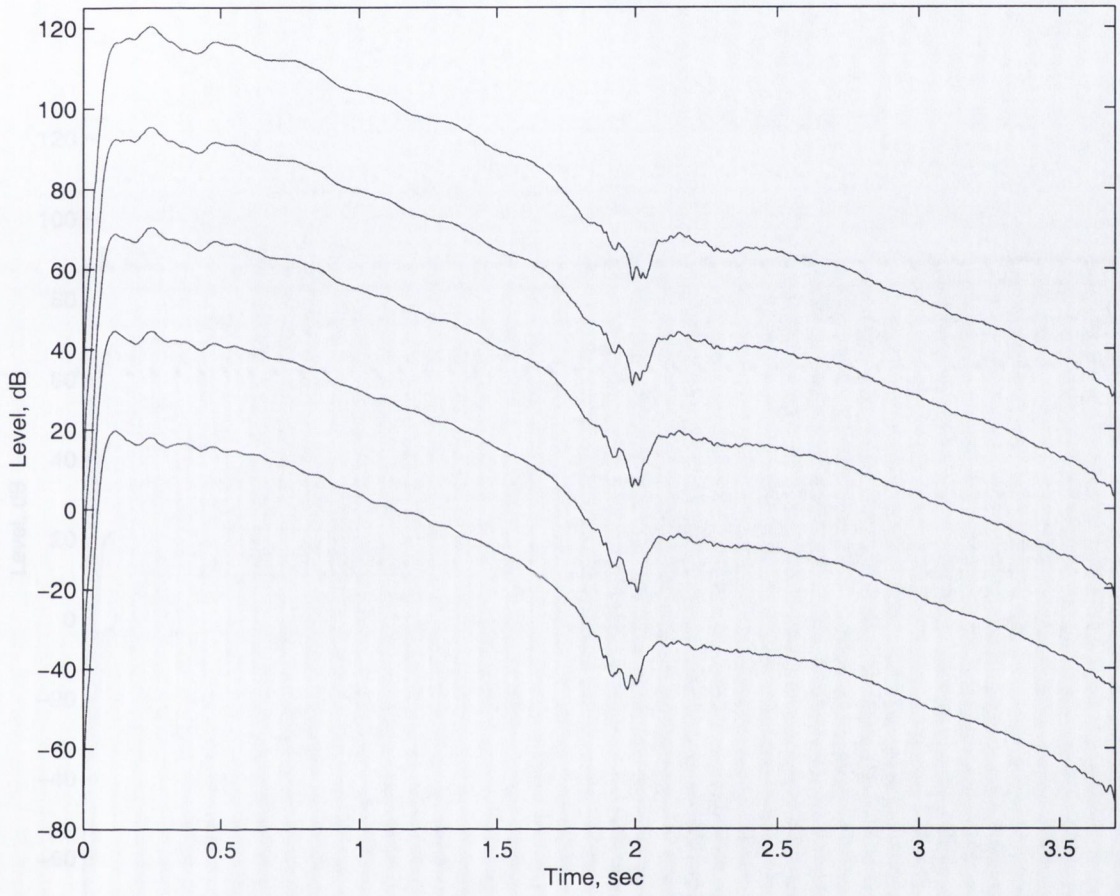


Figure 7.19: The temporal marginals for the fundamental of the vibraphone signal reverberated with the 2 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1.

Overall, in analysing the effect of gradually increasing superposition using the variable gain non-prime section in the manner outlined above, we were able to link a gradual increase in echo superposition to a gradual increase in fluttering and grainyness, depending on the particular partial chosen. The addition of the second comb filter section was shown to have a measurable and detectable influence on the reverberation, especially for the longer times of 2 and 3 seconds. The accuracy of the EWD in predicting these artifacts was well borne out by listening tests. The presence of the undesirable reverberation artifacts of grainyness and fluttering for Schroeder's algorithm was found to be more dependent on overall reverb

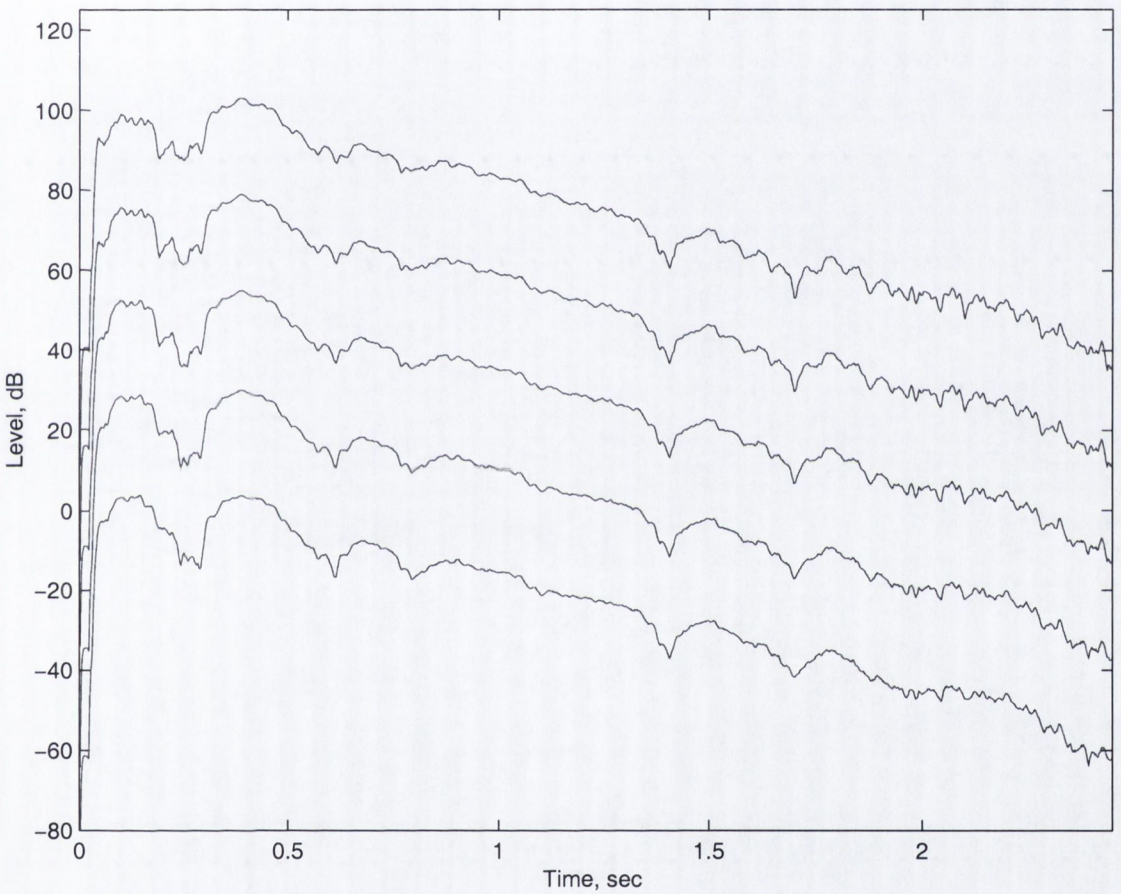


Figure 7.20: The temporal marginals for the partial at $4f_0$ of the vibraphone signal reverberated with the 2 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1.

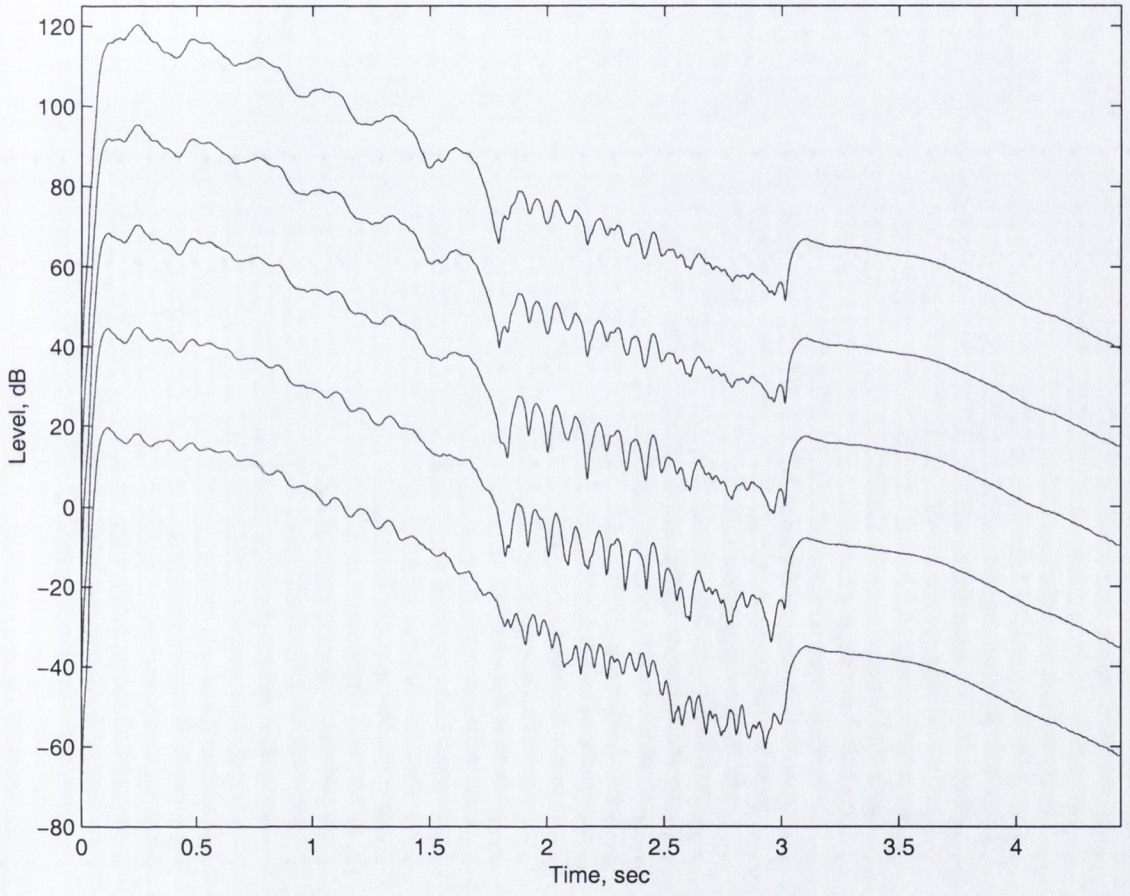


Figure 7.21: The temporal marginals for the fundamental of the vibraphone signal reverberated with the 3 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1.

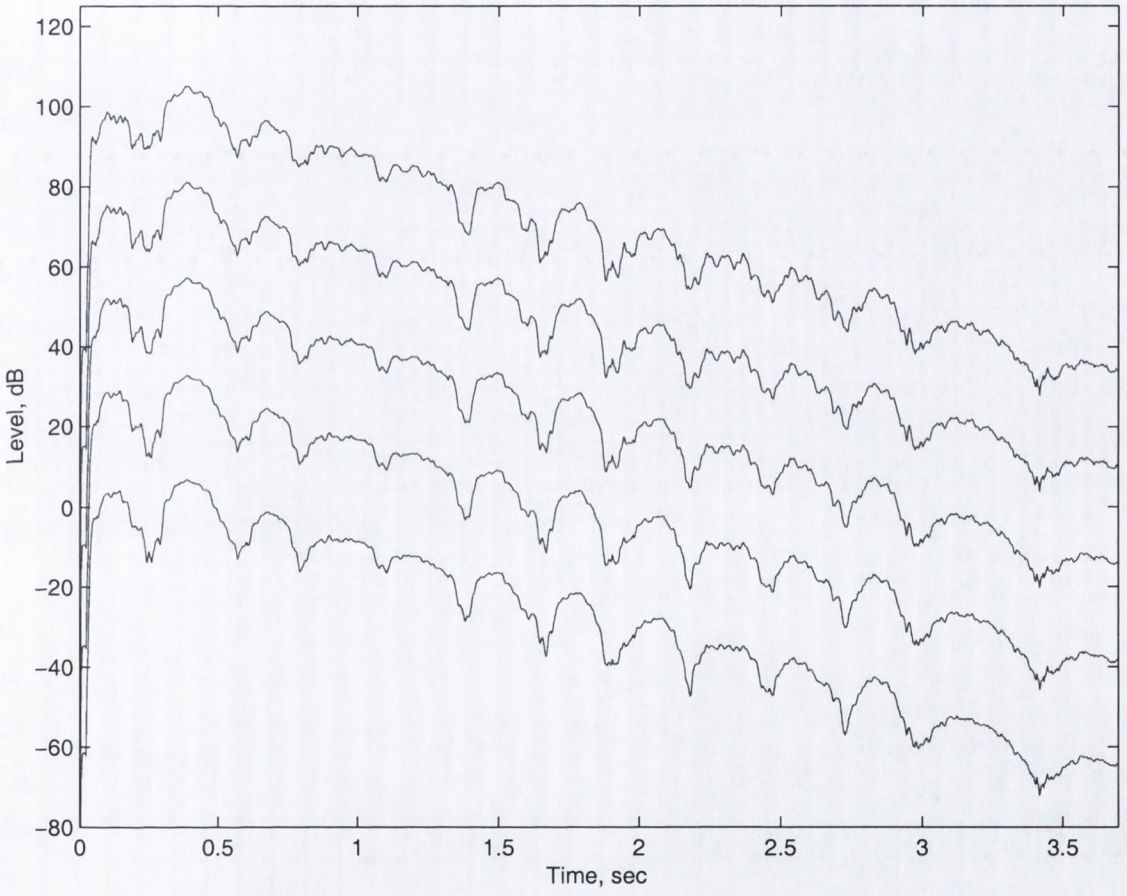


Figure 7.22: The temporal marginals for the partial at $4f_0$ of the vibraphone signal reverberated with the 3 second Schroeder reverb with variable gain section. The value of gain, G, from bottom to top is 0, 0.25, 0.5, 0.75 and 1.

time than comb filter delay. It should also be noted that although we present results for the vibraphone signal only, a correspondence between the EWD temporal marginals and the perception of audible artifacts was noted for all the signals exhibiting these effects. Therefore, the EWD proved very useful indeed in characterising the modulations noted in these artifacts and in predicting their appearance in reverberated signals.

7.4 Spectrogram Reverberation Analysis

For the purposes of comparison with the EWD, the spectrogram was also used to analyse many of the reverberated signals presented here. However, whether or not the temporal marginals of the signal partials exhibited the effects of reverberation 'distortion' depended strongly on the choice of analysis window used. In general, there was much less correspondence between the spectrogram and the auditory perception of the reverberation artifacts discussed here than there was for the EWD. Furthermore, as demonstrated in Chapter 5 Section 5.2.3 it is not possible to simultaneously model both temporal and spectral masking with the spectrogram. In fact it is not even possible to model either critical band masking or temporal masking individually, since neither the time nor frequency resolutions of the spectrogram may be made frequency dependent.

In order to illustrate these shortcomings, Figures 7.23, 7.24 (a) and 7.24 (b) present a comparison of the EWD and spectrogram for the (unreverberated) violin signal used in Section 7.3. The spectrogram windows chosen are exactly the same as those used in Chapter 5, Section 5.2.3 to compare the spectrogram and EWD for a complex tone signal. Figure 7.24 (a) and (b) display the spectrograms computed using the temporal window shape and gamma window respectively, again both at centre frequency 1 kHz. It is very evident that neither spectrogram models the spectral masking clearly shown in Figure 7.23 for the EWD. The spectrogram therefore fails to hide masked spectral detail. The partial modulations shown by the EWD and spectrogram computed with the temporal window shape correspondence well below about 1000 Hz. However, for increasing frequency the spectrogram continues to register signal partials which should be spectrally masked.

In Figure 7.24 (b) a gamma window is chosen as the spectrogram analysis window. Using this window now gives us the correct spectral resolution at 1000 Hz but not at any other frequency. Also, the temporal resolution is equal to the gamma window time domain response and does not therefore model temporal masking. Thus neither temporal nor spectral masking effects are properly modelled. In this work it is intended to study the perceptable effects of reverberation degradation. It is therefore important that the

time frequency distribution chosen for such work incorporate the main features of auditory processing. The spectral and temporal masking properties of the ear are two features which are very relevant to auditory processing. It has been clearly demonstrated that the spectrogram cannot model these phenomena. Therefore, for these reasons, the EWD was deemed a superior method of investigating reverberation degradation than the spectrogram.

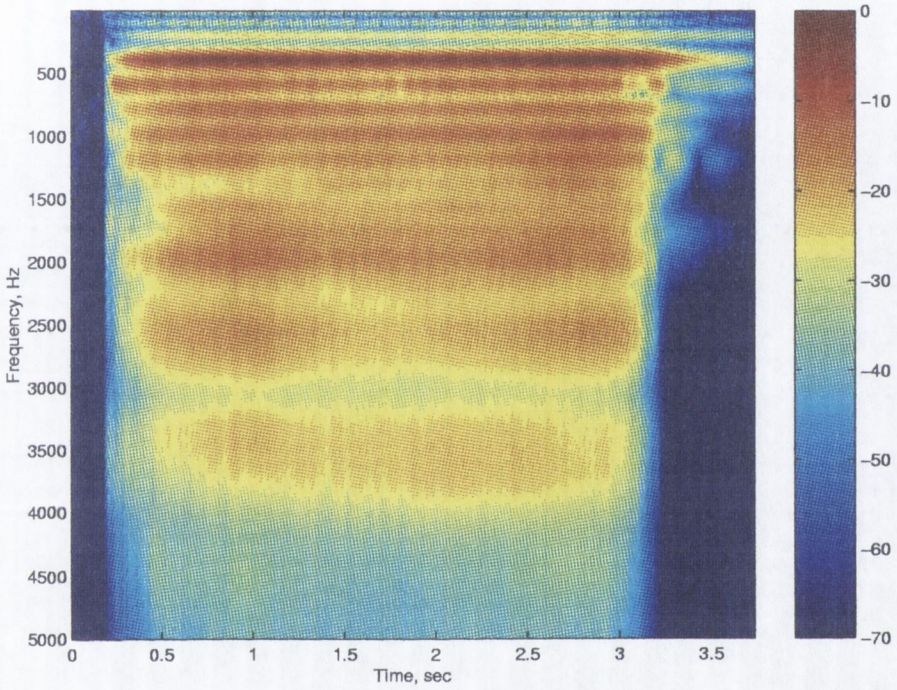
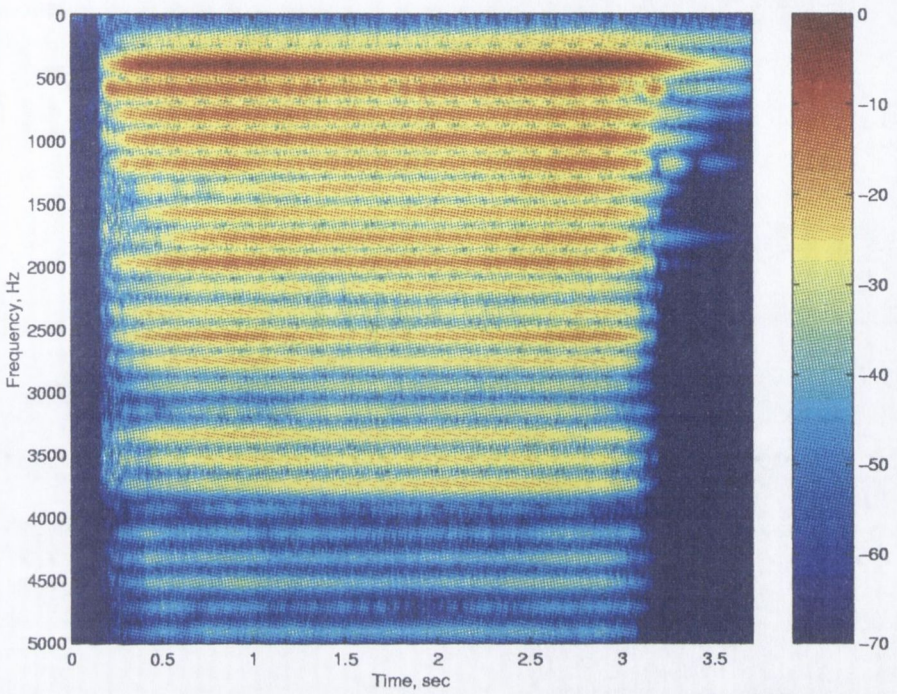


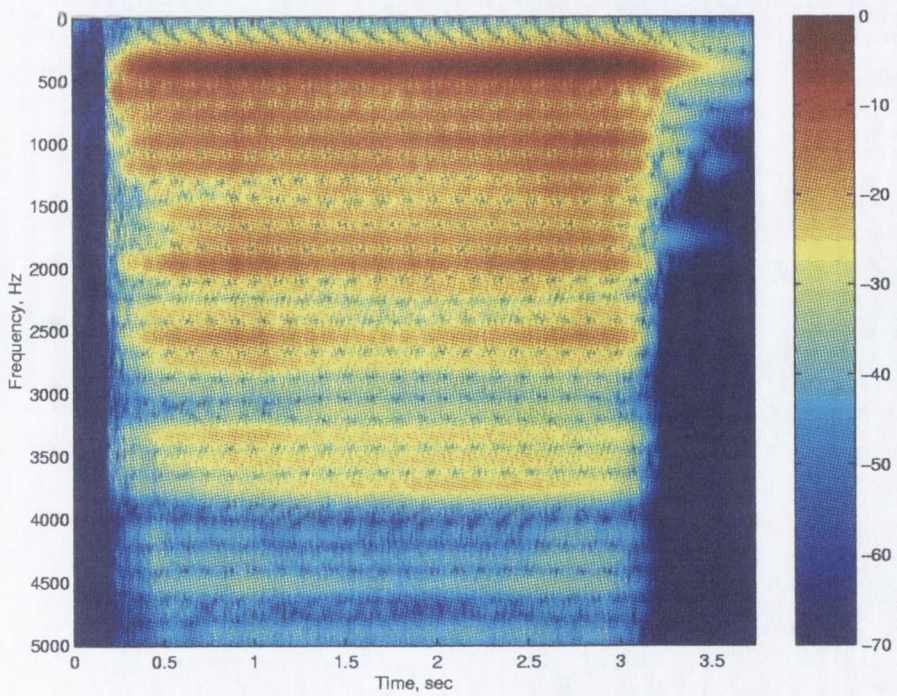
Figure 7.23: EWD of an unreverberated violin signal. Amplitude is in dBSPL.

To summarise, in this chapter two distinct artifacts of the Schroeder reverberator were investigated in detail, namely grainyness and fluttering. In general, the characterisation of these artifacts by the EWD was consistent across a wide range of signals. Grainyness was characterised by a rapid and random temporal fluctuation in the amplitudes of certain reverberated signal partials, whereas fluttering was caused by a slower, almost periodic partial modulation. Overall, these artifacts and other facets of signal perception such as the 'hearing out' of upper harmonics were mirrored in the EWD time-frequency representations to a high degree. However, the results presented should only be viewed as a pilot study, since these experiments were conducted by the author only. Further objective experimentation is needed to fully evaluate the effectiveness of the EWD at mirroring the perception of artifacts of reverberation.

The EWD allowed a link to be formed between the physical properties of reverberated signals and their resulting perception. The main benefit of the EWD analysis presented



(a)



(b)

Figure 7.24: The spectrograms of an unreverberated violin signal using (a) the temporal window shape centred at 1000 Hz and (b) the gamma window centered at 1000 Hz as the analysis window. Amplitude is in dBSPL.

was that, through comparison with natural sounding real reverberation, a more objective understanding of the causes of synthetic reverberation degradation was arrived at. A preliminary investigation of the modulation spectra of signal partials demonstrated that it may well be possible to use the EWD to objectively measure reverberation degradation.

The mathematical framework provided by Cohen's class and the partial power estimation techniques of Pielemeier allowed an accurate analysis and comparison of the temporal marginals of the signal. Again, the lack of cross-term or lobing artifacts exhibited by the EWD was crucial to the accuracy of the analysis. The success and accuracy of the EWD in characterising the reverberation artifacts and in reflecting signal perception are seen as further strong evidence of its utility as a perceptually relevant audio signal analysis tool.

Chapter 8

Summary and Conclusions

The objective of this work was to develop a new approach to auditory modelling and audio time-frequency analysis using a member of Cohen's class which incorporates the frequency dependent temporal and spectral resolutions of the ear. This is a departure from the conventional time-frequency design approach of simply striving for the best time and frequency resolutions possible while keeping cross-term interference to a minimum. To the best of the author's knowledge, no successful attempt at auditory modelling has been previously made using a Cohen class distribution. This is possibly due to the problems of cross-term interference, windowing artifacts, clutter and high computation and memory requirements usually associated with Cohen's class of bilinear distributions. To date, standard window functions rather then psychoacoustic models of masking have generally been the norm for the design of smoothing kernels for members of Cohen's class.

In this thesis is it demonstrated that the computation time and memory requirements of Cohen's class may be significantly reduced if the modal distribution design approach of Pielemeier is adopted. Furthermore, it is demonstrated how a Cohen's class distribution may achieve auditory temporal and spectral resolutions by using frequency dependent separable kernel. The frequency smoothing kernel windows are designed using a novel approach which utilises a baseband version of the popular gammatone filter function as a smoothing window function. This method of achieving auditory spectral resolution is both simple and accurate and is also applicable to asymmetric filter shapes. The response of the EWD at a particular time yields the instantaneous excitation pattern of a signal. Integration of the EWD along the time axis to extract the EWD frequency marginal yields a (temporally averaged) excitation pattern which is shown to be equivalent to that produced by the conventional approach to excitation pattern generation. Temporal masking is incorporated into the design of the EWD using a frequency dependent temporal smoothing window based

on a parameterisation of the temporal window model of masking. Again, this is a novel approach to kernel design. Therefore, the EWD simultaneously and independently models both temporal and spectral masking. It is clear that the use of Cohen's bilinear class of TF distributions, rather than the use of linear methods such as the wavelet transform or spectrogram, simplifies greatly the task of simultaneously achieving the auditory temporal and spectral resolutions required.

The EWD's kernel is not level dependent. This work demonstrates that this is not a serious limitation for the modelling of spectral masking, at least for moderate signal levels. However, temporal masking is strongly dependent on signal and masker level. Therefore, the accuracy of a level independent model of temporal masking may be somewhat limited, especially for signals with abrupt changes in level, where temporal masking effects become more noticeable. As the approach taken in this thesis to auditory modelling was a novel one, level dependency was not incorporated at this stage due to time constraints. However, because of the flexibility afforded by Cohen's class, kernel level dependency may readily be achieved, if desired. Furthermore, since the kernel is separable, time and frequency resolutions may independently be set. This means that revisions to psychoacoustic models of temporal or spectral resolution may easily be incorporated into the EWD. Changing the EWD's temporal resolution will have no effect on its frequency resolution and *vice versa*. This is not the case for filterbank based models of auditory processing or auditory models based on linear time-frequency distributions. The EWD has been shown to have a computational complexity which is comparable to filterbank models of auditory processing. Therefore, the EWD may in the future provide a more flexible auditory modelling approach which does not incur an excessive computational load.

The prospect of an inverse EWD is very interesting. It is well known that a signal may be recovered up to a constant phase factor from its WD [28]. It is also possible, however, to *synthesise* signals from modified Wigner distributions [22], pseudo-Wigner distributions [192], [91] and smoothed Wigner distributions [92]. TF signal synthesis allows a very flexible approach to signal filtering to be adopted. For example, the removal of unwanted signal components can be achieved by graphically removing or 'masking' noise components in TF and reconstructing the signal. This is equivalent to time-varying filtering, since the areas in TF where the noise mask is applied are completely at the discretion of the user and may change with time. Also, signals can be synthesised by graphically designing an imagined TF distribution and synthesising the resulting signal.

In this work, algorithms for the inverse WD, PWD and SPWD were implemented and tested by the author. Due to time limitations, these algorithms were not applied to the

EWD and were therefore not included in this thesis. However, there is a strong possibility that an inverse EWD algorithm may be designed. Since the EWD removes masked signal detail, one interesting application area for an EWD analysis/synthesis algorithm would be the removal of masked signal information from audio signals. This could have application as a front end for automatic speech recognition algorithms, for example. Another possible application area would be the analysis/synthesis of audio equalisation filters. However, accuracy and lack of noise and other perceptable degradation is crucial to the success of inverse EWD algorithms for such applications. It is still not certain whether existing inverse TF algorithms can meet these critera.

Using a Cohen's class distribution means that the possible detrimental effects of cross-term interference must be tackled. However, this work shows that, interestingly, no extra smoothing beyond that provided by the EWD smoothing kernel is required to suppress cross-term interference below auto-term excitation level. The wide variety of signals analysed by the EWD throughout the thesis clearly demonstrate that cross-term interference is not manifest for typical broadband signal analysis tasks such as music or speech analysis. Cross-terms are shown to occur only for very specific, worst-case scenarios where the signal to be analysed consists solely of narrowband components spaced very far apart in frequency. Furthermore, the numerous examples presented in this work serve to highlight the lack of clutter present in the EWD. This lack of clutter is due to the fact that the windows used in the design of the EWD kernel are themselves smooth and free from artifacts such as sidelobes over a very high dynamic range. The care taken in the design of these windows might seem excessive at a first glance, yet it is necessary to ensure that the dynamic range of the EWD is sufficient to encompass that of the human auditory system. Thus, the EWD is in itself only limited by the limitations of the models of temporal and spectral resolution adopted. These limitations are discussed in this thesis and are not deemed serious for its intended application. Furthermore, the separable kernel approach allowed by Cohen's class means that it is relatively straightforward to incorporate more complex models of auditory resolution such as temporal resolution level dependency, if required.

In order to demonstrate the practical application of the EWD to real problems in acoustics, two interesting and important areas of audio research were investigated, namely that of 3-D surround-sound design and artificial reverberation analysis. For the former, the EWD was used to analyse HRIR data in the hope that it might provide a novel approach to reducing inter-subject variability in these measurements. This was carried out by utilising the mathematical properties of Cohen's class and extracting the frequency marginals of the EWDs of individual HRIR measurements. Each frequency marginal extracted gave the

excitation pattern, or critical-band smoothed HRTF spectrum, containing only unmasked HRTF spectral detail. This data was used in a 2-speaker surround-sound implementation designed to produce horizontal surround-sound panning. An informal comparison of the EWD with existing PCA panning models found the EWD to be very successful at the task of HRTF smoothing and it outperformed PCA smoothing in terms of both naturalness and localisation. However, a more extensive and objective comparison of EWD and PCA smoothing is requred to fully assess the advantages of EWD smoothing.

Inter-subject variability in HRTF measurements is generally much worse at high frequencies than at low. Such measurements often contain discontinuous high frequency spectral detail beyond the resolving power of the human auditory system. The smoothing performed by the EWD was thus ideally suited to the removal of such unwanted detail as, being psychoacoustically based, this smoothing removed only masked spectral detail. Thus, the EWD successfully reduced inter-subject variability at high frequencies while still preserving overall trends in the high frequency regions of the HRTF data. Spectral detail at low frequencies was subject to much less smoothing because of the narrower critical bandwidths at these frequencies. This was appropriate because of the lower levels of inter-subject variability in HRTF measurements exhibited at such frequencies. In a final analysis, the averages of the unsmoothed and EWD-smoothed HRTF data were compared in terms of naturalness and localisation. It was demonstrated that even minor roughness in HRTF (or DTF) data resulted in a noticeable degradation in naturalness once the final loudspeaker feeds were calculated. The smoothness and lack of clutter in the EWD was shown to be vital in ensuring that the EWD smoothed data was free from roughness. Such smoothness over a high dynamic range is necessary for serious audio applications.

The second EWD application area explored in this work was an analysis of reverberation data. This was conducted with the aim of achieving a better understanding of the causes of unpleasant audible artifacts of some artificial reverberation algorithms. With this in mind, the Schroeder reverberator was selected for investigation because its flaws are well documented in the literature. For long reverb times, Schroeder's algorithm can suffer from a grainy sounding initial response, a fluttering decay and also a metallic timbre. Also, the algorithm is quite sensitive to even minor changes in the feedback delays of its comb filter sections. Usually performance is best if the delays are incommensurate or mutually prime. Whether or not a Schroeder reverberated signal suffers from audible artifacts also depends on the signal itself. Because of this, a wide variety of signals were initially convolved with real reverberation responses and 'prime' and 'non-prime' Schroeder reverberation responses of various lengths. The Schroeder reverberation was found to lead to the presence of fluttering,

grainyness and a metallic timbre for some of the reverberated signals, whereas the real reverberation did not introduce these artifacts for any of the signals investigated. Grainyness and colouration worsened as the reverb length increased and was more noticeable for non-prime delays than prime. Fluttering also worsened as the reverb time increased. However, the effect of comb filter delay changes on fluttering was less consistent. Since the aim was to investigate and characterise the phenomena of grainyness and fluttering, the vibraphone signal was chosen as a good candidate signal for further analysis because the Schroeder reverberated vibraphone signal suffered from both of these artifacts.

Schroeder reverberation responses of various durations, with both prime and non-prime comb filter delays, were initially generated and were convolved with the vibraphone signal. EWD analysis revealed interesting temporal fluctuations in some of the signal partials. By locally extracting the temporal marginals of these partials, using the power estimation technique formulated by Pielemeier, it was discovered that fluttering and grainyness were manifest as two different types of partial modulations. Grainyness was characterised as a rapid, noise-like perturbation of signal partial amplitudes, whereas fluttering was characterised as a slower, almost periodic modulation. A preliminary investigation of the partial modulation spectra of the reverberated vibraphone signal also highlighted the possiblility of using such spectra to objectively measure reverberation degradation.

It is a well established fact that the quality of the Schroeder algorithm is quite sensitive to the choice of delays used in the parallel comb filter structure. In order to minimize the number of echo cancellations and superpositions, mutually prime or incommensurate delay times are generally preferred. To test the effect of gradually introducing 'non-primeness' into the comb filter delays discussed, a new Schroeder reverberator was designed consisting of a 4-comb filter section using mutually prime delays and a second, variable gain section comprised of two comb filter sections. The two delays of this section were deliberately chosen to be mutually prime with each other, but non-prime with the 4-comb filter section. Therefore, by varying the gain of this second filter, overall reverberation echo superposition and cancellation could be controlled. The vibraphone signal was again chosen as the candidate signal for reverberation. It was noted that as the gain of the non-prime section was increased, grainyness and fluttering worsened, especially for the longer reverb times. Again, this was very accurately mirrored in the signal partials extracted from the EWD. Due to limited time, these experiments were all conducted by the author only and are therefore open to bias. A more objective method of determining the correspondance between partial amplitudes extracted from EWD and their perception is required to fully determine the ability of the EWD to predict the perception of reverberation degradation.

Because the EWD is virtually cross-term and clutter free, and because the temporal marginals of the unreverberated signals did not exhibit the modulations present in the Schroeder-reverberated data, it reasonable to confidently assert that such modulations must be artificial reverberation artifacts alone. Thus, using the EWD, it is possible to link the distinct percepts of grainyness and fluttering with the physical partial modulations presented in the reverberated signals. Due to time constraints, no further investigation of reverberation was possible. However the results presented here are encouraging and clearly demonstrate the suitability of the EWD to such work. To the best of the author's knowledge, this work is an entirely novel approach to reverberation analysis. The field of audio and acoustics is a very rich and diverse one and the author feels that there are many other potential application areas which could benefit from EarWig analysis. The rest of this chapter is therefore devoted to a few suggestions.

8.1 Loudspeaker, Microphone and Equalisation Impulse Response Analysis

It is a well known and often perplexing fact that different brands of loudspeakers or microphones sound different, even though their manufacturers may all (usually truthfully !) claim that their responses are flat across the audible frequency range. Therefore, these audible differences must be due to differences in the phase responses of the loudspeakers or microphones or else transient effects of their impulse responses. Although the ear is sometimes assumed to be phase-deaf, this assumption only really holds true for steady state sounds. The ear is much more sensitive to phase changes in transient or short-duration sounds, which are common in music and speech. Therefore, the phase response of a loudspeaker can have an important effect on the subjective quality of the reproduced sound. Most modern 'hi-fi' speakers contain more than one transducer to cope with different frequency ranges, using for example a 'woofer' for the low-frequency range and a 'tweeter' for the high-frequency range. In the crossover region between woofer and tweeter, abrupt phase changes or phase distortion can occur which can produce audible effects in some circumstances [108].

In a similar manner to loudspeakers and microphones, studio equalisation units generally also have their own 'sound'. Gerzon [61] has argued that some of the audible effects of equalisation cannot be measured simply in the magnitude or the phase responses of these units. He suggests that low-level, transient effects of the equalisation impulse responses may also have a surprisingly noticeable effect. In particular, filter pre-responses, well away from the main impulse of the equalisation filter and in some cases up to 80 dB (!) down from the

impulse response peak, have been known to cause audible colouration. The audibility of pre-responses is inherently dependent on backward temporal masking. Gerzon notes that post-responses are not as audible as pre-responses—this is generally assumed to be because of the dominance of forward masking over backward masking.

Gerzon points out that joint time-frequency analysis is necessary to further investigate these phenomena. He argues that "... by time/frequency analysis using a carefully-smoothed version of the Wigner distribution, in future we may have the tools to see what filters and equalisers are doing in the time and frequency domains with more detail than was previously possible. Designers of the software packages for audio analysers need to both master the relevant mathematical tools and to design the required smoothing filters in the software to avoid the kind of discontinuity or oscillation behaviour we are looking for in the hardware audio filters and equalisers we want to analyse. In other words, the designer of the analyser software requires the same kind of skills required to design good audio equalisers."

While there has been some investigation of filter responses [158], [157] and loudspeaker impulse responses [80] using the smoothed Wigner distribution, the dynamic range and cross-term suppression achieved in such work does not even approach that deemed necessary by Gerzon, which should equal or better that of the auditory system. Also, such work completely ignores auditory masking effects. The EWD has been shown to clearly fulfil the criteria of cross-term suppression, dynamic range and lack of clutter and discontinuity considered necessary by Gerzon. It also has the great advantage of incorporating both spectral and temporal masking and thus providing the engineer with a more useful means of deciding whether physical impulse response features are perceptible or not. Therefore, the EWD is recommended as an ideal tool for pursuing this area further.

8.2 Musical Instrument Analysis and Synthesis

Time frequency analysis has seen much application to the analysis of musical signals and musical timbre. An excellent review of this work is presented in [147]. The modal distribution described in Chapter 5 has recently been applied to an analysis of violin vibrato [104] in an attempt to characterise the nature of the partial frequency and amplitude modulation responsible for this effect. The amplitude and frequency estimates extracted using the power and frequency estimators developed by Pielemeier [148] are used in an additive-synthesis model to generate synthetic violin signals. The synthetic signals are then compared subjectively with the original signals to estimate the effectiveness of the technique. It is a well known psychoacoustic fact that has been clearly demonstrated in this thesis that usually

only the first 5-7 harmonics of most musical tones may be 'heard out', higher harmonics being either partially or fully masked. However, the modal distribution analysis of the violin vibrato described above treats all harmonics equally as it does not incorporate any form of spectral masking. Furthermore, because of the temporal masking effects of the ear, much low-level or rapid modulations are likely to be inaudible. Therefore, the synthesis process adopted above may be argued to be ineffective as inaudible signal properties are recorded and synthesised with the same rigour as perceptible signal properties. The author therefore recommends a more perceptually motivated approach such as the EWD. Using the EWD, synthesis effort may be better focused on perceptible and important signal features. Evidence in support of this is contained in the analysis of reverberation carried out in this work, which demonstrated an excellent correspondence between the modulations present in the reverberated signal partials, as shown by the EWD, and their perception.

8.3 Automatic Artificial Reverberator Design

In Chapter 7, a study of some undesirable audible artifacts of artificial reverberation was conducted. The percepts of grainyness and fluttering were linked to distinctly different signal partial amplitude modulations introduced by the artificial reverberation process. These modulations were absent from the corresponding partial amplitudes of the signals reverberated with real reverberation responses. In the study presented in this thesis, only the Schroeder reverberator was investigated, its failings being well documented in the literature. It is therefore suggested that this study be continued to other, more sophisticated reverberation algorithms, with the aim of gaining a more complete characterisation of artificial reverberation degradation. If achieved, such knowledge might facilitate the automatic detection of undesirable artifacts of reverberation by computer analysis, rather than subjective assessment, as is currently employed. This may be of considerable benefit to the process of reverberator design as presently the 'tuning' of artificial reverberation parameters must be carried out by conducting time-consuming listening tests.

In some reverberation designs, the aim may be to achieve an artificial reverberator which mimics the perceptually important characteristics of a real performance space. In this instance, an iterative, computer-controlled, reverberation-tuning process could perhaps be designed. Such a process could be set to repeatedly tune the parameters of an artificial reverberation algorithm until the EWDs of artificial and synthetic reverberations matched to a sufficiently high degree. In this manner, much of the drudgery could perhaps be eliminated from reverberator design and the number of subjective listening tests could be

significantly reduced.

8.4 Virtual Auditory Display Robustness

Chapter 6 demonstrated that EWD based processing was very successful at reducing inter-subject variability in HRTF measurements. Here, PCA and EWD smoothed HRTFs were subjectively evaluated in terms of imaging naturalness and accuracy, both of which are crucial to the design of effective virtual auditory displays. However, another important issue central to auditory display design is imaging *robustness*. As demonstrated in Chapter 6, the recreation of localisation cues relies on a knowledge of the position and orientation of the listener's head. If this changes considerably then virtual imaging accuracy may suffer, often resulting in the virtual image veering towards the closest loudspeaker. The term robustness therefore refers to the ability of a given virtual auditory display to tolerate movements of the user's head.

In the work presented in Chapter 6, the EWD was primarily used to smooth HRTF data. As a further application of the EWD to 3-D surround sound design, a more detailed analysis of virtual auditory displays similar in spirit to the EWD analysis proposed above for the analysis of microphone, loudspeaker and equalisation responses is proposed. This could be carried out by measuring the response of a given virtual auditory display for a certain azimuth and elevation, using a dummy head microphone. A systematic and measured degree of (dummy) head movement could be introduced into these measurements to simulate listener head movement. An investigation and comparison of the EWDs of the resulting HRIR data could prove quite insightful. If a virtual auditory display was known to lack robustness for rotational movements of the head, for example, then an analysis of its responses with equivalent and known dummy head rotations, using a perceptually motivated approach such as the EWD, could help to identify the causes of poor imaging robustness.

8.5 Speech Processing and Auditory Scene Analysis

Two further applications of the EWD are in the areas of speech processing and auditory scene analysis. Given the ever increasing interest in automatic, robust speech analysis and speech recognition, any approach which intelligently reduces speech signal redundancy would be useful. It is proposed that pre-processing of speech using the EWD may provide a more perceptually relevant starting point for subsequent speech analysis, given that perceptually redundant speech features would be eliminated before higher order analysis was

commenced. The possibility of using an inverse EWD algorithm to re-synthesise a speech signal with masked features removed is also of great interest here.

A related, but more general, application area would be that of auditory scene analysis or ASA. The foundations of ASA have been established over the last 30 years by Albert Bregman in McGill University [24]. His work has focussed on auditory grouping and streaming phenomena in many domains and has involved much laboratory testing in order to establish the principles by which such grouping and streaming occurs. Subsequent work by other researchers, for example the work of Cooke [35], [37], [34], has focused on computational ASA. Here the objective is to automatically identify auditory groups and streams in a given audio signal, thereby allowing, for example, the automatic segregation of different instrument signals from a single channel audio mix.

The basis for all ASA analysis is a joint time-frequency representation (usually a spectrogram) as spectral and temporal features both influence grouping and streaming processes. In this context, and particularly given the perceptual rather than physical focus of ASA, it is suggested that EWD pre-processing would provide a much firmer foundation for subsequent grouping and streaming analyses which would include such issues as spectral and temporal feature proximity, similarity, continuity *etc.* The perceptual motivation of the EWD and the significant reduction of processing artifacts in EWD representations would suggest that EWD pre-processing would hold much promise for future explorations in computational ASA.

8.6 A Final Note

Interest in joint time-frequency signal analysis has increased dramatically over the last 20 years or so. Because of its simplicity, the spectrogram is still the most widely used time-frequency distribution, but as the proliferation of more advanced TF techniques increases, the dominance of the spectrogram is dwindling in favour of more modern, flexible and accurate TF methods. In tandem with this, the advent of new consumer technologies, such as mobile communications and digital audio, and the demand for a more immersive and realistic aural experience than provided by stereo panning, has fuelled an increased interest in the workings of the ear.

The advantages of joint time-frequency audio analysis over either time or frequency analysis alone are clear. A similar case may be argued for a simultaneous model of both temporal and spectral masking. With this in mind, it is appropriate to apply TF techniques to the task of simultaneously modelling spectral and temporal masking. However, for this

to be effectively achieved, the resulting TF distribution resolutions should accurately match those specified in the psychoacoustic literature over a wide dynamic range. Furthermore, issues such as cross-term interference and computational complexity must be tackled if bilinear TF methods are to be effectively employed. This thesis demonstrates that it is indeed possible to accurately incorporate psychoacoustic models of auditory resolution in a bilinear TF distribution and still overcome the difficulties normally associated with such distributions. As such, it is expected that the work presented here will be of considerable interest to researchers in the areas of TF signal analysis, acoustics or psychoacoustics. We look forward to the many future developments in the domain of audio engineering and hope that the EWD may be put to good use in facilitating some of these developments...

Bibliography

[1] HEAD Acoustics. *Binaural Analysis System Manual, app. B.* Herzogenrath, Germany.

[2] F. T. Agerkvist. Time-frequency analysis with temporal and spectral resolution as the human auditory system. In *Proceedings of the IEEE-SP International Symposium Time-Frequency and Time-Scale Analysis*, pages 425–428, 1992.

[3] F. T. Agerkvist. A time-frequency auditory model using wavelet packets. *Journal of the Audio Engineering Society*, 44(1-2):37–50, January-February 1996.

[4] J. B. Allen, J. L. Hall, A. Hubbard, S. T. Neely, and A. Tubis, editors. *Peripheral Auditory Mechanics.* Springer-Verlang, Berlin, 1986.

[5] J. B. Allen and S. T. Neely. Micromechanical models of the cochlea. *Physics Today*, pages 40–47, July 1992.

[6] R. A. Altes. Wide-band, proportional-bandwidth wigner-ville analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 67:1005–1012, 1990.

[7] T. H. Applebaum and B. A. Hanson. *Perceptually-based dynamic spectrograms*, pages 153–156. In Cooke et al. [36], 1993.

[8] University of Cambridge Auditory Perception Group. Auditory demonstrations and useful software. *http://hearing.psychol.cam.ac.uk/Demos/demos.html.*

[9] F. Auger, P. Flandrin, O. Lemoine, and P. Gonçalvès. Time-Frequency Toolbox for Matlab.
http://www.syntim.inria.fr/fractales/Software/TFTB/, May 1998.

[10] F. Auger, P. Flandrin, O. Lemoine, and P. Gonçalvès. Time-Frequency Toolbox for Matlab - Tutorial.
http://www.syntim.inria.fr/fractales/Software/TFTB/, May 1998.

[11] Various authors. Special Issue on Time-Frequency Analysis. *Proceedings of the IEEE*, 84(9), September 1996.

[12] Bang and Olufsen. Archimedes CD. 1992.

[13] P. Basile, F. Cutugno, P. Maturi, and A. Piccialli. *The time-scale transform method as an instrument for phonetic analysis*, pages 169–174. In Cooke et al. [36], 1993.

[14] S. Bech. *Perception of Reproduced Sound.* Audio Engineering Society, 1993.

[15] D. R. Begault. *3-D Sound for Virtual Reality and Multimedia.* Academic Press Professional, 1994.

[16] J. Blauert. *Spatial Hearing.* MIT Press, 1997.

[17] A. D. Blumlein. British Patent Specification 394,325 (Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems. *Journal of the Audio Engineering Society (Reprint)*, 6:91–101, 1933.

[18] B. Boashash and P. J. Black. An Efficient Real-Time Implementation of the Wigner-Ville Distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(11):1611–1618, December 1987.

[19] P. J. Boles and B. Boashash. The Cross Wigner-Ville Distribution - A Two Dimensional Analysis Method for the Processing of Vibroseis Seismic Signals. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 904–907, 1988.

[20] J. Borish. Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[21] F. Boudreaux-Bartels. *Mixed Time-Frequency Signal Transformations.* In Poularikas [156], 1996.

[22] G. F. Boudreaux-Bartels and T. Parks. Time-varying filtering and signal estimation using wigner distribution synthesis techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(3):442–451, 1986.

[23] R. N. Bracewell. The Fast Hartley Transform. *Proceedings of the IEEE*, 72(8):1010–1018, August 1984.

[24] A. S. Bregman. *Auditory Scene Analysis : The Perceptual Organization of Sound.* MIT Press, 1990.

[25] S. Buus, E. Schorer, M. Florentine, and E. Zwiker. Decision rules in detection of simple and complex tones. *Journal of the Acoustical Society of America*, 80:1646–1657, 1986.

[26] L. H. Carney and T. C T. Yin. Temporal coding of resonances by low-frequency auditory nerve fibres: single-fibre responses and a population model. *Journal of Neurophysiology*, 60:1653–1677, 1988.

[27] T. A. Claasen and W. F. Mecklenbräuker. The Wigner distribution - a tool for time-frequency signal analysis : Part 1: Continuous-time signals. *Phillips Journal of Research*, 35(3):217–250, 1980.

[28] T. A. Claasen and W. F. Mecklenbräuker. The Wigner distribution - a tool for time-frequency signal analysis : Part 2: Discrete-time signals. *Phillips Journal of Research*, 35(4/5):276–300, 1980.

[29] T. A. Claasen and W. F. Mecklenbräuker. The Wigner distribution - a tool for time-frequency signal analysis : Part 3: Relations with other time-frequency signal transformations. *Phillips Journal of Research*, 35:372–389, 1980.

[30] L. Cohen. Generalized phase-space distribution functions. *Journal of Mathematical Physics*, 7:781–786, 1966.

[31] L. Cohen. Time Frequency Distributions - A Review. *Proceedings of the IEEE*, 77(7):941–981, July 1989.

[32] L. Cohen. The Scale Representation. *IEEE Transactions on Signal Processing*, 41(12):3275–3292, December 1993.

[33] L. Cohen. *Time-Frequency Analysis : Theory and Applications.* Prentice Hall, 1995.

[34] M. Cooke. Homepage. *http://www.dcs.shef.ac.uk/ martin/.*

[35] M. Cooke. *Modelling Auditory Processing and Organisation.* PhD thesis, University of Sheffield, 1993.

[36] M. Cooke, S. Beet, and M. Crawford, editors. *Visual Representations of Speech Signals.* Wiley, 1993.

[37] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Accepted by Speech Communication*, June 2000.

[38] D. H. Cooper and J. L. Bauck. On acoustical specification of natural stereo imaging. *Journal of the Audio Engineering Society*, 28:368–370, 1980.

[39] P. Dallas, A. N. Popper, and R. R. Fay, editors. *The Cochlea.* Springer-Verlag, New York, 1996.

[40] C. d'Allesandro. *Auditory-based wavelet representation*, pages 131–138. In Cooke et al. [36], 1993.

[41] E. de Boer. On the principle of specific coding. *Journal of Dynamic Systems Measurement*, 95(3):265–273, September 1973.

[42] E. de Boer. Synthetic whole-nerve action potentials for the cat. *Journal of the Acoustical Society of America*, 58:1030–1045, 1975.

[43] E. de Boer. *Mechanics of the cochlea: modeling efforts*, pages 258–317. In Dallas et al. [39], 1996.

[44] E. de Boer. Classical and non-classical models of the cochlea. *Journal of the Acoustical Society of America*, 102(6):2148–2150, April 1997.

[45] E. de Boer. Cochlear models and minimum phase. *Journal of the Acoustical Society of America*, 101(4):3810–3813, December 1997.

[46] I. Debauchies, S. G. Mallat, and A. Willsky. Special issue on wavelets and multiresolution analysis. *IEEE Transactions on Information Theory*, 32, 1992.

[47] R. O. Duda. Sound localisation research. *http://www-engr.sjsu.edu/ duda/Duda.Research.html.*

[48] R. O. Duda. Modelling head related transfer functions. In *Annual Asilomar Conference on Signals, Systems and Computers (Asilomar, CA)*, 1993.

[49] D. A. Eddins and D. M. Green. *Temporal Integration and Temporal Resolution.* In Moore [107], 1995.

[50] D. A. Eddins, J. W. Hall, and J. H. Grose. Detection of temporal gaps as a function of frequency region and absolute noise bandwidth. *Journal of the Acoustical Society of America*, 91:1069–1077, 1992.

[51] F. Alton Everest. *The Master Handbook of Acoustics.* TAB Books, 1994.

[52] M. A. Fagelson and C. A. Champlin. Correlations between auditory-filter shape parameters measured at proximal center frequencies. *Journal of the Acoustical Society of America*, 98, November 1995.

[53] A. Farina. Aurora CD ROM. *http://www.ramsete.com/aurora/directory.htm.*

[54] J. M. Festen and R. Plomp. Relations between auditory functions in normal hearing. *Journal of the Acoustical Society of America*, 70:356–369, 1981.

[55] J. M. Festen and R. Plomp. Relations between auditory functions in impaired hearing. *Journal of the Acoustical Society of America*, 73:652–662, 1983.

[56] J. L. Flanagan. Models for approximating basilar membrane displacement. *Bell Systems Technical Journal*, 39:1163–1191, 1960.

[57] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12:47–65, 1940.

[58] M. Florentine, S. Buus, and R. P. Hellman. *A Model of Loudness Summation Applied to High-Frequency Hearing Loss*, pages 187–198. In Jesteadt [82], 1997.

[59] W. G. Gardener. A realtime multichannel room simulator. *Journal of the Acoustical Society of America, http://sound.media.mit.edu/papers.html*, 1992.

[60] W. G. Gardener. *Reverberation Algorithms*, pages 85–132. In [88], 1998.

[61] M. Gerzon. Why Do Equalisers Sound Different? . *Studio Sound*, pages 58–65, July 1990.

[62] M. A. Gerzon. Unitary (energy preserving) multichannel networks with feedback. *Electronics Letters*, 12(11):278–279, 1976.

[63] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[64] B. R. Glasberg and B. C. J. Moore. Effects of envelope fluctuations on gap detection. *Hearing Research*, 64:81–92, 1992.

[65] D. M. Green, M. J. McKey, and J. C. R. Licklider. Detection of a pulsed sinusoid in noise as a function of frequency. *Journal of the Acoustical Society of America*, 33:1446–1452, 1959.

[66] D. Griesinger. Practical processors and programs for digital reverberation. *Proceedings of Audio Engineering Society, 7th International Conference, Toronto, Canada*, pages 187–195, 1989.

[67] S. L. Hahn. *Hilbert Transforms in Signal Processing.* Artech House, 1996.

[68] J. W. Hall, M. P. Haggard, and M. A. Fernandes. Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, 76:50–56, 1984.

[69] W. M. Hartmann. *Signals, Sound and Sensation.* American Institute of Physics Press, 1997.

[70] J. E. Hawkins and S. S. Stevens. The Masking of Pure Tones and of Speech by White Noise. *Journal of the Acoustical Society of America*, 22:6–13, 1950.

[71] H. L. F. Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Gruddlage für die Theorie der Musik.* F. Vieweg, Braunschweig, 1863.

[72] F. Hlawatsch. personal communication, May 2000.

[73] F. Hlawatsch, G. Matz, H. Kirchauer, and W. Kozek. Time-frequency formulation, design, and implementation of time-varying optimal filters for signal estimation. *IEEE Trans. Signal Processing*, 48:1417–1432, May 2000.

[74] Z. Hou and C. V. Pavlovic. Effects of temporal smearing on temporal resolution, frequency selectivity, and speech intelligibility. *Journal of the Acoustical Society of America*, 96(3):1325–1340, September 1994.

[75] T. Houtgast. Psychophysical experiments on gating acuity. In *Symposium on Hearing Theory, IPO, Eindhoven, The Netherlands*, 1972.

[76] T. Houtgast. *Lateral Suppression in Hearing.* PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, 1974.

[77] D. A. Huffman. The generation of impulse-equivalent pulse trains. *IRE Transactions*, IT8:S10–S16, 1962.

[78] T. Irino and R. D. Patterson. A time-domain, level dependent auditory filter : The gammachirp. *Journal of the Acoustical Society of America*, 101(1):412–419, January 1997.

[79] T. Irino and M. Unoki. An Efficient Implementation of the Gammachirp Filter and its Filterbank Design. *ATR Technical Report*, pages 1–13, 1997.

[80] C. P. Janse and A. J. M. Kaizer. Time-Frequency Distributions of Loudspeakers : The Application of the Wigner Distribution. *Journal of the Audio Engineering Society*, 31(4):199–223, April 1983.

[81] H. Jeong and J.-G. Ih. Implementation of a New Algorithm Using the STFT with Variable Frequency Resolution for the Time-Frequency Auditory Model. *Journal of the Audio Engineering Society*, 47(4):240–251, April 1999.

[82] W. Jesteadt, editor. *Modelling Sensorineural Hearing Loss.* Lawrence Erlbaum Associates, 1997.

[83] P. I. M. Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Proceedings of the Symposium on Hearing Theory, IPO, Eindhoven, The Netherlands*, 1972.

[84] I. T. Jollife. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[85] J. M. Jot. *Etude et réalisation d'un spatialisateur de sons par modèles physiques et perceptifs (Design and implementation of a sound spatializer based on physical and perceptual models, in French).* PhD thesis, Telecom Paris, 1992.

[86] J. M. Jot, L. Cerveau, and O. Warusfel. Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model. In *The 103rd Convention of the Audio Engineering Society*, pages 1–27, 1997.

[87] J. M. Jot, V. Larcher, and O. Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. *Proc. of the Audio Eng. Soc. Conv.*

[88] M. Khars and K. Brandenburg. *Applications of Digital Signal Processing to Audio and Acoustics.* Kluwer Academic Press, 1998.

[89] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91(3):1637–1647, June 1991.

[90] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91(3):1637–1647, March 1992.

[91] W. Krattenthaler and F. Hlawatsch. Improved signal synthesis from pseudo-wigner distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 39(2):506–509, 1991.

[92] W. Krattenthaler and F. Hlawatsch. Time-frequency design and processing of signals via smoothed wigner distributions. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 41(1):278–287, 1991.

[93] C. Kyriakakis, P. Tsakalides, and T. Holman. Surrounded by Sound : Acquisition and Rendering Methods for Immersive Audio. *IEEE Signal Processing Magazine*, pages 55–66, January 1999.

[94] S. Launer, V. Hohmann, and B. Kollmeier. *Modelling Loudness Growth and Loudness Summation in Hearing-Impaired Listeners*, pages 175–185. In Jesteadt [82], 1997.

[95] H. Lehnert and J. Blauert. Principles of binaural room simulation. *Applied Acoustics*, 36:259–291, 1992.

[96] E. R. Lewis, E. Hecht-Poinar, R. F. Lyon, P. M. Narins, and C. R. Steele, editors. *Diversity in Auditory Mechanics.* World Scientific Publishing, 1997.

[97] L. S. Lim and A. V. Oppenheim. *Advanced Topics in Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[98] R. F. Lyon. The All-Pole Gammatone Filter and Auditory Models. *http://www.pcmp.caltech.edu/anaprose/dick/auditory/APGF*, 1996.

[99] R. F. Lyon. *All-Pole Models of Auditory Filtering*, pages 205–211. In Lewis et al. [96], 1997.

[100] R. F. Lyon and C. A. Mead. An Analog Electronic Cochlea. *IEEE Transactions on Acoustics Speech and Signal Processing*, 36:1119–1134, 1988.

[101] H. Margineau and R. N. Hill. Correlation between measurements in quantum theory. *Progress of Theoretical Physics*, 26:722–738, 1961.

[102] R. H. Margolis and A. M. Small. The measurement of critical masking bands. *Journal of Speech and Hearing*, 18:571–587, 1975.

[103] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity using a computer model of the auditory periphery: I pitch detection. *Journal of the Acoustical Society of America*, 89:2866–2882, 1991.

[104] M. Mellody and G. H. Wakefield. The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis. *Journal of the Acoustical Society of America*, 107(1):598–611, January 2000.

[105] B. C. J. Moore, editor. *Frequency Selectivity in Hearing.* Academic Press, 1986.

[106] B. C. J. Moore. *Characterization of Simultaneous, Forward and Backward Masking.* In [14], 1993.

[107] B. C. J. Moore, editor. *Hearing.* Academic Press, 1995.

[108] B. C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, 1997.

[109] B. C. J. Moore and J. L. Alcantara. Masking patterns for sinusoidal and narrow-band noise maskers. *Journal of the Acoustical Society of America*, 1040(2):1023–1038, August 1998.

[110] B. C. J. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753, 1983.

[111] B. C. J. Moore and B. R. Glasberg. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. *Hearing Research*, 28:209–225, 1987.

[112] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas. The shape of the ear's temporal window. *Journal of the Acoustical Society of America*, 83(3):1102–1116, March 1988.

[113] B. C. J. Moore, B. R. Glasberg, and A. Simpson. Evaluation of a method of simulating reduced frequency selectivity. *Journal of the Acoustical Society of America*, 91(6):3402–3423, June 1992.

[114] B. C. J. Moore, R. W. Peters, and B. R. Glasberg. Auditory filter shapes at low center frequencies. *Journal of the Acoustical Society of America*, 88(1):132–140, July 1990.

[115] J. A. Moorer. About This Reverberation Business. *Computer Music Journal*, 3(2):13–28, 1979.

[116] M. Mummert. *Sprachcodierung durch Konturierung eines gehörangepaßten Spektrogramms und ihre Anwendung zur Datenreduktion (Speech coding by contourizing an aurally adapted spectrogram and its application to data reduction). In German.* PhD thesis, Technische Universität München, 1997.

[117] A. H. Nuttall. Alias-Free Wigner Distribution Function and Complex Ambiguity Function for Discrete-Time Samples. *Naval Underwater Systems Center Technical Report*, pages 1–56, April 1989.

[118] A. H. Nuttall. Alias-Free Smoothed Wigner Distribution Function for Discrete-Time Samples. *Naval Underwater Systems Center Technical Report*, pages 1–28, October 1990.

[119] J. J. O'Donovan and D. J. Furlong. A joint time-frequency auditory masking model. In *Proceedings of the Irish Signals and Systems Conference*, pages 417–424, June 1998.

[120] J. J. O'Donovan and D. J. Furlong. A joint time-frequency model of auditory masking. In *Proceedings of the 136th Meeting of the Acoustical Society of America*, October 1998.

[121] J. J. O'Donovan and D. J. Furlong. Signal analysis-synthesis from a joint time-frequency model of auditory masking. In *Proceedings of the Irish Signals and Systems Conference*, pages 258–265, June 2000.

[122] F. Opolko and J. Wapnick. McGill University Master Samples (MUMS). *http://improv.music.mcgill.ca/resources/mums/html/mums.html*, 1987-1989.

[123] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing.* Prentice Hall, 1990.

[124] D. O'Reilly. *The Application of Digital Signal Processing to the Modelling and Alleviation of Auditory Impairment.* PhD thesis, Trinity College Dublin, 2000.

[125] A. J. Oxenham and B. C. J. Moore. Additivity of masking in normally hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America*, 98(4):1921–1934, October 1995.

[126] C. H. Page. Instantaneous power spectra. *Journal of Applied Physics*, 23:103–106, 1952.

[127] J. H. Patterson and D. M. Green. Discrimination of transient signals having identical energy spectra. *Journal of the Acoustical Society of America*, 48:894–905, 1970.

[128] R. Patterson and M. Allerhand. Time domain modeling of peripheral auditory processing. *Journal of the Acoustical Society of America*, 98(4):1890–1894, October 1995.

[129] R. D. Patterson. Auditory filter shape. *Journal of the Acoustical Society of America*, 55(4):803–809, April 1974.

[130] R. D. Patterson. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59(3):640–654, March 1976.

[131] R. D. Patterson. The sound of a sinusoid : Time-interval models. *Journal of the Acoustical Society of America*, 96(3):1419–1428, September 1994.

[132] R. D. Patterson and G. B. Henning. Stimulus variability and auditory filter shape. *Journal of the Acoustical Society of America*, 62:649–664, 1977.

[133] R. D. Patterson, J. Holdsworth, and M. Allerhand. *Auditory Models as preprocessors for speech recognition*, pages 67–83. In [167], 1992.

[134] R. D. Patterson and B. C. J. Moore. *Auditory filters and excitation patterns as representations of frequency resolution*, pages 123–177. In Moore [105], 1986.

[135] R. D. Patterson and I. Nimmo-Smith. Off-frequency listening and auditory-filter asymmetry. *Journal of the Acoustical Society of America*, 67(1):229–245, January 1980.

[136] R. D. Patterson and I. Nimmo-Smith. Off-frequency listening and auditory-filter asymmetry. *Journal of the Acoustical Society of America*, 67:229–245, 1980.

[137] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. Spiral VOS Final Report—Part A : The Auditory Filterbank. *MRC Applied Psychology Unit, Cambridge*, 1988.

[138] R. D. Patterson, I. Nimmo-Smith, D. L. Weber, and R. Milroy. The deterioration of hearing with age : Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America*, 72(6):1788–1803, December 1982.

[139] S.-C. Pei and I.-I. Yang. Computing Pseudo-Wigner Distribution by the Fast Hartley Transform. *IEEE Transactions on Signal Processing*, 40(9):2346–2349, September 1992.

[140] M. J. Penner. Effects of masker duration and masker level on forward and backward masking. *Journal of the Acoustical Society of America*, 179-182:56, 1974.

[141] M. J. Penner and E. Cudahy. Critical masking interval: A temporal analog of the critical band. *Journal of the Acoustical Society of America*, 65:1530–1534, 1973.

[142] M. J. Penner, C. E. Robinson, and D. M. Green. The critical masking interval. *Journal of the Acoustical Society of America*, 52:1661–1668, 1972.

[143] M. Pflueger and R. Hoeldrich. A nonlinear model of the peripheral auditory system. In *103rd Convention of the Audio Engineering Society*, pages 1–12, 1997.

[144] J. Pickles. *An Introduction to the Physiology of Hearing.* Springer-Verlag, New York, 1982.

[145] W. Pielemeier. personal communication, July 1998.

[146] W. Pielemeier and G.H. Wakefield. Time-frequency and time-scale analysis for musical analysis. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 421–424, 1992.

[147] W. J. Pielemeier. *A Time-Frequency Distribution with Parameter Estimators for a Musical Signal Model.* PhD thesis, University of Michigan, 1994.

[148] W. J. Pielemeier and G. H. Wakefield. A high-resolution time-frequency representation for musical instrument signals. *Journal of the Acoustical Society of America*, 98(4):2382–2396, April 1995.

[149] C. Plack. personal communication, May 2001.

[150] C. J. Plack. *Temporal Resolution in the Auditory System.* PhD thesis, University of Cambridge, 1990.

[151] C. J. Plack and B. C. J. Moore. Temporal window shape as a function of frequency and level. *Journal of the Acoustical Society of America*, 87(5):2178–2187, May 1990.

[152] C. J. Plack and A. J. Oxenham. Basilar-membrane nonlinearity and the growth of forward masking. *Journal of the Acoustical Society of America*, 103(3):1598–1608, March 1998.

[153] R. Plomp. The rate of decay of auditory sensation. *Journal of the Acoustical Society of America*, 37:277–282, 1964.

[154] R. Plomp and W. J. M. Levelt. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 78:548–560, 1965.

[155] M. A. Poletti. The development of a discrete transform for the wigner distribution and ambiguity function. *Journal of the Audio Engineering Society*, 84(1):239–252, July 1988.

[156] A. D. Poularikas, editor. *The Transforms and Applications Handbook.* CRC Press and IEEE Press, 1996.

[157] D. Preis and V. C. Georgopoulos. Wigner Distribution Representation and Analysis of Audio Signals : An illustrated Tutorial Review. *Journal of the Audio Engineering Society*, 47(12):1043–1053, December 1999.

[158] D. Preis and F. Hlawatsch. Wigner Distribution Analysis of Filters with Perceptible Phase Distortion. *Journal of the Audio Engineering Society*, 35(12):1004–1012, December 1987.

[159] R. M. Rao and A. S. Bopardikar. *Wavelet Transforms.* Addison Wesley Longman, 1998.

[160] A. W. Rihaczek. Signal energy distribution in time and frequency. *IEEE Transactions on Information Theory*, 14:369–394, 1968.

[161] C. E. Robinson and I. Pollack. Interaction between forward and backward masking: A measure of the integrating period of the auditory system. *Journal of the Acoustical Society of America*, 53:1313–1316, 1973.

[162] D. Ronken. Monoaural detection of a phase difference between clicks. *Journal of the Acoustical Society of America*, pages 1091–1099, 1970.

[163] S. Rosen and R. J. Baker. Characterising auditory filter nonlinearity. *Hearing Research*, 73:231–243, 1994.

[164] S. Rosen, R. J. Baker, and A. Darling. Auditory filter nonlinearity at 2khz in normal hearing listeners. *Journal of the Acoustical Society of America*, 103, May 1998.

[165] T. H. Schafer, R. S. Gales, C. A. Shewmaker, and P. O. Thompson. The Frequency Selectivity of the Ear as Determined by Masking Experiments. *Journal of the Acoustical Society of America*, 22:490–496, 1950.

[166] B. Scharf. *Critical Bands.* In Tobias [177], 1970.

[167] M. E. H. Schouten. *The Auditory Processing of Speech: from the auditory periphery to words.* Mouton de Gruyter, Berlin, 1992.

[168] M. R. Schroeder. Natural Sounding Artificial Reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223, July 1962.

[169] M. R. Schroeder. Digital simulation of sound transmission in reverberant spaces. *Journal of the Acoustical Society of America*, 47(2):424–431, 1970.

[170] M. J. Shailer and B. C. J. Moore. Gap detection and the auditory filter: phase effects using sinusoidal stimuli. *Journal of the Acoustical Society of America*, 81:1987, 1110-1117.

[171] M. J. Shailer and B. C. J. Moore. Gap detection as a function of frequency, bandwidth and level. *Journal of the Acoustical Society of America*, 74:1983, 467-473.

[172] D. Sinha and A. Tewfik. Low bit rate transparent audio compression using adapted wavelets. *IEEE Transactions on Signal Processing*, 41(12):3463–3479, December 1993.

[173] M. Slaney. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. *http://www.slaney.org/malcolm/pubs.html#AuditoryModeling*, 1993.

[174] M. F. Spiegel. Thresholds for tones in maskers of various bandwidths and for signals of various bandwidths as a function of signal frequency. *Journal of the Acoustical Society of America*, 69:791–795, 1981.

[175] J. A. Swets, D. M. Green, and W. P. Tanner. On the width of critical bands. *Journal of the Acoustical Society of America*, 34:108–113, 1962.

[176] M. Tate. *Principles of Hearing Aid Audiology.* Chapman & Hall, 1994.

[177] J. Tobias, editor. *Foundations of Modern Auditory Theory.* Academic Press, New York, 1970.

[178] N. P. McAngus Todd. A computational model of temporal integration. *Journal of the Acoustical Society of America*, 96(5), November 1994.

[179] N. F. Viemeister. Temporal modulation transfer functions based on modulation thresholds. *Journal of the Acoustical Society of America*, 66:1364–1380, 79.

[180] M. A. Viergever. *Cochlear mechanics - A review*, pages 63–72. In Allen et al. [4], 1986.

[181] J. Ville. Theorie et applications de la notion de signal analytique. *Cables et Transmissions*, 2A:61–74, 1948.

[182] M. J. Walsh. *3-D Sound System Design for the Generic Listener.* PhD thesis, Trinity College Dublin, 1997.

[183] D. L. Weber. Growth of masking and the auditory filter. *Journal of the Acoustical Society of America*, 62:424–429, 1977.

[184] J. C. Webster, P. H. Miller, P. O. Thompson, and E. W. Davenport. The Masking and Pitch Shifts of Pure Tones near Abrupt Changes in a Thermal Noise Spectrum. *Journal of the Acoustical Society of America*, 24:147–152, 1952.

[185] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America*, 85:858–867, 1989.

[186] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, 85:868–878, 1989.

[187] F. L. Wightman and D. J. Kistler. Hearing Development Research Lab HRTF Data. *http://www.waisman.wisc.edu/hdrl/index.html*, 2000.

[188] E. P. Wigner. On the quantum correction for thermodynamic equilibrium. *Physical Review*, 40:749–759, 1932.

[189] Y. Wu and Z. Zhu. Efficient Computation of the 2-D Discrete Pseudo-Wigner Distribution by the Fast Hartley Transform. *Proceedings of the IEEE 1995 National Aerospace and Electronics Conference. NAECON*, 1:320–323.

[190] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 83(2):824–839, March 1992.

[191] W. A. Yost. *Fundamentals of Hearing.* Academic Press, 1994.

[192] K. Yu and S. Cheng. Signal synthesis from pseudo-wigner distribution and applications. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(9):1289–1302, 1987.

[193] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, 33:248, 1961.

[194] E. Zwicker and H. Fastl. *Psychoacoustics.* Springer, 1999.

[195] E. Zwicker and E. Terhardt. Analytic expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68(5):1523–1525, November 1980.