COMPUTER MODELLING OF SOUND FOR TRANSFORMATION AND SYNTHESIS OF MUSICAL SIGNALS

Paul Masri

December 1996

A thesis submitted to the University of Bristol in accordance with the requirements for the degree of Doctor of Philosophy in the Faculty of Engineering, Department of Electrical and Electronic Engineering.

ABSTRACT

The purpose of this thesis is to develop a sound model that can be used as a creative tool by professional musicians. Post-production editing suites are used for compiling and arranging music tracks, and for creating soundtracks and voice-overs for the radio, television and film industries. A sound model would bring a new dimension of flexibility to these systems, allowing the user to stretch and mould sounds as they please.

Sound models already exist but they are limited both in their usability and in their scope for representation. All aspects of the model in this thesis use designer-preset global variables which are transparent to the user. Within this restriction and preserving manipulation flexibility, the aim of the thesis is to improve the range of sounds that can be modelled and the accuracy of modelling. These are dependent on the choice of model elements and the accuracy of the analysis-resynthesis system (which translates between the playable time domain waveform and the controllable model feature domain, making the model usable).

The basis of the model of this thesis is a deterministic-stochastic classification; the partials of the harmonic structure of pitched sounds are individually represented in the deterministic aspect, whilst the stochastic aspect models the remainder as broadband noise. Three studies were carried out to improve aspects of the analysis-resynthesis system. These focus on:

- the time-frequency representation, by which the analyser ‘sees’ detail in the sound;
- frame linking, which converts the instantaneous partial estimates into continuous trajectories – this is essential for synthesis quality and for musical manipulation;
- percussive note onsets, which are not represented in the existing models.

The standard time-frequency representation for sound modelling, the Short-Time Fourier Transform, has limited resolution and is inadequate for capturing the detail of rapidly changing elements. The first study examines the distortion it generates when it represents a nonstationary element and derives a method for extracting extra information from the distortion, thereby improving the effective resolution.

The fact that partials belong to a harmonic structure is not considered in the existing ‘Nearest Frequency’ method of frame linking; the result is audible scrambling of the higher frequencies. The second study proposes using the harmonic structure as the basis for linking. Although this is not a new concept, it is implemented in such a way that detail can be extracted from the harmonically weak start and end of harmonic regions, thereby improving synthesis quality.

The existing model assumes all sound elements are slow-changing, so abrupt changes are poorly represented and sound diffused upon synthesis. The third study finds a way of incorporating ‘attack transients’ into the model. The method pre-scans a sound for percussive onsets and synchronises both analysis and synthesis so as to avoid the previous problems. The crispness of synthesised attack transients clearly demonstrate the effectiveness of this method.

From many observations over the course of these studies, it became noticeable that the hard deterministic-stochastic classification was not capturing the ‘roughness’ of some sounds accurately. Further investigations revealed that detail is missing from the synthesised partials. A new basis for a sound model, termed here the Noisy Partial model, aims to rectify this by introducing the noisiness into the partials themselves. In this new classification, deterministic and stochastic appear as opposite extremes on a continuously variable scale. The new model promises a simplified structure and more efficient processing. Suggestions are made for investigating this further as a future work direction.

AUTHOR'S DECLARATION

Unless otherwise acknowledged, the content of this thesis is the original and sole work of the author. No portion of the work in this thesis has been submitted by the author in support of an application for any other degree or qualification, at this or any other university or institute of learning. The views expressed in this thesis are those of the author, and not necessarily those of the University of Bristol.

Paul Masri

COPYRIGHT

Attention is drawn to the fact that the copyright of this thesis rests with the author. This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author. This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purpose of consultation.

CONTENTS

LIST OF FIGURES...v

ACKNOWLEDGEMENTS vii

1. INTRODUCTION...1

1.1 Computer Music and the Field of Sound Modelling 4

1.2 Analysis-Resynthesis and its Applications...5

1.3 Overview of the Thesis...6

2. MUSICAL SIGNALS AND HOW THEY CAN BE MODELLED............................11

2.1 What is a Musical Signal?..13

2.1.2 Harmonic Structure 14

2.1.2.1 Short-Term Periodicity.. 14

2.1.2.2 Non-Harmonic Overtones.. 15

2.1.3 Vibrato and Tremolo..15

2.1.4 The Amplitude Envelope 16

2.1.5 Unpitched Sounds...17

2.1.5.1 Additive and Non-Additive Noise... 17

2.2 Review of Research into Sound and Instrument Modelling.........................19

2.2.1 Physical Modelling 19

2.2.1.1 The Wavetable Model 19

2.2.1.2 Digital Waveguide Modelling.. 20

2.2.1.3 Current Research Activity 21

2.2.2 Analysis-Resynthesis..22

2.2.2.1 The Sinusoidal Model 22

2.2.2.2 The Formant Model 23

2.2.2.3 Other Techniques.. 24

2.2.3 Granular Synthesis 25

2.2.3.1 Wavelet Analysis-Resynthesis 25

2.2.3.2 Sampling Granular Synthesis 26

2.3 The Initial Model..28

2.3.1 Model Overview...28

2.3.2 Deterministic Analysis..29

2.3.2.1 Time-Frequency Analysis – the STFT 31

2.3.2.2 Critique and Modification – Time-Frequency Representation 33

2.3.2.3 Peak Detection and Accurate Location 34

2.3.2.4 Critique and Modification – Peak Detection and In-Frame Validation......... 35

2.3.2.5 Frame Linking to Generate Partial Trajectories 36

2.3.2.6 Critique and Modification – Frame Linking 37

2.3.2.7 The Importance of Phase 38

2.3.3 Deterministic Synthesis...38

2.3.3.1 Additive Synthesis with Linear Interpolation.. 38

2.3.3.2 Critique and Modification – Additive Synthesis with Linear Interpolation.... 41

2.3.4 Calculation of the Residual..42

2.3.4.1 Critique and Modification – Calculation of the Residual............................ 42

2.3.5 Stochastic Analysis 44

2.3.5.1 Critique and Modification – Stochastic Analysis 45

2.3.6 Stochastic Synthesis 45

2.3.6.1 Critique and Modification – Stochastic Synthesis 47

2.3.7 The Model as a Whole 47

2.3.7.1 Critique and Modification – The Model as a Whole........ 47

2.3.7.2 Model Aspects Targeted For Improvement 48

3. TIME-FREQUENCY REPRESENTATION (TFR) 51

3.1 What is a TFR and How Can One be Chosen? 53

3.1.1 The World Is As You See It 53

3.1.2 The TFR Must Match the Sound Model 54

3.1.3 What is the Application for the Model? 55

3.1.4 What Features Define the Sound Model? 59

3.1.5 Observations on Model Development 60

3.1.5.1 Circular Reasoning 61

3.1.5.2 Every Model is Incomplete 62

3.2 A Brief Review of Alternative Approaches to Time-Frequency Representation 63

3.2.1 Short Time Fourier Transform (STFT) 63

3.2.2 Wavelet Transform (WT) and Bounded-Q Fourier Transform (BQFT) 63

3.2.3 The Cohen Class of TFR 65

3.2.3.1 Being Certain about the Uncertainty Principle 65

3.2.3.2 The Wigner Distribution and the Bilinear Form 66

3.2.3.3 Cohen’s Class and Higher Order Spectra 67

3.2.3.4 Cross-terms and Modifications to the Kernel 68

3.2.3.5 Application of the Cohen Class of TFR for Sound Representation 68

3.2.4 Parametric TFRs 69

3.2.4.1 Autoregression and Moving Average 69

3.2.4.2 Higher Order Parametric Modelling 70

3.2.4.3 Application of Parametric TFRs for Sound Representation 71

3.2.5 Model-Informed TFRs 71

3.2.5.1 Multiresolution Fourier Transform 71

3.2.5.2 Parametric Modelling of the STFT 72

3.2.5.3 Extraction of ‘Hidden’ Information from the FFT 73

3.3 Obtaining Second Order Information from Phase Distortion in the FFT 74

3.3.1 How Measurements were Made 76

3.3.2 Phase of an Unmodulated Sinusoid 76

3.3.3 Linear Frequency Modulation (LFM) 77

3.3.4 Exponential Amplitude Modulation (EAM) 78

3.3.5 Concurrent LFM and EAM 79

3.4 Performance of the Technique 81

3.4.1 Practical Application of LFM and EAM Analysis 81

3.4.1.1 Simulated Data 81

3.4.1.2 Real Musical Signals 82

3.4.2 Implementation Within the Sound Model 83

4. FRAME LINKING BY HARMONIC STRUCTURE 86

4.1 Nearest Frequency versus Harmonic Structure 88

4.1.1 Higher Partials Vary More Rapidly 89

4.1.2 The Smoothed Representation of the STFT...93
4.1.3 A Specialist Musical Model..95
4.2 Methodology and Implementation...96
4.2.1 Overview...96
4.2.2 Fundamental Frequency Estimation and Harmonic Structure Detection.................97
4.2.2.1 Two Popular Approaches 97
4.2.2.2 A Hybrid Method 101
4.2.3 Inter-Frame Correction and Harmonic Region Definition.....................................105
4.2.3.1 Estimate Assessment... 105
4.2.3.2 Estimate Correction... 111
4.2.4 Matching Peaks to Partials and Linking the Frames 118
4.3 Performance of the Technique..122
4.3.1 Subjective Results 122
4.3.2 Objective Performance...122
4.3.2.1 Matching the Higher Partials to Peaks in the Spectrum 123
4.3.2.2 Matching Partials to Noise Peaks 123
4.3.2.3 Problems with Reverberation and Extraneous Noises.. 124
5. IMPROVED SYNTHESIS OF ATTACK TRANSIENTS.......................................125
5.1 Problem Description and Solution Definition...126
5.1.1 The Disparity between Attack Transients and their Representation 126
5.1.2 Modifications to the Model to Incorporate Attack Transients 129
5.2 Detection and Location of Transient Events 130
5.2.1 Detection by Energy Distribution...130
5.2.1.1 Implementation.. 131
5.2.1.2 Time Resolution... 134
5.2.2 Detection by Attack Envelope..134
5.2.2.1 Implementation.. 134
5.2.2.2 Time Resolution... 136
5.2.3 Detection by Spectral Dissimilarity...137
5.2.3.1 Implementation.. 138
5.2.3.2 Time Resolution... 140
5.2.4 Comparison of Detection Methods 141
5.2.4.1 Reliability 141
5.2.4.2 Temporal Accuracy 143
5.2.4.3 Computation Speed 143
5.2.4.4 Summary.. 143
5.3 Synchronisation of Analysis and Synthesis 148
5.3.1.1 Region Definition... 148
5.3.1.2 Synchronised Analysis 148
5.3.1.3 Synchronised Synthesis.. 150
5.4 Results 151
5.4.1 Performance of the Technique..151
5.4.2 Impact upon the Model 152
6. OUTCOMES AND FUTURE DIRECTIONS 156
6.1 Phase Distortion Analysis (PDA)...159
6.1.1 Impact upon the Model 159
6.1.2 Critique and Future Directions..159

6.1.2.2 Alternative Techniques and Recent Advances 161

6.2 Harmonic Frame Linking 163

6.2.1 Impact upon the Model 163

6.2.2 Critique and Future Directions 164

6.2.2.2 Localisation in Time and Frequency 164

6.2.2.3 Alternative Techniques and Recent Advances 166

6.3 Attack Transients 167

6.3.1 Impact upon the Model 167

6.3.2 Critique and Future Directions 168

6.3.2.1 Alternative Techniques and Recent Advances 169

6.4 The Results Combined in a Single System 170

6.5 Noisy Partials – The Basis for a New Model of Sound 173

6.5.1 Observations about the Deterministic Plus Stochastic Model 173

6.5.1.1 Observations Regarding Capture of Feature Detail 173

6.5.1.2 Observations Regarding The Deterministic-Stochastic Classification 175

6.5.2 The Noisy Partial 177

7. APPLICATIONS OF THE SOUND MODEL 181

7.1 Applications in Musical Systems 182

7.1.1 Analysis-Resynthesis for Advanced Post-Production Editing 182

7.1.1.1 Post-Production Editing Suites 182

7.1.1.2 Enhancement Through a Sound Model 184

7.1.2 Generative Synthesis of Sound Model Features 185

7.1.2.1 Instrument Simulation 187

7.1.2.2 Architecture Simulation 187

7.1.2.3 System Unification 188

7.2 Applications in Ongoing Research 189

7.2.1 Instrument Timbre Modelling (ITM) Synthesis 189

7.2.2 Expressive Speech Synthesiser 191

8. CONCLUSIONS 193

8.1 The Global Aims of the Thesis 194

8.2 Summary of Achievements 195

8.2.1 The Specific Goals of the Thesis 195

8.2.2 Improved Time-Frequency Resolution 195

8.2.3 Improved Frame Linking 196

8.2.4 Improved Modelling of Percussives 197

8.3 The Future of Sound Modelling 198

8.3.1 Development of a New Model 198

8.3.2 Applications of the Model 198

APPENDICES 200

A. PARTIAL DOMAIN SYNTHESIS OF MUSIC 201

B. IDENTIFICATION OF NONSTATIONARY AUDIO SIGNALS USING THE FFT, WITH APPLICATION TO ANALYSIS-BASED SYNTHESIS OF SOUND 207

C. IMPROVED MODELLING OF ATTACK TRANSIENTS IN MUSIC ANALYSIS-RESYNTHESIS.......214

BIBLIOGRAPHY 219

LIST OF FIGURES

Figure 1.1 - Sound analysis-resynthesis system 2
Figure 2.1 – Venn diagram of the set of all sounds 13
Figure 2.2 – Visible periodicity depends on harmonicity of sinusoids 15
Figure 2.3 – Overview of the complete analysis-resynthesis system 29
Figure 2.4 – Deterministic analysis 30
Figure 2.5 – Sonogram of a cello note with one FFT highlighted 31
Figure 2.6 – Additive synthesis 39
Figure 2.7 – The life of an oscillator (whose lifespan is 5 frames) 40
Figure 2.8 – Residual errors due to differences between the partial peaks of the source signal and deterministic signal 44
Figure 2.9 – Stochastic analysis 45
Figure 2.10 – Stochastic synthesis 46
Figure 3.1 – Alternative representations of frequency modulation 54
Figure 3.2 – The importance of choosing an appropriate set of features 56
Figure 3.3 – Image Manipulation Techniques (as analogues for sound transformation) 57
Figure 3.4 – Amplitude and phase 'distortion' caused by nonlinearities 75
Figure 3.5 – Spectrum of an unmodulated sinusoid 76
Figure 3.6 – Spectra of sinusoids with linear frequency modulation 77
Figure 3.7 – Linear FM phase distortion at various frequency offsets from the maximum (Hamming window) 78
Figure 3.8 – Spectra of sinusoids with exponential amplitude modulation 78
Figure 3.9 – Exponential AM phase distortion at various frequency offsets from the maximum (Hamming window) 79
Figure 3.10 – Phase distortion across main-lobe for combinations of LFM and EAM 80
Figure 3.11 – Phase distortion mapping curves at '+1/8 bin' frequency offset (close-up of Figure 3.7 and Figure 3.9) 81
Figure 3.12 – Trajectory predictions for simulated data 82
Figure 3.13 – Trajectory predictions for a real data example (cello with vibrato) 83
Figure 4.1 – Linking two frames by nearest frequency 88
Figure 4.2 – Linking two frames by harmonic structure 89
Figure 4.3 – Analysis of a cello note (with vibrato) 90
Figure 4.4 – Distortion artifacts become linked if the hop-distance is too short 94
Figure 4.5 – Block diagram of complete harmonic linking process 96
Figure 4.6 – Fundamental frequency estimation by the Template method 98
Figure 4.7 – Fundamental frequency estimation by the Cepstrum method 99
Figure 4.8 – Variations in performance for different template shapes 100
Figure 4.9 – Fundamental frequency estimation and harmonic structure detection 102
Figure 4.10 – Effect of different template peaks widths on accuracy and template size 104
Figure 4.11 – Calculation of scaling factor, strength and resistance 106
Figure 4.12 – A stronger is more likely to alter a weaker frame 107
Figure 4.13 – Strength and resistance calculations for a real example 108
Figure 4.14 – Combining forward and reverse direction scores 110
Figure 4.15 – The Highest Scorer correction scheme (for previous example data) 112
Figure 4.16 – The Highest Scorer correction scheme (for a burst error example) 114

Figure 4.17 – Harmonic analysis of a speech sample (male voice, spoken)...................................... 117
Figure 4.18 – The Scan-and-Propagate correction scheme (for the burst error example) 119
Figure 5.1 – The ‘confused’ spectra surrounding an attack transient... 128
Figure 5.2 – Attack transients sound ‘diffused’ upon synthesis 128
Figure 5.3 – The sonogram reveals a large increase in energy throughout the spectrum for an attack transient (bass drum) 131
Figure 5.4 – Energy Distribution detector (for a sequence of three drums) 133
Figure 5.5 – An attack transient with the time domain envelope overlaid 134
Figure 5.6 – Attack Envelope detector (for a sequence of three drums) 135
Figure 5.7 – Generation of the dissimilarity matrix from the sonogram 138
Figure 5.8 – Spectral Dissimilarity detector (for a sequence of three drums) 140
Figure 5.9 – Comparison for ‘Tricky’ drum sequence with various background sounds 144
Figure 5.10 – Comparison for a legato clarinet melody.. 145
Figure 5.11 – Comparison for a staccato piano melody 146
Figure 5.12 – Comparison for ‘Please talk clearly’, a male spoken voice 147
Figure 5.13 – The four steps to incorporating the feature of attack transients................................. 149
Figure 5.14 – The modified model, incorporating attack transients.. 153
Figure 6.1 – Analysis in the Initial Model 158
Figure 6.2 – Synthesis in the Initial Model 158
Figure 6.3 – Changes to the Initial Model to incorporate Phase Distortion Analysis 159
Figure 6.4 – Changes to the Initial Model to incorporate Harmonic Frame Linking........................ 163
Figure 6.5 – Changes to the Initial Model's Analysis to incorporate Attack Transients 167
Figure 6.6 – Changes to the Initial Model's Synthesis to incorporate Attack Transients 168
Figure 6.7 –Harmonic Linking and Attack Transients combined within one system (Analysis only shown).. 170
Figure 6.8 – All three innovations within one system (Analysis only shown) 171
Figure 6.9 – Two-dimensional phase space representation of sound signals 174
Figure 6.10 – Sound component capture by the Deterministic-Stochastic classification.................. 176
Figure 6.11 – Noisy Partials generated by noise modulation of a sinusoid...................................... 179
Figure 6.12 – Synthesis of a Noisy Partial 179
Figure 7.1 – Arrangement of Audio Tracks in a Post-Production Editing Suite (example screen-shot courtesy of Soundscape Digital Technology Ltd.) 183
Figure 7.2 – Swapping the roles of generative and transformative synthesisers............................... 186
Figure 7.3 – Using a neural network to determine the mapping function 190

ACKNOWLEDGEMENTS

First and foremost, I wish to express my gratitude to Prof. Andy Bateman who has helped this thesis from its inception to its conclusion. He has shown a willingness to move into an unexplored research area for the department and has been an enthusiastic supporter of the work, both within the department and in our interactions with external business, media and research communities. Above all, I must personally thank Andy for his faith in me, and (particularly in the final write-up) for his patience!

For making this research project financially possible, my thanks go to Soundscape Digital Technology. In particular, I am grateful to Chris Wright whose faith in the project has been an enabler at a number of critical junctures. (Their kit, the SSHDR1 hard disk recorder, has also been a boon.) I wish them every success with future developments.

Within the department of electrical and electronic engineering, my thanks go to Prof. Joe McGeehan, whose behind-the-scenes support has given this project a home within the Centre for Communications Research. Not only has the CCR been a logistical, financial and technical support, I am glad to have been in the company of the staff and postgraduate students of the lab. Ross & co. had to put up with hearing the dulcet tones of Tori Amos singing “Friday morning”, just a few too many times before we invested in headphones. In 5.01, Will, Tim and Martin have had to put up with not-so-dulcet tones of despair or elation from me, despite the headphones.

In the closing stages, Dr. Will Peasgood has shown a selfless willingness to proof-read all aspects of the thesis (except this of course). His comments have been insightful and to the point. Strangely though, now it is finished he wants to read it for pleasure?!! Bizarre!

Of a more personal nature, my deepest thanks go to all those who have been of support to me over the course of this Ph.D. as friends and flatmates of various abodes. I am happy to know Jonathan Pratt and Simon Craske, who have been friends for many years, and who have also gone through /are going through the birthing process of delivering a thesis. In my penultimate flat, Herr. Dr. Hans F. Müller was great to have around for his humour and generosity, as was Lizzie for her warm spirit and kind heart.

In my current dwelling at “The Ferns”, I am very, very grateful to Fay and Jan, who have quickly become good friends. From the start, they have had words of enthusiasm and have shared my excitements at breakthroughs; they have also been an invaluable help with keeping me fed and generally alive – I am indebted to you both.

Silent thanks also go to those whose personal and spiritual guidance have brought me to this point. This is not a cop-out for forgetting people, but a way of expressing thanks without putting names in print – I know one or two would prefer it that way.

Finally, my family have also become my friends, and I would like to thank them for being who they are and for being special to me: Mum and Dad (Diane and George), Maria and Anita and Dave. Especial thanks to Dad, who has supported me through his regular contact and his unfailing prayers and blessings (and without whose computer, I might have had to write my thesis at college!)

CHAPTER ONE

INTRODUCTION

1. INTRODUCTION

This thesis serves as a summary of research activities over the past four years in the subject of computer modelling of musical signals. The primary research goal is the development of a sound model and the analysis-resynthesis tools to implement it, with the target set on perfect representation of sound.

In essence, the model is a set of parameterised sound features. The choice of features defines the scope of the model and their parameters define properties of the sound within that scope. If the feature set matches the properties of a particular sound, the model has in effect *captured* the sound's features. Therefore the model can be applied to existing sounds and the feature set can be considered a specialised representation of sound.

All aspects of this thesis revolve around the concept of the *representation* of sound. The time domain waveform, for example, is a perfect representation of an audio signal. It captures all the features of a sound, because it captures the air pressure fluctuations that were generated by the sound source(s) at the recording location.

If the aim of the model were simply the storage and reproduction of sound, then the time domain sample would be the ideal representation, because it captures all the features and it easily facilitates reproduction. The purpose of the presented study, however, is for musical application. The sound model should not simply capture features for reproduction (although it should be perceptually accurate), but should represent them in such a way as to facilitate modifications that have musical meaning.

In the time domain sample, the features are present but they are fused together, so that they are not easily accessible, and therefore not easily transformable. Within the sound model representation, the features must be distinguishable, so that individual features can be accessed and/or modified. Hence the model is implemented as an analysis-resynthesis system; see Figure 1.1.

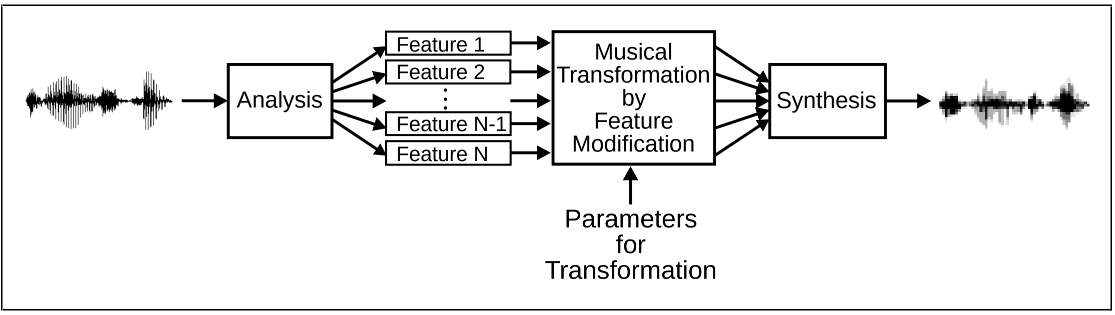

Analysis
Feature 1
Feature 2
Feature N-1
Feature N
Musical Transformation by Feature Modification
Synthesis
Parameters for Transformation

Figure 1.1 - Sound analysis-resynthesis system

The analysis process 'looks for' certain properties in the sound signal and compiles the results into a data set, in effect separating the features. The musician can then transform the sound in the model domain, by modifying the features. To hear the results, the synthesis process recombines the features, generating a time domain sample which can be played.

With the purpose of the analysis-resynthesis system defined as a tool for musical transformation, the search begins for the model components – the sound features – with which a sound is to be represented. This investigation begins in chapter two, which considers what features to look for in a sound and reviews previous studies to provide a basis for the investigations of this thesis. Further details of the chapter breakdown can be found in section 1.3 below.

Before examining the finer details of the model, the following discussions give an overview of how the field of Computer Music research has evolved and includes a summary of the role of analysis-resynthesis.

1.1 Computer Music and the Field of Sound Modelling

The research field known as ‘Computer Music’ is relatively young, having only started – in a co-ordinated sense – about 25 years ago. It is itself a synthesis, having arisen from scientific research applying signal processing techniques to musical sounds and from music research which began to use computers as its tools for automating composition analysis. Today, the subject remains just as eclectic, as can be witnessed at the annual International Computer Music Conference (ICMC), which attracts both the scientific and music communities and includes both technical presentations and concerts of electro-acoustic music[1].

The current areas of research activity span a wide range of disciplines, whose common thread is the application of computers to music. At one extreme this includes philosophical discussion of the aesthetics of electro-acoustic music, and at the other extreme machine recognition of (traditional) music, which applies psychoacoustics and the theories of timbre perception, to model rhythmic and melodic structures. At one end of the spectrum it includes methodologies for synthesising sound with all its micro-textural detail, whilst at the other end it includes interactive performance, in which a computer controls a synthesiser to provide an accompaniment to live jazz improvisation of human performers. Within composition, it encompasses, on the one hand, algorithmic composition (in which the computer generates the composition), and on the other, data structures and notation schemes, for developing new music representations that are better suited to the increasing degree of sonic flexibility than the traditional method of score notation.

This thesis concerns itself with the modelling of sound textures. It is the application of engineering skills in signal processing to model sound signals. However, in encountering the aesthetic world of music, it is not untouched; the discipline requires modelling that goes beyond the sterile technicalities of audio compression and reproduction, towards a representation that integrates sufficient understanding to facilitate musical transformations. As such, there is an essential subjective element to the investigations, where the human ear becomes the ultimate spectrum analyser, presiding over the electronic ears of automated signal representations.

[1] Electro-acoustic music means music generated electronically, but it has come to mean music that goes beyond the normal concepts of rhythm and melody and uses computers in some way to actuate or synthesise sound.

1.2 Analysis-Resynthesis and its Applications

> "These days ... people make more use of samples, attempting to push them further in terms of performance and expression. Modern synths and workstations have sounds based on samples, so why shouldn't their sound manipulation and performance capabilities be available on machines that actually record samples?"
>
> ——— *Future Music magazine [Evans,1993]*

The above quote was taken from a product review in a popular UK music technology magazine, which appeared shortly after this work began. It highlights the gap in music technology for tools that can bring the flexibility of synthesisers to sample-based systems.

Synthesisers gain their flexibility by having a parameterised architecture, where each parameter can effect a change in the timbre of the sound. Samplers conversely, are glorified record-playback devices with flexible editing and filtering controls, but no parameters for timbre control. Analysis-resynthesis bridges the gap by bringing parameterised control to samples. It has the same ability as a sampler to record and replay, cut and paste, reverse and filter, but it also gains the advantage of the synthesiser to manipulate the timbre by parametric control.

The immediate applications for analysis-resynthesis are obviously as the engine of a sample manipulation system, as might be found in a post-production editing suite. However, the fact that it bridges the gap between the disparate formats of synthesisers and samplers may lead it to become a universal sound description format, so that sound data would be communicated between systems in a flexible format, only being synthesised at the point of digital-to-analogue conversion.

This brief discussion indicates the importance of analysis-resynthesis modelling as the basis of a musicians' tool. The discussion is pursued in more depth in chapter seven, where the results of the intermediate chapters are also taken into account.

1.3 Overview of the Thesis

The theme of this thesis is the features that make up the sound model's data set. Under consideration are the choice of features, how well they represent an arbitrary sound, how usable they are for musical transformations and how they may be extracted from a source sound during the analysis phase.

Chapter 2

Since the data set is not an arbitrary mathematical description, but must be relevant to the musical nature of the sounds to be processed, the first step in the thesis is to observe properties of sounds that can be considered musical. Chapter two begins by exploring a possible definition for the term 'musical signal'. Since many people's views differ as to what is or is not a musical sound, this definition is not intended as an absolute, but merely as a basis from which to begin building up a model. (As the model matures, the choice of features can be refined until an arbitrary sound can be analysed and used musically under transformation.) This catalogue of observations establishes a foundation, from which model developments can evolve, as a process of matching signal processing techniques to perceived phenomena, and a benchmark against which achievements can be evaluated.

The chapter follows with a review of literature in the field of sound and musical instrument modelling. The review forms the technical foundation for the presented work, both by placing it in context and through its value as a source of experience and wisdom within this field. Indeed the initial model used for the thesis work, presented in the next section, draws largely from these references.

The Deterministic Plus Stochastic model [Serra,1989] is presented as the basis of the 'Initial Model', so the description begins with a summary of this model. Where changes have been made, these alterations are explained and justified. Each description is accompanied by a critique that aims to inform the reader of the merits and demerits of the techniques employed. In this way the major issues are brought into perspective. Briefly, they include:

- the lack of automation caused by the need to manually set many analysis parameters;
- the lack of resolution in the Short Time Fourier Transform (STFT);
- the inherent weakness of the 'Nearest Frequency' approach to frame linking, which is also a by-product of the lack of resolution in the STFT;
- the inability of the model to respond to fast changes in a sound, particularly attack transients;
- the lack of fusion between the deterministic and stochastic aspects of synthesised sounds.

Chapter two, therefore, provides a launching pad for the investigations of the thesis, providing both the technical and musical bases from which to begin the investigations.

Chapter 3

The feature set that is used in the thesis is based on a time-frequency representation (TFR) of sound signals. Therefore it is of the utmost importance that this representation be accurate. More specifically, the representation must be capable of presenting *the data that is required* accurately. The Fourier Transform (FT) and its algorithmic counterpart, the Fast Fourier Transform (FFT) have long been bemoaned for the lack of time-frequency resolution. Yet the STFT, which is composed of FFT frames of multiple time segments, is the accepted time-frequency representation for sound decomposition. For fast changing spectral components, which include the higher partials of even the most static pitched sounds, its representation can be rather inadequate.

Chapter three is concerned with the appropriateness of the time-frequency representation (TFR) for sound analysis. The chapter begins with a discussion promoting the necessity first to decide what is needed from a TFR, and then to choose or design a representation that best satisfies those needs. It is shown that this decision depends on the features to be extracted for the model, which in turn depends on the application for the model. The basis of the discussion is that, as model designers, our perspective on the problem colours what we can see from a representation. For example, the FFT presents the localised spectrum of a signal without ascribing a meaning to the data, yet we can choose to interpret a maximum as evidence of a partial, and in so doing, we can gain information and suffer from perceived distortion.

The chosen application for demonstrating the effectiveness of the model is time-stretch. The desired effects under transformation determine the features of greatest importance, and this in turn indicates the properties that must be easily extracted from a TFR.

The second part of the chapter presents a number of alternative mathematical distributions including the wavelet transform, higher order spectra (e.g. bilinear, Wigner) and parametric techniques (e.g. ARMA). This is followed by a catalogue of innovative TFR's that have been constructed from mathematical distributions, but do not rely on the distribution's 'plain image' for extraction of information. The merits and limitations of applying each as a TFR for sound analysis are examined.

One particular avenue is explored in more depth in the third part of the chapter: that of extracting higher order information (usually considered limited to higher order spectra) from the FFT, by application of model information. Briefly, knowledge that the FFT contains *all* information about the signal represented (not just information about stationary spectra) is combined with the expectation that peaks in the FFT are indeed partials, to yield data about the rate of change of frequency and amplitude, in addition to the usual frequency, amplitude, phase measurements. This is achieved by considering changes in phase across each peak as an information source, instead of labelling it 'distortion'.

The advantage of this extra information is an effective improvement in time-frequency resolution. In addition, the information could be usefully employed to assist in peak validation (deciding whether a maximum is a partial or a spurious FFT artifact) and peak linking between frames.

Chapter 4

Partial trajectories are the primary model feature for describing the pitched aspect of sounds. For this frame-based time-frequency analysis, each partial's instantaneous status is identified from peaks in the spectrum of each frame. In order to preserve continuity upon synthesis, there needs to be a method for linking the peaks between frames, to reconstruct the partials' trajectories. This is based on linking a peak in one frame to the peak of nearest frequency in the next.

Chapter four presents an argument for replacing the 'nearest frequency' linking method with one based on the harmonic structure. The first part of the chapter highlights deficiencies in the nearest frequency method that would be overcome (or improved upon) by changing to the harmonic method. One obvious motivation is that only peaks which are partials (i.e. those that belong to a harmonic structure) are desired, so a method based on that structure is an obvious choice.

The idea of harmonic linking is not new, but to date most of the methods employed have been rather primitive. The next section of the chapter describes a method that determines the fundamental frequency (where a harmonic structure exists), links peaks to the harmonic structure within each frame and then links the structure between frames, by virtue of the slow changing fundamental. The method has been designed to cope with the limitations of fundamental frequency detection algorithms, so that suspected miscalculations can be corrected. In addition it can track the structure when it is weak, such as at the start and end of a voiced phoneme in speech.

Chapter 5

One limitation of time-frequency representations already mentioned is that of resolution. The need to correctly separate partial peaks in the spectrum forces time resolution to be surrendered in exchange for frequency resolution. This makes the model suitable only for relatively slow changing spectra. This trait is further reinforced during synthesis, where the model features are slowly faded in and out between frames. For the majority of sounds, there is little or no problem (beyond a perceived increase in reverberation), however for sounds that include percussive attacks, the result upon resynthesis can be an audible diffusion or even a total dropout.

Chapter five begins by examining this problem in terms of both the perceived effect and its signal processing cause. Problems exist both in analysis, where sudden changes in the waveform cause large scale distortion in the FFT spectrum, and upon synthesis, where the slow fades soften the hard edge of an attack transient.

The solution, described in the next section of the chapter, uses a detection scheme to locate attack transient onsets, and then restrains the analysis and synthesis around these locations, to avoid the problems. The method enables a new feature – the attack transient – to be incorporated into the model without much impact to the established model structure.

Chapter 6

Chapter six provides an opportunity to reflect on the work presented and to consider where it can lead. Each of the innovations of chapters three to five are individually assessed and critiqued, in much the same way that chapter two approached the Deterministic Plus Stochastic model. The impact on the model is examined in terms of changes required to the analysis-resynthesis structure and the effect on computational load. Then the specifics of the methodology and its implementation are examined, highlighting achievements and pointing out weaknesses. The effect of combining all three studies is also appraised, indicating the synergies that can be gained and the areas that require further investigation.

Following on naturally from the critiques, future research directions are proposed for each of these areas – suggestions that are made by virtue of experience gained during the work of this thesis. One such example is a major direction for new work, whose need became evident over the course of the described studies, although it is not a direct extension to any of the aforementioned investigations. This is presented in the final part of the chapter.

The discussion questions the validity of the deterministic-stochastic classification and suggests a reshaping of the model that should improve sound quality, simplify computation and data storage, and reduce the data set. The basis of the idea is that the noisiness of sounds is often not additive, but results from small scale chaotic variations in the partials themselves.

The discussion proposes that the hard classification between deterministic and stochastic be replaced by a smooth transition, where only the extremes are perfectly sinusoidal partials and totally additive noise. In between lies a spectral component that is neither one nor the other – a noisy partial, whose noisiness becomes a model parameter. The proposal is justified through observations of physical sound generators and evidence from signal processing. Means for implementation with its implicit reshaping of the model are discussed and potential routes for proceeding with the research are suggested.

Chapter 7

Chapter seven discusses practical applications of the sound model in two parts. The first part looks at how it could work to improve musical systems. Already indicated in section 1.2, there is a difference in the roles and methods of sound creation between synthesisers and samplers. This is a result of the different representation formats and synthesis architectures. The sound model provides a common platform for all sound description. With intelligent implementation it could not only enhance the sound quality and flexibility of the existing systems, but also open up possibilities for new (and more intuitive) ways of creating and working with sound. Applications are presented for both live performance tools and studio-based post-production editing consoles.

The second part of the chapter presents applications of the sound model in new research. Two examples are discussed, which are ongoing projects at the University of Bristol's *Digital Music Research Group*. The first is a scheme applying neural networks to enable the sound model to become an instrument model, capturing and facilitating manipulation of playing controls as well as sound features. The second applies a simplified form of the analysis-resynthesis system to speech synthesis; its flexibility for manipulation is ideal for creating text-to-speech systems that can emulate natural intonation.

Chapter 8

Chapter eight concludes by reviewing the goals of sound model development. The specific objectives of this thesis are restated and the achievements of each of the studies are summarised in this light, to demonstrate how they have helped to move sound modelling toward those goals. The closing discussion takes a look at the final model in comparison with the current state-of-the-art and then looks forward to where new research is leading.

Appendices

Appendices A-C provide the reader with copies of the author's publications to-date in this subject area.

CHAPTER TWO

MUSICAL SIGNALS

AND HOW THEY CAN BE MODELLED

2. MUSICAL SIGNALS AND HOW THEY CAN BE MODELLED

The first part of the chapter looks at what is meant by the term 'musical signal'. Various properties of sounds are explored to identify which ones constitute a musical sound, in the context of the work of this thesis. As these properties are examined, their signal form is also presented, the first step in developing a model for musical signals.

The second section is a review of the prior history of modelling musical instruments and sounds. This covers a range of literature but it is presented in categories, based on the broad approach to modelling of each study.

From this basis, the third and final part of the chapter describes 'the Initial Model', which is to be the starting point for all developments described in later chapters. The Initial Model is based on the innovations of previous sinusoidal models (described in the review section, 2.2.2.1) because these have been shown to provide the closest correlation between model features and observed properties of sounds. One particular variant is focused on, which demonstrates an enhanced flexibility for sound transformation. The Initial Model is described in some detail, its aims and its functionality, so that the investigations of the later chapters are placed in a clear context.

2.1 What is a Musical Signal?

First of all, why do we need a definition of a musical signal in order to model sound? The answer is: because we want to transform sounds in ways that have musical meaning. Therefore we need to be able to extract features from the audio signal that have musical value.

Building the Model by Expanding the Definition

Just as the word 'sound' is specific from the word 'audio', in that it implies a listener and therefore it is a perceptual description of a signal, so too 'musical' specifies a usage of sounds that is artistic and creative. It is true that any description limiting which sounds can be used in music will not be universally acceptable. However, in the context of developing a model, such restrictions make it easier to begin classifying features.

Within this section, the description is initially narrow, enabling one class of sounds to be included. Upon consideration of the next property, that description widens. In this way, with ever decreasing steps, the description of 'musical signal' approaches parity with 'sound signal' and a set of features is compiled that approaches the total domain of audio signals. This process is presented graphically in Figure 2.1 for the properties considered in this section.

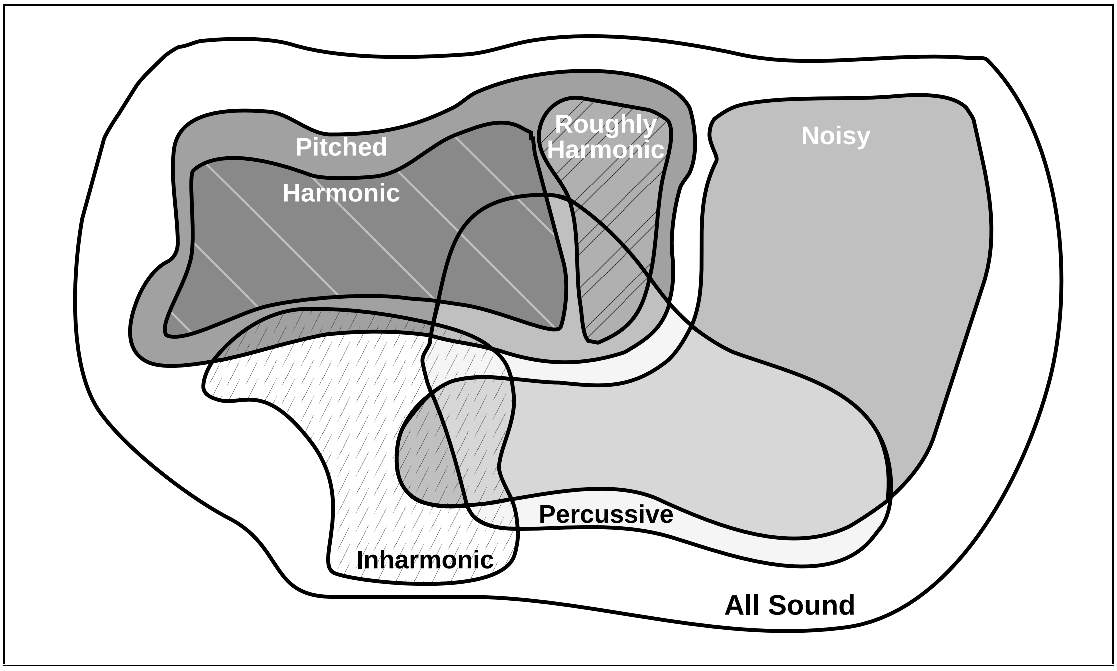


Figure 2.1 – Venn diagram of the set of all sounds

Our Improved Understanding of Sound

The first experiments to understand the composition of sound are credited to Hermann von Helmholtz (whose work is translated in [von Helmholtz,1954] – it was originally published in 1877). His work effectively established the notion of the sinusoidal partial and the concept that timbre – the tonal equivalent of hue – was related to the strengths of the harmonic resonances. Indeed the results of these experiments on the steady-state portion of sounds persisted to the latter half of this century, in which synthesis of the steady-state spectrum alone was hoped to be sufficient for reproducing instrument timbres. The poor results of steady-state synthesis marks the start of the recent investigations into the composition of sound and how it may be synthesised.

Much of today's research into sound decomposition comes from the standpoint of creating flexible tools for musicians. However the refinements to the model also reflect improved understanding of the composition of sound. In the past 20-30 years understanding of sound has progressed at a rapid pace (compared to previous history). Yet even today our understanding (our model) is somewhat limited.[1]

2.1.2 Harmonic Structure

The most important component of conventional music is melody, the sequence of notes or pitches. Pitch is perceived in signals from the rate of repetition of the audio waveform; the higher the frequency of repetition, the higher the pitch.

By application of Fourier theory – more specifically the Fourier Series – any waveform that is periodic can be described as the superposition of harmonically related sinusoids. These sinusoids, whose frequencies are all integer multiples of a fundamental frequency, sum to generate a waveform whose period is that of the fundamental. In musical parlance, these sinewaves are called *partials*, where the index of the partial is its frequency as a ratio of the fundamental frequency. (i.e. the fundamental, at frequency f_0, is the 1st partial; the first harmonic, at frequency $2f_0$, is the 2nd partial; etc..)

2.1.2.1 Short-Term Periodicity

Fourier theory requires that a periodic signal be *eternally* periodic for it to be truly described by the Fourier Series. Sound waveforms can change rapidly, but there is sufficient repetition of waveform shape to indicate short-term periodicity. This is termed *quasi-periodicity*. If the rate of change of the waveform is slow compared to the fundamental frequency, say by an order of magnitude, then the signal can be analysed in short-time sections where the approximation of periodicity is locally good.

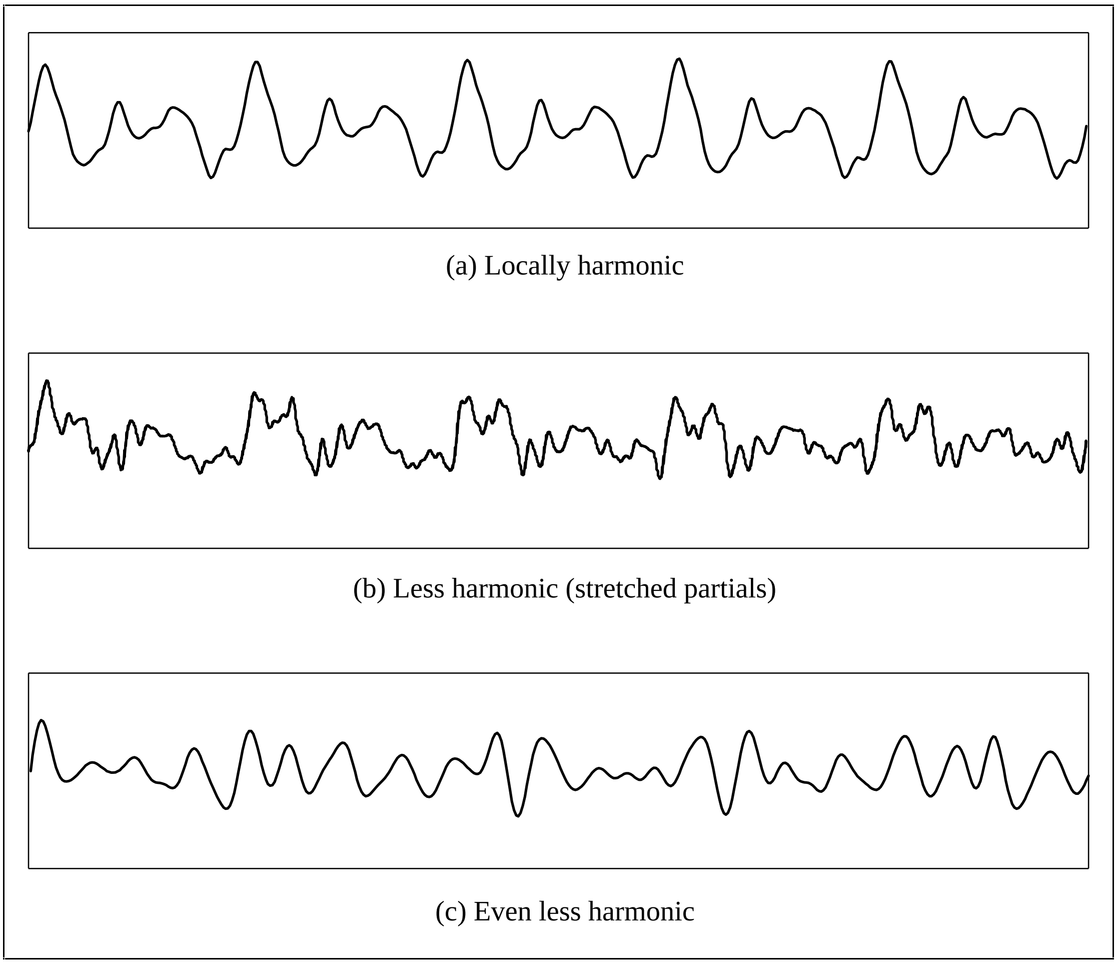
In terms of frequency, periodicity is described as 'stationarity', because the sinusoidal components are stationary, with no variation in frequency or amplitude. For quasi-periodic signals, the components are not true sine waves. However they appear locally stationary – 'quasi-stationary' – and within this thesis, they are termed locally or instantaneously sinusoidal for intuitive simplicity.

With the local approximation to stationarity, the harmonic model of the Fourier Series can be approximated for each short-time section. Linking the partials between sections then yields 'partial trajectories'.

[1] For example, within this dissertation it becomes clear that we have incomplete understanding of what happens in the first few milliseconds of a percussive note onset. Are there rapid variations in the resonances that are too fast to track? Is the concept of the partial (defined in section 2.1.2 following) not valid at that time? Maybe the spectral elements are 'unfocused partials' that become focused as steady-state conditions are approached.

2.1.2.2 Non-Harmonic Overtones

Partials other than the fundamental are often called harmonics. A more accurate term is ‘overtone’, because not all overtones have strictly harmonic frequencies. The most commonly cited example is the piano which has ‘stretched’ partials; the frequency ratio of overtones is slightly greater than harmonic, giving the impression of a stretched harmonic structure [Pierce,1992]. However the perception of pitch is not lost, and although the waveform is never truly harmonic (even locally), there is a *loose periodicity*. If the degree of inharmonicity is small, the sense of pitch remains strong, but as inharmonicity is increased that sense is weakened. Similarly, for a small degree of inharmonicity, periodicity is apparent visually in the waveform, but the pattern weakens as the inharmonicity increases; see Figure 2.2.


(a) Locally harmonic

(b) Less harmonic (stretched partials)

(c) Even less harmonic

Figure 2.2 – Visible periodicity depends on harmonicity of sinusoids

In application to a musical sound model, instantaneous sinusoidal components can only be considered partials of a sound source if their frequencies are *roughly harmonic*. This necessitates that the analysis process extract the features of each partial separately.

2.1.3 Vibrato and Tremolo

Vibrato and tremolo are common devices for introducing expressive modulation of a note. In the signal they appear as oscillatory modulation in frequency and amplitude respectively. Stringed instruments are better adapted to vibrato through variation of the ‘active’ string length, whereas tremolo is easier on brass instruments through breath control. (It is also possible for string vibrato to introduce some tremolo at the same rate, although the oscillations are not necessarily in phase.) It is interesting to note that perceptually the presence of oscillation is more dominant than the type of oscillation, and to the untrained ear vibrato and tremolo are not easily distinguished.

Oscillatory modulations have been shown to be important in sound synthesis, where vibrato particularly aids in the perceptual 'fusing' of partials. That is, the partials no longer sound like separate sine tones once vibrato is introduced [Pierce,1989]. It is no coincidence that the phase and rate of modulation is common to all partials of a note. Indeed this property has been used successfully during auditory scene analysis to group partials into streams, where each stream corresponds to one sound source within a composite sound [Stainsby,1996].

For the purposes of extracting musical features, it is important that the oscillatory modulations be detected separately from the frequency components of the harmonic structure; for musical relevance they modulate the partials, but do not form part of the harmonic structure themselves. This places a tight constraint on the time-scale for spectral analysis: the analysis must be capable of resolving partials at the bottom of the audio range, say 40-50 Hz, but it must also track their modulation which could be faster than 10 Hz. This example is considered further, in the question of how to choose a time-frequency representation, in section 3.1.

2.1.4 The Amplitude Envelope

In addition to the varying spectral profile, the time domain amplitude envelope also plays an important role in the perception of timbre. This is particularly true for the initial transient of a note, termed the 'attack' [Risset & Mathews,1969]. The truth of this can be validated through two simple experiments:

- Play only the steady-state portions of sounds (without additional perceptual clues such as vibrato) with a gradual fade in. The spectral envelope is the only perceptual cue, so it becomes difficult to identify the source instrument; it is even possible to confuse whole families of instruments, such as strings and woodwinds;
- Play just the initial transient of sounds (say, the first 50ms). This alone is often sufficient to identify an instrument, especially if the note onset is usually fast.

Amplitude envelopes were first incorporated into synthesisers with the ADSR (Attack-Decay-Sustain-Release) four-stage profile, for the frequency modulation (FM) synthesisers of the early 1980's. The importance of the attack signature of instruments was recognised in the Sample-plus-Synthesis (S+S) synthesisers of the late '80s, which used a sample for the attack and FM (or other) synthesis for the remainder of the sound. This development was primarily because of the difficulties of synthesising the complexities of the attack.

During the initial transient there are many rapid changes, so that it can sound like a noise burst (e.g. for trumpet [Risset,1991]). However, the process is not truly random. In traditional instruments, the initial transient is the phase during which resonances are building up, but the steady-state condition of standing waves has yet to be established. Because the attack is very short and rapidly changing, it is difficult to study. As a result it is a difficult property to incorporate into a sound model.

2.1.5 Unpitched Sounds

So far the description ‘musical signal’ has implied only sounds with a pitch, but there are musical instruments that produce unpitched sounds. These include ‘noisy’ sounds and rhythmic instruments. Many percussion instruments provide short transients without pitch, so they are ideally suited where rhythm is desired without affecting the melody. More prolonged unpitched sounds include the vortex noise of the breath (which accompanies many instruments), sibilants and fricatives (‘s’, ‘sh’, ‘f’, ‘h’) and whispered speech. All these vocal-based examples are termed ‘unvoiced speech’, because there is no (pitched) resonance of the vocal cords. Natural world examples of unpitched sounds exist in the sound of rain, the ocean or the rustle of leaves, which are increasingly used for ambiance in music. Other examples of unpitched sounds include tape hiss and static, which are not (usually) part of the music but are nevertheless present in audio recordings.

‘Unpitched’ sounds have no *absolute* pitch, but they do possess a *relative* pitch. One noise sounds higher in pitch than another, but neither can be identified as having a specific pitch on a musical scale. Whereas pitched sounds are composed of discrete narrowband spectral components – the instantaneous sinusoids – unpitched sounds exist over a range of frequencies, and can be considered as band-limited noise. (A noise in a higher frequency band sounds higher in pitch than a noise from a lower frequency band.) The essential difference between pitched and unpitched sounds is that the former is narrowband and the latter is wideband, but there is no definite cut-off point; narrowband noise possesses a weak sense of pitch.

2.1.5.1 Additive and Non-Additive Noise

The above discussion considers sound sources which are exclusively unpitched in contrast to the previously explored pitched sounds. However in most cases, there is a combination of pitched and unpitched elements from a single sound source: the breath noise with the tone of a flute, the bow noise with the note from a cello, the combination of voiced and unvoiced aspects of speech and singing.

Noise-like waveforms can be modelled as additive white Gaussian noise (AWGN) which has been band-limited and shaped (filtered). In the nonstationary case, chaotic sounds present short-time noise-like waveforms and can be modelled as *dynamically* filtered AWGN. However this assumes that all perceived ‘noisiness’ is additive, where the noisy aspect of the sound is added to the pitched aspect, which is only true to a degree.

The sound of breathiness from a flute is the addition of the musician’s breath noise to the note from the flute. But the note is produced as a result of the breath’s excitation of an air column. That is, the process by which the pitched aspect of the sound is generated, is itself dependent on a chaotic signal, and cannot therefore be free of (small-scale) chaotic behaviour. Similarly, the drag-slip oscillations of a bow against a string generate the excitation energy that is transmitted along the string. So any chaotic variations in the rate or degree of drag must affect the sound generation mechanisms. The noisiness of the resultant sound is therefore an intrinsic part of the variations in the waveform and not some external source added to it.

In summary…

The definition of ‘musical signal’ includes both pitched and unpitched, gradually varying and percussive sounds. Pitched sounds are considered locally periodic and can be treated as periodic in the short term. Spectrally, such sounds have a harmonic structure, where the partials are instantaneous sinusoids, whose frequencies are at roughly harmonic ratios. Thus partials are dynamic and semi-independent, but modulation such as vibrato and tremolo affect all partials (of one instrument) in the same way. The attack portion, the initial transient of a percussive note onset, is chaotic (but not truly random) behaviour, resulting from rapidly varying resonances before the stability of standing waves is established. Unpitched sounds possess a noise-like quality and can be considered as dynamic wideband spectral components. Most sounds are a combination of pitched and noisy aspects, which can only be considered additive to a degree.

2.2 Review of Research into Sound and Instrument Modelling

The following literature review presents an overview of the various approaches that have been pursued, within the fields of sound and instrument modelling. The approaches have been categorised into three broad classifications: physical modelling, analysis-resynthesis and granular synthesis. Within each of these areas, the aim has been to present the current research activity and the history that has led to the latest innovations.

2.2.1 Physical Modelling

A physical model is a *causal model*; it encapsulates the musical instrument – the cause of the sound – rather than the sound itself. The term implies that the physics of the instrument is simulated, however it is increasingly used to denote any synthesis architecture that is specific to a particular instrument or instrument family, especially if there is parametric control that appears to correlate with changing some physical property. In such cases the term ‘physical modelling’ is somewhat a misnomer – ‘instrument modelling’ might be more appropriate.

An example of a true physical model is the CORDIS system developed at ACROE [Cadoz & Luciani & Florens,1984]. The initial version of this model was applied to strings and was intended for real-time use by musicians. The model consisted of many interconnected ‘cells’, each cell possessing mass, elasticity and damping properties. The cells were connected in a three-dimensional lattice so that vibrations could be modelled in the transverse, longitudinal and rotational axes, with the minor simplification of no interaction between these modes. The excitation mechanism (e.g. hammer, plucking) was modelled much more simply as a time-conditional displacement or force.

Vibrational simulation occurs as a result of propagation of the forces and displacements of the individual cells. In order to model the modes of vibration with sufficient accuracy, it is necessary to have in the order of a hundred cells per string. With the spatial equivalent of the Nyquist criterion, a string which is one hundred cells long would allow simulation up to the fiftieth partial. That is, there need to be at least two sample points per standing wave period along the string.

Such physically accurate models are of theoretical importance because they help us to learn more about the physical mechanisms of vibration generation and propagation. However the requirement to simulate hundreds of cells, each requiring the solution of computationally burdensome force and displacement equations at each time sample, has hindered the practical (musical) use of such models.

2.2.1.1 The Wavetable Model

At the same time as the CORDIS system was being developed, Kevin Karplus and Alex Strong developed a synthesis method that was extremely simple and efficient for instrument modelling [Karplus & Strong,1983]. This was based on dynamically modifying the contents of a wavetable in such a way as to simulate the acoustic response of a musical instrument.

The synthesis method was based on wavetable synthesis, in which a sample is cyclically output. Regardless of the contents of the sample waveform (called the 'wavetable'), a tonal sound is generated as a result of the repetition. The contents of the wavetable influence the relative strengths of the harmonics. However, the lack of variation in the waveform from one period to the next makes the sound perceptually dull.

Karplus and Strong set about to find ways of modifying the synthesis that would inject life into the sounds. This was predominantly achieved through dynamic modifications to the wavetable itself – as soon as a sample was read for outputting, that point in the table would be modified, so that each period of the output sample would be different from the last. Strong discovered that by recursively smoothing the waveform, the sound was similar to that of a plucked string. Further investigations showed that this was due to the way that the harmonics decayed, highest first, that was similar to the decay of vibrations in a real plucked string. Karplus extended the method to drum-like sounds, by partially randomising the wavetable during the synthesis. Further experiments also yielded modifications that improved control over the rate of decay.

The Karplus-Strong method became a basis for physical modelling, despite having no explicit reference to the physics of musical instruments and no formal model. Its popularity was probably driven by the fact that in addition to evoking the characteristics of musical instruments, it was simple to implement and required very little processing power.

2.2.1.2 Digital Waveguide Modelling

In their 1983 paper [Karplus & Strong,1983], Karplus and Strong noted that their generic design could also be described from a digital delay-line standpoint. The digital delay-line is the basis of the digital waveguide, which became the next revolution in physical modelling. This is probably no accident, since Julius Smith, whose 1992 paper [Smith,1992] placed the digital waveguide on the map[2], had also worked on extensions to the Karplus-Strong algorithm [Jaffe & Smith,1983].

The difference between the methods lies in its application. Whereas Karplus and Strong proposed direct synthesis from the wavetable, Smith incorporated the waveguide as an element within a more traditional physical model. The digital waveguide is, in essence, a digital delay coupled with a filter. Its role in the physical model is to simulate the lossy propagation of waves through resonant objects. (The delay simulates the propagation delay and the filter simulates frequency-dependent losses.)

The waveguide approach trades the model of many simple components for the model of few semi-complex ones. Instead of hundreds of elements to represent the points along a string, there are two or three[3]. Instead of each element responding to the forces of its neighbours thereby effecting the propagation of a wave, each element encapsulates the wave motion directly and the results are only known at the points where there is some external interaction.

[2] This is the most oft-quoted reference for digital waveguide modelling and it provides the mathematical and physical basis of the technique. However, some work by Smith and others using digital waveguides predates this paper [Smith,1987, 1991; Cook,1991].

[3] A string could be modelled (for transverse waves) with a single waveguide. However it is split into sections to enable placement of excitation points (i.e. point of bowing, plucking, striking) or other points of interaction.

In fixing the modelled elements by their interconnections instead of their physical co-ordinates, it is possible to vary their physical attributes in real time. This begins with variations in exciter position (e.g. bowing point), which is essential for realistic performance, to altering the dimensions and material construction, which is impossible or unfeasible in conventional instruments but allows for realistic-sounding modifications to the physical modelling synthesiser.

2.2.1.3 Current Research Activity

1D to 2D to 3D

The model, as put forward by Smith, enabled synthesis of strings (for guitar, piano, violin, cello, etc.) and air-columns (for oboe, clarinet, trumpet, saxophone, etc.) with a one-dimensional digital waveguide. Recently this has been extended to two dimensions for synthesis of plates and membranes (for drums, gongs and cymbals) [Van Duyne & Smith,1993]. Further work continues to extend this to three dimensions [Van Duyne & Smith,1996], where closed spaces can be modelled. (Work is ongoing in this area within the University of Bristol's *Digital Music Research Group*, though there are no publications to date.)

Introducing Chaos

The waveguide is linear, whereas real resonators only approximate linearity to an extent. So in digital waveguide based models there is the need for nonlinear elements, which serve to make the models' mathematics better resemble observed instrument behaviour. Since the waveguide is implemented as a delay with linear feedback, nonlinearities can be introduced by modifying the feedback filter. The problem with introducing nonlinearities into the loop is that stability is no longer guaranteed. The intention is that the generated chaos will be small-scale, causing random perturbations around a point of stability. Xavier Rodet has provided a method for introducing noise at frequencies around the partials, by causing controllable instabilities with a low pass feedback loop [Rodet,1994]. Other recent contributions include [Chafe,1995; Radunskaya,1996].

New Exciters

The digital waveguide models the resonant structure within an instrument. The element that drives the model is the exciter; through this element it is possible to enable the subtle (or not-so-subtle) nuances of a real player through expressive controls. Exciter modelling has progressed in the last couple of years from modelling the basic action that enables 'normal' playing (e.g. lips against trumpet [Rodet & Depalle,1992b; Rodet & Vergez,1996], string on bow [Smith,1993]) toward more complicated techniques (e.g. slap-tonguing and multiphonics for woodwind instruments [Scavone,1996], sul ponticello or spiccato for string instruments [Jaffe & Smith,1995]).

Computational Cost

Despite the added complexity in the structure that has grown around the digital waveguide, a central focus in the implementation of these models remains computational simplicity. Often the new innovations are accompanied by an explanation of how few 'multiply' instructions are required. Without a doubt, it is the relatively low processing load of digital waveguide synthesis (coupled with the real-time controllability) that has made commercial physical modelling synthesisers a reality – the Yamaha VL1 was launched in January 1994.

2.2.2 Analysis-Resynthesis

'Analysis-Resynthesis' describes the systems based on sound modelling without recourse to the instrument or sound source. Each use of the system involves the analysis of a source sound, after which many syntheses are possible. Analysis is the process of feature extraction, where the model parameters are estimated from the source material. Direct resynthesis should reproduce the source sound, whilst synthesis of modified parameters should yield sound transformations.

2.2.2.1 The Sinusoidal Model

The original model, and the most enduring one, is the sinusoidal model. Based on the concept of the harmonic structure and its mathematical decomposition through the Fourier Series, the inspiration for the sinusoidal model is to estimate the time-varying frequencies and amplitudes of the partials.

The earliest methods (pre-1930) attempted steady-state analysis; it was believed that the spectral envelope of the partials would hold the key to the musical quality (the timbre) of the tone [Risset & Mathews,1969]. Besides, the mechanical or electro-mechanical equipment of the day was incapable of analysing the time-varying spectral content. Unfortunately, tones synthesised on this basis sound dull because of the lack of any variation.

In 1969, Jean-Claude Risset and Max Mathews used computers with audio sampling capabilities, for the first thorough study of the time-varying properties of musical instrument sounds [Risset & Mathews,1969]. The work focused primarily on trumpet tones, for which a pitch-synchronous analysis yielded the period-by-period frequencies and amplitudes of the partials. Risset and Mathews achieved much in terms of laying the foundations for modern analysis-resynthesis techniques. However their work was more computer-assisted than automatic and the sinusoidal model did not advance much in the next 10-15 years. The solution eventually came from the related field of speech modelling.

Robert McAulay and Thomas Quatieri published a paper in 1986 describing a sinusoidal analysis-resynthesis method for speech coding, with the intention that it could be used for speech compression [McAulay & Quatieri,1986]. The method went a step further than the filter-bank energy decomposition methods (commonly known as phase-vocoders), in ascribing a meaning to spectral peaks. The analysis was frame-based, using the Short Time Fourier Transform (STFT – to be described in section 2.3.2.1) for spectral estimation. Within each frame, the peaks were detected and their frequency, amplitude and phase determined. The frames were then linked together by matching the peaks in one frame with those in the next. In this way, the trajectories of the sinusoidal components were tracked.

McAulay and Quatieri's Sinusoidal model (also known as the MQ method) has been the most important advance in automatic analysis-resynthesis. Many current systems are based on this spectral decomposition (e.g. Lemur [Fitz & Haken & Holloway,1995], implementation on IRCAM's sound workstation (ISPW) [Settel & Lippe,1994], Kyma [Haken,1995]).

Possibly the most important extension to the model is the Deterministic Plus Stochastic decomposition, developed by Xavier Serra [Serra,1989], also known as Spectral Modeling Synthesis [Serra & Smith,1989]. Although McAulay and Quatieri had observed the potential for their model as a tool for transformation (particularly for time-stretching and pitch-shifting), the practical realisation generates much distortion. The Sinusoidal model is indiscriminate in its spectral peak-picking, and therefore synthesises both partials and non-partials as though they were time-varying sinusoids. Serra introduced a dualistic classification that aimed to separate these two types. The former he called 'deterministic' and the latter 'stochastic'.

The deterministic aspect was analysed and synthesised very much like the Sinusoidal model, with the exception that the peak detection and frame linking processes were more strict about which peaks could be classified as partial peaks. All non-deterministic spectral elements were then assumed to be stochastic – noise-like. So these were modelled simply, as a spectral envelope and synthesised by simulating a white noise source filtered by the time-varying envelope. The final resynthesised sound was an additive combination of the two aspects.

The primary advantage of this model is its flexibility. The features have a correspondence to musical properties of sounds and they can be transformed without introducing distortion. In addition, the resynthesis quality is good. These qualities have made this an oft quoted reference for current work.

Nevertheless, the Deterministic Plus Stochastic classification does not fully represent musical sounds. Some important features are not captured accurately and there is an over-reliance on user-knowledge to operate the system. The work presented in this thesis has used the Deterministic Plus Stochastic model as its starting point, with the aim of improving computer modelling of sound features and their means of analysis, whilst retaining the flexibility that is essential for use as a musical tool. The final part of this chapter, section 2.3, describes the Initial Model for the studies, and therefore contains descriptions of the innovations introduced by McAulay & Quatieri and Serra.

2.2.2.2 The Formant Model

An alternative, but closely related modelling technique was also derived from speech modelling and applied to sound modelling in general. The vocal and nasal tracts can be considered as resonators that amplify certain regions of the spectrum. The resonances produced are termed 'formants', because they form peaks in the spectral envelope; these usually span several partials. The shape of the mouth, placement of the tongue, etc. move these formants in a predictable way, so that the sound of a particular phoneme from a particular person will have formants in roughly the same location each time it is uttered.

The sound is generated by the expulsion of air, as in whispered speech or unvoiced phonemes (e.g. 's', 'f', 'h' – the sibilants, fricatives and aspirants), or periodic vibration of the vocal folds (also known as the vocal cords) as in the singing voice or voiced phonemes (e.g. 'a', 'e', 'l', 'n' – the vowels and semi-vowels). The spectrum of the noise or tone is filtered by the formants to generate the sound that is heard.

This essentially physical model has been used for speech synthesis since the 1970's, where it is known as linear predictive coding [Makhoul,1975]. The spectral envelope is estimated by autoregression (AR) modelling and the pitch period is estimated for voiced sounds. At synthesis, the envelope is implemented as a filter and it is excited by white noise (to simulate unvoiced phonemes) or a pulsed source whose repetition rate is the desired pitch (to simulate voiced phonemes).

An implementation of the voiced aspect of this model was realised for the singing voice by Xavier Rodet, Yves Potard and Jean-Baptiste Barrière at IRCAM and entitled CHANT [Rodet & Potard & Barrière,1984]. This includes a modified synthesis method, termed FOF (Forme d'Onde Formantique) which translates as Formant Wave Functions. In FOF synthesis, each pitch period impulse is individually filtered by the spectral envelope at that period. The responses are time-aligned and summed to generate the full sound. Furthermore, the filter for each formant is implemented separately and phase-aligned to avoid interference (that can be generated by other methods). Note that although only a pulsed source is used, when the interval between pulses is irregular, aperiodic sounds are generated; therefore the method is not restricted to pitched harmonic sounds.

The method is considered here as an analysis-resynthesis technique because it was proposed by the authors as a means for synthesis of sound in general. There is a parallel between the formant resonances of the vocal tracts and the soundboards of many instruments (e.g. piano, guitar, violin), whose function is to amplify and radiate the sound, but which inevitably impose their own resonance. CHANT and FOF synthesis remain popular and a source of inspiration to new techniques (e.g. synthesis from spectral surfaces [Bargar & Holloway et al.,1995], a granular technique [Clarke,1996]).

2.2.2.3 Other Techniques

A number of alternative techniques have also been tried for their ability to extract features from the sound waveform. Some recent examples include novel approaches like neural networks [Ohya,1995] and nonlinear chaos functions [Mackenzie,1995]. At present these techniques are at a stage comparable to the Sinusoidal model, in which direct resynthesis is good, but transformation is difficult. This suggests that the features are well-captured, but that they do not have good correspondence with musical properties.

Parameter-Fitting Methods

A further category of analysis-resynthesis technique uses an existing synthesis method, for which it estimates its parameters from a source sound. This differs from techniques such as the Sinusoidal model, since the model is fixed and the data extraction method is adapted, instead of adapting the model (and synthesis method) to better represent the data. As a result, the synthesis parameters do not naturally reflect musical properties and additional interfaces are required, if the translation is to be made.

Frequency Modulation (FM) synthesis is perhaps the most popular commercial synthesis method. It was developed in 1973 by John Chowning [Chowning,1973] as a means for generating rich spectra without the computational expense of additive synthesis. The biggest drawback with FM is the difficulty of programming: the parameters of modulation index and carrier frequency bear no direct relevance to the timbre (see Figure 3.2a on page 56 and surrounding discussion). Therefore it has been difficult to find parameters that will yield more than a mimicry of real instrument sounds. Andrew Horner has recently applied genetic algorithms to the problem, to find the best match between FM parameters and a source harmonic sound [Horner,1996].

Another method in the same vein, also developed by Horner's team uses genetic algorithms to find a minimal set of sample wavetables that can describe an evolving sound, through variations in their absolute and relative weights. Recent results show good correspondence for sounds represented by just three wavetables [Chan & Yuen & Horner,1996]. The only drawback with these methods at present (for reproduction) is the limitation to purely harmonic sounds. They give a subjective quality that is similar to the synthesis of the deterministic aspect only, from the Deterministic Plus Stochastic model; i.e. lacking in roughness.

2.2.3 Granular Synthesis

Like many of the analysis-resynthesis techniques described above, granular synthesis is a form of additive synthesis, in which the sound elements, called 'sonic grains', are superposed to generate the composite sound. However unlike the analysis-resynthesis techniques, which are based on a continuous representation of sound using long-duration evolving trajectories, each sonic grain of granular synthesis is of fixed spectrum and short duration (typically 1-50ms). (An introduction to granular synthesis can be found in [Roads,1988]. More recent references can be found in the Computer Music Journal and the proceedings of the annual International Computer Music Conference.)

The origin of granular synthesis is the concept of 'acoustical quanta', put forward by Dennis Gabor in 1947 [Gabor,1947] as the basis of a theory of hearing. In granular synthesis, the detail of the spectrum is built up by the co-synthesis of multiple grains, and the evolution of the spectrum is defined by the waveform content of the sequenced grains. Each grain is a wavelet, which consists of a waveform with a tapered envelope. The simplest waveform is the sinewave, although complex waveforms (such as generated using FM or waveshaping synthesis [Dodge & Jerse,1985] or taken directly from a sampled sound [Truax,1994]) can also be used.

Depending on the type of waveform used and the method of control of the grains, granular synthesis can be used for analysis-resynthesis, traditional generative synthesis (where sound is generated from a specific architecture, like FM) or sound transformation (directly from the samples). Therefore 'Granular Synthesis' has come to be a catch-all phrase for sounds synthesised using additive synthesis of wavelets. For example, FOF synthesis (section 2.2.2.2 above) is sometimes referred to as a granular synthesis technique.

2.2.3.1 Wavelet Analysis-Resynthesis

The grain defined as a sinewave with a quasi-Gaussian envelope is termed a 'Gabor wavelet'. It can be considered the elemental grain, since it defines the smallest location on the time-frequency map and other grains can be considered as the superposition of multiple Gabor wavelets. Because of its simplicity, it requires many grains to build up a rich sound, hence the development of the more complex grains mentioned above. However its simplicity also means it can be used to represent an arbitrary waveform, the purpose of analysis-resynthesis.

Using the Gabor wavelet as an 'analysing wavelet', the evolving spectrum of a sound can be investigated. This is in essence what is done by the FFT of a windowed sample (to be explained in section 2.3.2.1) if the grain size remains constant as the frequency is changed, or by the Wavelet Transform (WT) if the grain size is inversely proportional to frequency. Analysis using the wavelet transform and synthesis by granular synthesis enables perfect reproduction of source sounds and has been used for sound transformation [Kronland-Martinet,1988]. Indeed, wavelet analysis-resynthesis is a rival approach to the sound decomposition and modelling approaches of this thesis. (The primary reasons for using the work of section 2.2.2.1 as the basis for this thesis work and not the wavelet transform are the resolution issues of the wavelet transform {discussed in section 3.2.2}, and the fact that analysed wavelets have no direct correlation to musical properties {see the beginning of section 2.1, above}.)

Matching Pursuit

A variant on wavelet-based analysis is the Matching Pursuit method [Mallat & Zhang,1993]. Instead of restricting the grain duration to a fixed law (i.e. FFT) or a constant-Q law (i.e. WT), the grain duration is free to vary such that a best-fit match may be obtained. A dictionary of wavelets specifies the range of possible wavelets. So far only sinusoidal waveforms with Gaussian envelopes have been explored. In the time-frequency space, the method matches the grain from the dictionary with the signal that minimises the residual energy (defined as the energy of the signal after subtracting the grain). The method repeats iteratively until the residual energy is below an acceptable threshold. The advantage of this method is that short grains can be used where there are rapid changes and long grains where there is quasi-stationarity, regardless of frequency.

A recent development, termed High Resolution Matching Pursuit, trades in some of the frequency resolution for an improved response at note onset boundaries [Gribonval & Depalle et al.,1996]. The authors observed that the Matching Pursuit method could introduce energy to the time-frequency space where there was none originally, if there is an overall reduction in residual energy. This is most prominent upon synthesis at note onsets where a pre-echo effect is generated. The refinement disallows energy creation, thereby forcing an alternative grain choice that does not minimise the energy as much, but does not generate energy either.

2.2.3.2 Sampling Granular Synthesis

Sampling granular synthesis, in which the grain waveform is a segment from a sampled sound, is not a modelling technique, but it has been used to good effect for sound reproduction and transformation [Truax,1994] – the aims of sound modelling. In this method, as the synthesis advances in time, so its grain waveform advances through the source waveform also. The most popular uses for granular synthesis within electro-acoustic music have been time-scaling, evaporation and coalescence. In time-scaling, the rate of advancement through the source differs from that of the synthesis; in evaporation, the grain density reduces so that the sound appears to disintegrate; coalescence is the opposite of evaporation.

Since the sample contains all information about the source sound at a particular time, it is possible to synthesise one grain at a time. However, under transformation problems arise with phase misalignment of periodic wave sections. The solution has been found by synthesising many versions of the same grain distributed randomly around each time location. Effectively the individual misalignments are blurred by the artificial reverberation. For rich sounds (especially those of many sources) or noisy sounds, the reverberant effects are masked by the volume of detail or the stochastic nature of the sound. For simple, pitched sounds, including voiced speech, the results are less faithful, but they have been used to good effect in a musical context [Truax,1990].

2.3 The Initial Model

Throughout this thesis, references are made to the 'Initial Model', described in this section. Modifications are made in the following three chapters with respect to the Initial Model and their impact is assessed in comparison with the performance of the Initial Model. The basis of the Initial Model is the Deterministic Plus Stochastic model, which is itself a variant on the Sinusoidal model (see section 2.2.2.1 above).

The following is a detailed description of the analysis-resynthesis system used to implement the Initial Model. The first step is a system overview that presents the overall signal flow and establishes the purpose of each element within the system. The description then descends to the elemental level in which every 'block' in the system is examined in some detail. The descriptions include not only *what* is done (the function of the block) but also *why* it is done (the purpose) and why it is done in this particular way. Finally we return to the model as a whole, for a more informed discussion of the system capabilities.

The Initial Model is *based on*, but is not exactly the same as, the Deterministic Plus Stochastic model. Each section contains a description of Serra's implementation and a critique highlighting its strengths and weaknesses. Where the weaknesses are localised and the solution is simple, changes are incorporated into the Initial Model. These are described with justifications. (The Initial Model also simplifies some aspects where major effort is to be concentrated in later chapters, so that the comparisons are easier to make. This is particularly the case where the Deterministic Plus Stochastic analysis-resynthesis system uses a complex set of rules whose effectiveness is difficult to assess or is signal dependent.)

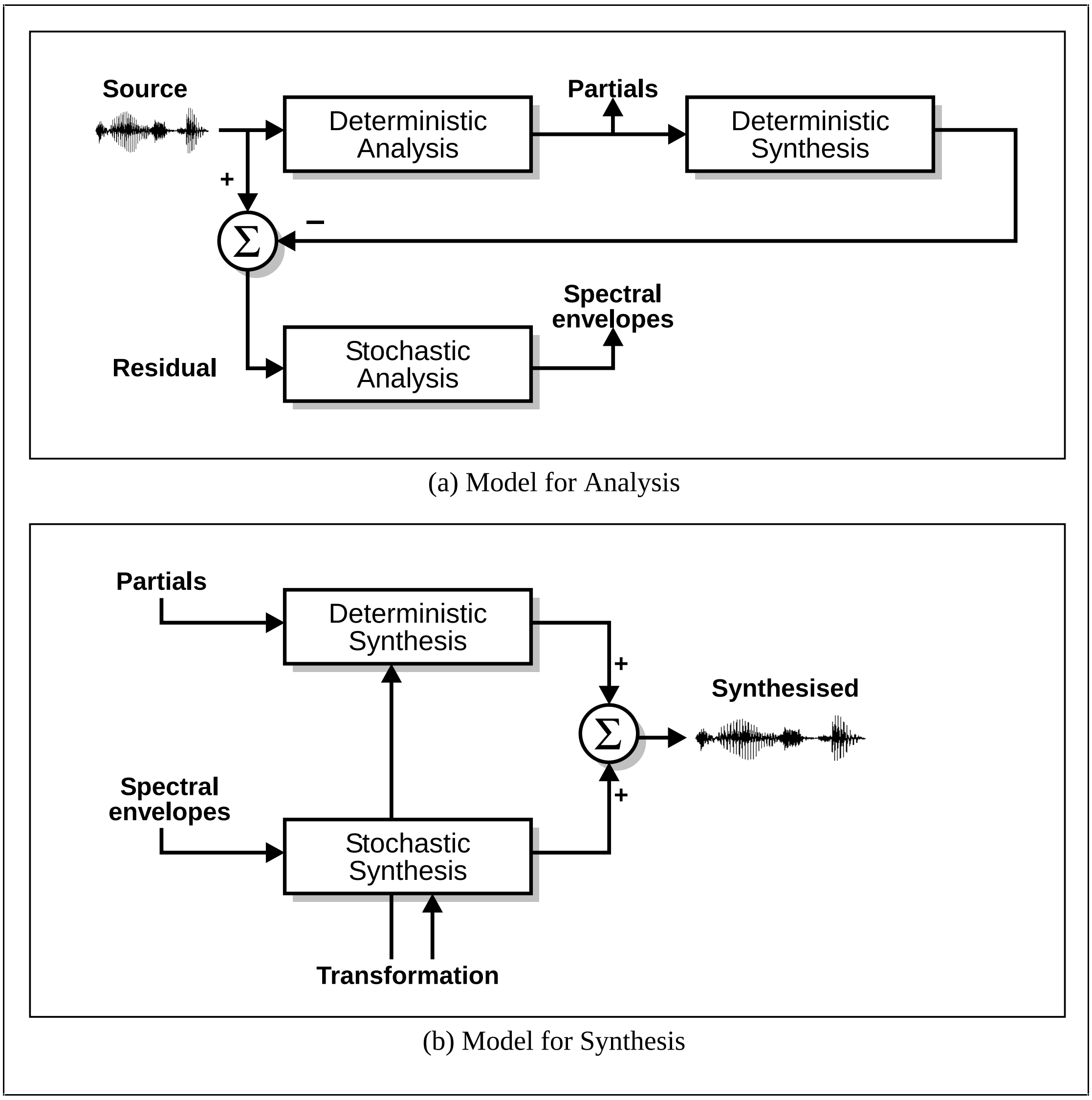
The critiques serve an important role, in identifying current areas of weakness in sound modelling. Each of the three following chapters targets one of the problem areas, in which the problem is discussed, a solution identified and an algorithm designed, to prove the solution through implementation. (The innovations of chapters three to five are similarly critiqued in chapter six, so that areas of future work can also be proposed.)

The Initial Model description has been written with two aims in mind:

- to acquaint the unfamiliar reader with the model and the terminology that pervades it;
- to describe the modifications made in translation from Serra's Deterministic Plus Stochastic Model to the Initial Model of this thesis, since the Initial Model will be used as a benchmark for later chapters.

2.3.1 Model Overview

Figure 2.3 presents a familiar display of the overall model. The model makes the assumption that sounds are composed of a deterministic aspect plus a stochastic aspect, which combine additively. The deterministic aspect includes instantaneous sinusoids, whilst the stochastic includes band-limited, shaped additive white Gaussian noise (AWGN). The source sound is first presented to the deterministic analysis, which aims to extract the trajectories of all sinusoidal components, whilst rejecting wideband noise and localised narrowband noise. This



(a) Model for Analysis

(b) Model for Synthesis

Figure 2.3 – Overview of the complete analysis-resynthesis system

aspect is resynthesised without transformation and subtracted from the original sound. The residual is assumed to contain only stochastic data. The stochastic analysis models the residual with a time-varying spectral envelope.

Synthesis of the deterministic aspect is through additive synthesis (i.e. superposition) of multiple sinusoidal oscillators, whose parameters are the analysed trajectories. The stochastic aspect is synthesised by filtering AWGN with the dynamic spectral envelope. Finally, both aspects are combined additively – 'mixed' – to render the synthesised sound.

If any transformations are desired, the appropriate modifications are made prior to synthesis. In some cases, they can be implemented within the synthesis engine, through additional control parameters.

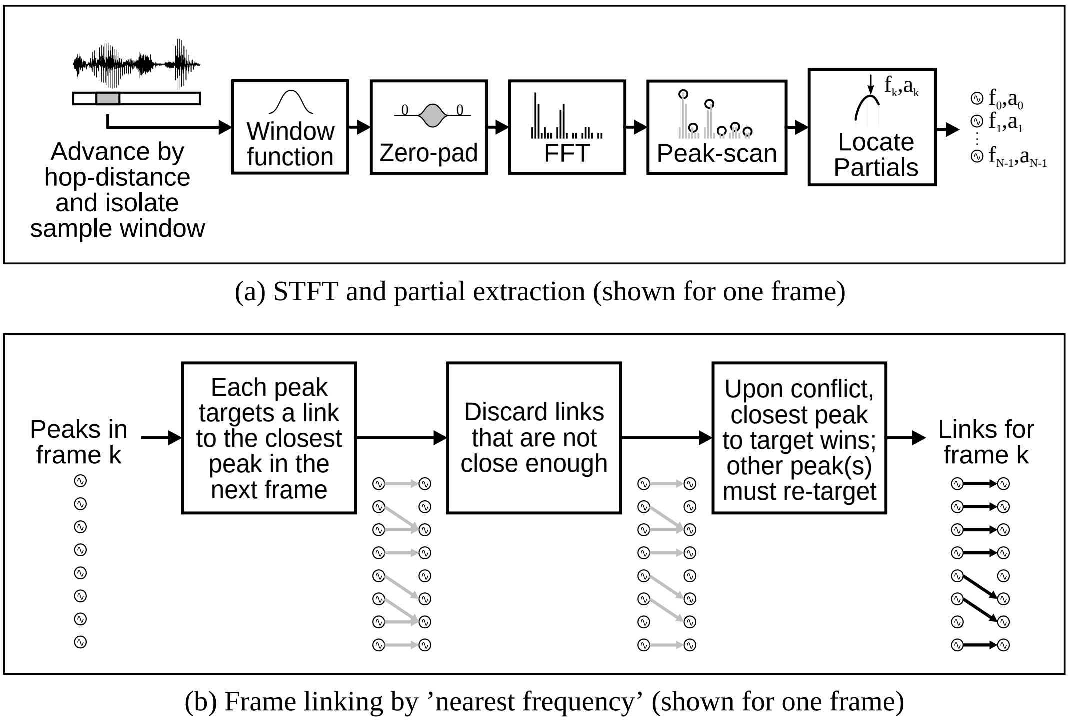
2.3.2 Deterministic Analysis

The aim of the deterministic analysis is to detect and extract the components of the source signal that are largely predictable. From the preceding discussion of section 2.1, these are seen to be the trajectories of the partials. Over the short term, each partial is approximately sinusoidal, so its instantaneous spectrum is a narrow spike located at the frequency of the sinusoid and scaled to its amplitude. Displayed on an idealised joint time-frequency graph, each one appears as a narrow ridge that runs in the time direction with variations in frequency and amplitude. The aim of the deterministic analysis is to extract the frequency and amplitude information for each ridge and compile them into a trajectory database.

The analysis process described below is organised on a frame-by-frame basis, in which each time frame is analysed independently, and then the results are compiled and linked to generate the continuous trajectories. Within a single (near-instantaneous) time frame, the partials are accompanied by less stable spectral components, which can also appear as narrow peaks in the spectrum. These can result from noise, momentary localised energy or representation artifacts (e.g. side-lobes).

In the Sinusoidal model, *all* components are extracted and synthesised. In this situation, the unstable and spurious components interact with one another additively. Often two or more spectral components will beat, causing modulation in the synthesised signal. For an untransformed resynthesis, the results can actually be better than the deterministic synthesis of partials alone, because the interactions recreate the subtler variations in frequency, amplitude and bandwidth of components. However upon transformation, where the phase relationship is disturbed (e.g. time-stretch), the beating effects become destructive and the synthesis can include significant audible distortion. Hence the Sinusoidal model does not possess the desired flexibility in its representation.

The deterministic analysis aims to exclude the non-partial components, sacrificing some extra sonic detail for a consistent quality of synthesis. This approach is more useful in a musical


Advance by hop-distance and isolate sample window
Window function
0
0
Zero-pad
FFT
Peak-scan
f_k,a_k
Locate Partials
f_0,a_0
f_1,a_1
f_{N-1},a_{N-1}

(a) STFT and partial extraction (shown for one frame)

Peaks in frame k
Each peak targets a link to the closest peak in the next frame
Discard links that are not close enough
Upon conflict, closest peak to target wins; other peak(s) must re-target
Links for frame k

(b) Frame linking by 'nearest frequency' (shown for one frame)

Figure 2.4 – Deterministic analysis

model of sound because the partials are not merely audio components, but musical features.

The process and its results are summarised graphically in Figure 2.4.

2.3.2.1 Time-Frequency Analysis – the STFT

The Short-Time Fourier Transform (STFT) is a spectral estimator that generates a time-frequency representation of a signal on a frame-by-frame basis. Each frame is calculated using the Fast Fourier Transform (FFT). See an example sonogram generated using the STFT in Figure 2.5.

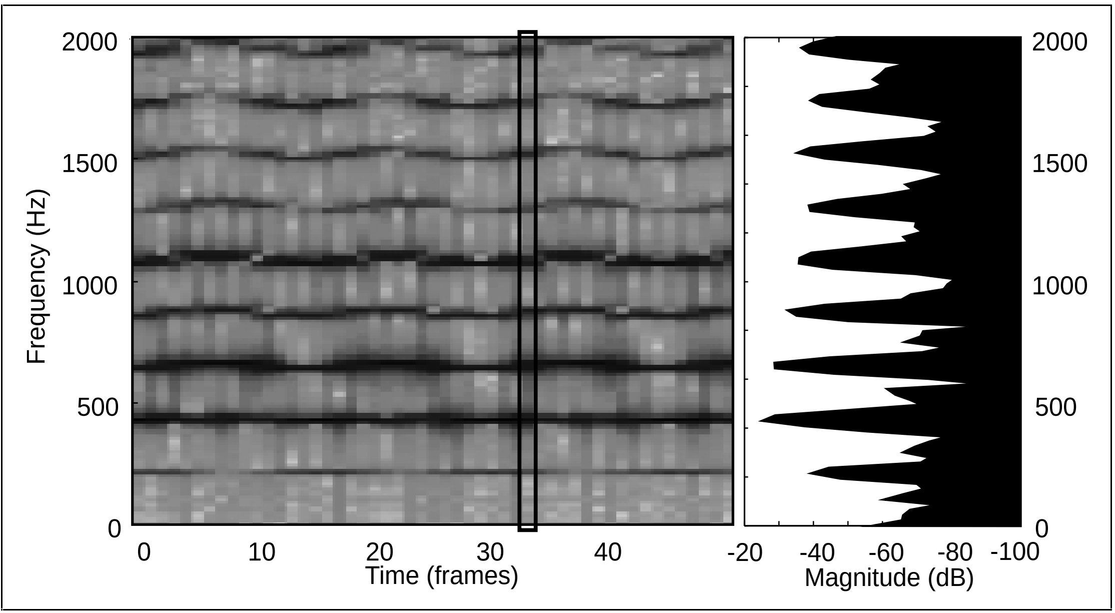


Figure 2.5 – Sonogram of a cello note with one FFT highlighted

Construction of the FFT

(Readers versed in the theory of the FFT can skip these paragraphs.)

The FFT is the algorithmic form of the Discrete Fourier Transform (DFT); the DFT is created from the FT as follows (see Figure 2.4). Firstly, for analysis on a digital system, the signal must be sampled. Sampling at regular intervals is mathematically equivalent to multiplying the time domain signal by a comb function – a series of equally spaced delta functions. In the frequency domain, this equates to repeating the signal spectrum every F_s, where F_s is the sampling rate. To avoid aliasing, the Nyquist theorem requires that the signal be low-pass filtered to exclude all frequencies above $F_s/2$.

The FT expects the signal to exist for all time, but a finite section of the signal is required for analysis. This is mathematically explained as multiplying the time domain signal by a rectangular 'window function', that isolates a small portion of the signal. In the frequency domain, the signal spectrum is convolved with the window function spectrum. For a truly periodic signal, the delta function 'spikes' in the frequency domain at the harmonic frequencies are each replaced with a copy of the window function spectrum, scaled according to the harmonic's amplitude. (The sinc pulse shape of the rectangular window's spectrum is composed of a main-lobe with a finite width and multiple side-lobes.)

Implicit in the process of windowing is the assumption that the signal outside the window is an infinite series of repetitions of the signal within the window (so as to satisfy the FT assumption of a stationary spectrum – each component existing unchanged for all time). Just as sampling the time domain was equivalent to periodic repetition of the spectrum in the frequency domain, so this implicit periodic repetition of the time domain window is equivalent to sampling the frequency domain, where the frequency samples, the 'bins', are spaced at $1/T_w$, where T_w is the duration of the window. This is termed the Discrete Fourier Transform.

When the signal under analysis is (locally) periodic with a period that is an exact fraction of the window duration, the main-lobe (the central peak of the window function spectrum) of each harmonic lies directly on an FFT bin, and the zero crossings in-between the side-lobes line up with the surrounding bins. Therefore all the energy of each harmonic is localised at its associated bin. When a component lies at some other frequency, different parts of the window function spectral curve appear on different bins, giving the impression that the component is spread over a wide frequency range (or that there are multiple harmonic components). This is termed spectral leakage. (An alternative explanation is that there is a phase discontinuity at the window boundaries in the time domain and this is 'explained' in the frequency domain with additional components.)

For readers who require a more thorough introduction to the FFT, almost any text book introducing the subject area of signal processing and the frequency domain will suffice. Nevertheless I have included some references: basic theory [Meade & Dillon,1986], in depth [Burrus & Parks,1985].

Window Function Selection

To minimise spectral leakage, an alternative window function can be chosen whose main-lobe width and side-lobe suppression are most appropriate for the application. Serra used the Kaiser window which has a parameter for trading off these two properties, but this implies that the parameter must be selected for each sound that is to be analysed. For simplicity, in the Initial Model (and in fact throughout this dissertation) the Hamming window has been used, which is somewhat a standard for signal processing applications. A comprehensive comparison of the most popular windows is presented by Harris in [Harris,1978] (and Nuttall provides a correction to Harris's paper in [Nuttall,1981]).

Time-Frequency Resolution

As a consequence of applying the time domain window function, the FT is able to represent a near-instantaneous spectrum. However, as the time resolution improves from infinite duration to the desired window length, so the frequency resolution falls from the continuous spectrum of the FT to the finite discrete frequency resolution of the windowed DFT. (This results from the frequency sampling effect described above.)

There is a trade-off between temporal resolution and spectral resolution. The shorter the time domain analysis window, the better the time resolution, but the worse the frequency resolution and vice versa. In the current application the greatest importance lies in distinguishing partials, so frequency resolution is favoured. This inevitably leads to some distortion as the assumption of true localised periodicity is violated and the approximation to a stationary spectrum becomes poorer within each window. For the less dynamic portions of a sound signal (and for the lower partials), this is nevertheless adequate.

Gaining Resolution by Zero Padding

The resolution chosen by this criterion is still far inferior to the frequency resolution of the ear, so zero padding is applied. By padding the time frame with zeros, the length of the FFT array is larger. Hence the FFT frequency resolution is improved. However, no extra data has been analysed, so the *true* resolution cannot be better; this process gives better *apparent* resolution through frequency interpolation of the unpadded signal's spectrum. This is sufficient to elicit the curve of the main-lobes to facilitate accurate frequency location of partials.

Parameter Choices for the STFT

Excluding the Kaiser window function parameter, there are three parameters that influence the representation of data by the STFT: window-length, hop-distance and zeropad-length. The necessity for a suitably long window-length has already been discussed.

The hop-distance is the time offset of the window between frames. The temporal resolution is defined by the window-length, so the choice of a small hop-distance would not provide greater resolution of dynamics – the dynamics are effectively smoothed by the averaging process inherent in analysing a finite length window. However it is important that the hop-distance be small enough that the peak linking process (section 2.3.2.5 below) can detect continuity of the time-frequency ridges. A commonly used value is a quarter of the window-length.

The frequency resolution desired from zero padding is discussed further in section 2.3.2.3 below), where it is used in combination with a less computationally intensive method of spectral interpolation.

2.3.2.2 Critique and Modification – Time-Frequency Representation

The Short-Time Fourier Transform is built upon the standard for spectral analysis – the FFT algorithm – and it has of itself become a standard for representing the joint time-frequency domain. Therefore its use within both deterministic and stochastic analyses is easily justified. However it is not the only means for time-frequency representation; nor is it necessarily the best method for sound analysis. In the transition from the idealised Fourier transform to its application as the STFT for real, nonstationary signals, the modifications have generated various side-effects. It is limited in its time-frequency resolution, such that rapid changes in a sound signal cause large scale distortion of the representation, and polyphonic signals that are close in pitch simply cannot be analysed because there is insufficient resolution to detect the presence of such closely spaced frequency components. Also, the window function has a finite spread in time and frequency, so that the outcome is a smoothed representation in both time and frequency axes.

Chapter three discusses the possibility of moving to an alternative time-frequency representation and considers what factors need to be examined in making such a choice. Key to this discussion is the need to be well-informed of the assumptions that are being made and how the signal is being conditioned. The discussion continues with a presentation of the techniques currently available and presents a novel method for eliciting 'hidden' information from the FFT, by rejecting the assumption of stationarity and applying foreknowledge about the type of signals being analysed.

Since time-frequency representation is being examined in some detail in chapter three, the Initial Model retains the STFT as the benchmark method. This is in part because the following sections of Serra's Deterministic Plus Stochastic model assume the use of the STFT.

2.3.2.3 Peak Detection and Accurate Location

Having determined that instantaneous sinusoids are represented as peaks in the FFT spectrum, a simple peak detection algorithm is employed which notes a 'hit' whenever the following situation is true:

$$\left|X_r(k)\right| \geq \left|X_r(k-1)\right| \quad \text{and} \quad \left|X_r(k)\right| \geq \left|X_r(k+1)\right| \qquad (2.1)$$

where $X_r(k)$ is the k^{th} bin of the FFT spectrum for frame r.

As mentioned above, the device of zero padding can be used to gain extra spectral resolution. If the signal were padded sufficiently to enable accurate location of a peak's maximum with the peak detection equation (2.1), then the resulting FFT array would be in excess of one million points. Such an FFT calculation is computationally prohibitive (even for a *workable* non real-time system).

Serra's solution was to use an alternative form of spectral interpolation. He found that quadratic interpolation – fitting a parabola to the maximal bin and its neighbours (using a dB scale for amplitude) – gave a good approximation to the actual curve. For good resolution at a reasonable computational cost, the methods were combined: zero padding with a factor of between 4 and 8 followed by quadratic interpolation.

Peak Validation

The above discussions have recognised the presence of spurious peaks due to noise and spectral leakage, so it is necessary to have a decision process that can distinguish 'good' peaks from 'bad' peaks. Within this model there are two stages at which bad peaks can be weeded out: at the peak detection stage, within the reference of an FFT frame (using frequency information alone), and at the frame linking stage, referencing between frames (i.e. using time information also). If some of the task is performed at the peak detection stage, there is reduced ambiguity for the frame linking stage.

Within the peak detection stage, Serra employs a simple threshold to remove peaks that are considered insignificant. The threshold is based on the *relative* amplitude of a peak which is calculated as:

$$h\left(k_p\right) = \left|X\left(k_p\right)\right|_{dB} - \frac{\left|X\left(k_{v+}\right)\right|_{dB} + \left|X\left(k_{v-}\right)\right|_{dB}}{2} \qquad (2.2)$$

where $h()$ is the measure of relative height of a peak,
$X()$ is the FFT array,
k is bin index,
subscript p represent the peak (local maximum),

subscripts *v*+, *v*- represent the adjacent valleys (local minima) either side of the peak, $| \, |_{dB}$ means magnitude measured in dB.

Two thresholds are applied to the relative peak height: a global threshold and a local one. The global threshold is a fixed value, while the local threshold varies according to the data in each frame. The local threshold is defined as a fixed offset below the maximal peak in that frame. Any peak whose relative height exceeds either or both of the thresholds is validated. Normally the global threshold is lower and is effectively the active threshold. However, if the overall frame energy is low, the local threshold drops below the global threshold and ensures that the quiet partials are validated. Serra justifies this: "Thus, in a quiet passage softer peaks are detected, mimicking the auditory system."

In addition, a pre-emphasis function is applied to the source sound, which could be implemented as a time domain filter on the source sample, or as a window function in the frequency domain. The function is based on a perceptual 'equal loudness curve', so as to boost the amplitudes of partials for peak detection, within the most sensitive auditory range.

2.3.2.4 Critique and Modification – Peak Detection and In-Frame Validation

The peak detection method is simple, resulting from the fact that sinusoids are simply represented in the FFT spectrum as peaks. One correction is made to the detection inequalities for the Initial Model:

$$|X_r(k)| > |X_r(k-1)| \quad \text{and} \quad |X_r(k)| \geq |X_r(k+1)| \qquad (2.3)$$

where $X_r(k)$ is the kth bin of the FFT spectrum for frame r.

The only change from equation 2.1 is the '>' which replaces a '≥'. This avoids a 'double hit' in the (rare) occurrence of two maximal bins having the same amplitude.

Peak Validation

One of the most challenging problems in the analysis process is identifying which spectral peaks are 'valid'. At the simplest level, this amounts to selecting peaks that arise from 'main-lobes' and rejecting those from 'side-lobes'. Within the Deterministic model, it is also necessary to reject peaks which result from localised noise, either being a maximum in a wideband noise spectrum or a momentary, random energy localisation. These latter peaks are valid spectral components, but they do not satisfy the criteria for being deterministic, so it is desirable to class them as 'bad' peaks.

Serra intentionally and justifiably defers the more complex distinction between partial peaks and noise-related peaks to the peak linking stage, where peak validation has a reference in time as well as frequency. His strategy for peak selection within each frame is minimal: a simple amplitude threshold aimed at retaining the more prominent main-lobe peaks and rejecting the less prominent side-lobe peaks.

The reasoning given for not using an absolute amplitude threshold is that "not all peaks of the same height are equally relevant perceptually". This would be a reasonable argument for using the pre-emphasis function based on sensitivity of the human ear, but not for rejecting an absolute threshold. Indeed the perceptual phenomenon of masking would appear to favour a fixed absolute threshold (for a simple implementation), or an adaptive threshold that acts locally in time and frequency about the largest peaks (for greater accuracy) [Zwicker & Zwicker,1991].

The choice of a relative measure of peak height is also perplexing, because this reduces the importance of partials that are within additive (or other) noise. However, no justification was provided for the chosen measure of relative amplitude.

The flaw in this choice is further emphasised in the employment of the 'equal loudness' pre-emphasis function[4]. If peaks were selected on an absolute amplitude basis, this function would give preference to frequencies around 4kHz, to which the human ear is most sensitive. However the peak height is calculated as a measure of *relative* height and so the pre-emphasis function has little or no impact.

The application of a peak height threshold goes some way to rejecting side-lobe peaks, because side-lobes are much less prominent than main-lobe peaks, but any strategy that is based on a threshold alone is bound to reject quiet, valid peaks also. For a direct reproduction system, it could be argued that valid peaks which fall below the threshold are perceptually insignificant, but for a system whose aim is transformation, it is important to retain as accurate a representation as possible. For example, a transformation might boost initially quiet partials; if these have already been rejected, there is nothing to boost.

For the Initial Model, the relative amplitude threshold is replaced with an absolute one and the pre-emphasis function is removed. Also the threshold is very low (-80dB for 16-bit normalised samples). This removes only the magnitude ripples along the noise floor, which result from the quietest side-lobes and quantisation errors, and that could otherwise be detected as peaks. The intention is to enable detection of every spectral component above the noise floor. This will inevitably also accept some side-lobe peaks and noise-related peaks, but the aim is to filter these out with the subsequent peak validation strategies. One such strategy appears as a by-product of the investigation in chapter three, and another as part of the process developed in chapter four.

2.3.2.5 Frame Linking to Generate Partial Trajectories

The final stage of the deterministic analysis is the generation of partial trajectories, achieved by linking the detected peaks between frames (see Figure 2.4b). This linking process is significant conceptually, because it recognises the importance of the continuation of spectral elements. In this way, it is a departure from a general audio model, to a more specific *musical sound* model – audio features are not only detected, but *a priori* expectations about pitched sounds are imposed. This significance is implicitly recognised in Serra's thesis: as the model matures from the 'sinusoidal' (which captures all spectral components) to the 'deterministic' (which is specific to stable partials), there is increased importance in the validation process of good and bad peaks.

[4] Note that the equation Serra provides does not match the curve presented (see [Serra,1989,pp.41-42]).

Serra's method of peak linking is based on frequency alone, the algorithm favouring a link between peaks of consecutive frames that are close in frequency. There is a parameter for the *maximum frequency deviation* between frames, which is proportional to the frequency of the peak (in the earlier of the two frames). This mimics the logarithmic frequency sensitivity of the ear and is accurate to the variations of partials within pitched sounds[5].

On a frame-by-frame basis, each peak attempts to link to the closest peak (by frequency) in the next frame. Where there is a conflict – multiple peaks attempting to link to the same peak of the next frame – the link requiring the smallest frequency change is favoured and the other peaks must reselect. This method is termed the 'nearest frequency method' throughout this dissertation.

Partial trajectories are formed by a 'birth' at a peak that has not been linked to from the previous frame; they continue following the links from that peak to the one in the next frame and so on until a peak is reached that has no link to the following frame, upon which the trajectory is 'killed'. The concept of trajectory births and deaths was first introduced by McAulay and Quatieri [McAulay & Quatieri,1986], but Serra introduced further heuristic rules and conditions in the shift from Sinusoidal model to Deterministic model.

These include limiting the number of concurrent trajectories, placing a lower threshold on the length of a trajectory – short trajectories are considered unstable and rejected – and placing a limit on the closeness of a new trajectory to the existing ones. In addition, the concept of the 'guide' was introduced, a virtual trajectory that only becomes real after all the validations. This concept facilitated the tracking of 'sleeping partials', so that partial trajectories could dip below the peak amplitude threshold for a few frames without being rejected for being too short.

2.3.2.6 Critique and Modification – Frame Linking

The described frame linking method can be viewed as a two-stage process. The first stage is the nearest frequency linking strategy which actually achieves the links; the second stage is the set of heuristic rules that constrain the formation and acceptance of partial trajectories.

The nearest frequency method of peak linking is based on the assumption that changes between FFT frames are small. This is usually only true for the lower partials of steady-state sounds. The higher partials are always more dynamic (in order to preserve the harmonic frequency ratios) and therefore harder to track. Also, rapid changes in frequency represent a significant violation of the FFT stationarity assumption, causing distortion to the peaks. This causes the frame linker to make errors, so it is not unusual upon synthesis for sounds to appear to have been low-pass filtered, as the short unstable trajectories of the higher partials are rejected.

In the transition from a Sinusoidal model to the Deterministic one, the second 'control' stage was added. Of its rules, the lower limit on trajectory lengths is probably the only one that can be intuitively justified (on the basis that partials are stable, continuous sinusoids). Unfortunately, no reasoning was provided for the other rules, which appear somewhat arbitrary and situation dependent, thereby making it difficult to assess. Serra himself acknowledged that

[5] Higher partials must change more rapidly in order to remain (roughly) harmonic.

"the peak continuation algorithm is the part of this dissertation that is most open for further work."

Chapter four examines the flaws of the nearest frequency scheme and proposes, as its replacement, a method based on the harmonic structure. The reasoning for this is justified on three counts: ability to cope with imperfections in the STFT, the trajectories of the higher partials, and the concept that partials belong to a harmonic structure. An implementation of a harmonic linking method is described, to validate the premise.

In the Initial Model, the nearest frequency linking method is retained, for the purpose of benchmark testing. However the second control stage is abandoned.

2.3.2.7 The Importance of Phase

There is still debate as to whether we are 'phase deaf'; i.e. whether the relative phase between sinusoidal components is perceptually significant. For many sounds, the phase relationship between partials appears to be inconsequential, to the extent that waveforms which, visually, are markedly different, nevertheless sound identical. Contemporary wisdom favours the concept of 'critical bands', within which the relative phase of components is important. If two sinusoids lie within a critical band, which is defined as approximately one third of an octave wide, their relative phase is significant; if not, then it isn't. There is the complicating factor that phase is important in timing and hence rhythm. The relative phase of a number of components can combine to place a time domain spike at a particular location, which would otherwise be displaced or not occur to the same degree.

Serra takes the approach that phase is not important. For analysis of single source sounds, this is justifiable, since the lower (and usually most significant) partials are widely spaced, not falling within a critical band until the fifth partial[6]. In the Deterministic model, frequency and magnitude alone are extracted from the STFT peaks, which facilitates simpler computation, but it has an impact on the calculation of the residual as discussed in section 2.3.4 below.

2.3.3 Deterministic Synthesis

2.3.3.1 Additive Synthesis with Linear Interpolation

In the deterministic synthesis, all components are dynamic sinusoids, instantaneously approximating to elements within the Fourier Series. Therefore synthesis is a process of superposition of the waveforms of each sinusoid, known as Additive Synthesis (see Figure 2.6 overleaf). The simplest implementation (conceptually) is to generate a virtual sinusoidal oscillator for each partial and to sum their outputs at each sample.

[6] The majority of energy is in the first five to eight partials for many sounds. Phase errors in the lower partials do not matter because none are within the same critical band, whilst phase errors in the higher partials are less likely to have a significant perceptual impact because of their small relative energy.

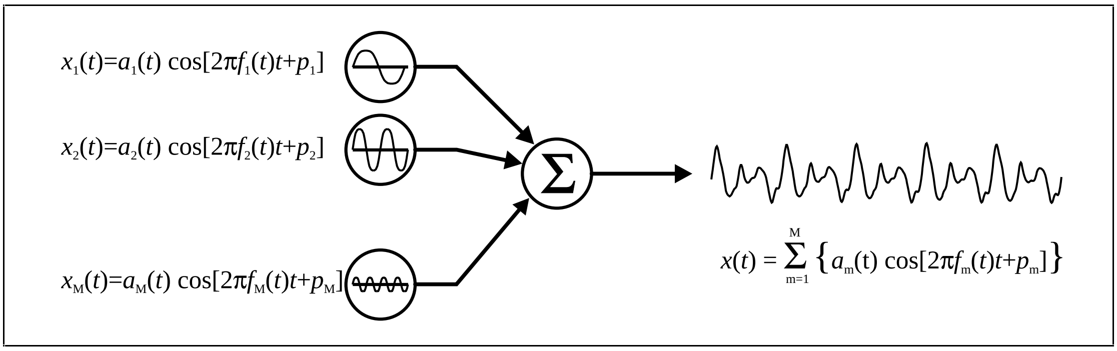

$x_1(t)=a_1(t)\cos[2\pi f_1(t)t+p_1]$
$x_2(t)=a_2(t)\cos[2\pi f_2(t)t+p_2]$
$x_M(t)=a_M(t)\cos[2\pi f_M(t)t+p_M]$
Σ
$x(t)=\sum_{m=1}^{M}\left\{a_m(\mathrm{t})\cos[2\pi f_m(t)t+p_m]\right\}$

Figure 2.6 – Additive synthesis

Since the status of each partial is known only at the frame boundaries (corresponding to the centre sample of each FFT frame in the analysis), the oscillator properties can only be updated at these points. In order to facilitate a smoothly evolving output, however, it is necessary to ensure that there are no discontinuities in the frequency, amplitude or phase trajectories. The simplest solution is to linearly interpolate the frequency and amplitude between frames.

In the Sinusoidal model, phase is important, so at each frame update point, each oscillator's phase must *arrive* at the desired value. This requires that phase compensation be introduced to predict drift errors and compensate for them, or to make the frequency trajectory more complex (e.g. cubic interpolation). The Deterministic model greatly simplifies this, since the relative phase of the oscillators is unimportant. Instead as each partial is 'born', its oscillator phase can be initialised to some arbitrary value. Thereafter only the frequency and amplitude are updated and the only constraint on phase is that it must be continuous.

Figure 2.7 (overleaf) graphically summarises the process for initialising and updating an oscillator, and below that is a pseudo-code implementation (not optimised) suitable for a digital signal processor. Each oscillator has five variables: absolute frequency, amplitude and phase and incremental frequency and amplitude. The absolute values describe the instantaneous state of the oscillator at all times, while the incremental values regulate the frequency and amplitude interpolation between frames. It is necessary to initialise all values at a partial birth (when the new oscillator element is 'created'); thereafter only the incremental values need updating at the frame boundaries. The incremental values represent the rate of change of frequency/amplitude per sample, so that the absolute frequency/amplitude linearly interpolate to the correct values of the next frame, over the duration of one frame.

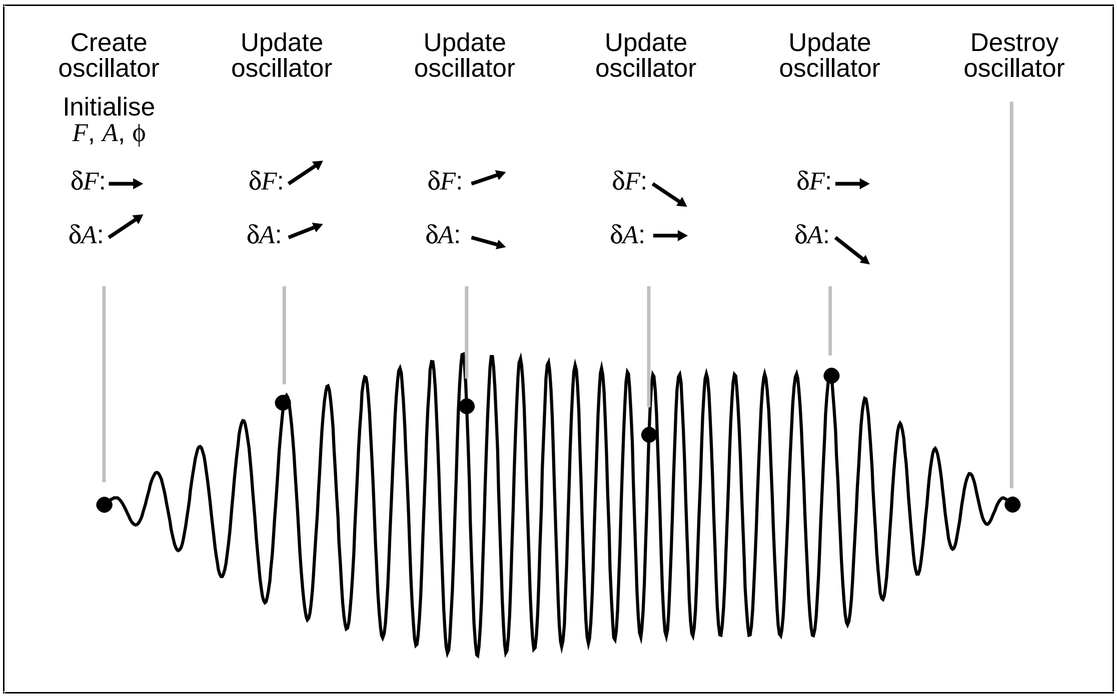


Figure 2.7 – The life of an oscillator (whose lifespan is 5 frames)

```
For each frame {
    Initialise $\phi_i$, $F_i$, $A_i$, $\delta F_{i,}$ $\delta A_i$ for each oscillator
    For each sample in frame {
        Initialise $x_{sum} = 0$
        For each oscillator {
            $x_{sum} = x_{sum} + \cos[\phi_i] * A_i$
            $\phi_i = < \phi_i + F_i >_L$
            $F_i = F_i + \delta F_i$
            $A_i = A_i + \delta A_i$
        } [Next oscillator]
        Output $x_{sum}$
    } [Next sample]
} [Next frame]
```

where ϕ_i is the instantaneous phase index
F_i is the instantaneous frequency index (the phase-increment)
A_i is the instantaneous amplitude scalar
δF_i is the frequency increment (to achieve linear interpolation)
δA_i is the amplitude increment (to achieve linear interpolation)
x_{sum} is the instantaneous sample value (the additive synthesis output)
cos is the cosine function, whose values are indexed from a wavetable
L is the length of the cosine wavetable (covering 360°)
$<>_L$ means modulo L – in this case it keeps the phase index in the range 0-360°

2.3.3.2 Critique and Modification – Additive Synthesis with Linear Interpolation

Fast Additive Synthesis

Serra's implementation of additive synthesis and the implementations within this thesis (the Initial Model and the modifications of the later chapters) use the algorithm described above. It is well recognised that for commercial systems, additive synthesis can be prohibitively expensive computationally. However such considerations are not as important for model development systems, where model integrity has a higher priority than implementation details. Nevertheless there are some schemes for reducing computational load that are worth mentioning here.

One approach has been to reduce the data size, so that partials with similar properties actually share the same data. This method, termed Group Additive Synthesis, often combines the amplitude variations of several partials, or restricts the frequency trajectories to only truly harmonic frequencies [Kleczkowski,1989]. Unfortunately, inherent in this approach is a trade-off between synthesis quality and processing workload. Recently, alternative approaches have been proposed and are currently being refined, which could be described as Fast Additive Synthesis algorithms. They are transparent (in the sense that the data itself is not compromised and the synthesis sound quality is maintained), relying on innovative processing techniques to achieve the reduction in computational load. Important contributions include:

- Xavier Rodet and Phillipe Depalle's IFFT synthesis [Rodet & Depalle,1992a; Freed & Rodet & Depalle,1993],
  where the FFT spectrum of each output frame is constructed in the frequency domain and synthesised by inverse FFT. This is a reworking of the familiar overlap-add STFT synthesis [Serra,1989,ch.2], but within the more flexible scenario of an analysed (and possibly transformed) spectrum. Further refinements that achieve amplitude and phase continuity for chirps (i.e. linear frequency interpolation) are presented in [Goodwin & Rodet,1994; Goodwin & Kogon,1995].
- Desmond Phillips, Alan Purvis and Simon Johnson's multirate optimisation [Phillips & Purvis & Johnson,1994, 1996],
  which makes use of the observation that "the computational cost of an oscillator is proportional to its update rate", to devise a synthesis filter bank of half-band filters, that allocates "near-optimal" sample rates to oscillators. That is, the number of samples per oscillator cycle is fixed (within a range) (instead of fixing the traditional 'number of samples per second' sample rate), so that lower frequency oscillators are calculated less often.

Initial Phase

Although phase has been judged not strictly important, in the Initial Model the *initial* phase value is provided for each oscillator, so that after the fade in (over one frame), the relative phases will match that of the original sound. This has enabled visual confirmation of the algorithm from the time domain waveform, during development.

2.3.4 Calculation of the Residual

Having extracted the stable sinusoidal components, the partials, from the source sound, the stochastic analysis operates on the residual. However, the partials have been detected and not removed from the source sound, so it is necessary to generate the residual signal. This is effectively the subtraction:

$$\textit{Residual} = \textit{Source} - \textit{Deterministic} \tag{2.4}$$

The source signal is available in the time domain, as a sample, whereas the deterministic is available as a set of features in the model domain. In order to subtract the two they must be in a common domain. If the deterministic aspect were synthesised to the time domain, the waveforms could be subtracted. However this necessitates that the phases of the partials be synthesised correctly, whereas the decision has been made to discard phase information[7]. So the signals are subtracted in the magnitude-only time-frequency domain.

The deterministic signal is synthesised to the time domain and converted to the time-frequency domain using the STFT (with the same parameters as the deterministic analysis). The source signal is converted to the time-frequency domain, also using the STFT with the same parameters. In this common domain their magnitude spectra are subtracted to generate the residual signal.

2.3.4.1 Critique and Modification – Calculation of the Residual

Justifiable Computation

The seemingly exorbitant amount of processing required to subtract the two signals can be justified, if placed within the context of the whole system. Firstly, the STFT's are both calculated with the same parameters as the deterministic analysis, so the STFT of the source sound has already been calculated (during the deterministic analysis). Secondly, the stochastic analysis (to follow) is performed in the time-frequency domain, so conversion to this domain would have been required anyway.

One saving could be made if the principles of IFFT Fast Additive Synthesis (see section 2.3.3.2) were applied, to simulate construction of the STFT directly from the deterministic signal. This was not, however, implemented in the Initial Model (largely because the algorithm had not been developed sufficiently to simulate linear interpolation at that time).

Residual Contains Error Signal or Stochastic Signal?

The deterministic aspect of the model aims to capture specific components of the source sound, intentionally rejecting other components. Therefore the residual signal is not merely an error signal, although it does include all the components that the deterministic aspect ignores or fails to capture.

However neither is the residual signal truly stochastic (random), in the mathematical sense. This results from Serra's *adapted* meaning of the term 'deterministic': "A deterministic signal

[7] The deterministic representation, using magnitude information only, has the same partial energy localisation as the source signal, but the absence of phase information renders time domain subtraction useless.

is traditionally defined as anything that is not noise (i.e. a perfectly predictable part, predictable from measurements over any continuous interval). However in the present discussion the class of deterministic signals considered is restricted to sums of quasi-sinusoidal components… By contrast, in the sinusoidal model, each sinusoid modeled a peak in the spectrum (not always a sinusoidal component)… In more musical terms the deterministic component models the partials of the sound, not just any energy."

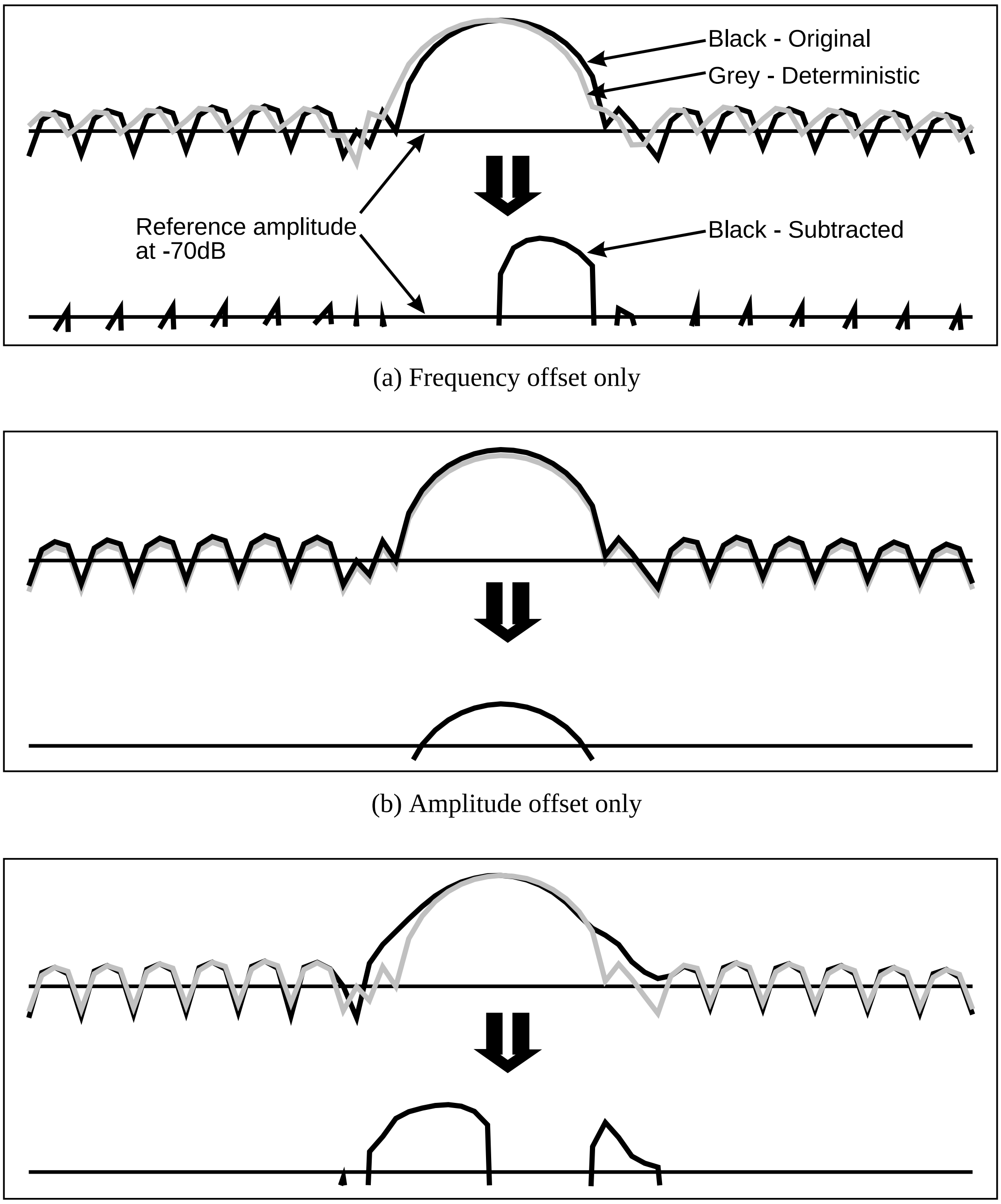
Given this adapted meaning of 'deterministic', perhaps 'stochastic' should be redefined as 'non-partial'. In fact the residual signal contains a combination of non-partial elements and an error signal. Anything that is not captured by the deterministic analysis is, by default, included in the residual. Therefore any partials that are not captured, due to failures in the peak detection, peak validation or frame linking stages, will appear in the residual. Similarly, any non-partial spectral peaks that are interpreted as a partial trajectory, or a part of one, will impact the residual. The impact could either be the deletion of truly stochastic elements, or the creation of (negative magnitude) components. This latter scenario is the worst because a specific component will appear in both deterministic and stochastic aspects of the model. (As with the Deterministic Plus Stochastic model, the Initial Model places a threshold at zero on the subtraction, so that negative magnitude components cannot be created.)

Inaccuracy in Calculation and its Impact on the Model

When the deterministic analysis correctly identifies a partial trajectory, there is still room for error in the residual, due to inaccuracies in the estimation process. A slightly erroneous *frequency* value will result in the source signal's spectral peak being slightly offset from the deterministic signal's one, so that upon subtraction they will not cancel properly; see Figure 2.8a (on the next page). A slightly erroneous *amplitude* estimate will result in an incomplete subtraction; see Figure 2.8b. Both of these scenarios are highly likely, especially where there is significant distortion to the FFT spectrum, due to nonstationarity.

Even if the deterministic analysis were to perfectly extract a partial's trajectory, its synthesis would be a piecewise linear approximation, since both amplitude and frequency components are linearly interpolated between frames. As a result, the time-frequency spectrum of the deterministic signal would not match the source signal's spectrum exactly. This type of error manifests as a difference in the width of the associated spectral peaks – the source signal's peaks are wider because its trajectories are further from quasi-stationarity and therefore generate more distortion to the FFT representation. This is depicted in Figure 2.8c. (The fact that the source peaks are significantly wider belies a level of detail beyond failing to track the maxima of ridges in the STFT, the implications of which are discussed in section 6.5.)

The impact of these errors upon the model's representation of a sound is significant, in that features are misrepresented and therefore transformations will not be effected correctly. Fortunately, upon synthesis (direct resynthesis or synthesis with small transformations), the impact of these errors is minimal. This is due to the processes of stochastic analysis and synthesis (described in the following sections) which smooth out the localised details of the residual, both in time and frequency.



(a) Frequency offset only

(b) Amplitude offset only

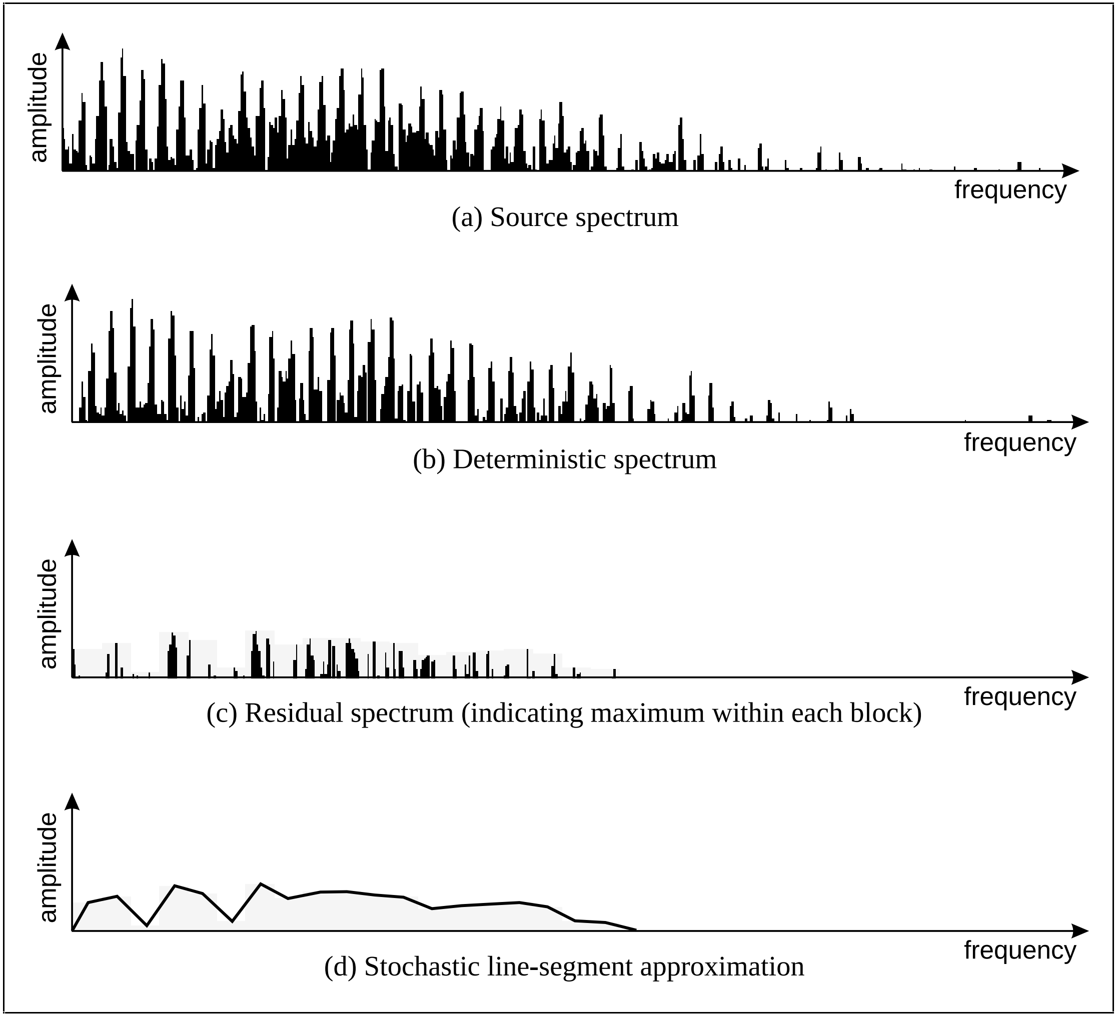
(c) Difference in peak width only

Figure 2.8 – Residual errors due to differences between the partial peaks of the source signal and deterministic signal

2.3.5 Stochastic Analysis

Serra makes the claim that just as phase is unimportant perceptually for the deterministic aspect of the model, so too it is “irrelevant to maintain the exact frequency characteristics (i.e. the exact magnitude spectrum)” for the stochastic aspect. Instead the *general* spectral shape is preserved.

For each magnitude spectrum, a line-segment approximation is made to the envelope. The FFT spectrum is split into equal-sized blocks and the maximum magnitude is found within each block. These are compiled into the envelope for that frame. The size of each block (measured in number of bins) is a model parameter, which should be smaller (allowing greater detail) for more complex sounds. See Figure 2.9.



(a) Source spectrum

(b) Deterministic spectrum

(c) Residual spectrum (indicating maximum within each block)

(d) Stochastic line-segment approximation

Figure 2.9 – Stochastic analysis

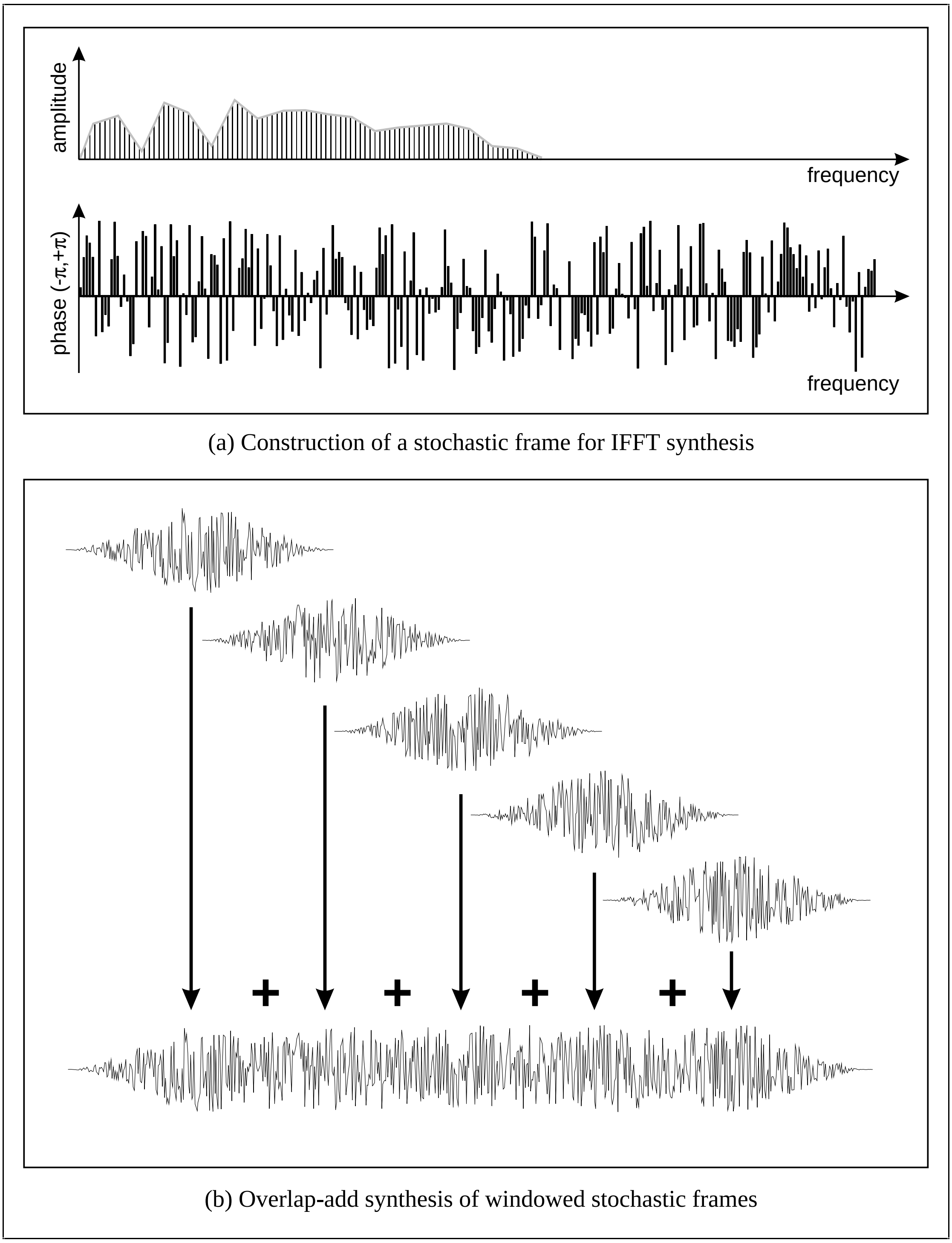
2.3.5.1 Critique and Modification – Stochastic Analysis

The stochastic aspect of the model must represent not only truly noise-like elements, but all non-partial elements and the error signal due to deficiencies in the deterministic aspect. Therefore it is inevitable that the results of stochastic analysis and synthesis will be imperfect. Within this thesis, the focus has been on improving the deterministic analysis, so that it can better detect and track partials. Consequently this should produce an improvement in the stochastic aspect. Hence the Initial Model implements the stochastic analysis as described by Serra.

2.3.6 Stochastic Synthesis

The aim of the stochastic synthesis is to generate dynamically filtered additive white Gaussian noise (AWGN). The spectrum of a rectangular window of AWGN has constant magnitude and random phase. So the spectrum of a rectangular window of *filtered* AWGN has the spectral envelope of the filter in the magnitude spectrum and a random phase spectrum. This principle is applied in a piecewise manner to construct the stochastic synthesised sound.

For each frame (see Figure 2.10a), the magnitude FFT spectrum is generated from line segments based on the associated envelope array. A random phase FFT spectrum is generated. The IFFT is computed to yield a rectangular window of filtered AWGN. Each frame's output is windowed (with a simple window function, such as the Hanning) and the outputs are overlapped and added such that a dynamic filter response emerges (see Figure 2.10b). The



(a) Construction of a stochastic frame for IFFT synthesis

(b) Overlap-add synthesis of windowed stochastic frames

Figure 2.10 – Stochastic synthesis

final stage of this is the familiar overlap-add synthesis, that was used in the first STFT-based implementations of McAulay & Quatieri's Sinusoidal model, where the hop-distance, window-length and window function are chosen such that the window envelopes add to unity.

2.3.6.1 Critique and Modification – Stochastic Synthesis

The IFFT Overlap-Add synthesis is an elegant method for recreating a smoothly evolving, dynamic noise spectrum. In practice it can suffer from pre-echo and reverberation as the instantaneous noise spectrum is spread over a finite time (future and past). Where the original sound contains rapid spectral changes, such as at a percussive onset, the audible effect of diffusion is amplified. Chapter five solves the problem for sudden attacks by forcing an abrupt rising edge on the amplitude envelope. However the effect of diffusion in other circumstances remains. This is particularly noticeable for speech, which sounds more reverberant and hence less distinct.

2.3.7 The Model as a Whole

In summary, the model features are partial trajectories and spectral noise envelopes, organised in time-frames. The model itself is an abstract entity – a concept – but it is made manifest through the analysis-resynthesis system.

A sound sample is analysed by the system to generate the feature set specific to that sound. Synthesis of the features yields a reproduction of the original sound. Yet the power of this model is its flexibility as a tool for musical transformation. The features, particularly the partials, represent elements that can be intuitively grasped in a musical sense. Indeed the model enables (with considerable ease) transformations that are difficult or impossible by direct manipulation of the samples.

For example, time scaling (commonly known as time-stretch) is a complex operation in the time domain, but in the model domain it is a simple scaling of the interval between frames. The frequencies and amplitudes of the deterministic aspect are synthesised at the new frame rate with faster or slower linear interpolation and the stochastic spectral envelopes are interpolated between frames (so that the synthesis hop-distance is preserved).

2.3.7.1 Critique and Modification – The Model as a Whole

McAulay and Quatieri laid the foundations for the sound analysis-resynthesis model, by applying the principles of the Fourier Series with the known properties of the human voice, to create a practical implementation. Serra's Deterministic Plus Stochastic model has taken this realisation into the domain of musical tools, by focusing on aspects that would yield flexibility for musical transformation. The question now is how well this model (specifically the Initial Model) satisfies the aims of this thesis.

Feature Encapsulation and Model Usability

How well does the Initial Model live up to the aims of the thesis (stated at the beginning of section 1.3)? How much of the timbre space (see Figure 2.1) does the model's feature set encapsulate? How well do the features match those musical properties? How accurately can the model reproduce a source sound (using direct resynthesis)? How usable is the model for musical transformation? How easily does the model extract its features?

In terms of the sound properties explored in the first part of this chapter, the model has a very clear definition of the elements that compose the harmonic structure, and is therefore well placed to model pitched sounds. The stochastic aspect is something of a catch-all feature set, whose features most closely match the unpitched noise-like sounds, like the sibilants and fricatives of speech. However there are limitations in both aspects, particularly in terms of the dynamics of sound. Both deterministic and stochastic aspects are good at extraction of features for slowly evolving spectra, but neither have the capability to adapt to rapid changes. In particular, a major failing of the model is its inability to target percussive attack transients, which are important features in any rhythmic music.

In terms of the match between feature parameters and musical properties, the partials describe the frequency and amplitude trajectories of the elements of the harmonic structure, so there is very good correlation. This in turn makes the deterministic very usable for musical transformation of pitched elements. The stochastic spectral envelopes do capture noisiness, but there is much less correlation between model parameters and musical properties. Inasmuch as noise sources tend to derive from chaotic turbulence that is filtered by whatever medium it passes through, the model's instantaneous magnitude spectrum is a fair approximation. However localised chaotic variations are not as well represented. This is particularly true when the chaos is not additive, but a part of a more long-term stable component. Nevertheless the representation does lend itself to musical transformation.

For direct resynthesis (i.e. reproduction), the quality is high for many musical instruments, where there is a sustained pitch within each note. A melody on an oboe for example. Instruments that have a complex onset transient or feature abrupt changes are not as well represented. Often the synthesis of these elements appears diffused or even dissolved.

The question of 'ease of extraction' is both an issue of effectiveness of the algorithms and an issue of system usability (for the musician). The major limitation to the model's ability to extract features is the resolution of the STFT. This affects the accuracy with which components (particularly deterministic) are extracted and obscures detection of closely spaced components (in frequency or time) so that polyphony and attack transients are not represented in the model.

In terms of system usability, the analysis and synthesis are automated. However there are many parameters that need to be set. In Serra's implementation, a number of these were source dependent, which required the musician to have an intimate knowledge of both sound structure and the analysis-resynthesis system.

2.3.7.2 Model Aspects Targeted For Improvement

From the above comments and the observations of all the critique sections of this chapter, the model achieves much in its ability to represent musical sounds and to provide a system for musical transformation. Nevertheless there are some weaknesses that could be targeted for improvement. At a global level, they include:

- synthesis of the attack portion of sounds that have a percussive onset;

- perceptual fusion of the deterministic and stochastic aspects of the model;
- capacity for sounds of multiple sources (including chords from one instrument);
- automated operation (including the selection of system parameters).

Of these the first three require direct modification of the model, whilst the fourth is more a product of the system implementation, although that too can be aided through modifications to the model. In the construction of the Initial Model from the Deterministic Plus Stochastic model and in the studies to follow, an emphasis has been placed on finding parameters that can be globally set, independent of the sound to be analysed so that the system becomes truly automatic. Hence, for example, the window-length is fixed so that the lowest partials can be separated and the Hamming window function replaces the Kaiser.

Time-Frequency Representation (TFR)

The synthesis of attack transients, like the ability to resolve closely spaced components in a polyphonic sound, is a problem of the time-frequency representation. The TFR directly influences how data is viewed and therefore what features are made visible. Chapter three sets its target on discovering why the STFT is limited, what alternatives exist and how more resolution might be squeezed from the representation. Key to this is the need (of the designer) to be well-informed of what happens to a signal when it is analysed and what assumptions are being made.

Frame Linking by Harmonic Structure

The process of generating partial trajectories is governed not only by the available resolution of the TFR, but also by the strategies employed for peak validation and frame linking. As the description of section 2.3.2.6 revealed, this is a major area for improvement. Chapter four proposes replacing the ‘nearest frequency’ strategy of peak linking with a method based on the harmonic structure. This stand is defended in the first part of the chapter and the rest of the chapter describes an implementation. Central to the presented algorithm is room for upgradeability. Therefore, although the system cannot deal with polyphonic sounds at present, the algorithm includes techniques that can be upgraded to polyphonic linking, when the capability is developed.

Attack Transients

The problem of attack transients is specifically targeted in chapter five. The time scale of a percussive attack is much shorter than the chosen time scale of the STFT. So even if there is an incremental improvement in the resolution of the TFR, there will still be the need to improve the model’s ability to represent attack transients. Chapter five incorporates attack transients into the model for the first time, by introducing a detection scheme and making minor modifications to the system structure, so that analysis can pick out features more clearly and synthesis can respond to the rapid changes in the spectrum.

Perceptual Fusion

The issue of perceptual fusion of the two aspects of the model (deterministic and stochastic) was not realised at the outset of the thesis work, but became apparent through listening to the synthesised sounds and certain key observations of signals within the system. There is a slight subjective mismatch between the deterministic and stochastic aspects, particularly noticeable for speech, where the tonal aspects and the noisiness appear to come from separate sources. Couple this with the observations made during calculation of the residual (see figure 2.9c, section 2.3.4.1) and the observation of musical sound properties (section 2.1.5.1) that not all noise is additive. The result is a suspicion that there is extra detail in the partials, that is missing from the model, and that that detail is probably in the form of small scale chaotic variations. Chapter six, which collates the results of chapters three to five and presents future work suggestions, examines this point in more detail (section 6.5) and presents evidence to support this theory. Finally suggestions are put forward for a proposed approach to solving the problem, as a direction for future work.

CHAPTER THREE

TIME-FREQUENCY REPRESENTATION (TFR)

3. TIME-FREQUENCY REPRESENTATION (TFR)

"The world is as you see it"

——— *Yoga Vasishtha (Ancient Indian Scripture)*
(Translated by [Venkateshananda,1984])

The Fourier Transform (FT) is the fundamental tool for spectral analysis. Indeed the terms 'Fourier transform' and 'frequency spectrum' are virtually interchangeable. (Strictly speaking the transform is not the spectrum, but a means of conversion to it, from the time domain. The spectrum is a mathematical construct and the Fourier Transform is an analytical tool.)

When applying the FT in a practical situation, various constraints must be applied that inevitably modify the representation of the signal under analysis. These are the mathematical means of forcing a real signal to appear to be ideal. Therefore what is represented is the spectrum of that ideal signal, not the spectrum of the actual signal. Only if there is a close enough correspondence between the two is it possible to infer the properties of the actual signal from the representation of the ideal.

The most prominent assumption that accompanies the FT is that the signal has a constant spectrum. In the practical case of quasi-stationarity, the real signal only approximates the ideal over short durations. Section 2.3.2.1 described the STFT and the constraints that are imposed on sound signals to achieve its representation. The most restrictive constraint is the limit to resolution in time and frequency – the STFT does not allow representation of the near-instantaneous spectrum with good frequency resolution. This is known to be inadequate for rapidly varying sections of sound; nevertheless, the STFT remains the most popular means for time-frequency representation of time-varying signals.

This chapter discusses the possibility of basing sound analysis upon an alternative time-frequency representation (TFR) and how such a choice should be made. It is shown that to make an informed choice it is necessary to understand the basis of the assumptions that have historically been attached to TFRs. The discussion begins by looking at the model designer and how his/her expectations can colour what data is retrievable from a signal. In this light, a clear definition of the model is sought, from which a TFR may be objectively chosen.

The latter sections of the chapter provide a review of time-frequency distributions and the current areas of active research, and time-frequency representations that have made innovative use of such distributions to elicit more information. The last part of the chapter looks at a particular innovation that was developed as part of the thesis, in which extra 'hidden' information is extracted from the FFT, thus improving its effective resolution.

3.1 What is a TFR and How Can One be Chosen?

3.1.1 The World Is As You See It

The foundation of the sound analysis process and hence the transformation and resynthesis processes also, is the time-frequency representation (TFR). The time-frequency representation provides the ground-truth – all subsequent processing relies on it and what it presents is taken to be true. Yet there is no single, unshakeable definition for time-frequency representation – any number of mathematical distributions can be used to describe the time-frequency domain legitimately. In fact upon closer examination, the way in which it is described within an analysis-resynthesis system depends more upon the model designer and the model definition, than the mathematics itself.

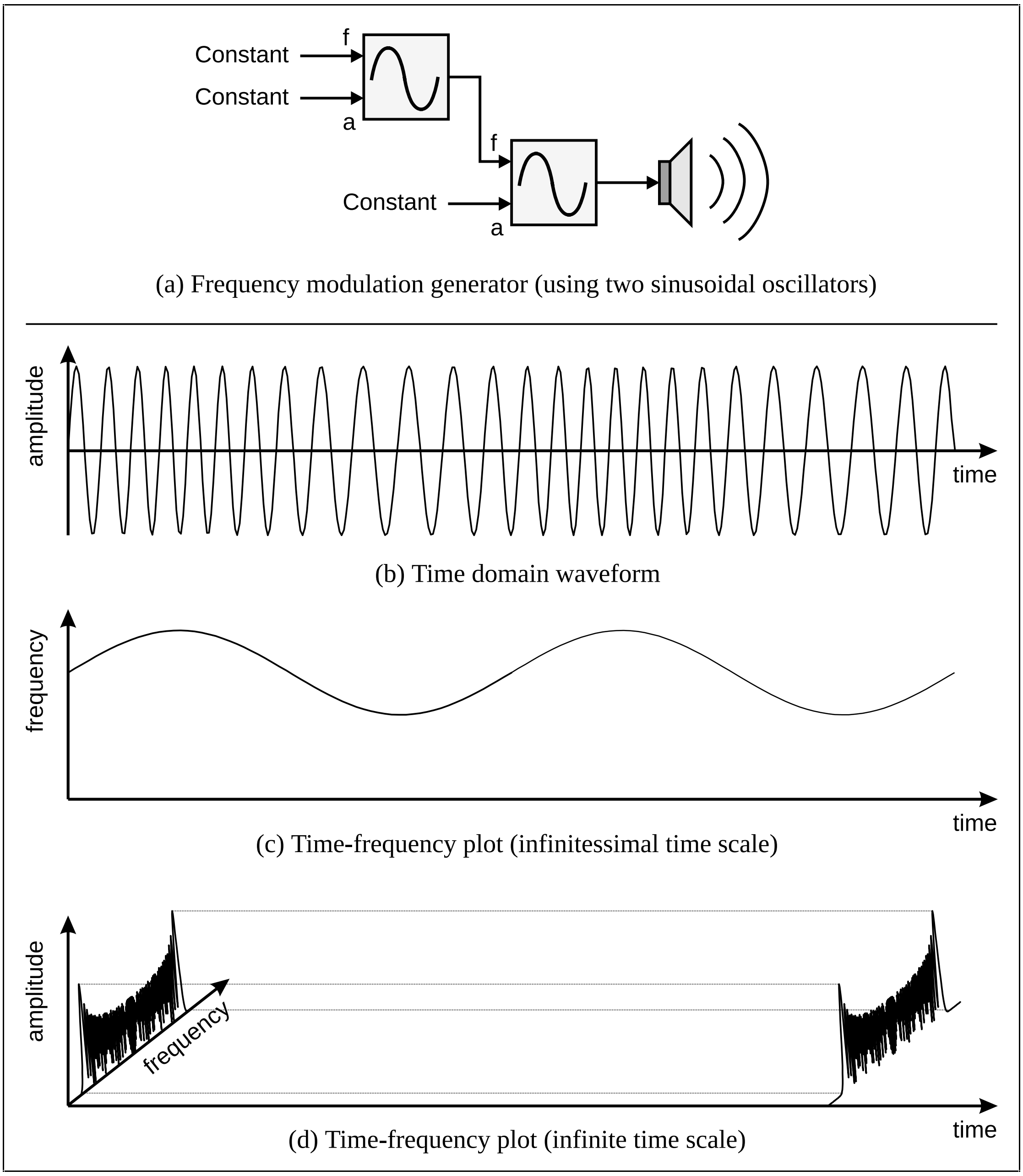
Take the example of a sinusoidal oscillator whose frequency is modulated by another sinewave (see Figure 3.1a-b). If the rate of modulation is slow, the sound heard is a simple tone with an oscillating pitch. In signal terms this would be described as a single, dynamic sinusoid whose time-frequency representation is shown in Figure 3.1c. If the rate of modulation is fast, the sound heard is a rich static tone, such as from an FM synthesiser [Chowning,1973]. This would be described as a static, complex spectrum, as shown in Figure 3.1d.

The two descriptions are very different, but all that was altered was the modulation rate. Which description is correct? If they are both correct, then at what rate of modulation should the representation switch from one to the other? The answer is that both forms of description are mathematically correct, at all rates of modulation. Neither form changes the content of the signal; they merely *represent* the signal in different ways. It is *our* perspective, as observers, that places the judgment on the correctness of a particular description.

In this example, what is changing is the time scale of the variations. The representation we choose depends upon how rapidly the human ear can react to changes – i.e. the time scale of the ear determines our choice. When the time scale of a signal change is slow compared to the time scale of the ear, then we favour the time-varying spectrum. When it is fast compared to the ear, then we favour the static, but more complex spectrum.

Taking this argument to its extremes, an entire piece of music could be represented by a single frequency spectrum; similarly it could be represented by the highly complex oscillations of a single particle. For musical purposes, neither description is very useful (although both are accurate). The former is the frequency domain description (infinitely long time scale), the latter is the time domain description (infinitesimally short time scale). What is required for sound analysis is an intermediate time scale that matches the natural time scale of our hearing system.

By requiring an intermediate time-scale representation, both time and frequency variables are present in the description. In seeking a two dimensional description from a one dimensional signal, the illustrated ambiguity of the signal's content is introduced. The time domain sample and its frequency domain equivalent are mathematically defined by a unique (one-to-one) mapping, whereas there are an infinite variety of valid joint time-frequency maps.



(a) Frequency modulation generator (using two sinusoidal oscillators)

(b) Time domain waveform

(c) Time-frequency plot (infinitessimal time scale)

(d) Time-frequency plot (infinite time scale)

Figure 3.1 – Alternative representations of frequency modulation

The procedure for choosing the appropriate TFR for the analysis-resynthesis system can be seen as finding the optimal match between the algorithmic signal representation and the perception of the human auditory system.

3.1.2 The TFR Must Match the Sound Model

The above discussion demonstrates that our perception of sound colours what we wish to see from a TFR. Similarly, when we observe a time-frequency representation, our preconceptions about signal content condition what can be extracted from the representation.

Take the example of deterministic analysis within the Deterministic Plus Stochastic model. Each frame of the STFT is an FFT of a short time window on the source waveform. The FFT, of itself, places no interpretations on the signal, nor does it condition it. Yet during the analysis, the peaks in the spectrum are interpreted as the instantaneous status of the partials of the sound. There is the *expectation* that the peaks have a meaning. In imposing a meaning on a representation, there are inevitable restrictions, which take the form of assumptions. In this case the assumption is that the windowed signal is periodic, and approximates stationarity over the duration of the window. The consequence is that when the analysed signal does not fit the assumptions, there is perceived distortion.

The extraction of model features from any TFR involves the application of assumptions and expectations. Indeed the assumptions and expectations can be considered as the implicit form of the model in the analyser. The aim of this discussion is not to create a TFR that is free from model conditions, but to demonstrate that it is necessary to be aware of how the signal is being conditioned. The 'innovative' TFRs of sections 3.2.5 and 3.3 make use of these conditions to extract extra information, instead of being limited by them.

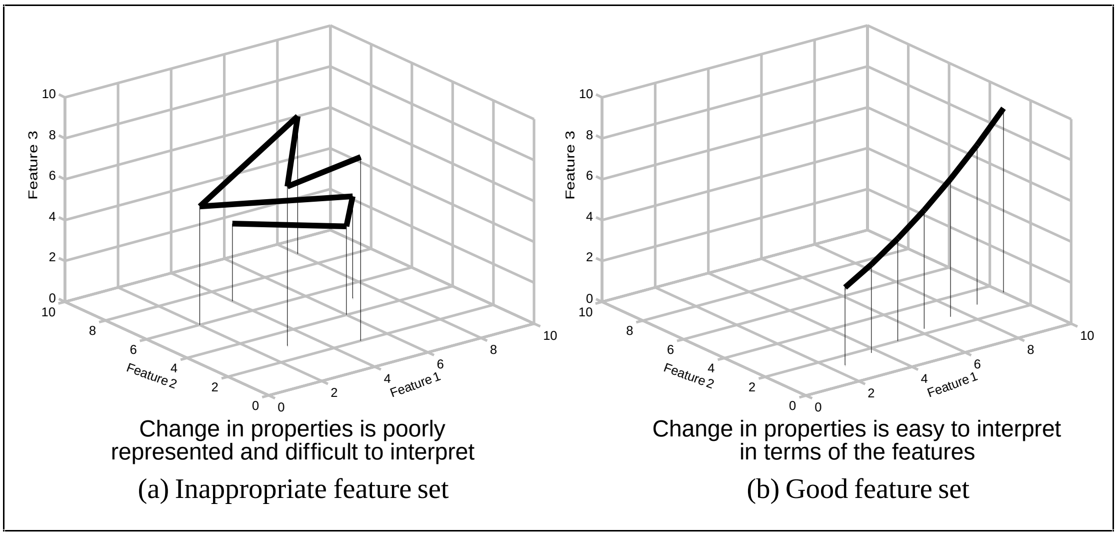
In summary, the mathematical transform or distribution that is the basis of the TFR, must be matched to the ear (in terms of time scale, resolution, etc.), so that the data presented is relevant to a sound model. This is because the analysis of sound is the extraction of perceptual features, not simply an encoding scheme for an arbitrary audio signal. In addition, by applying the model to a time-frequency distribution, assumptions and expectations are implicitly included. These can place conditions upon the extraction of information and can influence the reliability of the analyser, so it is important that a) the features are easily accessible from the representation (e.g. partials from peaks), and b) the conditions are known so that the analyser can detect unexpected situations, or even make use of the data in those situations.

In order that the TFR can be chosen well and applied effectively, it is obviously necessary to have a clear definition of the model, its assumptions and expectations. In turn the features that are to be extracted depend on the application of the model. The next two sections examine, in the context of this thesis, what is the application for the model, and hence, what features should define the model.

3.1.3 What is the Application for the Model?

As has been intimated already, the role of analysis-resynthesis is more than encapsulation and resynthesis of sound, for such tasks could be achieved by the FFT, the Wavelet Transform, or any number of reversible (invertible) transforms. (Even the time-domain sample fulfils these criteria.) The true application of analysis-resynthesis is *musical transformation*, or more precisely, *musically-useful transformation*.

Once again this brings into focus the question of what is musical. Since analysis-resynthesis requires some source material, it may be assumed that the choice of source material is musically important, and therefore that some of its musical properties should remain. Of those that are transformed, they may be modified within a scope that is considered musical. In a model whose features are well-chosen a musical modification should therefore be simply represented. Progressive transformation of one property of the sound should appear as a simple movement in feature space. Figure 3.2 demonstrates this pictorially by looking at a hypothetical three-dimensional feature space.



(a) Inappropriate feature set

(b) Good feature set

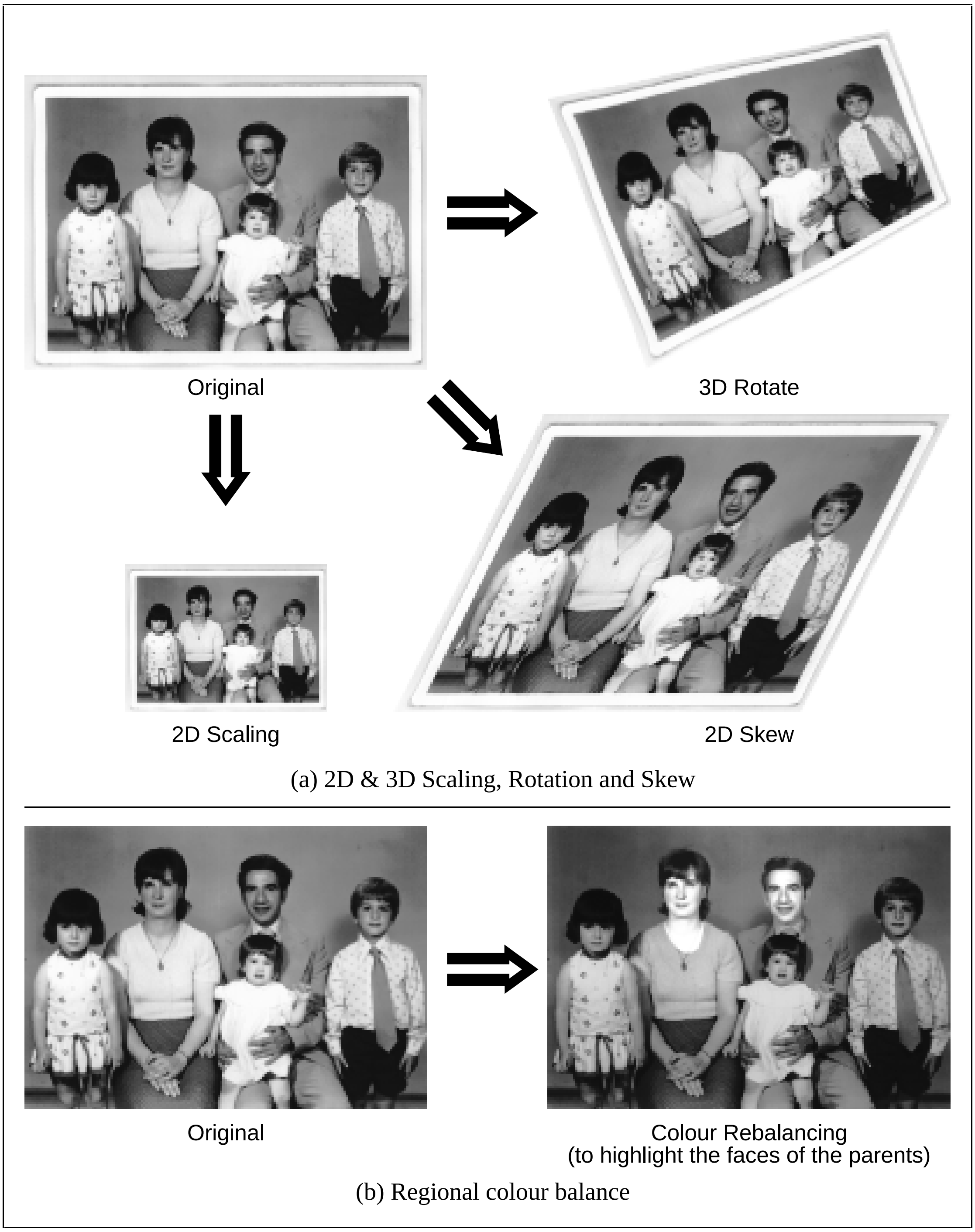
Figure 3.2 – The importance of choosing an appropriate set of features

This definition of a musically-useful transformation provides some basic guidelines but it is ultimately open ended. As a scientific definition it is inadequate and vague, but this imprecision is intentional, because it is flexible enough to allow the model to develop. There is the expectation that as the model reaches maturity and as musicians explore its capacity for transformation, the definition will be refined.

At present, analysis-resynthesis is in its infancy and musicians (outside this specialised research community) have not had access to these tools. So there is very little knowledge about what *could* be done to sounds that might be musically useful. Nor have musicians been given the opportunity, until now, to imagine beyond the conventional music tools, what tools they might require. Once there is more experience with such tools and this highly flexible way of working with sound, then the definition of 'musically useful' can begin to be clarified.

In the meantime, some clues can be gleamed from the parallel field of image processing, which despite dealing with the more complex case of two dimensional data, has made significant progress in generating tools that are 'artistically useful'. To illustrate current and potential applications of sound analysis-resynthesis, some image-based analogues are presented.

Just as there is generative sound synthesis, where sounds are created from scratch (such as in analogue or FM synthesis), so too there are painting tools (such as pencil, brush, airbrush, line, arc, fill) that enable images to be generated from a 'clean sheet'. Analysis-resynthesis uses source material which is modified, so it is more similar to the modern photo-based art packages, such as Adobe's Photoshop™.



(a) 2D & 3D Scaling, Rotation and Skew

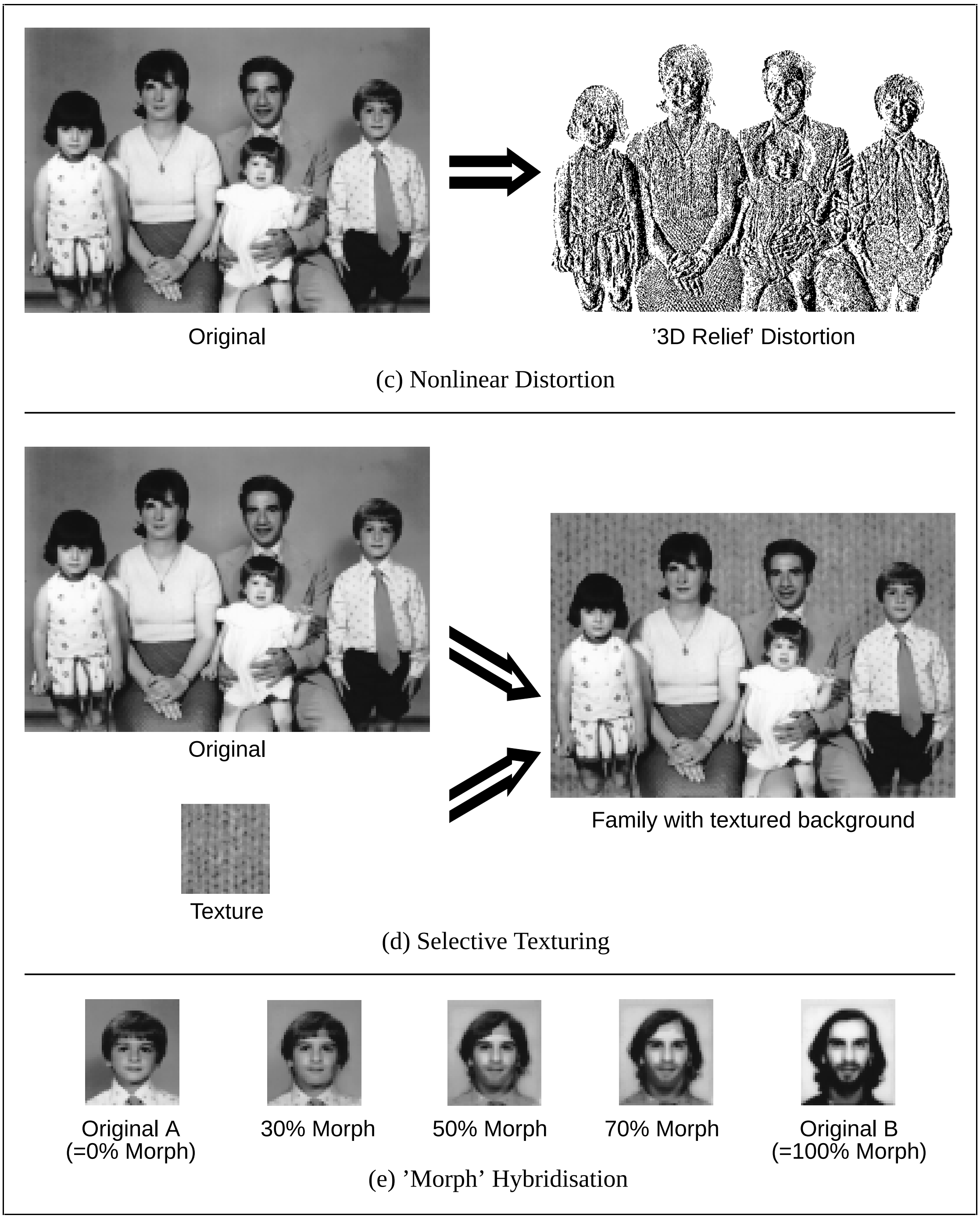
(b) Regional colour balance

Figure 3.3 – Image Manipulation Techniques (as analogues for sound transformation)

The image modification tools include (examples shown in Figure 3.3):

- stretch, compression, skew and rotation in two and three dimensions (Figure 3.3a);
- regional colour balance, which can enhance certain image features (Figure 3.3b);
- nonlinear distortion, to create effects like ‘embossing’ (Figure 3.3c);
- texture mapping, where an image of a textured surface modulates the source image or part of it, say to give the appearance of a painting on cloth (Figure 3.3d);

- morphing between images, which ‘pulls’ defined regions of one image toward the corresponding regions of a second image and interpolates the region contents also, generating a hybrid image which is more interesting than a cross-fade (Figure 3.3e).



(c) Nonlinear Distortion

(d) Selective Texturing

(e) ’Morph’ Hybridisation

Figure 3.3 – Image Manipulation Techniques (as analogues for sound transformation)

The equivalents in sound are (respectively):

- time-stretch, pitch-shift and formant-corrected pitch-shift;
- sound source extraction/suppression;
- nonlinear distortion;

- cross-synthesis;
- timbre morphing.

Time-stretch, more formally known as time-scaling, allows linear temporal compression/expansion of a sound, but retains its pitch and timbral character. Pitch-shift, more accurately frequency-scaling, allows pitch modification, whilst retaining the temporal properties and some of the timbral character. Pitch-shift is the counterpart of time-stretch because one can be obtained from the other by resampling.

Formant-corrected pitch-shift is most relevant to the human voice which does not sound natural when pitch-shifted, because the formant structure is also scaled. By retaining the spectral envelope of the formants, individual partials can be frequency-scaled to produce a pitch-shift that retains the timbral character of the original voice.

Instrument extraction/suppression relies on being able to distinguish one sound source from another in a ‘mix’ (a composite sound), and altering their ‘mixing levels’ (amplitude weightings).

Nonlinear distortion is a broad term that defines a mapping function. It could be applied to the time domain amplitude, in which case it generates extra harmonics; this is commonly known as waveshaping [Roads,1985]. It could also be applied to individual features, say the degree of variation in a partial’s trajectory, to flatten or enrich the timbre.

Cross-synthesis imposes the spectral envelope of one sound (c.f. the texture) onto the partials of another (c.f. the source image).

Timbre morphing requires more than one source sound, from which a hybrid sound is created through interpolation of the analysed musical features.

All of these sound transformations are related in that they can be implemented from information about the partial structure and the accompanying stochastic energy. The simplest of these are time-stretch and pitch-shift, so these two have been used as the benchmark for testing the hypotheses of this dissertation. They are also the answer to the question “What is the application for the model?” within the context of this thesis.

3.1.4 What Features Define the Sound Model?

With the application of the model defined as time-stretch, this section looks at what features are important. As indicated by Figure 3.2 above, the purpose of a transformation is to alter one feature (or a small collection of features) in a continuous manner, whilst preserving the other features. In the case of time-stretch, it is desirable to alter the rate of evolution of a sound (including its tempo and duration), whilst preserving its pitch, rhythm and timbral qualities.

The preservation of pitch can be seen as the requirement that at any particular point in the synthesised waveform, the period will be the same as the original waveform period, at the corresponding location in the sound. The timbre is a complex entity that can itself be described as a multidimensional space [Grey,1975]. Some aspects of the timbre are inevitably changed by altering the rate of evolution of the sound, but much of the character of a sound can be preserved by retaining the instantaneous magnitude spectrum. To preserve rhythm there is the need to keep the original relative timing between emphasised events. Also, there is the requirement to preserve the sense of emphasis, so that the rhythmical structure is neither fused (upon speed up) nor diffused (upon slow down).

These requirements appear to be upheld by the concept of the Deterministic Plus Stochastic model, in which the deterministic aspect ensures preservation of pitch and together both aspects represent the variations in the instantaneous magnitude spectrum. Therefore (at least at this stage[1]) there is no urgent need to replace this model. Indeed, by confirming the use of the model, it allows developments in the implementation to continue with greater confidence that the underlying basis is well-founded. It should be noted at this point that there is a distinction between the model and the analysis-resynthesis system that realises it, so acceptance of the model's feature set is not automatically an acceptance of the methods used to implement it.

The features of the Deterministic Plus Stochastic model are time-varying sinusoids (that are expected to be roughly equally spaced throughout the spectrum for single source sounds, and no closer than 50Hz from one another) and time-varying spectral envelopes (whose spectral accuracy is less important). The deterministic features require frequency resolution sufficient to enable detection of sinusoids 50Hz apart throughout the spectrum and to estimate the instantaneous frequency to the resolution of the ear. Also, both types of feature require sufficient time resolution to characterise rapid changes. These requirements determine the objectives in choosing or designing a TFR. The problem with the current TFR of the Initial Model, the STFT, is that by satisfying the former requirement it compromises the latter.

With the aims of the TFR clarified, section 3.2 proceeds to present and evaluate a number of time-frequency distributions that are popular for signal processing or are currently being developed, in the search for a suitable alternative to the STFT. Beforehand, the final part of this section presents two personal observations of a cautionary nature about model development. Although model implementation could proceed without these warnings, their purpose is to equip the designer with the ability to see *outside* the model.

3.1.5 Observations on Model Development

Often it is possible to stare in great detail at a problem without seeing a solution, optimising the system by parameter tweaking, to minimise undesirable effects. At these times, if the designer stands back and looks at the system from a wider viewpoint, viewing not only the specific problem but observing the workings of the system as a whole, then there is sometimes the opportunity to make a fundamental change that could avoid the problem altogether.

The two cautions that are presented in this section focus on observations of my own approach to model development, as well as that of others. This discussion is included because of the great value I have discovered in applying these approaches. They were the inspiration for the scheme presented in section 3.3 and ultimately were the source of the investigation presented in this chapter.

[1] For the purposes of this chapter, the Deterministic Plus Stochastic model of sound is considered adequate. However it is brought into question elsewhere in the thesis (see section 6.5).

3.1.5.1 Circular Reasoning

How is a model created in the first place? Observations are made of the system of interest. From a number of observations, patterns are detected and hypotheses made that either describe the behaviour or suggest an explanation for its cause. The model is the formal construction of such hypotheses. The model is validated by comparing observations of real behaviour with the model's predictions, for data used to generate the hypothesis (bug-fixing) and for new data (true testing). Through an iterative process, where the discrepancies are fed back to modify the hypotheses, the model can be refined.

This whole process rests on the observations of the system's behaviour. In the case of sound modelling, beyond the initial observation of quasi-periodicity, almost all observations are made from a TFR. However the TFR already views the audio signal from a particular perspective and therefore it already imposes its own implicit model upon the data[2]. The hypotheses that result are therefore influenced by the TFR model and cannot fail to be based upon it. It is no surprise then, if the resulting model confirms the use of the original TFR.

The process of reasoning by which the TFR is chosen has become a circular argument. A proves B given C. B proves C given A. C proves A given B.

The result is a closed system which is only capable of modification within its own parameters. That is not to say that a good model cannot be found, but simply that it will be limited in scope. The only way to widen that scope is to break out of the loop and take a wider (or an alternative) perspective.

There is a danger of circular reasoning in developing the sound model further. Ever since the STFT was first applied as the TFR for sound analysis, it has been the standard method for examining the time-varying spectrum of sound signals. Subjective results from the analysis-resynthesis system may indicate that certain features are not being well-represented and there is the temptation to look at the STFT representation for evidence of this deficiency. However, if part of the fault lies with the STFT's inability to represent those features well, the solution will never improve beyond small increments. As an example, the deterministic aspect of the model is aimed at partials, which are considered to be stable sinusoids. The stochastic aspect is intended to capture all non-partials, yet it captures noise which is also only slowly-varying. The inability of the stochastic aspect to represent rapidly varying elements or abrupt changes in the sound is a direct consequence of relying on the STFT for its analysis.

[2] Note that here the TFR is not the time-frequency distribution, the mathematical transform alone, but the applied form of the distribution in which assumptions are imposed and there are expectations of how the results should be interpreted.

3.1.5.2 Every Model is Incomplete

> "So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality."
>
> ——— *Albert Einstein (sourced from [Kosko,1994,p.3])*

The aim in designing a model is that it should be a complete representation of reality within the domain of the system of interest. However, real systems do not exist in isolation to the rest of reality. They interact with every aspect of reality to some degree, directly or indirectly. Taking this argument to its logical conclusion implies that it is *impossible* to develop a model that perfectly represents the world (since the model itself must exist in its own world model). A less esoteric and more practical conclusion is that every model is incomplete.

Models can be useful tools for understanding small portions of the world. Once the boundaries of a limited world-view are defined, the model provides a simplified description of the world within those boundaries. As long as the input data complies with the model restrictions, the model can mimic the behaviour of the real system. However, the input data are real signals generated externally to the system, so they are not constrained by the model, and the assumptions of the model are bound to fail under some circumstances.

The aim of this discussion is to show that models can be useful if their boundaries are understood and well-defined. Furthermore, model deficiencies need not result in total failures.

Ideally, with an adaptive model the effective domain is modified with experience. Also, 'graceful degradation' can be incorporated, such that the errors are small for inputs that lie outside, but near the model's domain. Finally, where a model is neither adaptive nor graceful, then a stated restriction can be applied to its use.

Ultimately the former solutions are desired for a sound model, but during the current development phase, while we are trying to gain better knowledge of sound signals ourselves, the latter option is our only satisfaction. Hence the model is restricted to musical signals only, where *musical* is a defined subset of *audio*, and represents the limit of our current understanding.

Having established the importance of the TFR within the modelling process, and having defined the features that need to be accessible from a TFR, the remainder of the chapter presents various candidate TFRs. Section 3.2 presents a review of 'standard' methods for spectral representation and current developments in extending those, and section 3.3 presents original work that provides an extension to the STFT.

3.2 A Brief Review of Alternative Approaches to Time-Frequency Representation

The STFT, which is the standard TFR for sound analysis-resynthesis, has desirable properties: it displays partials simply (as peaks in each frame); it displays multicomponent signals without inter-component distortion (save additive effects); and it enables easy calculation of frequency, amplitude and phase for each frame. Its only limitation is its time-frequency resolution and this is the primary impetus in the search for a suitable alternative.

Within this section, various techniques and transforms are explored for their suitability as TFRs for sound analysis, given the role of the TFR in sound analysis. Since the STFT is the standard time-frequency transform for sound representation and much nonstationary signal analysis, it seems appropriate that the STFT is both the starting point of the review and the benchmark against which alternative techniques are evaluated.

It should be noted that the purpose of this review is to inform the reader of the wide scope and the differing approaches to signal representation. Therefore the bulk of detail for implementing such schemes (particularly the rigours of the mathematical foundations) are omitted; instead many references are provided from which the necessary information can be obtained.

3.2.1 Short Time Fourier Transform (STFT)

The details of how the STFT is implemented within the sound analysis-resynthesis system were explained in section 2.3.2.1. In that section it was discovered that the transform’s time-frequency resolution is limited and that any increase in frequency resolution is at the expense of time resolution.

3.2.2 Wavelet Transform (WT) and Bounded-Q Fourier Transform (BQFT)

The window-length of the STFT is fixed over the entire frequency range, such that the time-frequency ‘cells’ are uniformly spaced in both time and frequency. The wavelet transform (WT) [Rioul & Vetterli,1991] takes the alternative approach of making the window-length inversely proportional to frequency, so that each window contains the same number of periods of a frequency.

In explaining how the FFT is created, it is common to think of the signal being windowed and then convolved with sinewaves at the frequencies of the bins, and indeed this is how it is implemented. Mathematically, the same outcome results from convolving the signal with windowed versions of the sinewaves. This latter explanation is more appropriate for the WT, where the customary terminology describes the windowed sinewave as the *analysing wavelet*. Since every wavelet contains the same number of periods of the sinewave, each wavelet is a dilated or time scaled version of the others, where the dilation or scale depends on the frequency.[3]

The same mathematical laws that govern resolution for the STFT apply to the WT (discussed below in section 3.2.3.1), so the narrower higher frequency wavelets have a wider bandwidth, such that the WT maps the spectrum logarithmically. The logarithmic approach appears to be more in tune with the way in which we perceive the world: in vision, the photoreceptors of the eye respond logarithmically to visual stimulus [Deutsch & Deutsch,1993], in sound, loudness is logarithmically calibrated as a function of signal amplitude, and in music, the pitch scale is logarithmically related to frequency[4]. For image analysis, the WT is a popular method of representing textures [Porter & Canagarajah,1996; Chang & Kuo,1993], and for pitch (or note) tracking in music recordings, the WT can provide a constant profile for all pitches.

The logarithmic law seems intuitively to be a better match to the way frequencies change – how can a low frequency change faster than a high frequency? If a low frequency appears to change rapidly, shouldn't the signal be described as having a high frequency content? The intuitive "Yes!" answer belies an implicit signal model (c.f. the FM example of section 3.1.1). Both the STFT and the WT contain all information about a signal and are invertible. The question for signal analysis is which one is better suited to extracting the musical features from sound.

The answer to this is not straightforward: the STFT provides the linear frequency spacing that is essential for distinguishing the equally spaced harmonics of a pitched sound, yet the WT enables capture of fast changing elements, like attack transients, as well as steadier components.

The Bounded-Q Fourier Transform (BQFT) [Chafe & Jaffe et al.,1985] is a hybrid of the WT and the FFT, developed as a compromise that provides *locally linear* spacing. The spectrum is divided into broad bands such that higher frequency bands are calculated to favour good time resolution, as with the WT. However, within each band the frequency spacing is linear, as with the STFT. The BQFT was implemented in a sinusoidal model [Masri & Bateman,1994] with the aim of providing an analysis scale that automatically adapts to the pitch of the analysed instruments – higher pitched sounds appear in the upper bands and their harmonics are more widely spaced.

The following results were observed (+ bullets signify advantages of the BQFT over the STFT, – bullets signify disadvantages):

+ Automatic TFR scaling, because each pitch is analysed at ideal frequency spacing for the first few (and often the most important) partials;

+ The automation enables analysis of multiple simultaneous pitches (chords or multiple instruments), where each pitch is analysed optimally (for the first few partials);

[3] The differences between STFT and WT can also be explained in frequency domain terms, where each time-frequency cell is the output of a linear FIR filter. The filters of the STFT are equally spaced in frequency, whereas the filters of the WT have constant Q and are therefore logarithmically spaced.

[4] inasmuch as the perceptual properties of loudness and pitch can be related to the physical properties of amplitude and frequency.

+ The initial transients of percussive pitched instruments are captured within the sinusoidal analysis; e.g. the onset of piano notes;

+ The noisiness that results from rapid changes in the higher frequency partials is captured within a sinusoidal analysis; (in fact many tonal instruments do not require a stochastic aspect to the model for faithful reproduction when using the BQFT);

– Above the first few partials, the lack of frequency resolution leads to erroneous detection. Upon resynthesis this is audible, although very quiet, as high frequency susurrus artifacts (similar to a radio with bad reception or a person making the sound “srsrsrsr”);

– Some sounds, particularly the human singing voice, have a wide bandwidth with significant energy in the higher frequencies, so improved enunciation is compromised by the ‘closed’ sound after resynthesis (i.e. the sound is effectively low-pass filtered by the analysis-resynthesis system, losing the higher harmonics);

3.2.3 The Cohen Class of TFR

3.2.3.1 Being Certain about the Uncertainty Principle

The STFT, WT and BQFT have different emphases on time or frequency resolution, for different parts of the spectrum, but all of them are limited to the same fixed resolution (in terms of the time-bandwidth product) by the same mathematical law. This law is often quoted as the Heisenberg uncertainty principle[5], and almost any technical paper on the subject of time-frequency analysis will mention it at some point (e.g. [Classen & Mecklenbrauker,1984; Fineberg & Mammone,1992]). The law is encapsulated in the inequality:

$$\Delta t \cdot \Delta\omega \geq \tfrac{1}{2} \qquad (3.1)$$

where Δt and $\Delta\omega$ are the effective duration and effective bandwidth (in angular frequency) of a signal. To avoid confusion, Cohen provides the clearer definition of Δt and $\Delta\omega$ as [Cohen,1989]:

$$(\Delta t)^2 = \int \left(t - \bar{t}\right)^2 |s(t)|^2 \, dt \qquad (3.2)$$

$$(\Delta\omega)^2 = \int \left(\omega - \overline{\omega}\right)^2 |S(\omega)|^2 \, d\omega \qquad (3.3)$$

[5] The energy localisation restriction that limits time-frequency resolution in this case looks similar to the Heisenberg uncertainty principle. However the Heisenberg uncertainty principle relates to a statistical uncertainty, whereas time-frequency representation is a signal representation problem in which there may be no actual uncertainty as to signal content [Cohen,1989]. In fact both equations are a special case of the Cauchy-Schwarz inequality, a property of every linear time-invariant system that results from making linear approximations to an essentially nonlinear world [Kosko,1994,pp.107-114]. For simplicity, and in keeping with the majority of publications, within this dissertation the inequality is referred to as the Heisenberg uncertainty principle or the linear uncertainty principle.

where $s(t)$ is the signal under analysis (notionally a monocomponent signal[6]),
$(\Delta t)^2$ is the standard deviation of the function, $s(t)$,
$(\Delta\omega)^2$ is the standard deviation of its Fourier transform, $S(\omega)$,
$\bar{t}$ is the mean time,
$\bar{\omega}$ is the mean frequency.

The Fourier transform and its variants achieve the minimum energy localisation permitted by equation 3.1[7], i.e. $\Delta t \cdot \Delta\omega = \frac{1}{2}$.

The uncertainty principle is a fundamental limitation of all linear, time-invariant systems. Within the context of time-frequency representation, it can be understood as a consequence of trying to generate a linear representation of potentially nonlinear signals. From its equation, the STFT can be seen to be linear, time-invariant:

$$STFT(t,\omega) = \int s(u)\, w(u-t)\, e^{-j\omega u}\, du \tag{3.4}$$

where *t* is time, *u* is offset (centre of transform time-window), ω is (angular) frequency,
s(t) is the time domain signal under analysis,
w(*t*) is the window function
STFT(*t*,ω) is the Short Time Fourier Transform.

The equation is linear because the integral includes a single, linear function of the signal, *s*(*t*). It is time-invariant because a time-shifted version of *s*(*t*) would result in time shifting of the time-frequency representation *and no other effect*. Time-invariancy is a desirable property because the results of the transform are identical for signals with the same content, regardless of when they occur. Linearity means that there is no conflict between components, a positive attribute (as will become evident), but it does limit the usability of the transform.

In the representation, the STFT is optimal for signals with components of linearly increasing phase. That is, for signals containing linear phase (stationary frequency) components, the representation is a delta function – a single, narrow spike – that takes up the minimum space on the time-frequency map – a single STFT cell. (This assumes that the default rectangular window has been used and the component frequency equals the analysis sinewave's frequency, so there is no spectral leakage.) This results in the familiar assumption or rule, that the signal must be stationary, for optimal use of the STFT.

3.2.3.2 The Wigner Distribution and the Bilinear Form

The Wigner distribution [Wigner,1932] was first introduced in 1932 in the context of studying quantum statistical mechanics [Cohen,1989; Jeong & Williams,1992]. In 1948, Ville applied it in the context of signal processing [Ville,1948]. The Wigner-Ville distribution is:

[6] A monocomponent signal contains one spectral component. It is usually assumed that this component is deterministic, and is therefore a sinusoidal signal.

[7] The bin spacing of the STFT equals the reciprocal of the length of the FFT array in the time domain, which appears to give a cell size of $\Delta t \cdot \Delta\omega = 1$, however the shape of the window function (in time and frequency) ensures that the minimum energy localisation can be achieved.

$$WV(t,\omega)=\frac{1}{2\pi}\int s^*\left(t-\frac{1}{2}\tau\right)e^{-j\omega\tau}\,s\left(t+\frac{1}{2}\tau\right)d\tau \qquad (3.5)$$

where t is time, τ is delay (time-lag), ω is (angular) frequency,
s(t) is the time domain signal under analysis,
$s^*(t)$ is its complex conjugate,
$WV(t,\omega)$ is the Wigner-Ville time-frequency distribution.

The Wigner-Ville distribution is bilinear, time-invariant because it has two linear functions of the signal, $s(t)$. As a result it is able to provide an optimal representation for quadratic phase (linear frequency modulation) signals, of which stationary frequency is one case. In terms of the representation, this means that nonstationary signals exhibit reduced distortion and the time-frequency resolution is better than that prescribed by the linear uncertainty principle.

3.2.3.3 Cohen's Class and Higher Order Spectra

In 1966, Leon Cohen developed an equation that could be used to describe all bilinear time-frequency distributions [Cohen,1966, 1989], including the spectrogram and the Wigner-Ville distribution:

$$P(t,\omega)=\frac{1}{4\pi^2}\iiint e^{-j\theta t-j\tau\omega+j\theta u}\,\Phi(\theta,\tau)\,s^*\left(u-\tfrac{1}{2}\tau\right)s\left(u+\tfrac{1}{2}\tau\right)du\,d\tau\,d\theta \qquad (3.6)$$

where t is time, u is offset (centre of transform time-window), τ is delay (time-lag),
ω is angular frequency, θ is Doppler (frequency-shift),
$s(t)$ is the signal to be represented,
$\Phi(\theta, \tau)$ is the 'kernel' (see next section),
$P(t, \omega)$ is the representation domain.

From this equation, the spectrogram, which is the magnitude-squared function of the STFT and is therefore bilinear in $s(t)$, can be shown to be linearly related to the Wigner-Ville distribution. (In fact the spectrogram is a smoothed version of the Wigner-Ville distribution, an alternative explanation for the time-frequency averaging that causes reduced resolution.)

Cohen's generalised form for representing all bilinear time-frequency distributions, known as the Cohen Class, enabled similar extensions to higher order spectra. Initially, the Wigner distribution was 'upgraded' to the Wigner bispectrum [Gerr,1988], the third order (trilinear) Wigner distribution, and later the generalised form for higher order spectra was developed [Boashash & O'Shea,1991; Boashash & Frazer,1992]. Called the *Generalised Wigner-Ville Distribution (GWVD)*, it enables design of distributions to *optimally* represent signals whose frequency trajectory could be described by any arbitrary polynomial.

The process by which a signal is represented is shown to be a mapping followed by a FFT. The mapping converts the frequency modulation (up to the targeted order) onto the linear phase plane (zero frequency modulation), where it can be analysed by Fourier techniques. This can be understood in terms of autocorrelation. The spectrogram can be generated as the magnitude FFT of the autocorrelation of a signal. In the autocorrelation, a time-shifted version of the signal is correlated with the original signal, generating peaks at the signal's periodicities. Bilinear distributions work in the same way, except the correlation is between a time-shifted, frequency-shifted version of the signal and the original signal. Therefore peaks arise where locally periodic elements have changed with linear frequency modulation.

Just as the bilinear achieve a better resolution than the linear, so the higher order spectra achieve better resolution than the bilinear, the accuracy increasing with every order (for an arbitrarily nonlinear frequency trajectory). The process can be extended to an arbitrary order, but at each increase in order there is an increase in system complexity and computation. Perhaps more importantly for applications to sound analysis, there are also extra interactions between the additive spectral components that can obscure the 'real' spectrum.

3.2.3.4 Cross-terms and Modifications to the Kernel

For a monocomponent signal, the bilinear and higher order spectra have been proven capable of extracting detail from highly complex frequency trajectories. However in the Fourier Series based model of sound, a sound signal contains many spectral components that are additively combined – a multicomponent signal. In the bilinear and higher order spectra, as the signal is mapped to stationary frequencies (by the autocorrelation function), these components interact with one another. The situation is examined for the bilinear case.

The expansion of the mathematical expression $(x+y)^2$ gives a term in each of the variables, x^2 and y^2, and a term in both, $2xy$. In an analogous way, the expansion of $s^*\left(u-\tfrac{1}{2}\tau\right)s\left(u+\tfrac{1}{2}\tau\right)$ yields terms in each of the spectral components (auto-terms) and terms in each combination pair (cross-terms). In the time-frequency representation of the signal, the auto-terms appear, as expected, at the time and frequency locations of the actual signal components, whereas the cross-terms appear halfway between the corresponding components. This has two undesirable effects: there appears to be energy in the signal at times and frequencies where there is none, and some cross-terms lie on auto-terms, obscuring or distorting them.

Although the expression $s^*\left(u-\tfrac{1}{2}\tau\right)s\left(u+\tfrac{1}{2}\tau\right)$ yields the cross-terms, it is the kernel, $\Phi(\theta,\tau)$, that determines the representation. Through modifications to the kernel, many authors have sought to suppress the cross-terms. Perhaps the most widely recognised solution is the exponential kernel developed by Choi and Williams [Choi & Williams,1989], known commonly as the Choi-Williams distribution. From these studies it was recognised that just as there is a trade-off between time and frequency resolution in the STFT, so too there is a trade-off between cross-term suppression and auto-term resolution. In the case of the spectrogram, the cross-terms are totally suppressed, and the auto-term resolution is the worst. (Similar kernel modifications have been applied to higher order spectra to reduce the cross-terms interference [Fonollosa & Nikias,1992].)

3.2.3.5 Application of the Cohen Class of TFR for Sound Representation

A few studies have compared the various distributions for application to multicomponent nonstationary signal representation. Jones and Parks have compared the spectrogram with the Wigner and Choi-Williams distributions [Jones & Parks,1992]. They conclude that for nonstationary, multicomponent signals the spectrogram outperforms both of the others (if the window-length is chosen to 'match' the signal).

Equation 3.6 facilitates development of an infinite number of bilinear time-frequency distributions, through kernel design. However, only a small proportion satisfy certain 'desirable' properties for an energy distribution. These include that the energy should never be negative (it can be with the Wigner distribution), that the marginals[8] are satisfied, and so on. The possibility for violation of such properties demonstrates a mismatch between the mathematics and the real world – only when the properties are satisfied can a physical interpretation truly be considered valid.

In Cohen's review paper of 1989 [Cohen,1989], he concludes "The enigma of these distributions is that they sometimes give very reasonable results and sometimes absurd ones… The fact that these distributions cannot be used in a consistent manner is one of the main areas that needs much further theoretical development."

Nevertheless, new distributions continue to be developed through kernel design, e.g. [Jeong & Williams,1992], with some promising results. In conclusion, it would appear that bilinear and multilinear distributions offer the desired improvement in time-frequency resolution, but that further kernel development (and better understanding of the underlying physical analogues) is required before these TFRs can be applied with confidence to sound modelling.

An informative introduction to the Cohen class, its history, derivation and properties as well as some personal insights can be found in [Cohen,1989]. The mathematical rigours are also presented in [Boashash,1990] with graphs comparing distributions and consideration of multicomponent analysis, with application to underwater sound analysis.

3.2.4 Parametric TFRs

The spectral estimators considered above are all nonparametric techniques: they generate their spectral representations through mapping functions and transforms that do not alter or interpret the data in any way, but simply change the form (trans-form) of the data. As such, they do not bias the outcome of an estimate (although subsequent processing in the analysis-resynthesis system may place interpretation upon features in the TFR).

There also exist parametric techniques for spectral estimation. These trade in the information preservation, and incorporate foreknowledge about the signals to be analysed, to gain better resolution of the desired features. This knowledge takes the form of a parameterised signal model. Estimation is then a process of matching the model parameters to the data.

3.2.4.1 Autoregression and Moving Average

The most popular parametric models are autoregression (AR), moving average (MA) and a combination of the two (ARMA). In brief, ARMA is based on a z-plane pole-zero model of the spectrum: AR models the poles of the signal spectrum (in the z-plane), which define narrow peaks, and MA models the zeros, which define localised minima. The assumption is made that

[8] The word 'marginal' derives from probability theory to indicate the distribution of a single variable. In this context, the total energy at a particular time, $|s(t)|^2 = \int P(t,\omega)\,d\omega$, and the total energy at a particular frequency, $|S(\omega)|^2 = \int P(t,\omega)\,dt$, are the marginals.

the signal has been generated by feeding a white noise input to a system with a transfer function of the form:

$$H(z) = \frac{B(z)}{A(z)} \tag{3.7}$$

where

$$A(z) = \sum_{k=0}^{p} a[k] z^{-k} \tag{3.8}$$

and

$$B(z) = \sum_{k=0}^{q} b[k] z^{-k} \tag{3.9}$$

where *H*(*z*) is the system transfer function,
A(*z*) is the z-transform of the AR branch (for *p* poles),
B(*z*) is the z-transform of the MA branch (for *q* zeros).

AR performs best (i.e. provides the most concise parameter set) where the resonances of a system or the harmonics of a signal are of interest, whereas MA is best at describing a broadband noise spectra. Where there are sinusoids embedded in white noise, ARMA is the best choice (because AR degrades in the presence of noise). In practice, the most widely used technique is AR, where parameter fitting requires solution to a set of linear equations, whereas MA and ARMA both require solutions to a set of (usually highly complex) nonlinear equations [Kay,1988,p.153].

Model Order

The order of a parametric model (i.e. the number of poles and/or zeros) defines the level of detail of the spectral estimate. If the model order is too small, then a highly smoothed spectral estimate results. If it is too high, then extra spurious peaks can appear. The need to find a suitable order adds a complication to the process. Several techniques have been developed, relying on some measure of closeness of the model to the actual spectrum at each iteration [Kay,1988,p.234; Makhoul,1975; Marple,1987,p.229].

Further Reading

For the interested reader, Kay [Kay,1988] and Marple [Marple,1987] provide a general grounding in linear parametric modelling techniques. Kay approaches the subject from a largely theoretical basis, comparing and contrasting different implementation techniques for statistical performance and computational efficiency. Marple’s book is geared more toward applications, providing an easier route to implementation. Makhoul [Makhoul,1975] provides a more readable introduction to AR and ARMA, from the perspective of Linear Prediction, in terms of system function rather than mathematical rigour.

3.2.4.2 Higher Order Parametric Modelling

Just as the above linear parametric techniques can improve over the time-frequency resolution of the FFT, so it is also possible to develop higher order parametric techniques that improve over the higher order spectra. The key to parametric model design is once again the kernel. The Cohen class of distributions can be made parametric by making the kernel a function of the signal [Cohen,1989]. Similarly, the higher order GWVDs could be made parametric by including the signal as a parameter of the kernel.

3.2.4.3 Application of Parametric TFRs for Sound Representation

The linear parametric techniques, described above are best known to computer musicians in the form of Linear Predictive Coding (LPC) [Markel & Gray,1976]. This is the basis of the well established speech model, in which voiced sounds are synthesised by filtering excitation functions through the model filter of the speech spectrum. For voiced sounds the excitation would be a pulse train at the pitch period; for unvoiced, it would be white noise. In this application, the model order is deliberately low, so that the spectral envelope (i.e. the formant structure) will be modelled instead of the individual partials. As such, this has potential for improving stochastic analysis and synthesis within the Deterministic Plus Stochastic model (where the dynamic response would be faster, because the estimates would require less data points). However this is not explored further within this thesis.

Despite the promise of better spectral resolution using a shorter analysis window, there are several drawbacks to the AR-MA techniques. In exchanging the generality of the nonparametric estimators for a model-based technique, the performance depends highly on the accuracy of the model. These techniques are linear and, like the FFT, assume that the signal is locally stationary. Also they assume that the signal is composed of pure sinusoids, possibly with Gaussian white noise. Because of the estimate's reliance on the model, it is sensitive to violations of the assumptions. Whereas the FFT would exhibit distortion to the peaks, for small violations the partial information would still be largely valid. The parametric methods are less resilient and even small violations can cause unpredictable results. Furthermore, because there is an interpretation associated with the parameters, the errors do not manifest as obviously as distortion, so it is not straightforward to discover whether an error has occurred.

Ultimately, a parametric approach would be desirable for time-frequency representation of sounds, because of its promise of increased time-frequency resolution. However, present techniques use theoretical signal models that are too restrictive in practice. It would appear that a greater improvement could be attained by first developing a suitable higher order Wigner distribution and then applying parametric techniques.

3.2.5 Model-Informed TFRs

The philosophy of parametric time-frequency representation – applying a model to achieve a more accurate estimate – can also be applied to modify existing TFRs, by using expectations about sound signals.

3.2.5.1 Multiresolution Fourier Transform

It has been established that the STFT can provide good time resolution (GTR) or good frequency resolution (GFR) but not both simultaneously. Using a GFR-STFT for sound representation, partials appear as well separated ridges running horizontally (along the time line). With a GTR-STFT, the partials cannot be distinguished, but vertical lines (parallel to the frequency axis) indicate the note onset points. By intelligent combination of these representations, it should be possible to enhance the resolution in both time and frequency.

One method, termed the Multiresolution Fourier Transform, was developed for improving note onset detection, for automatic score transcription [Pearson & Wilson,1990]. This uses the *min* function, so that energy peaks are represented only where they appear in both STFTs. Thus the note onsets appear well localised in time and frequency. Another closely related method was independently developed in the context of the Cohen class [Loughlin & Pitton & Atlas,1992]. This aimed to achieve 'correct marginals' from the spectrogram (which, as observed earlier, is the only bilinear distribution to attain total cross-term suppression). The proposed method calculates the minimum cross entropy between two resolution scales of the spectrogram.

It should be noted that although the apparent resolution of these techniques is better than the linear uncertainty principle, the gain in resolution is a result of applying expectations about sound signals to TFR development. For a general (non-sound) signal, peaks could be generated in the combined TFR, which result entirely from smearing in the individual GFT-TFR and GTR-TFR that happen to overlap, and therefore have no physical significance of themselves.

In application to sound representation, the results appear good where there are energy boundaries, such as percussive onsets or stable partials. Therefore the results should be good for note onset localisation. Where tracking of gradually evolving phenomena is required, there is no improvement: the broadband peaks in the GTR-TFR do not enable better frequency tracking of the partials. In fact, because the GTR-TFR varies across individual periods of stable waveforms, its minima break up the otherwise continuous ridges of the GFR-TFR, thereby degrading the combined representation. (This is quite visible in the example figure provided by [Loughlin & Pitton & Atlas,1992].)

3.2.5.2 Parametric Modelling of the STFT

In order to detect partials from the STFT, it is necessary to have sufficient frequency resolution, that the peaks are separated. A rule of thumb separation is a minimum of four bins, which corresponds to a window-length of four periods. Once the presence of a partial has been established, accurate location is simpler, because its shape in the frequency domain is largely predictable. This is especially true if the partial trajectory is a good approximation to local stationarity.

A new method which is currently being developed [Depalle & Tromp,1996], aims to detect the presence of partials using a long enough window-length to separate partials. The analysis then continues at a shorter window-length (1½ periods of the waveform), using a least-squares-error method to iteratively improve the match between the estimated partial locations and the actual locations. The method oscillates between improving amplitude and phase location whilst fixing frequency, to improving frequency location whilst fixing amplitude and phase. Initial results suggest that five iterations is sufficient to locate the partials accurately, even in the presence of noise.

3.2.5.3 Extraction of 'Hidden' Information from the FFT

Despite the assumption of stationarity that accompanies the FFT, it is nevertheless an invertible transform. Whether the analysed signal is periodic or not, the inverse FFT (IFFT) of the FFT of a signal is the signal itself:

$$\mathrm{IFFT}\Big\{\mathrm{FFT}\big[s(t)\big]\Big\} = s(t) \tag{3.10}$$

Therefore, the FFT *fully* describes the signal. That is, the FFT not only describes linear phase signals (as the assumption of stationarity implies), but also captures the higher order information. With a change to the assumptions employed when extracting data from the FFT, it is possible to uncover 'hidden' information [Masri & Bateman,1995]. A method for determining second order information is presented in the following section.

3.3 Obtaining Second Order Information from Phase Distortion in the FFT

A stationary sinusoid appears in the FFT spectrum as the spectrum of the window function, translated in frequency and scaled, so that the maximum of the main-lobe coincides with the frequency and amplitude of the sinusoid. (The phase also corresponds to the sinusoid's phase at the centre of the window.) The magnitude and phase shape for a nonstationary sinusoid is more complicated, but is nevertheless totally predictable, due to the unique mapping between the time domain and the frequency domain.

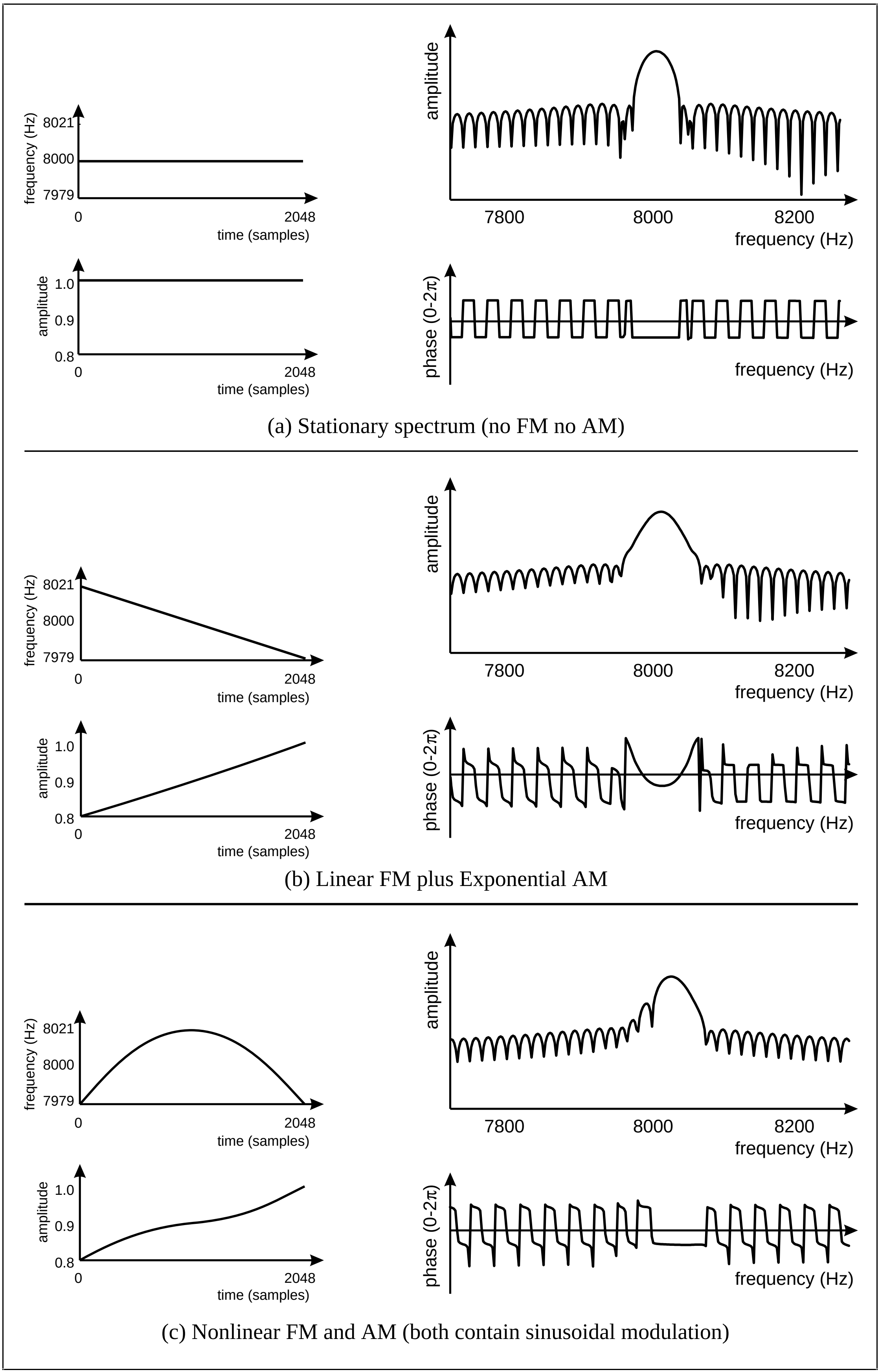
The common approach to partial extraction is to capture the frequency and amplitude information from the maximum of the main-lobe and to disregard the rest of the peak and its side-lobes. In the case of a locally stationary sinusoid, this is justifiable, because the shape of the lobes is totally predictable – it is equal to the shape of the window function and contains no more useful information. In the practical case of sinusoids that *approximate* local stationarity, some distortion to the shape is expected.

However the FFT is an invertible transform and therefore contains all the information about a signal's trajectory – stationary and nonstationary. What has been labelled as 'distortion' is actually a source of information.

It remains true that the FFT provides an *optimal* representation for stationary signals. That is, the sinusoidal components appear *simply* as narrow peaks in the FFT spectrum. The representation of nonstationary signals, although complete as stated, is not as straightforward.

The most obvious effect of introducing higher-order (nonlinear) elements into the phase equation of a sinusoidal component is that its amplitude peak in the FFT spectrum becomes wider. Once the nonlinear elements are increased beyond a certain point the shape of the peak begins to change becoming flatter. The phase spectrum of the FFT across the main-lobe and the side-lobes also changes as nonlinear elements are introduced. See Figure 3.4.

Although phase is not as readily interpreted as amplitude, it has been shown to carry much, if not *most*, of the signal information [Oppenheim & Lim,1981]. It has been implied in papers promoting joint time-frequency distributions, that frequency-only distributions are incapable of representing the timing information in a nonstationary signal and examples have been cited where two radically different signals have the same spectra. However, the spectra that are displayed in these examples are the amplitude spectra only [Boashash,1990]. If the phase spectra had been observed, and its 'message' interpreted, it would have yielded the timing information.



(a) Stationary spectrum (no FM no AM)

(b) Linear FM plus Exponential AM

(c) Nonlinear FM and AM (both contain sinusoidal modulation)

Figure 3.4 – Amplitude and phase 'distortion' caused by nonlinearities

3.3.1 How Measurements were Made

The investigation presented here has been confined to observation of the phase around the maximum of the main-lobe. It will be shown that this is sufficient to extract linear frequency modulation (LFM) and exponential amplitude modulation (EAM) information.

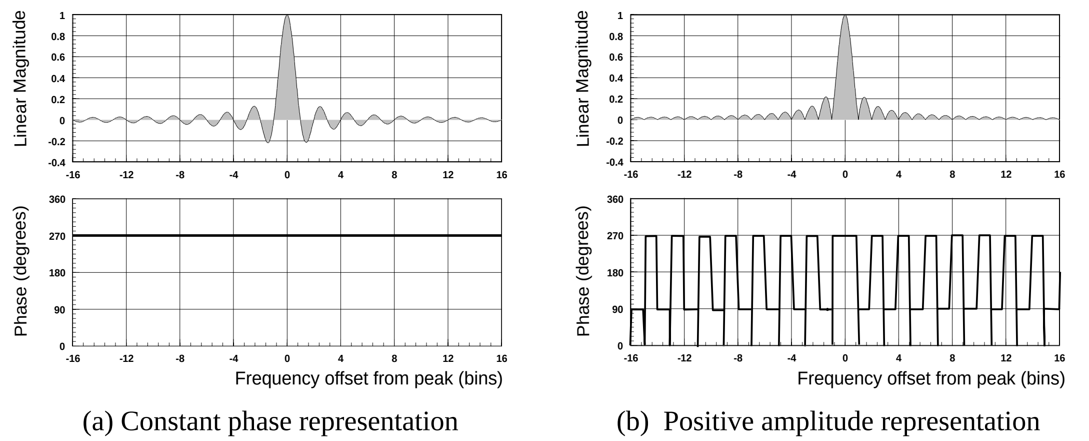
To minimise the additive interference of other sinusoids that are close in frequency in the practical multicomponent scenario, the phase measurements must necessarily be made close to the maximum of the main-lobe where the component of interest has its greatest energy concentration. Hence the zero-padded FFT was used, where the interpolated spectrum could be measured at fractions of a bin. The data for all figures were generated using an FFT zero-padded by a factor of 16.

As discussed in section 2.3.2.1, the window function plays an important part in the shape of the main-lobe and the significance of side-lobes. Experiments on the rectangular, triangular, Hamming and Hanning windows suggest that the *degree* of phase distortion is dependent on the window function, but that the *form* of the distortion is not. Hence the presented technique could be applied to any window function, but the measurements would need recalibration. The majority of results are presented for the Hamming window function, which was used in the author's analysis algorithm.

In all cases, the measurements were found to be invariant of the frequency and amplitude of the modulated (carrier) sinusoid. Also, the modulation is described in absolute terms; i.e. not relative to the modulated signal.

3.3.2 Phase of an Unmodulated Sinusoid

For an unmodulated sinusoid, the phase is constant across the main-lobe and all the side-lobes as shown in Figure 3.5a. However its amplitude oscillates about zero, so for an FFT whose amplitudes are all represented as positive, the phase will appear to be shifted by 180° between alternate zero-crossings (see Figure 3.5b).



(a) Constant phase representation

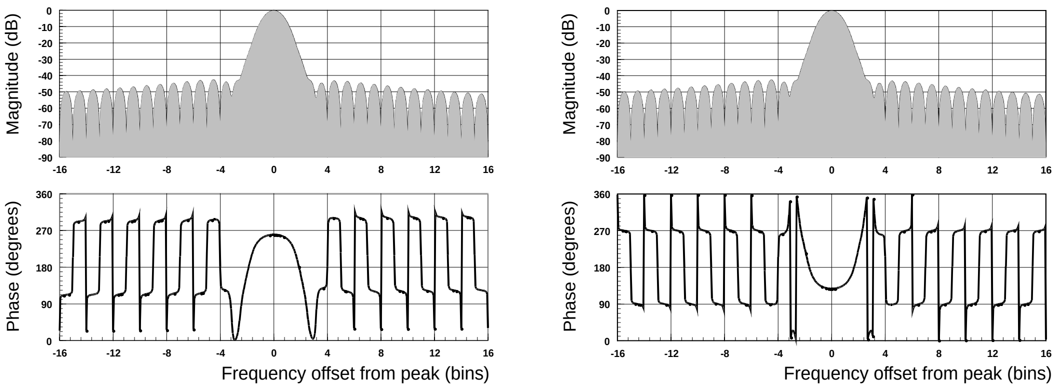
(b) Positive amplitude representation

Figure 3.5 – Spectrum of an unmodulated sinusoid

As modulation is introduced, the phase either side of the maximum is altered. In the following sections, the symbol ΔΦ represents the phase offset with respect to the phase at the maximum, and ΔF represents the frequency offset (measured in bins) from the maximum at which the measurement is taken. The values of modulation, $\frac{df}{dt}$ for LFM (linear frequency modulation) and $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$ for EAM (exponential amplitude modulation), measure a fixed rate expressed as a frequency/amplitude change per frame, where the frame length is the length of the FFT's time domain analysis window.

3.3.3 Linear Frequency Modulation (LFM)

For rising chirps (linearly increasing frequency), Figure 3.6a shows that the phase offset, ΔΦ, either side of the maximum is negative and that the degree of offset increases with distance from the peak. The reverse is true for falling chirps (linearly decreasing frequency), where the shape of the curve is vertically mirrored – ΔΦ is positive but the degree of offset is the same as for rising chirps; see Figure 3.6b. It should also be noted that the curves are symmetrical about the maximum; $\Delta\Phi|_{\Delta\mathrm{F}=k} = \Delta\Phi|_{\Delta\mathrm{F}=-k}$.



(a) Rising frequency = +1 bin per frame

(b) Falling frequency = -1 bin per frame

Figure 3.6 – Spectra of sinusoids with linear frequency modulation

Values of ΔΦ were measured at a number of fixed offsets from the maximum, for varying chirp rates, $\frac{df}{dt}$, and the results are displayed in Figure 3.7. From these curves it can be noted that there is strong similarity between the curves, indicating that measurements taken anywhere within the main-lobe could be used in a practical implementation. Also it should be noted that the curves represent a non-unique mapping between $\frac{df}{dt}$ and ΔΦ. This restricts the useful range over which $\frac{df}{dt}$ could be estimated from ΔΦ.

The amplitude spectrum shows a widening of the main-lobe which increases with the magnitude of $\frac{df}{dt}$, and is accompanied by a skew that is dependent on the sign of $\frac{df}{dt}$.

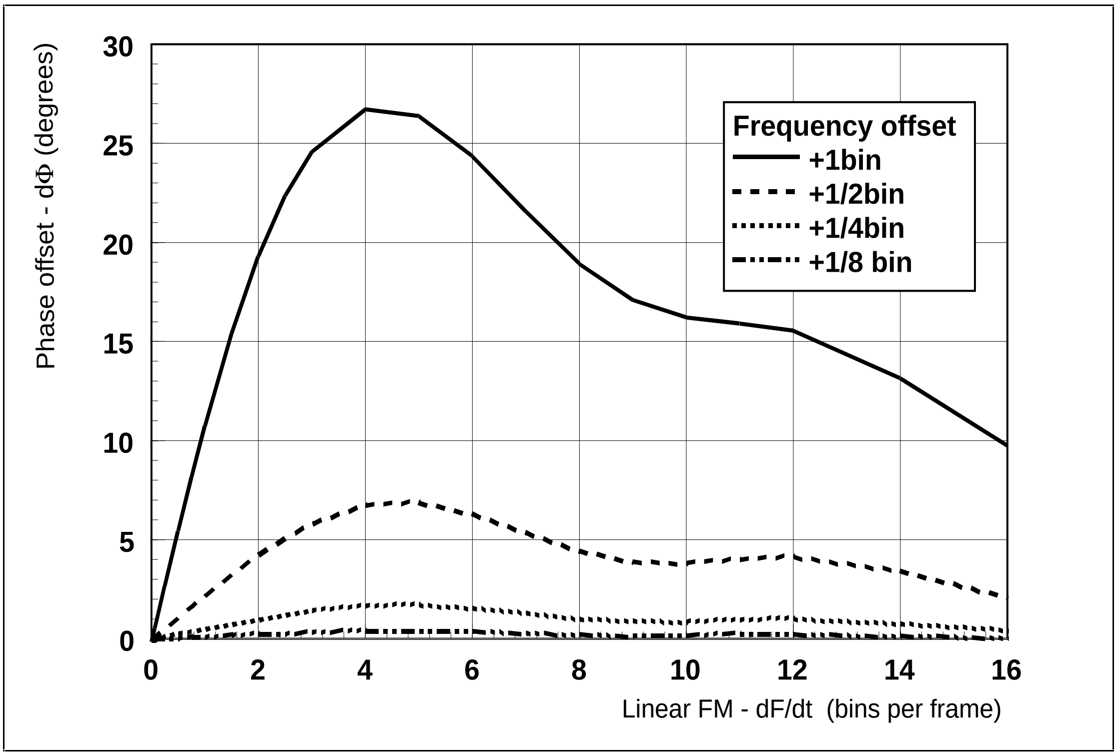
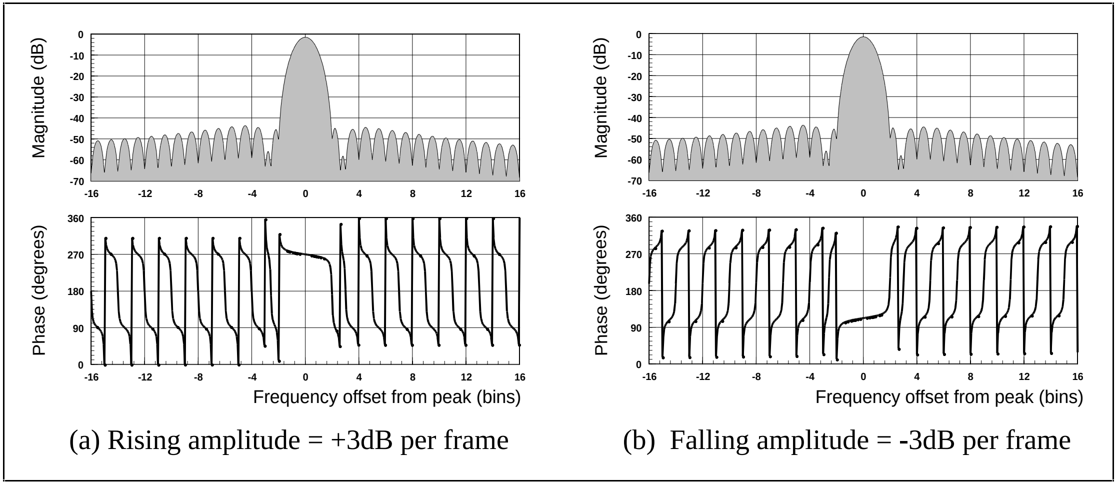


Figure 3.7 – Linear FM phase distortion at various frequency offsets from the maximum (Hamming window)

3.3.4 Exponential Amplitude Modulation (EAM)

A constant relationship was discovered between ΔΦ and exponential amplitude modulation, not linear amplitude modulation. That is there is a fixed law relating ΔΦ and $\frac{d(\log A)}{dt}$ expressed in dB per frame.

Whereas the phase offsets for LFM were even symmetrical about the maximum, in the case of EAM, there is odd symmetry (i.e. the sign is opposite on either side of the maximum). See Figure 3.8. For exponentially increasing amplitude, the phase at a positive frequency offset from the maximum is negative, whilst at a negative frequency offset, it is positive (Figure



(a) Rising amplitude = +3dB per frame

(b) Falling amplitude = -3dB per frame

Figure 3.8 – Spectra of sinusoids with exponential amplitude modulation

3.8a).

Values of $\Delta\Phi$ were measured at a number of (positive) fixed offsets from the maximum, for varying modulation rates, $\frac{d(\log A)}{dt}$, and the results are displayed in Figure 3.9. Unlike LFM, there is a unique mapping between $\Delta\Phi$ and $\frac{d(\log A)}{dt}$. Furthermore, it appears to be a linear relationship. Experimental evidence was so strong that the remainder of this discussion assumes that the relationship is truly linear. All the curves of Figure 3.9 bear this out, once again indicating that a practical implementation could use measurements at any offset within the main-lobe.

As for LFM, the amplitude spectrum shows a widening of the main-lobe which increases with the magnitude of $\frac{d(\log A)}{dt}$, but it is independent of the sign of $\frac{d(\log A)}{dt}$.

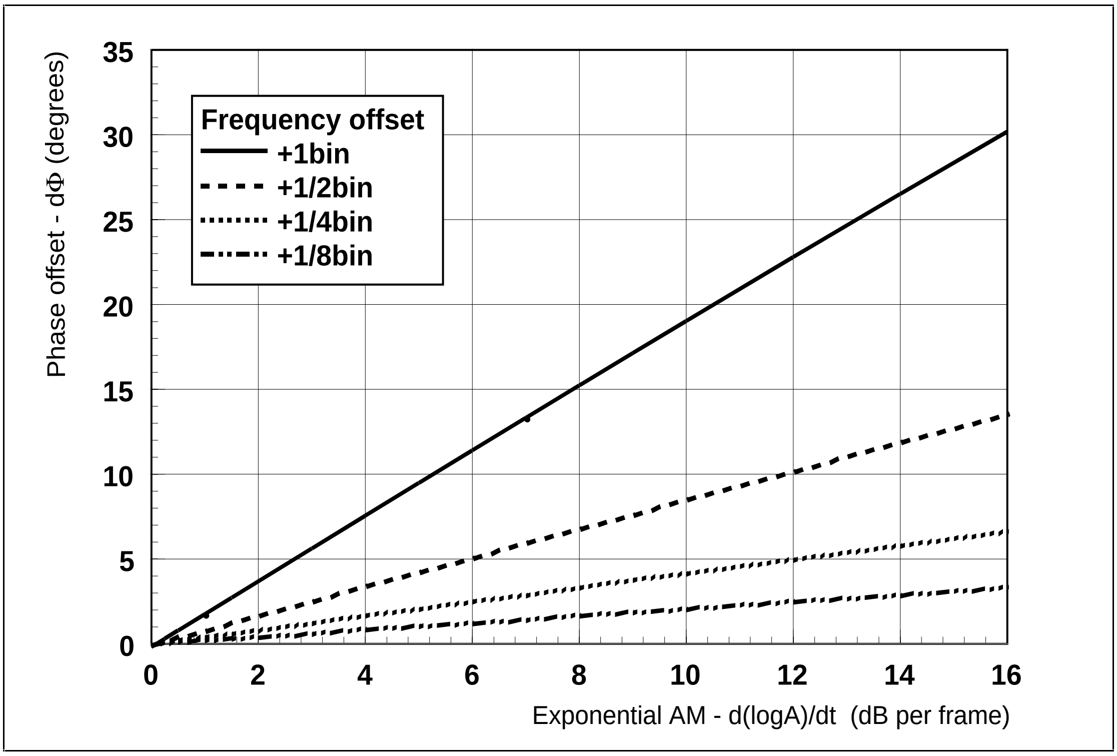


Figure 3.9 – Exponential AM phase distortion at various frequency offsets from the maximum (Hamming window)

3.3.5 Concurrent LFM and EAM

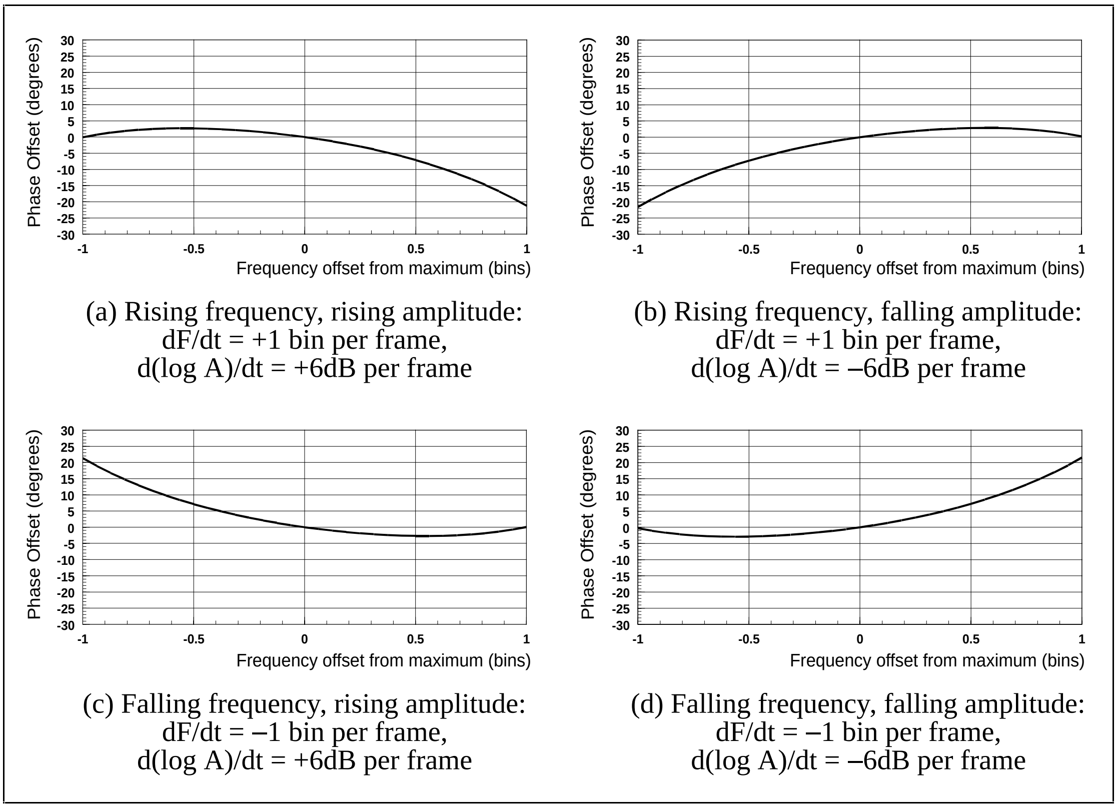
Perhaps surprisingly, the phase offsets due to LFM and EAM are additive. At any frequency offset from the maximum (tested in the range –1 to +1 bin for the Hamming window), the total phase offset is the sum of the offsets due to each type of modulation:

$$\Delta\Phi|_{Total} = \Delta\Phi|_{LFM} + \Delta\Phi|_{EAM} \qquad (3.11)$$

The four graphs of Figure 3.10 (overleaf) display combinations of rising and falling LFM and EAM.

Because of the symmetries about the maximum – LFM with the same sign and EAM with opposite signs – it is possible to identify the values of $\frac{df}{dt}$ and $\frac{d(\log A)}{dt}$ from two measurements of

ΔΦ, one taken on either side of the maximum. For example, if two measurements are taken at equal distances either side of the maximum, then the phase offsets due frequency and amplitude modulation are respectively, the sum÷2 and the difference÷2 of the measured offsets.



(a) Rising frequency, rising amplitude:
dF/dt = +1 bin per frame,
d(log A)/dt = +6dB per frame

(b) Rising frequency, falling amplitude:
dF/dt = +1 bin per frame,
d(log A)/dt = –6dB per frame

(c) Falling frequency, rising amplitude:
dF/dt = –1 bin per frame,
d(log A)/dt = +6dB per frame

(d) Falling frequency, falling amplitude:
dF/dt = –1 bin per frame,
d(log A)/dt = –6dB per frame

Figure 3.10 – Phase distortion across main-lobe for combinations of LFM and EAM

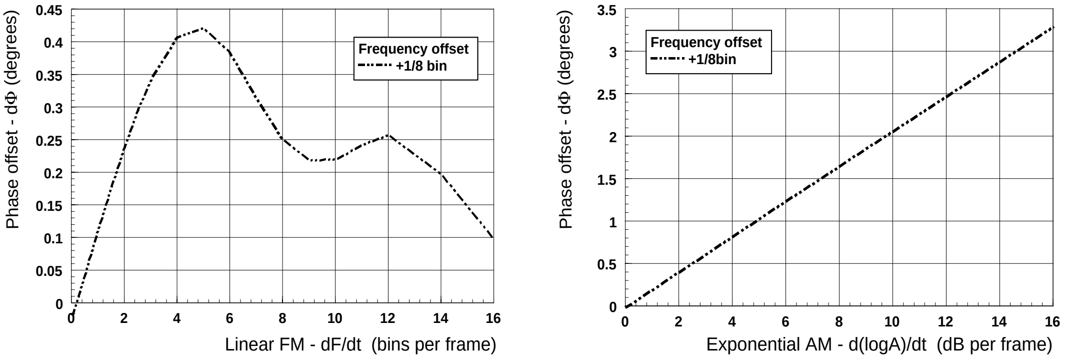
3.4 Performance of the Technique

The technique of Phase Distortion Analysis (PDA) has been described and a method has been presented for implementing it as a trajectory estimator. In this section the technique is put to use and its performance assessed. Section 3.4.1 tests the effectiveness of the estimator in practical situations and section 3.4.2 looks at how the estimator can be usefully employed within the sound analysis-resynthesis system.

3.4.1 Practical Application of LFM and EAM Analysis

3.4.1.1 Simulated Data

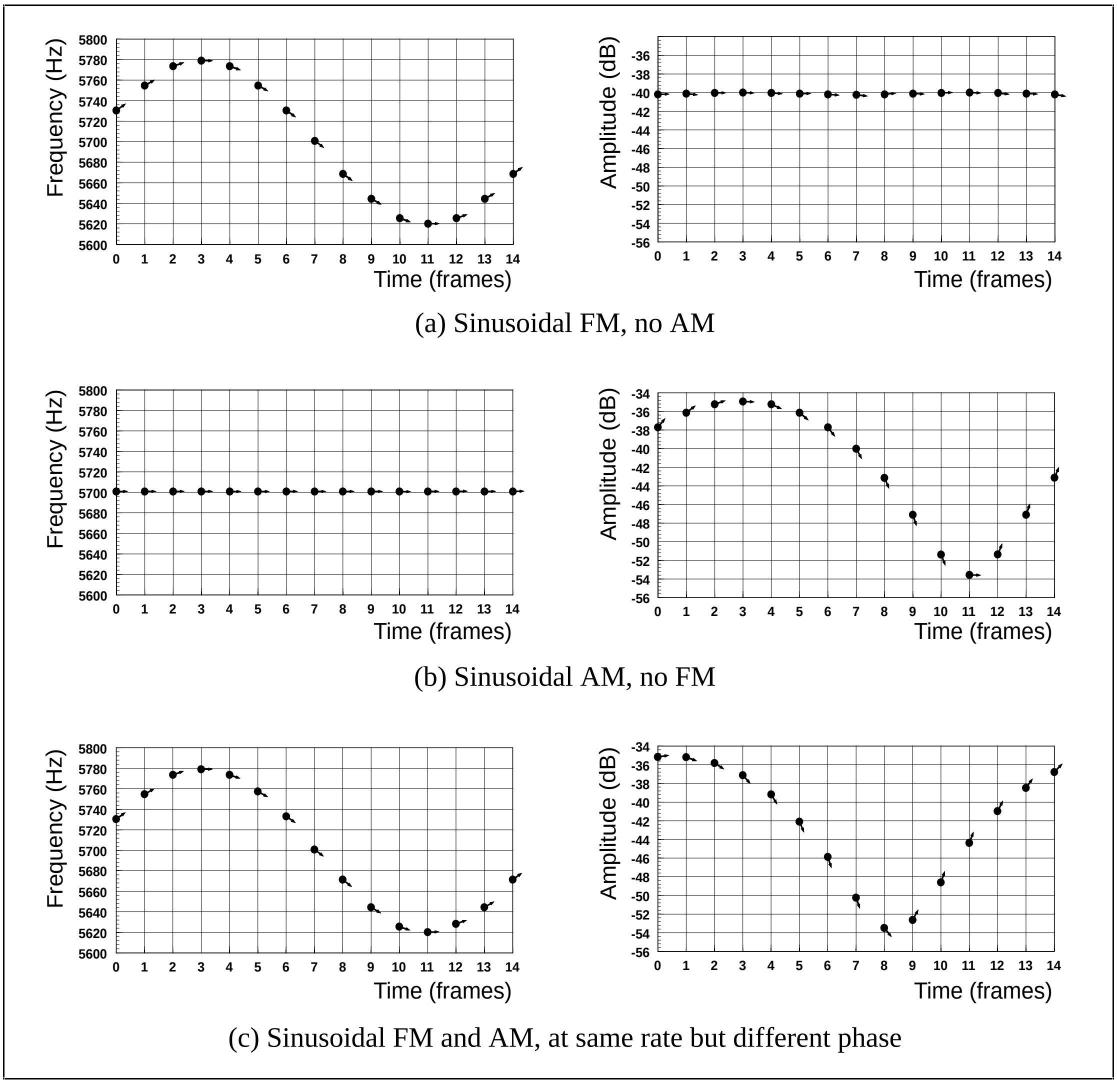
The effects of LFM and EAM have so far been investigated in isolation from possible sources of interference. This section shows that the method is robust when higher order components are present, *if* the frequency/amplitude trajectories are a close approximation to LFM and EAM within the span of a frame. This is similar to the previous approximation of stationarity, but allows for an extra order of freedom. The following section (3.4.1.2) applies the method to real musical signals where there are multiple components and the inherent risk of corrupted estimates due to additive interference. As stated earlier, the need to minimise these latter effects requires that measurements of $\Delta\Phi$ be taken close to the maxima. Hence the measurements of this and the following section were made at $\frac{1}{8}$th bin from the maxima (see Figure 3.11, based on Figure 3.7 and Figure 3.9), which requires eight-fold zero-padding.



(a) Phase distortion for Linear FM (b) Phase distortion for Exponential AM

Figure 3.11 – Phase distortion mapping curves at '+1/8 bin' frequency offset (close-up of Figure 3.7 and Figure 3.9)

Figure 3.12 shows three examples of simulated audio signals. The points indicate the instantaneous estimates of frequency and amplitude (measured from the maxima), and the arrows from them indicate the instantaneous estimates of the modulation (measured from the phase offsets). The arrows are not strictly pointers showing the expected destination of the trajectory, but estimates of the instantaneous rate of modulation, the gradient of the graph.



(a) Sinusoidal FM, no AM

(b) Sinusoidal AM, no FM

(c) Sinusoidal FM and AM, at same rate but different phase

Figure 3.12 – Trajectory predictions for simulated data

The examples display sinusoidal FM and AM, where the FFT window is short enough to isolate approximate line segments of the frequency/amplitude curve. Consequently, the arrows approximate tangents to the curves. Figure 3.12a is the analysis of sinusoidal FM (with parameters comparable to vibrato of a musical instrument), where the amplitude is constant. Figure 3.12b is the analysis of sinusoidal AM (comparable to realistic tremolo), where the frequency is constant.

Figure 3.12c shows a combination of FM and AM. The rate of modulation of each is the same, but the phase has been offset by 60° to demonstrate that the technique is not reliant on correlation between frequency and amplitude. Note that the amplitude modulation does not *appear* to be sinusoidal because a logarithmic (dB) scale is used.

3.4.1.2 Real Musical Signals

The two graphs of Figure 3.13 show the technique applied to harmonics of real audio: a cello note with a large amount of vibrato. Figure 3.13a tracks the 1st harmonic centred about 1050Hz, where the frequency modulation is slight, and Figure 3.13b tracks the 13th harmonic centred about 7350Hz, where the modulation is more pronounced.

These graphs show the results for partials that were quite pronounced, so the additive interference from neighbouring partials was slight. In the case of quieter partials, the interference is stronger and can render the modulation estimations unworkable. The following section looks at the impact this has upon immediate practical implementation and section 6.1.2 discusses potential solutions as options for future work.

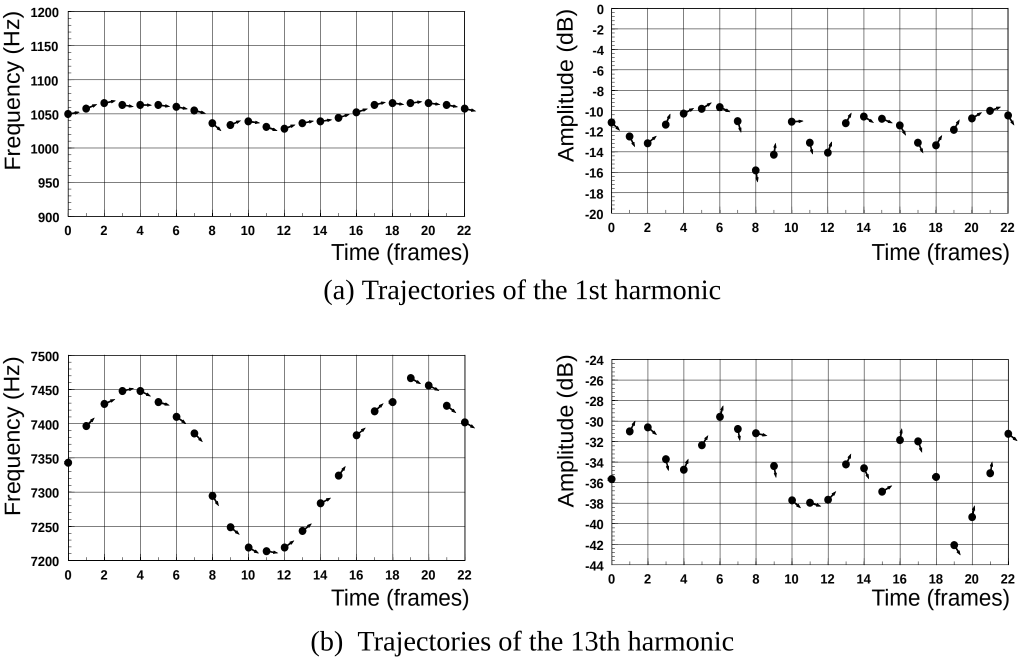
3.4.2 Implementation Within the Sound Model

The ability to extract second order information from the FFT spectrum relaxes the requirement that partials approximate stationarity within each analysis window, to the requirement that their trajectories approximate LFM+EAM (which includes stationarity as one valid case). The technique effectively increases the overall time-frequency resolution of the FFT – it enables greater temporal resolution, since trajectory variations can be described in greater detail, without a reduction in frequency resolution. This is much needed, since the time resolution is already compromised in the Initial Model, in favour of good frequency resolution for the separation of partial peaks.

Improved Peak Validation

In the deterministic analysis, peaks in the FFT spectrum are interpreted as partials. Yet peaks can also occur from the side-lobes (of a partial peak) or from the momentary energy localisation of a stochastic source. A cursory observation of side-lobe phase (see Figure 3.6 and Figure 3.8) indicates that side-lobes are easily distinguishable from partial peaks, because the phase change across a side-lobe peak is much more dramatic. By the same token, stochastic energy localisations are an unstable phenomenon of short duration, suggesting that their phase profiles would also be largely different from partial peaks.

PDA was implemented in the in-frame peak validation stage, as part of the peak detection



(a) Trajectories of the 1st harmonic

(b) Trajectories of the 13th harmonic

Figure 3.13 – Trajectory predictions for a real data example (cello with vibrato)

process – as the location of peaks is refined by interpolation, the frequency and amplitude trajectories are also estimated. As a validation tool, phase distortion values that suggest impossible values (i.e. phase offsets greater than the curve in Figure 3.11a) or $\frac{df}{dt}$ / $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$ values that are above an arbitrary threshold cause rejection of a peak. This was found to be very effective at removing side-lobes.

Unfortunately the method proved unreliable, because it rejected valid peaks whose estimates had been corrupted by additive interference. This occurred during the less stable portions of the waveform where it affected the less prominent peaks. As mentioned earlier, solutions to this are discussed in chapter six.

Improved Frame Linking

Chapter four discusses a frame linking strategy based on the harmonic structure to replace the nearest frequency scheme of the Initial Model. However both methods would benefit from in-frame estimates of the frequency trajectory. The nearest frequency scheme could try to link a peak in one frame to the peak in the next that lies closest in frequency to its trajectory. In this way the *maximum frequency deviation* threshold could be reduced for higher partials, since large shifts between frames would be incorporated as a result of following the estimated trajectories.

In the harmonic structure case (where peaks are validated by whether they belong to the harmonic structure) the frequency trajectory estimates could aid in confirming which peaks are partials, especially for the higher partials where slight errors in the fundamental frequency estimate become significant.

Improved Deterministic Synthesis

The implication for sound synthesis is increased detail. At present, each frame holds the frequency and amplitude of each partial, and these values are linearly interpolated over a frame's duration. With the addition of frequency and amplitude gradients at each frame, the values could by interpolated with a cubic function. If this enables the deterministic synthesis to track the partials of the source sound more accurately, then less error will be present in the residual (see Figure 2.8 in section 2.3.4.1). Therefore the technique should yield an improvement in sound quality. (This depends on whether the extra computation is justifiable.)

In summary…

Section 3.3 has shown that Phase Distortion Analysis can reveal second order information about nonstationary sinusoids. It is robust if the frequency and amplitude laws are a close approximation to Linear Frequency Modulation and Exponential Amplitude Modulation, and therefore relaxes the stipulation that frames of the STFT include stationary components only. In the practical situation, where additive interference from the more prominent peaks affect the phase profile across less prominent neighbouring peaks, the present implementation has not been resilient enough (for the less stable portions of the sound).

These results suggest that the principle of PDA is valid and could possibly be extended to higher orders for encapsulation of more complex frequency/amplitude trajectories. They also suggest that the implementation, based on the minimum two phase offset measurements, is not sophisticated enough to cater for additive interference. Chapter six (section 6.1.2) examines this further and proposes some solutions.

CHAPTER FOUR

FRAME LINKING BY HARMONIC STRUCTURE

4. FRAME LINKING BY HARMONIC STRUCTURE

In the deterministic aspect of the model, the time-frames of the time-frequency representation are initially analysed separately. In order to preserve the sense of continuity upon synthesis, the next step is to link the frames together. The deterministic features in each frame are the peaks. By linking the peaks between frames, the long term paths of the partial trajectories are identified. The linking process is also important within the model (prior to synthesis), because storage of partials is more musically meaningful than storage of peaks, and therefore provides a basis for the more profound transformations. A by-product, within the analysis process, is that by virtue of the frame linker's potential to spot trends over time, it can aid in peak validation (the identification of which spectral maxima correspond to partials).

As described in chapter 2 (section 2.3.2.5), the Deterministic Plus Stochastic model has traditionally employed a simple 'nearest frequency' method, based on linking peaks between neighbouring frames that are closest in frequency. The method is simple and intuitively pleasing, and seems to reflect the expectation of stable sinusoidal elements with slowly changing frequencies.

Unfortunately the reality is that the spectrum can include spurious or volatile components and that the stable elements themselves are not always slow changing. As will be shown, the nearest frequency method is adversely affected by these circumstances, making it unable to track the partial trajectories.

It is the premise of this chapter that a more sophisticated linking procedure is essential, and one which matches the physical observation of a harmonic structure (in pitched sounds). This is discussed in section 4.1, with justifications on three counts. To this end, a method of 'harmonic linking' was developed, which is presented in section 4.2 and assessed in section 4.3.

4.1 Nearest Frequency versus Harmonic Structure

The following three subsections compare the capabilities of the two approaches with one another, in the light of the circumstances in which they operate. These circumstances include (respectively): properties of the sound signal, limitations of the time-frequency representation and the purpose for creating a music analysis-resynthesis model.

Before delving into the discussions, the following provides an outline of how the 'harmonic structure' method works, as compared to the 'nearest frequency' method:-

The basis of the nearest frequency method is simply:

- For each peak in each frame, attempt to link to a peak in the next frame that is closest in frequency (within a certain range); see Figure 4.1.

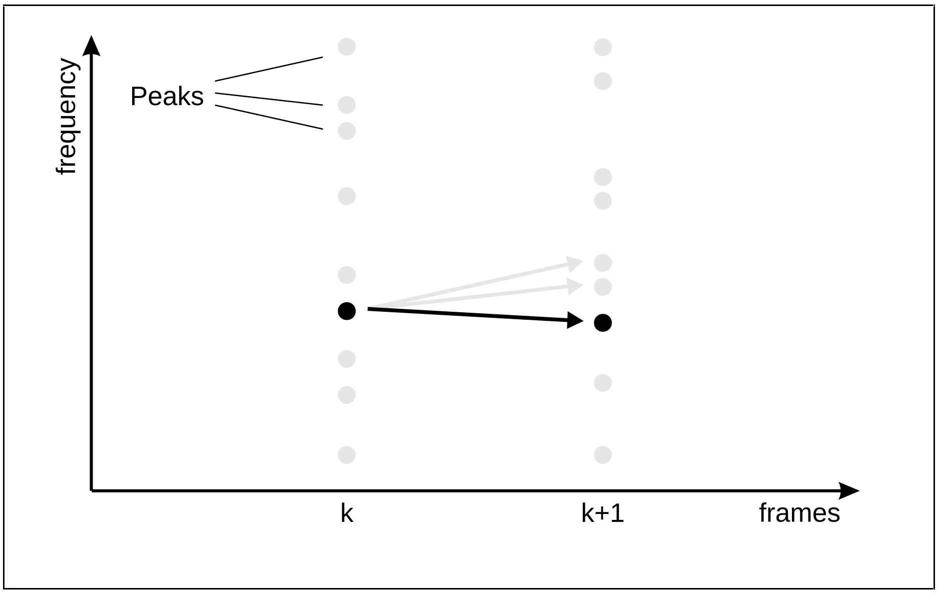


Figure 4.1 – Linking two frames by nearest frequency

The basis of the harmonic structure method is:

- Identify which peaks correspond to which partials in the harmonic structure, within each frame; see Figure 4.2a;
- Link the fundamental frequency only between frames on a nearest frequency basis (or by another simple method). This automatically links the harmonic structures between frames and therefore the peaks corresponding to harmonics are also automatically linked; e.g. the peak in one frame tagged 'Partial #4' is linked to the peak in the next frame tagged 'Partial #4'; see Figure 4.2b.

The 'nearest frequency' method is therefore indiscriminate between peaks, attempting to make links where it can. The 'harmonic structure' method conversely, attaches a meaning to each peak, by identifying its place (or not) within the harmonic structure, so that links are made to the structure as a whole.

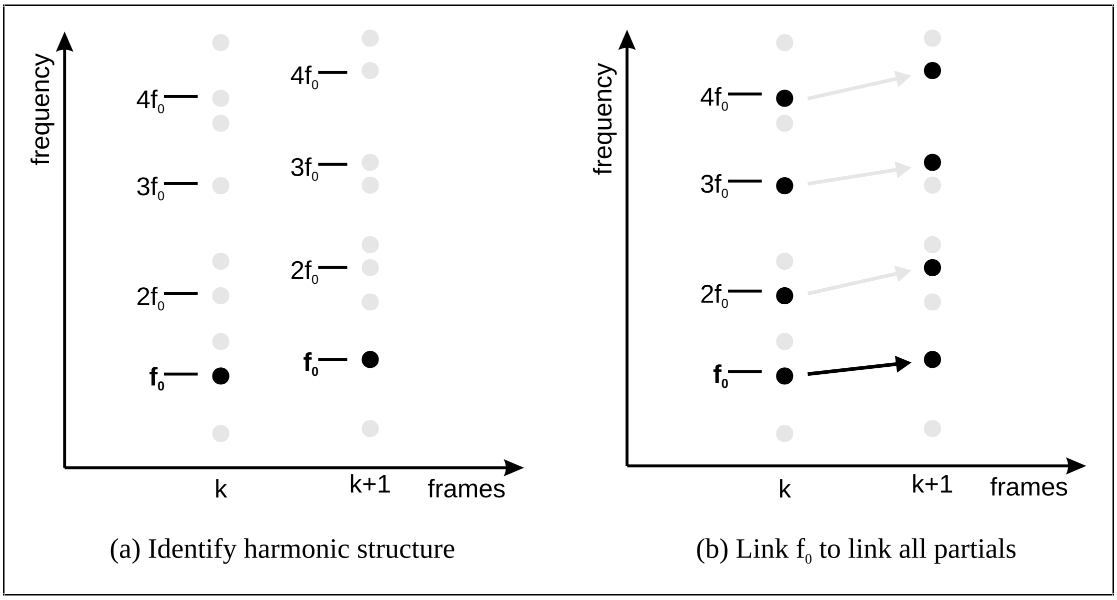

frequency
4f0
3f0
2f0
f0
k
k+1
frames
(a) Identify harmonic structure
(b) Link f0 to link all partials

Figure 4.2 – Linking two frames by harmonic structure

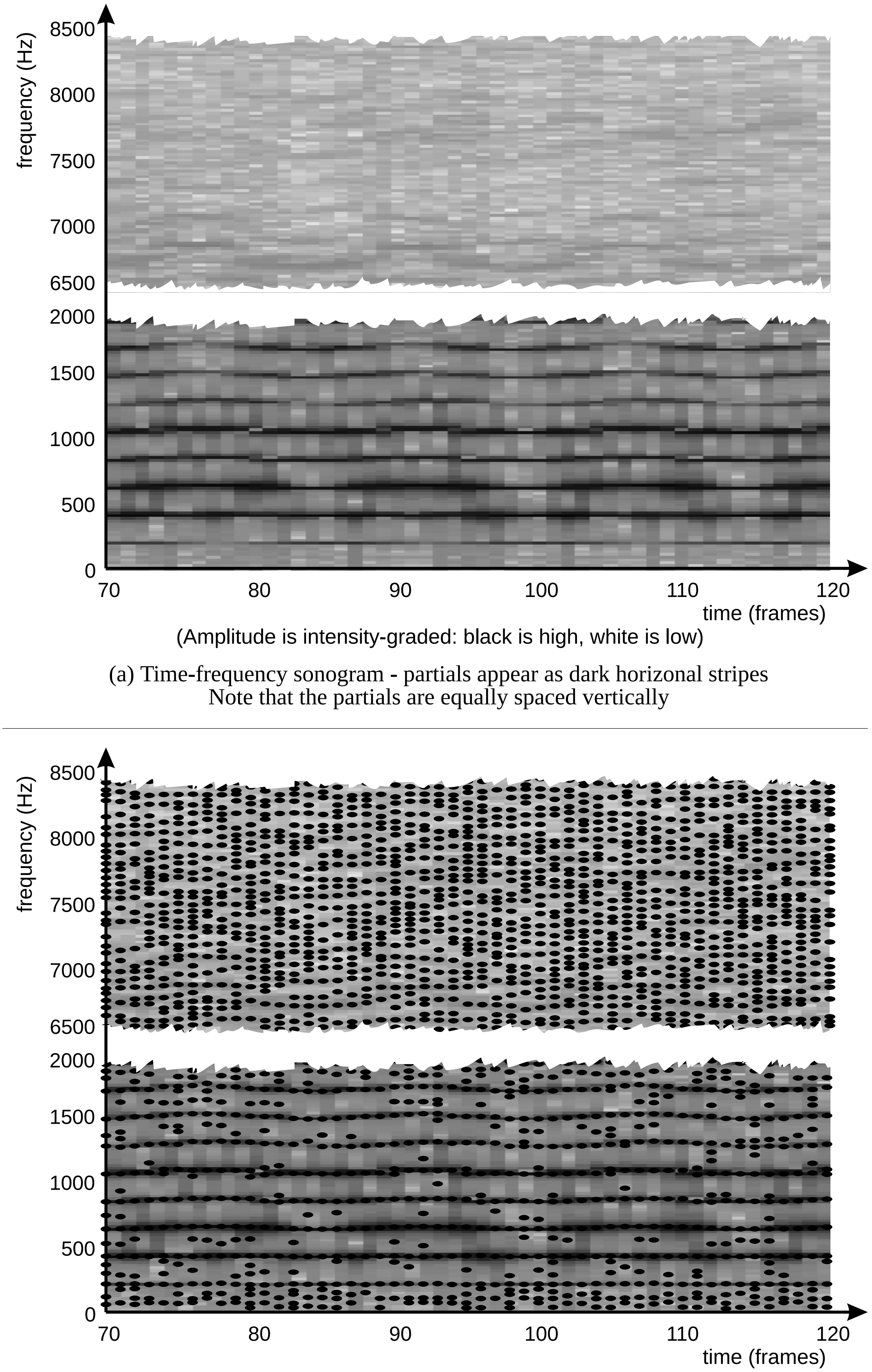
4.1.1 Higher Partials Vary More Rapidly

The first observed property of sound that makes it musical is that pitched sounds possess a harmonic structure (see section 2.1.2). The constraint of partials to a harmonic structure means that the frequencies of the partials are at a fixed integer multiple of the fundamental frequency (or in practice they remain close to this ideal). Therefore, any changes in the fundamental frequency are magnified for the harmonics, proportional to their partial index. See Figure 4.3a-b overleaf.

It is rare for the fundamental frequency to be stationary. Even during the steady-state portion of a sound from a traditional instrument, there is often some vibrato, and even when the pitch appears steady, there is often some slight variation. Since the fundamental frequency is almost always in flux, by implication the higher harmonics are almost always varying quite dramatically.

The proportionately greater variations of the higher harmonics were recognised by Serra – in his nearest frequency peak linking strategy, there is a parameter for 'maximum frequency deviation (between frames)' that is proportional to frequency. Unfortunately, without further constraint, this increases the span of potential links at higher frequencies and consequently also increases the potential for erroneous links. The result upon synthesis is the familiar "ssrsrsr" sound – the high frequency artifacts – that can accompany deterministic synthesis. See Figure 4.3c.

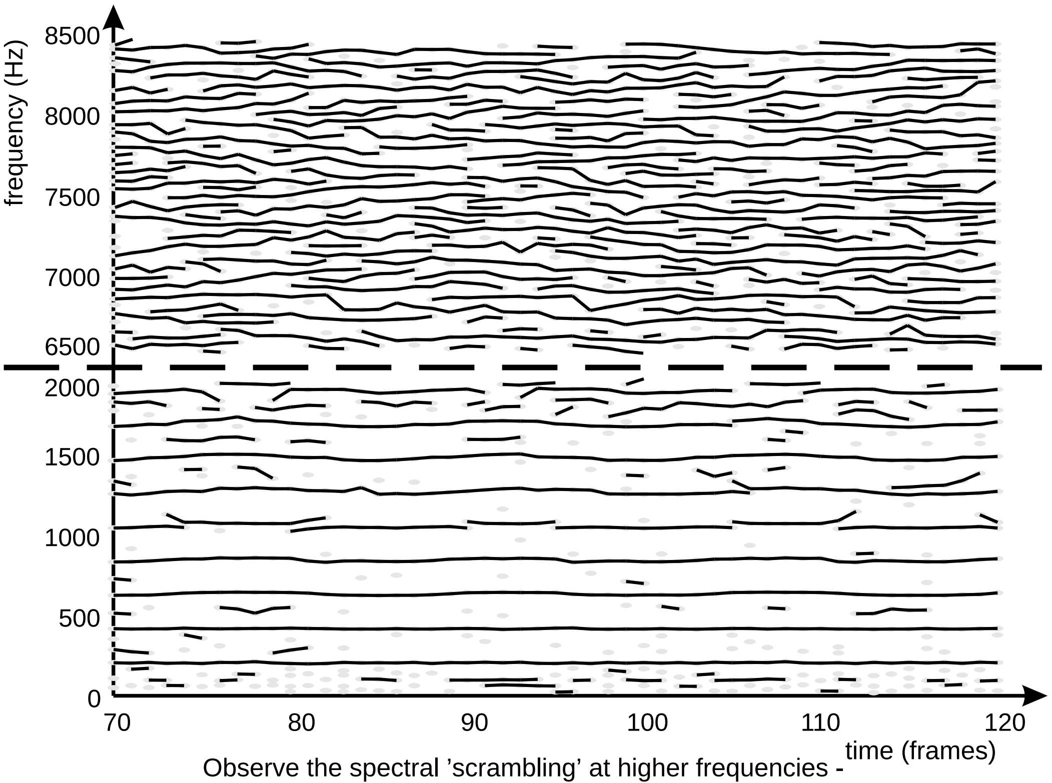
A solution can be found with the harmonic linking approach, that attaches the partial peaks to the harmonic structure (Figure 4.3d) and then links the structure between frames via the fundamental frequency (Figure 4.3e). The fundamental frequencies are linked simply, because they exhibit relatively little frame-to-frame variation. The constraint of the partial peaks to link to peaks tagged with the same partial index enables the necessary large deviations in frequency, but in a controlled manner that is accurate to the source sound.



(a) Time-frequency sonogram - partials appear as dark horizonal stripes
Note that the partials are equally spaced vertically

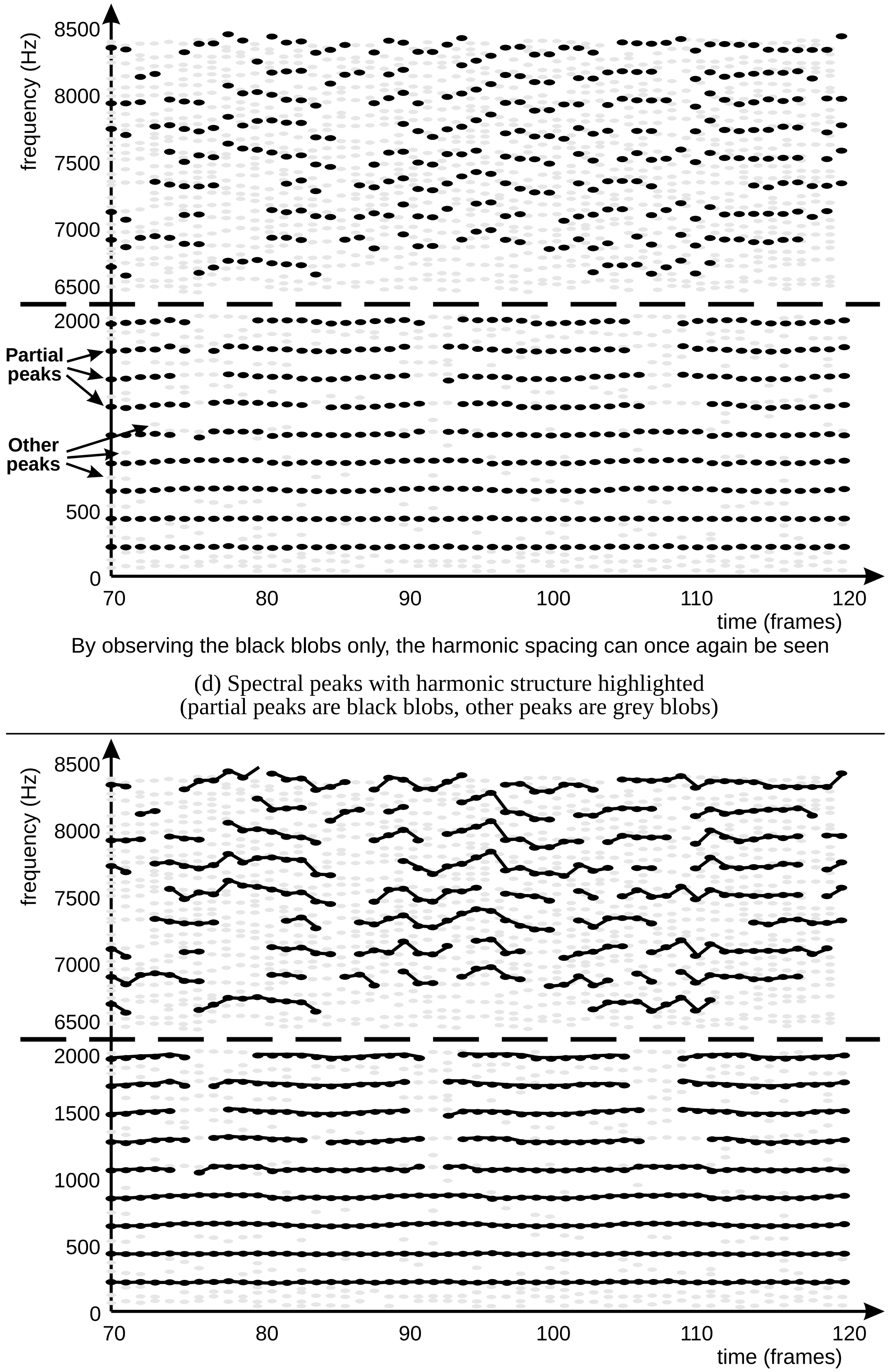
(b) Spectral peaks (as black blobs) overlaid on the sonogram

Figure 4.3 – Analysis of a cello note (with vibrato)



(c) Peak links by Nearest Frequency method
(links are black lines, peaks are grey blobs)

Figure 4.3 – Analysis of a cello note (with vibrato)



By observing the black blobs only, the harmonic spacing can once again be seen

(d) Spectral peaks with harmonic structure highlighted
(partial peaks are black blobs, other peaks are grey blobs)

The partial trajectories are equally-spaced throughout the spectrum

(e) Peak links by Harmonic Structure method
(links are black lines connecting partial peaks; the other peaks are discarded)

Figure 4.3 – Analysis of a cello note (with vibrato)

4.1.2 The Smoothed Representation of the STFT

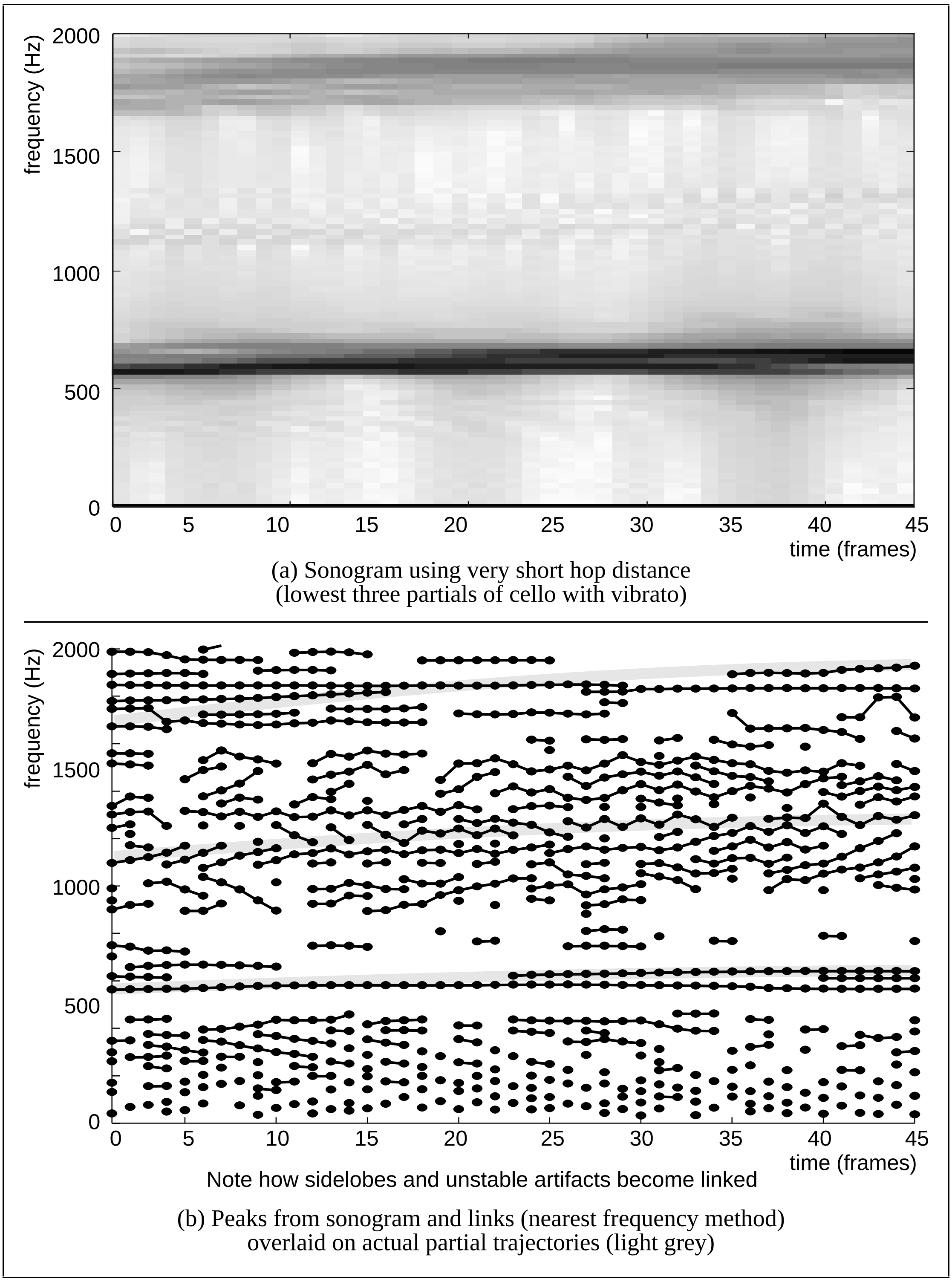
It is well established that the STFT yields magnitude spectra that are *smoothed* representations of the time-frequency map [Boashash,1990; Jones & Parks,1992]. Therefore small scale variations along a partial trajectory will be averaged within each FFT. Such small violations of the assumption of stationarity within each frame, cause only a slight widening of the peaks in the spectrum, for which the analysis process is tolerant. It is as though the STFT applies a low-pass filter to the partial trajectories, erasing the rapid small-scale variations and preserving the less rapidly varying trajectories. In such an apparently ideal situation, the partials have become easier to track and the simple, nearest frequency scheme ought to suffice.

The nearest frequency scheme was introduced by McAulay & Quatieri [McAulay & Quatieri,1986] when they created the Sinusoidal speech model, and adopted by Serra [Serra,1989] for his Deterministic Plus Stochastic musical sound model. The method has been widely accepted and the basic algorithm (described at the start of section 4.1) has also been refined. Depalle, Garcia & Rodet [Depalle & Garcia & Rodet,1993] provide a mathematically elegant algorithm, that uses a Hidden Markov Model (HMM) to track partial trajectories. This latter example also extends the technique to make use of amplitude and gives preference to a continuation of frequency and amplitude momenta (i.e. the best match is sought on the basis of frequency, f, amplitude, a, $\frac{df}{dt}$ and $\frac{da}{dt}$, instead of frequency alone).

All of these variations on the nearest frequency method rely to some extent on the low-pass filtering effect of the FFT on partial trajectories to provide smooth trajectories which are easy to track. Unfortunately the distortion of peaks in the FFT can be more severe than a mere widening effect, and at this point the low-pass filter analogy loses meaning. Fast changing elements cause problems, because they cause a large-scale violation to the stationarity assumption, so their peaks become significantly distorted in shape and can include multiple maxima. Other short term effects, including localised noise, can complicate the picture further by introducing apparent peaks of their own, which are not constrained to any particular expected peak shape or width. In this scenario, the nearest frequency method and its variants are prone to error.

Given the rapidity of change of the fast changing elements, it might seem logical to minimise the frame-to-frame difference by reducing the hop-distance – the distance that the analysis window advances between frames in the time domain. However the problem lies not with the separation between frames, but the length of the frame’s window. A shorter hop-distance minimises the difference between the spectra of each frame, but the contents of each frame are influenced by the duration of the analysis window. In other words, shortening the hop-distance does not reduce the distortion, but it does make the distortion look similar between frames, since their data content is more similar. Below a certain point, a reduction in the hop-distance will only serve to make spurious peaks *look* as if they are stable and connectable between frames. Figure 4.4 shows an example of this.

In these difficult circumstances the nearest frequency method can generate erroneous trajectories that track the side-lobes or spurious noise peaks, or link a true partial peak in one frame to a spurious peak in the next and so on. Upon synthesis, the sound becomes partially scrambled.



(a) Sonogram using very short hop distance
(lowest three partials of cello with vibrato)

Note how sidelobes and unstable artifacts become linked

(b) Peaks from sonogram and links (nearest frequency method)
overlaid on actual partial trajectories (light grey)

Figure 4.4 – Distortion artifacts become linked if the hop-distance is too short

A peak linking method based on the harmonic structure would be no more capable of distinguishing which maximum of a peak with multiple maxima is the correct one. However the general shape of the spectrum should still make it possible to detect the harmonic structure and identify the fundamental frequency. In the process of matching peaks to the harmonic structure, one and only one maximum from a multiple maxima peak would be selected as the correct one. There would be no guarantee that the selected maximum is the ‘correct’ one, but its frequency would be guaranteed to be in the vicinity of the true partial (because of proximity to the ideal harmonic structure). Therefore the general trend of the partials would be preserved, even if the individual variations were compromised. This method also avoids the problem of extra trajectories associated with the other maxima of the same peak, since only one maximum is selected per partial. The result should be that audible artifacts are reduced and the trajectories of the synthesised partials, although simpler than the partials of the analysed sound, still follow roughly the same paths.

In addition, there is greater certainty that peaks corresponding to partials are selected, so the peak validation stage of deleting short trajectories is not necessary. Some partials may only exist for two or three frames, but they are included in the synthesis because they are more likely to contribute, than detract, from resynthesis sound quality.

4.1.3 A Specialist Musical Model

The Deterministic Plus Stochastic model is a model of *musical sounds* – the deterministic aspect aims to describe the trajectories of partials, and 'partial' is a musical construct. With this stated aim, there is every justification to move from a generalised spectral analyser toward a more specialised musical signal analyser.

The deterministic analyser already goes part of the way (in the initial model) by excluding short trajectories, on the basis that they are unstable and should be classed as stochastic. With the proposed shift toward a method based on the harmonic structure, a constraint is inevitably imposed on the types of signals that the model can handle, as the trade-off for improved linking. Even so, the compromise is small – with the almost exclusive exception of the sine wave, all periodic waveforms would be captured. (Sine waves, in particular, would not be catered for, because they are solitary spectral components, independent of any structure. However they rarely feature in musical sounds, and almost never naturally; i.e. only from synthetic sources.)

From this viewpoint, the harmonic structure approach better fulfils the specialist musical model philosophy, than the nearest frequency strategy. Put another way, what the harmonic scheme gains in its ability to track partials is more important, musically, than what it loses in its ability to model all waveforms.

4.2 Methodology and Implementation

The short term aim in this investigation is to design an implementation that is robust, flexible and automatic. Initially, single source sounds are the target, although this is seen as a stepping stone not the goal, with the intention that the method should be extendible toward polyphony – the longer term aim.

This section describes the development of a frame linking strategy solely based on the harmonic structure. As with any such method the following three elements must be included:

- Ability to detect the presence of a harmonic structure;
- Method for estimating the fundamental frequency;
- Method for matching peaks to the harmonic structure.

The first element acts as the on-off switch to the linking process, on a frame by frame basis. The second locates the structure, when it is present. The third attaches the peaks to the structure, thereby effecting their frame-to-frame links. The third element is also important as part of the peak validation process, because it involves selecting and rejecting peaks for the identified harmonic structure.

Although these elements can be viewed conceptually as separate entities, their implementation is often closely intertwined. Therefore the following sections, which describe the steps of the implementation, are organised chronologically, instead of conceptually.

4.2.1 Overview

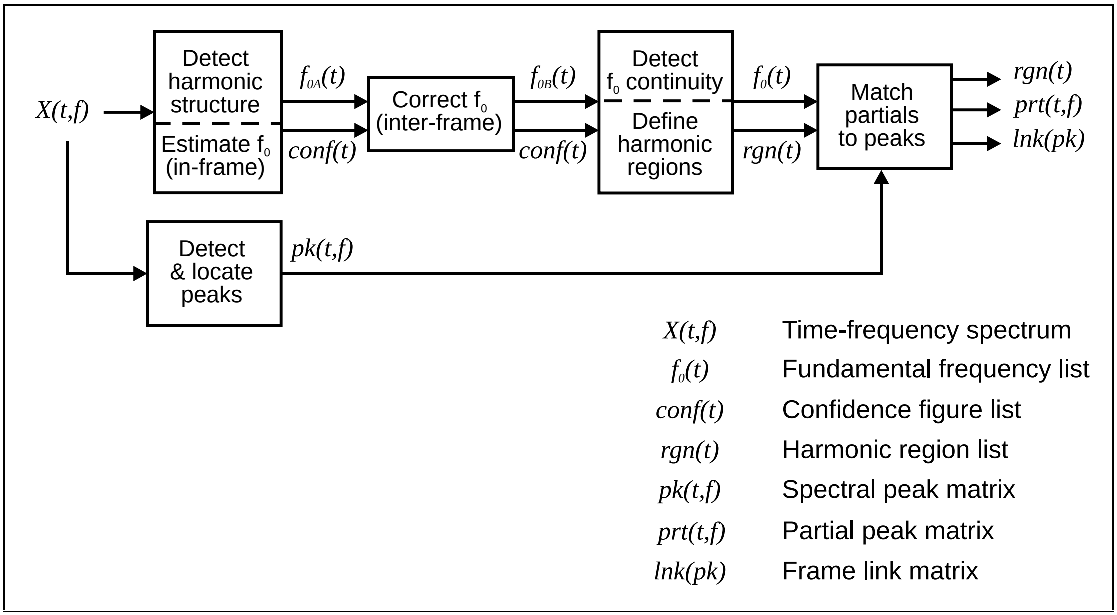

X(t,f)
Detect harmonic structure
Estimate f0 (in-frame)
f0A(t)
conf(t)
Correct f0 (inter-frame)
f0B(t)
conf(t)
Detect f0 continuity
Define harmonic regions
f0(t)
rgn(t)
Match partials to peaks
rgn(t)
prt(t,f)
lnk(pk)
Detect & locate peaks
pk(t,f)
X(t,f) Time-frequency spectrum
f0(t) Fundamental frequency list
conf(t) Confidence figure list
rgn(t) Harmonic region list
pk(t,f) Spectral peak matrix
prt(t,f) Partial peak matrix
lnk(pk) Frame link matrix

Figure 4.5 – Block diagram of complete harmonic linking process

Figure 4.5 is a block diagram of the proposed process. Initially, the harmonic structure is detected and the fundamental frequency estimated within each frame. Attached to the estimate is a confidence figure that is based on the prominence of the harmonic structure within the spectrum and how strongly it suggests the resultant fundamental frequency value. The next phase compares the results between frames, making corrections on the basis that highly confident estimates can promote changes to less confident estimates. The entire sound is then organised into 'harmonic regions', where each region denotes continuity of the fundamental frequency. Finally the peaks of each frame are mapped to the partial indices based on how well they approximate to a multiple of the fundamental frequency.

4.2.2 Fundamental Frequency Estimation and Harmonic Structure Detection

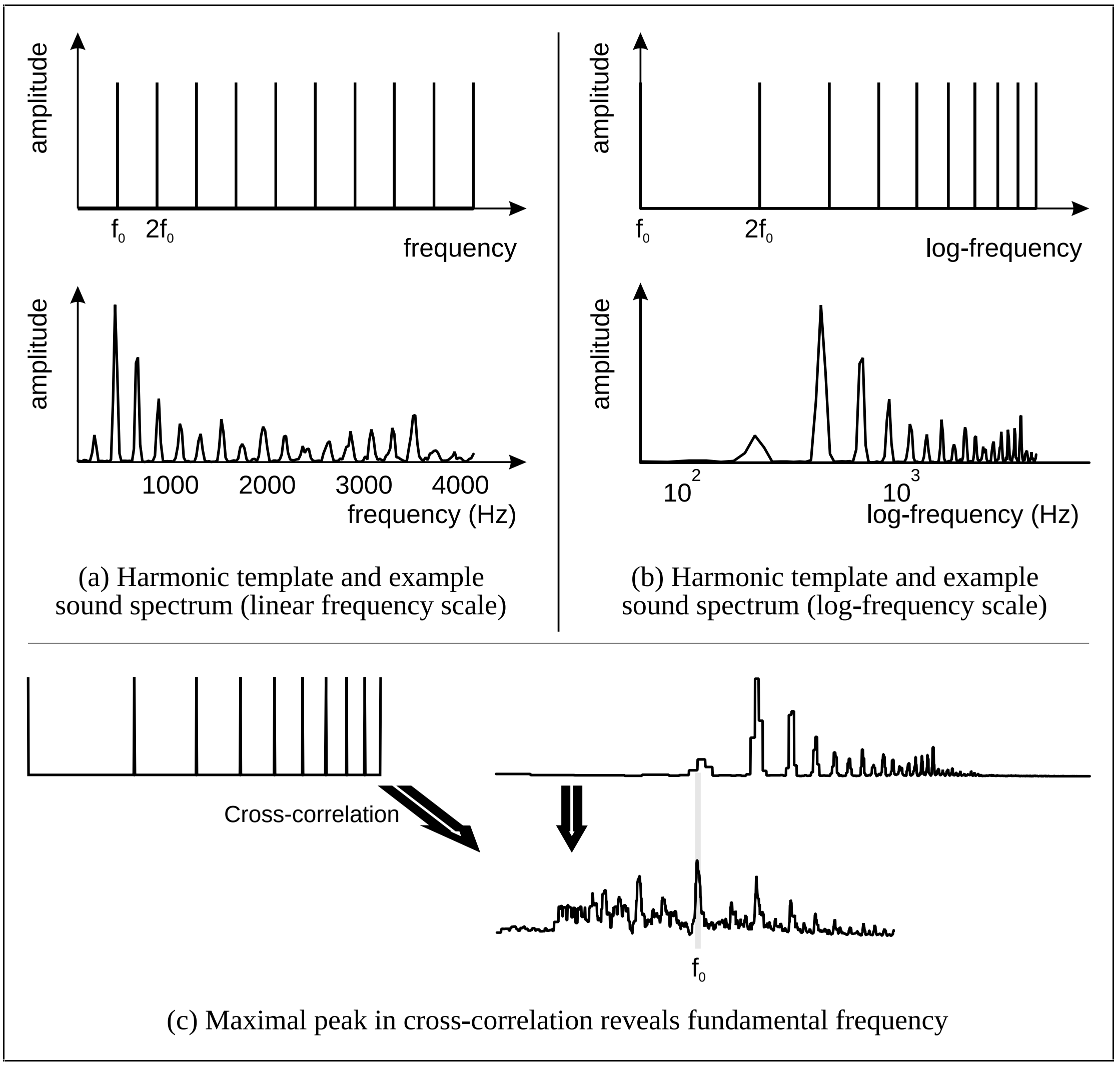
There are a vast number of publications describing methods for estimating the fundamental frequency (e.g. [Noll,1964; Goldstein,1973; Hermes,1988; Dogan & Mendel,1992; Doval & Rodet,1993; Sharp & While,1993; Petroni et al.,1994]), sometimes inaccurately called pitch estimation (which has an associated, but not identical, aim). Instead of providing assurance that this is a well understood subject, it suggests that fundamental frequency estimation is not straightforward. There is no standard method, although certain approaches are popular, and new papers continue to be published every year.

4.2.2.1 Two Popular Approaches

This section introduces a couple of popular approaches with desirable properties. Note that time domain methods, although quite accurate, were rejected because they would not enable an upgrade path to polyphony – if there are two pitches present in a sound, time domain techniques (such as auto-correlation) cannot cope.

The Template Method

One approach uses the shape of the ideal harmonic structure as a template and tries to fit it to the presented spectrum. The original method scaled the template (along the frequency axis) and calculated its correlation with the real spectrum at each scale [Martin,1982]. However a much more efficient method is possible using the log-frequency spectrum, because scaling on a linear axis is equivalent to translation on a logarithmic one [Hermes,1988]. Thus the shape and size of the harmonic template can be fixed and the correlation method reduces to cross-correlation. In both implementations, there is a peak where the template coincides with a harmonic structure. This method is hereon referred to as the 'template' method. See Figure 4.6.



(a) Harmonic template and example sound spectrum (linear frequency scale)

(b) Harmonic template and example sound spectrum (log-frequency scale)

(c) Maximal peak in cross-correlation reveals fundamental frequency

Figure 4.6 – Fundamental frequency estimation by the Template method

The Cepstrum Method

An alternative approach observes that the (linear frequency) magnitude spectrum of a harmonic structure has regularly spaced peaks, where the spacing is dependent on the fundamental frequency. By treating the magnitude spectrum as though it were a time domain signal and calculating its spectrum, the cepstrum is generated[1]. In this distribution, peaks represent the periodicity of the magnitude spectrum [Noll,1964]. Therefore, regularly spaced peaks in the *spectrum* cause a peak in the *cepstrum* at the ‘quefrency’ (c.f. frequency) corresponding to their spacing, which in turn is indicative of the fundamental frequency of the time domain signal. See Figure 4.7.

[1] There appear to be two different definitions for the term ‘cepstrum’. One, as defined in this text, is common to all publications regarding pitch or fundamental frequency determination and musical transformation [Noll,1964; Martin,1982; Pabon,1994; Todoroff,1996]. The other is defined as the IFFT of the logarithm of the FFT [Makhoul,1975; Tohyama & Lyon & Koike,1993]. Throughout this thesis, the former definition is assumed.

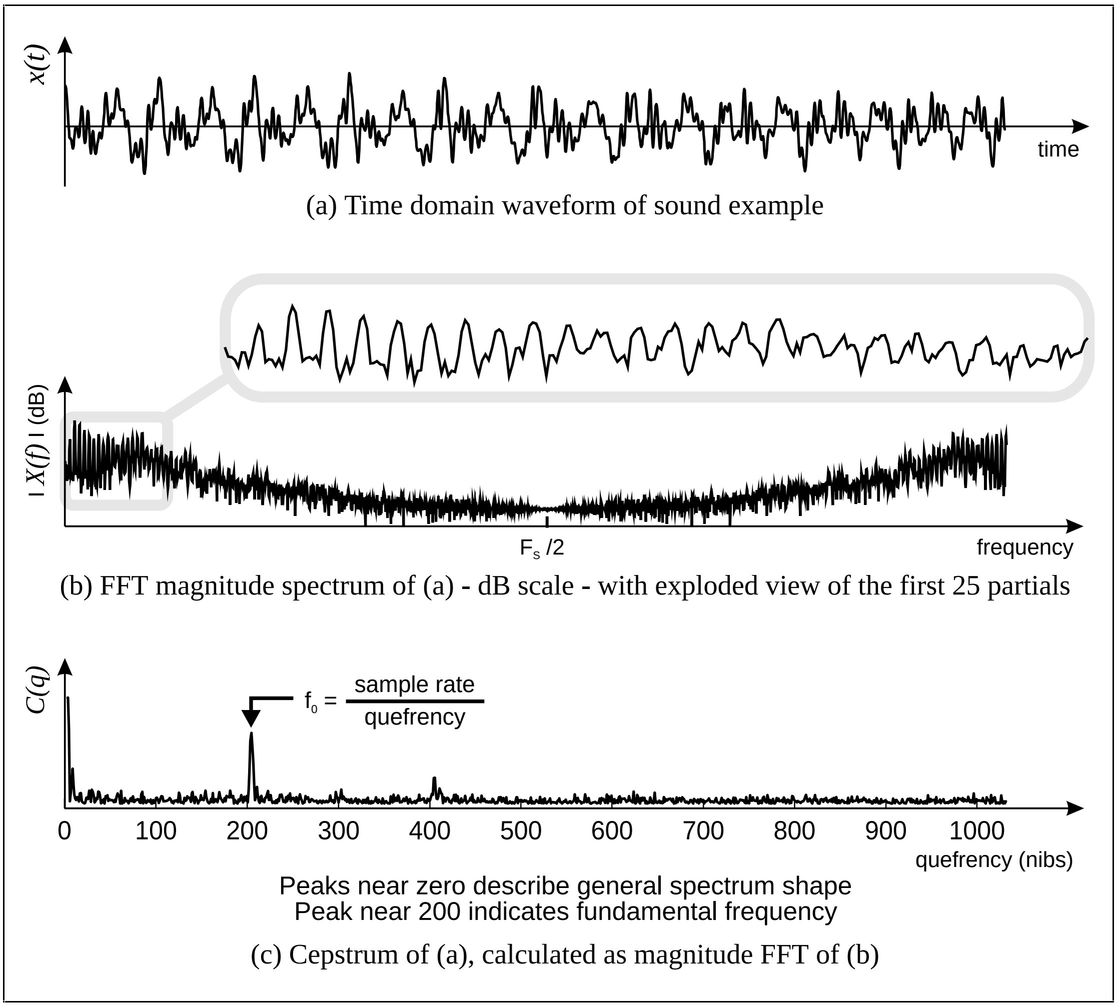


Figure 4.7 – Fundamental frequency estimation by the Cepstrum method

Assessment of the Template Method

The template method produces a peak when each partial matches up with the template, but it also produces peaks when the template matches up with every second or third peak, or when every second or third template spike matches up with a partial. Spurious peaks in the spectrum can also encourage these errors when they are unfortunately located. If the global maximum is taken, it can sometimes correspond to multiples ($2f_0$, $3f_0$, $4f_0$, etc.) or fractions (½ f_0, ⅓ f_0, ¼ f_0, etc.) of the true fundamental, f_0.

The likelihood of such errors can be conditioned by altering certain parameters, such as the shape of the template's envelope (see Figure 4.8) to give greater weighting to certain partials. In tests on a number of sound examples, the most effective shape was found to be where the amplitudes of the template spikes decrease linearly with partial order. (On the log-frequency scale the envelope appears as a logarithmic roll-off.) This emphasises the lowest partials, which are generally contain the most energy and are most prominent with respect to the surrounding FFT spectrum. Furthermore it was found that by restricting the template comb to between five and ten spikes, there are sufficient elements for estimation when the fundamental frequency peak is small and not too many elements that estimation becomes affected by higher frequency non partial energy.

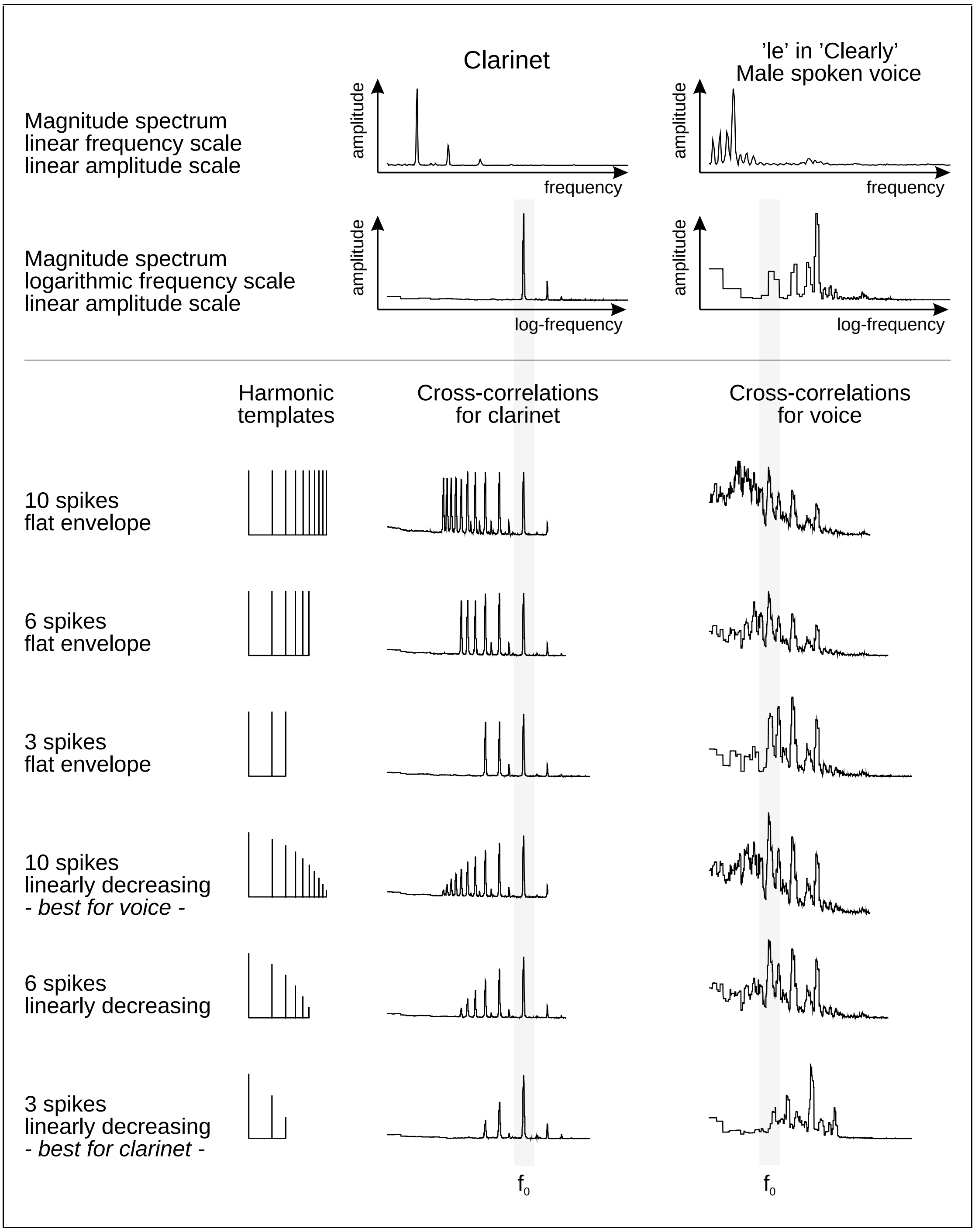


Figure 4.8 – Variations in performance for different template shapes

Given a focus on the correct maximum, the template method appears to provide a good estimate of the fundamental frequency It should be noted that estimation of the fundamental frequency, for overtones that are not precise harmonics, has no fixed mathematical definition. Therefore the degree of correctness of this or any other detector cannot be absolutely gauged. (Any number of statistical measures could be legitimately used, even though they might give conflicting results.) In this case the 'goodness' of the algorithm was calibrated against a best fit by eye, which seems to be as good a measure as any other.

The greatest shortfall of the template method is the lack of an indicator for whether a harmonic structure exists, because the cross-correlation will produce a maximum even for a random spectrum. Several measures of confidence were tried but with no success.

Assessment of the Cepstrum Method

The cepstrum is the spectrum of the magnitude spectrum, so its low quefrency components correspond to the overall envelope of the frequency distribution. The higher quefrencies correspond to more localised spectral components (i.e. narrow spectral peaks). Spectral peaks that are unrelated generate low level spikes of similar magnitude throughout the mid to high quefrency cepstrum. Since these are present in almost all audio signals, they generate an apparent 'cepstral noise floor'. Conversely, a set of spectral peaks that are harmonically related produce a single prominent spike in the cepstrum, from which the fundamental frequency can be calculated.

In the practical situation where partials are only *roughly* harmonic and there are additional spurious peaks, the 'single prominent spike' splits into a cluster of closely spaced spikes, that are nevertheless prominent above the noise floor. Because of the loose locality of the harmonic structure peak in the cepstrum, it is difficult to make a clear estimate of the fundamental frequency, although its neighbourhood can be clearly identified.

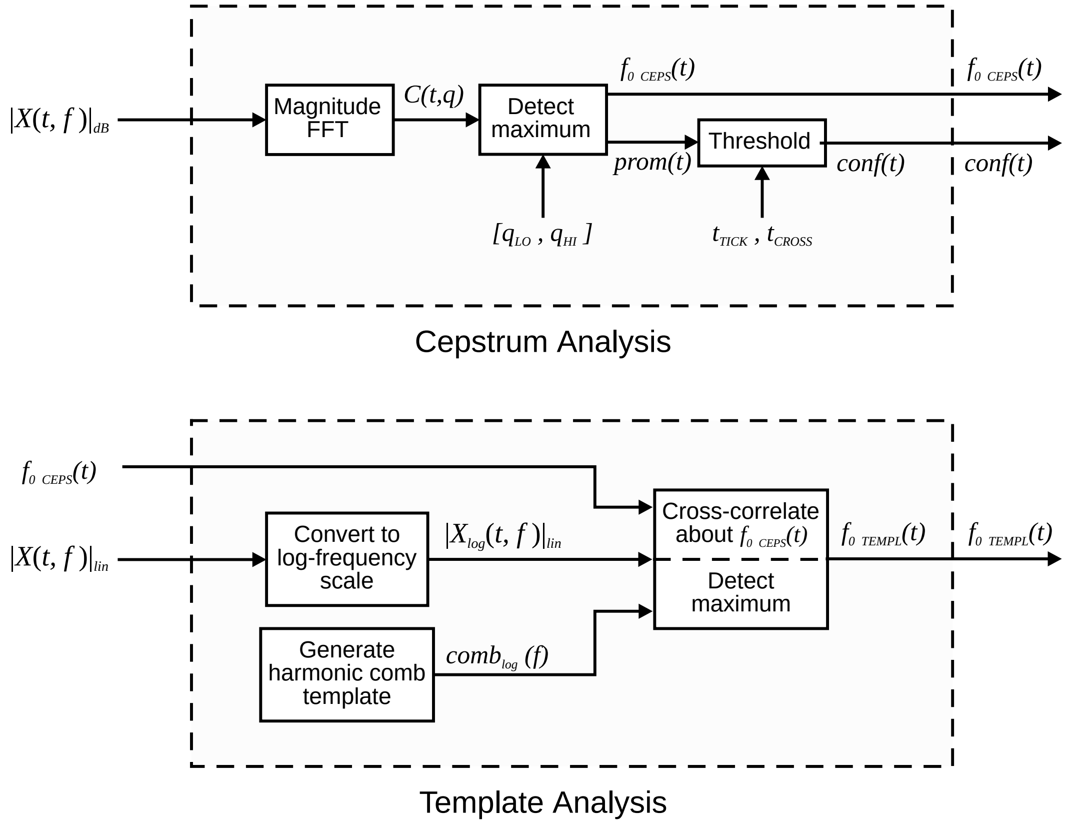
Since the cepstrum is a spectral distribution, it not only displays a peak for the spacing of the harmonic elements, but also at integer multiples of this spacing – 'rahmonics' (c.f. harmonics) of the harmonic structure. Therefore the cepstrum, like the template method, is prone to estimating multiples or fractions of the true fundamental frequency, but it seems to be less susceptible than the template method.

Experiments show that using a dB scale for the magnitude spectrum (from which the cepstrum is calculated) provides the best results, probably because it brings high and low magnitude peaks onto a comparable scale, thus accentuating the periodicity of the shape of the harmonic structure.

The greatest asset of the cepstrum method is the ease of generating a confidence figure, which is a measure of the presence of a harmonic structure. This has been based on the prominence of the maximum harmonic spike above the background 'noise' (details presented in the next section). Unfortunately the working range of the cepstrum is limited at low quefrency (high fundamental frequency) by the more dominant components describing the spectral envelope and at high quefrency (low fundamental frequency) by the cepstral equivalent of the Nyquist upper limit. In practice, the high quefrency limit corresponds to the low frequency threshold of hearing (between 40 and 50Hz), but the low quefrency constraint limits fundamental frequency estimation to a maximum of around 630Hz. (These figures are based on a 2048 sample window-length for spectrum FFT and cepstrum FFT, and 44.1kHz sample rate.)

4.2.2.2 A Hybrid Method

The method that was implemented consists of a hybrid of the two described approaches. It aims to take advantage of the cepstrum's ability to locate the neighbourhood of the fundamental and to provide a confidence measure, and the template method's ability to accurately locate the fundamental frequency, given the correct maximum to look at. See Figure 4.9.



$|X(t,f)|_{dB}$	(Log-)magnitude spectrum
$C(t,q)$	Time-quefrency cepstrum
$[q_{LO}, q_{HI}]$	Quefrency range
$f_{0\ CEPS}(t)$	Fundamental frq (cepstrum est.)
$prom(t)$	Prominence measure
t_{TICK}, t_{CROSS}	Tick,Cross thresholds
$conf(t)$	Confidence figure
$|X(t,f)|_{lin}$	(Linear) magnitude spectrum
$|X_{log}(t,f)|_{lin}$	Log-frq (linear) magnitude spectrum
$comb_{log}(f)$	Harmonic comb template
$f_{0\ TEMPL}(t)$	Fundamental frq (template est.)

Figure 4.9 – Fundamental frequency estimation and harmonic structure detection

First of all the cepstrum is calculated directly from the magnitude FFT spectrum of a frame. The maximal ‘nib’ (c.f. bin) within the usable quefrency range is stored as the initial estimate of the fundamental frequency. A measure of its prominence is also calculated and stored. The estimate is passed to the template method which only performs the cross-correlation in a region centred around that frequency. The location of the new maximum gives the refined fundamental frequency estimate. The following provides greater detail of the implemented algorithm with values that were found to be useful.

Cepstrum Calculation

For a sound recorded at 44.1kHz sampling rate, and analysed with an FFT window-length of 2048, the cepstrum calculation uses the (dB-scale) magnitude of every FFT bin, and is therefore also a 2048-length FFT. Although the upper half of the FFT is the negative frequency bins and is therefore a mirror image of the lower half (in the magnitude spectrum), by using the whole FFT it is guaranteed that there will be no discontinuities at the array ends. Hence spectral leakage is avoided without the need for a window function (that would otherwise affect the data itself and would reduce the amplitude particularly of the lower partials).

The prominence of the peak is calculated:

$$p = \frac{\max\limits_{l=q_{LO}}^{q_{HI}} \left\{ \left| C(l) \right| \right\}}{\operatorname*{mean}\limits_{l=q_{LO}}^{q_{HI}} \left\{ \left| C(l) \right| \right\}} \tag{5.1}$$

where p is the prominence measure,
q_{LO}, q_{HI} are the lower and upper limits of the valid quefrency range,
$|C(l)|$ is the magnitude of the lth nib of the cepstrum array.

Useful parameter values for the valid quefrency range are:

q_{LO} = 70 nibs,
q_{HI} = 1024 nibs,

which yields a fundamental frequency (f_0) range of:

$$\mathrm{f}_0 \in \left(43.07, 630.0\right) \mathrm{Hz}.$$

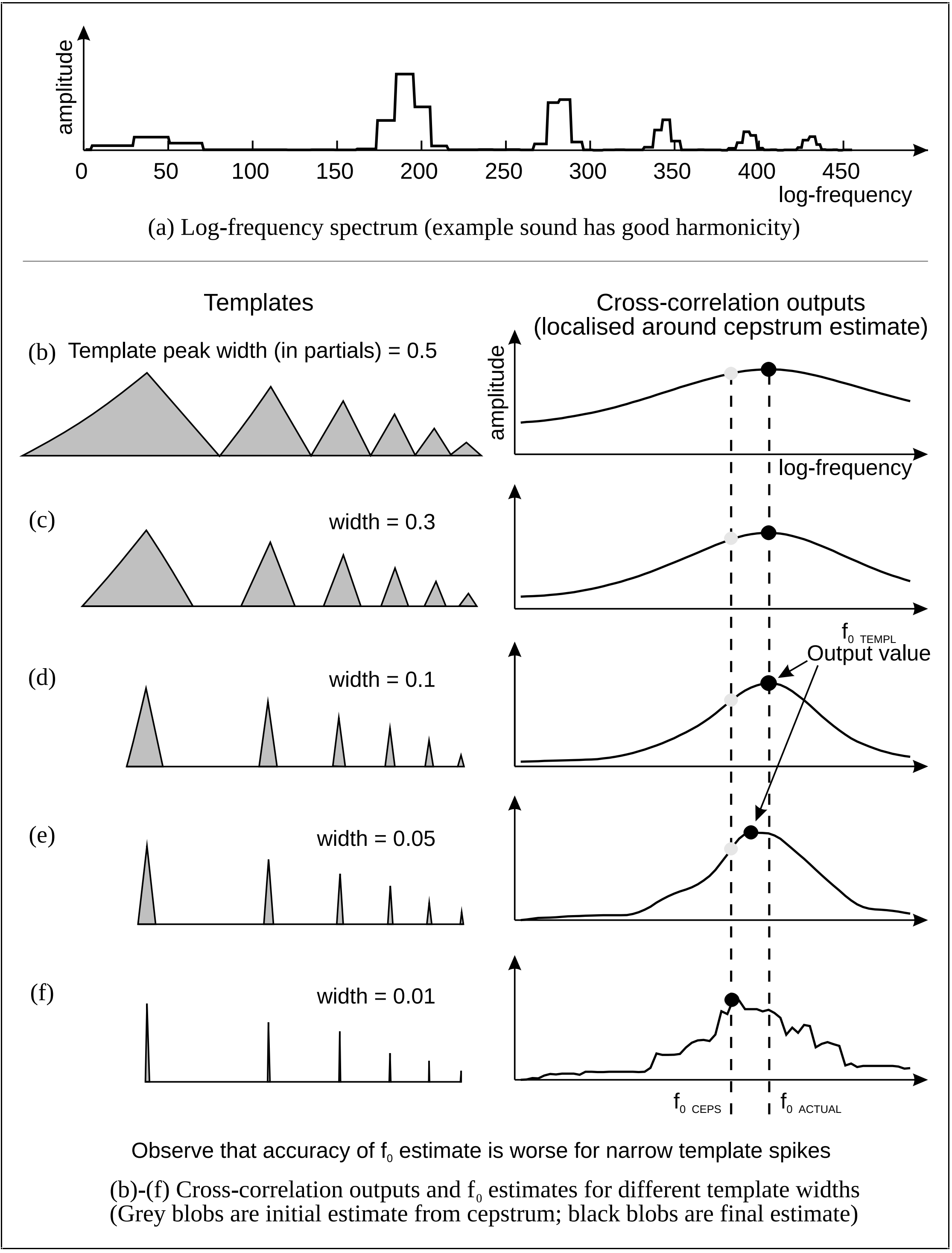
Confidence Figure

The peak prominence value is reduced to a confidence figure by application of two thresholds: T_{TICK} , T_{CROSS}. A prominence above T_{TICK} is denoted ✓, and represents a definite harmonic structure. A prominence below T_{CROSS} is denoted ✗, which says that the presence of a harmonic structure is unlikely. In between these thresholds is the status of uncertainty, denoted **?**, to be resolved at a later stage where inter-frame information can help. Based on a number of sound examples that were hand scored (with reference to the time domain waveform and time-frequency sonogram), the following values were obtained:

T_{TICK} = 7.7
T_{CROSS} = 4.8

Cross-Correlation with Harmonic Template

For each frame, regardless of the confidence figure, the cepstrum's fundamental frequency estimate is passed to the template cross-correlation method, which performs a local search for a maximum. The localised search reduces the computational cost of the template method and removes the possibility of locating multiples or fractions (unless the error has already been made by the cepstrum).



(a) Log-frequency spectrum (example sound has good harmonicity)

Observe that accuracy of f_0 estimate is worse for narrow template spikes

(b)-(f) Cross-correlation outputs and f_0 estimates for different template widths
(Grey blobs are initial estimate from cepstrum; black blobs are final estimate)

Figure 4.10 – Effect of different template peaks widths on accuracy and template size

The harmonic template is constructed for the desired number of partials, with the triangular peaks of Figure 4.10 at each partial location. The idealised delta function 'spikes' of the template (shown in previous figures) were widened to account for inexact harmonic ratios, giving them a finite capture range. The triangular shape places the greatest weighting at the idealised central location, so as to bias the fundamental frequency estimate toward it (especially when the harmonicity is good). In experiments, the best width was found at around 0.3 partials' separation. Shorter widths tend to produce separate spikes in the cross-correlation output for the contributions of each partial, if the harmonicity is poor. Wider peaks have the desired smoothing effect of agglomerating the partials' contributions into a single maximum, but they require a larger template array and hence more computation. (This is primarily due to the width of the first partial – this is visible in Figure 4.10.)

The log-frequency spectrum array is generated by sampling the magnitude FFT spectrum at the desired resolution over the desired range. The data that is included ranges from a little below the f_{0CEPS} to the same amount above it times the number of partials in the template, i.e. [f_{0CEPS} – search span , (f_{0CEPS} + search span)×n^{o} template partials]. The 'search span' defines how far above and below f_{0CEPS} in which to search for a better approximation to the true f_0 value. The cross-correlation of the template array with the log-frequency spectrum array is restricted to f_{0CEPS} ± search span, and the new f_0 estimate, f_{0TEMPL}, is taken as the location of the maximum on that curve.

Useful values for the template calculation parameters are:

Array resolution	=	150 points per octave = 8 cents
Search span	=	1.5 bins = $1.5\times(44100\div2048) \approx 32.3$ Hz

4.2.3 Inter-Frame Correction and Harmonic Region Definition

With the awareness that both template and cepstrum fundamental frequency estimation methods are prone to failure by locating multiples or fractions of the true value, a two-stage procedure was planned from the outset. The first stage in the estimation process, described above, is 'in-frame', considering the peak frequency ratios instantaneously; this provides an f_0 estimate and a confidence figure. The second stage, described here, is 'inter-frame', using the more confident estimates to adjust the less confident ones, based on the expectation of continuity between frames. (Refer to Figure 4.5 above.) This method implicitly assumes that the stronger the harmonic structure, the more reliable the f_0 estimate. The continuity between frames is also the basis for defining harmonic regions, within which partials are matched to peaks and implicitly linked.

4.2.3.1 Estimate Assessment

Each time-frame considers its own f_0 estimate to be correct, and attempts to influence its neighbours, where there is the possibility of connectivity. That is, if a neighbour's f_0 estimate is close to an integer multiple or fraction of the current frame's, then it will try to scale the neighbour's estimate to achieve connectivity; see Figure 4.11a overleaf. The maximum frequency deviation between frames for connectivity and the maximum amount of scaling are specified by the parameters (values determined by experimentation):

Maximum multiple:	$f_{0\,\mathrm{NEIGHBOUR}} = f_{0\,\mathrm{CURRENT}} \times 6$
Minimum fraction:	$f_{0\,\mathrm{NEIGHBOUR}} = f_{0\,\mathrm{CURRENT}} \div 6$

Maximum frequency deviation of scaled $f_{0\ \mathrm{NEIGHBOUR}}$ = 10%

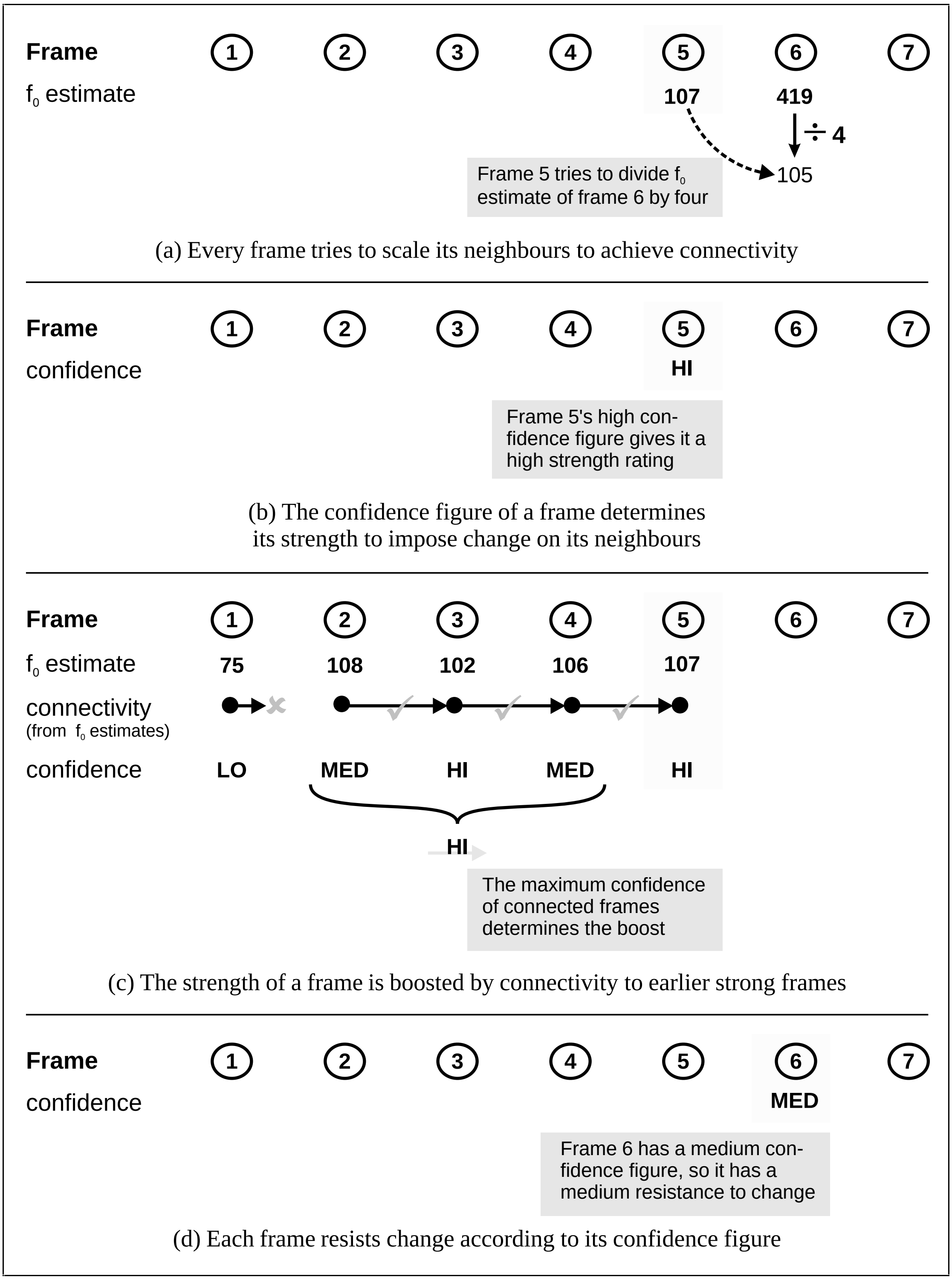
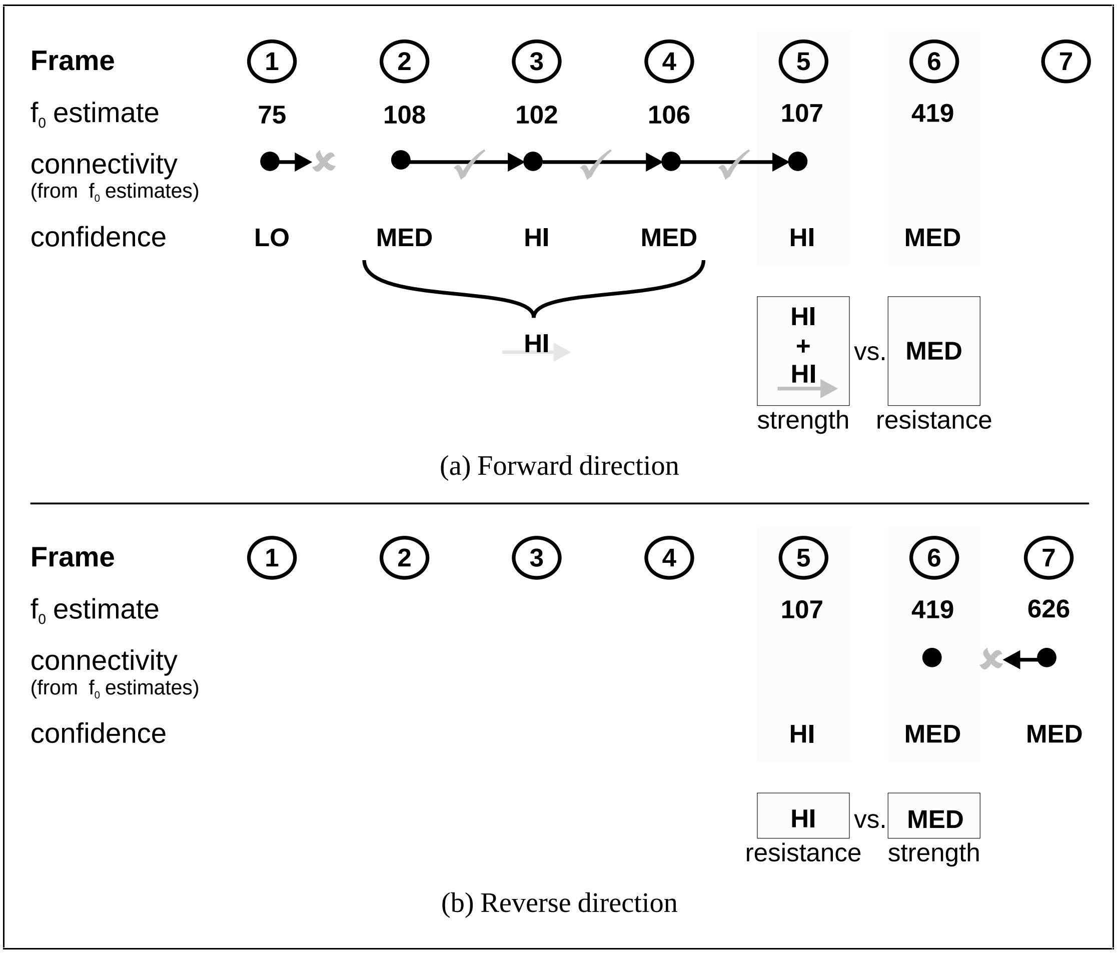

Frame
1
2
3
4
5
6
7
f0 estimate
107
419
÷ 4
105
Frame 5 tries to divide f0 estimate of frame 6 by four
(a) Every frame tries to scale its neighbours to achieve connectivity
Frame
confidence
HI
Frame 5's high confidence figure gives it a high strength rating
(b) The confidence figure of a frame determines its strength to impose change on its neighbours
Frame
f0 estimate
75
108
102
106
107
connectivity
(from f0 estimates)
confidence
LO
MED
HI
MED
HI
HI
The maximum confidence of connected frames determines the boost
(c) The strength of a frame is boosted by connectivity to earlier strong frames
Frame
confidence
MED
Frame 6 has a medium confidence figure, so it has a medium resistance to change
(d) Each frame resists change according to its confidence figure

Figure 4.11 – Calculation of scaling factor, strength and resistance

The amount of influence a frame has is a direct function of its confidence figure (see Figure 4.11b), in conjunction with the historical trend (see Figure 4.11c). The resistance of a frame to being changed by its neighbours is also a function of its confidence; see Figure 4.11d. Therefore a strong frame and a weak frame that are adjacent will both attempt to influence each other, but the strong frame will exert a greater influence and the weaker frame will be more susceptible; see Figure 4.12.



(a) Forward direction

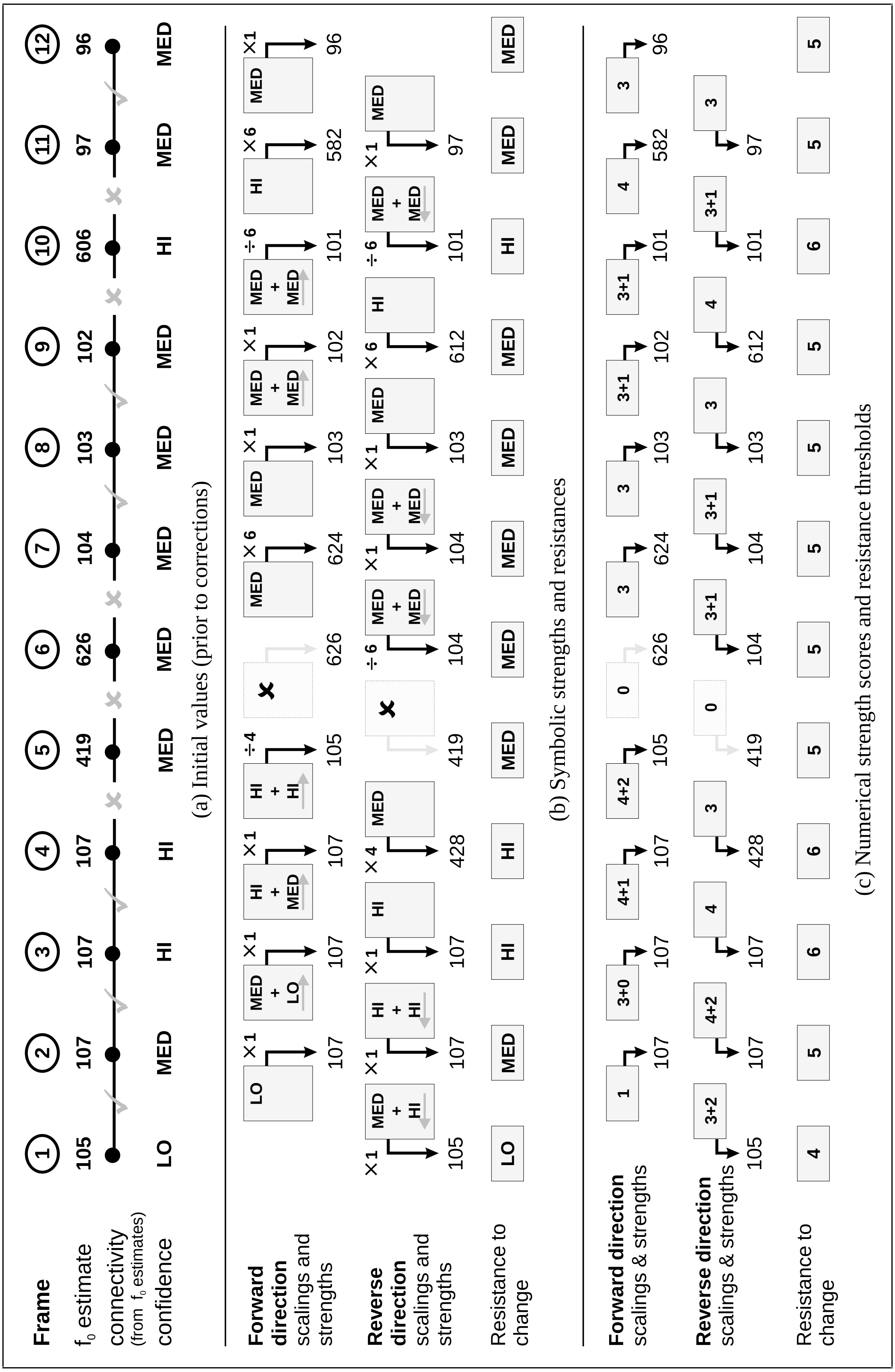
(b) Reverse direction

Figure 4.12 – A stronger is more likely to alter a weaker frame

To implement the algorithm, the frames are first scanned in the forward direction, each frame exerting its influence on the next frame, and then in the reverse direction, each frame attempting to change the previous frame. In this way, it is possible to take into account the previous history of a fundamental frequency trajectory (in both forward and reverse directions), so that a frame is given greater strength to impose a change on the following frame if it already has connectivity to previous strong frames.

Figure 4.13 overleaf shows twelve frames of a real example. Figure 4.13a shows the initial values, obtained from the in-frame estimates of section 4.2.2 above. The first row of Figure 4.13b shows how each frame is being influenced by preceding frames (in the forward direction): the strength with which they try to impose change and the new f_0 value to which they try to change it. The next row shows influences in the reverse direction. The third row shows each frame's resistance to change, which is directly based on its confidence figure. (Note that frames 5 and 6 are unable to influence one another, because neither can achieve connectivity through scaling the other's f_0 estimate by a integer factor.)

In the algorithm, the strengths are calculated as scores, where the score is a function of the confidence figure. This value is boosted by the strength of historical connectivity. Similarly, each frame has a resistance threshold based on the confidence figure – a frame will be altered if the score imposed by its neighbours is *greater than or equal to* its resistance. The values for these were determined from examples which were hand-scored:



(a) Initial values (prior to corrections)

(b) Symbolic strengths and resistances

(c) Numerical strength scores and resistance thresholds

Figure 4.13 – Strength and resistance calculations for a real example

Strength scores based on confidence of immediate neighbour

HI	4
MED	3
LO	1

Boost scores based on highest confidence of earlier frames (to which the immediate neighbour has f_0 connectivity)

HI	+2
MED	+1
LO	+0

Resistance thresholds based on confidence of each frame itself

HI	6
MED	5
LO	4

Figure 4.13c displays the same information as Figure 4.13b but in numerical form using the key above.

See Figure 4.14 (on the next page), which shows how forward and reverse direction influences are combined for the example data of Figure 4.13. This example illustrates many of the possible situations. If both neighbours either side of a particular frame are trying to exert the same change on it (e.g. frame 2 in Figure 4.14b), then their scores are added, making the change more likely. If neighbours disagree (e.g. frame 4 in Figure 4.14b), then the one with the stronger score wins. If they disagree and their scores are equal (e.g. frame 9 in Figure 4.14b), then influence is in the forward direction – an arbitrary rule.

If a frame's f_0 estimate is close to the next frame's, so that they have connectivity without needing any scaling, then it will try to change the next frame "not at all", which has a locking effect, preventing a weaker frame on the other side of the neighbour from making an alteration; i.e. it is trying to scale its neighbour by a factor of 1. (e.g. reverse direction onto frame 7 in Figure 4.14b.)

If the frequency difference between two frames is greater than the deviation threshold, even after scaling, then a score of zero is allocated; i.e. no influence. (e.g. forward direction onto frame 6 in Figure 4.14b.)

Having determined the combined pressure on each frame from its neighbours, the strength scores are compared with the frame's own resistance threshold. Only if the score is greater than or equal to the threshold can the change be made. Figure 4.14c summarises which frames would be altered.

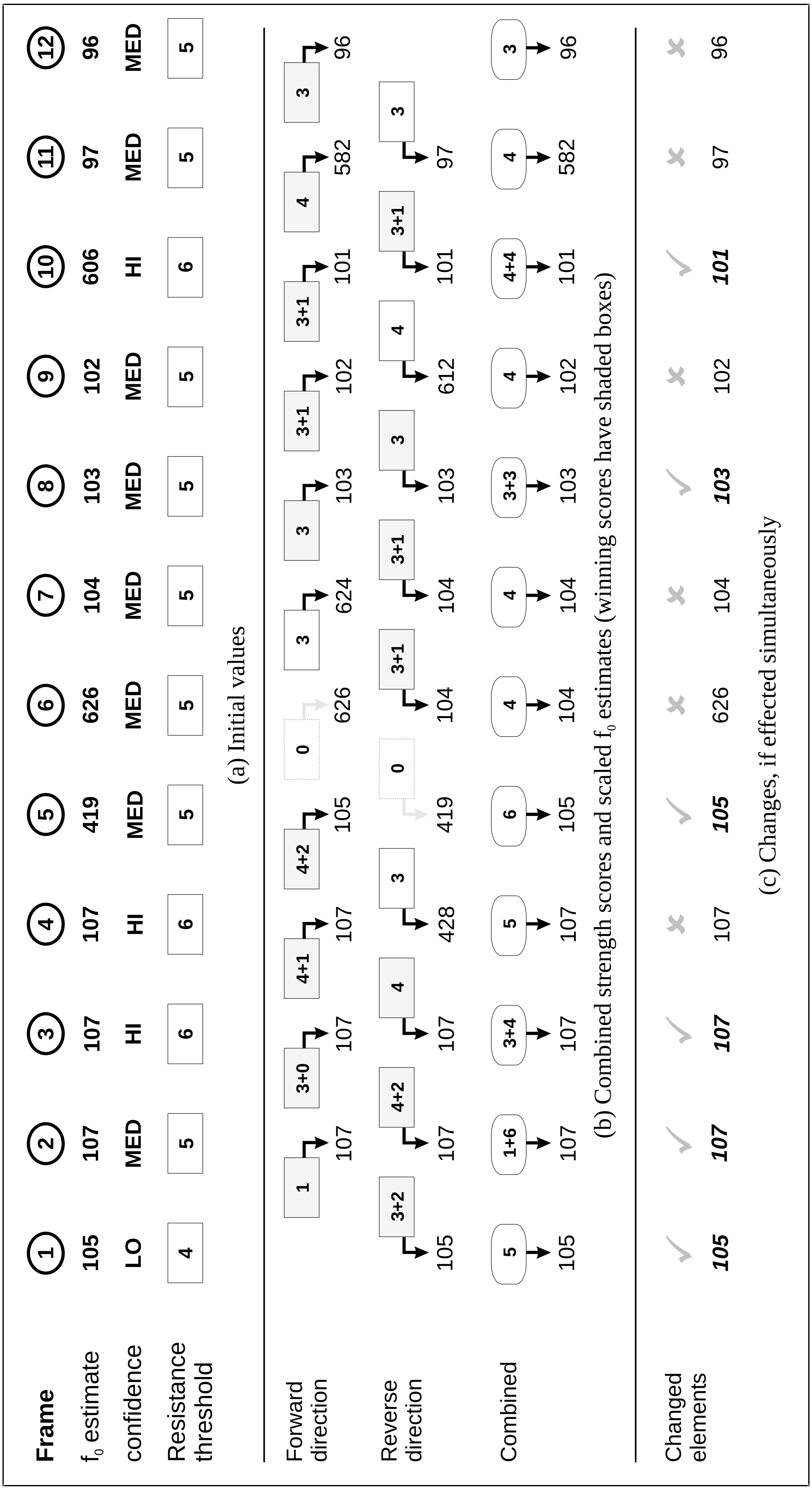


Figure 4.14 – Combining forward and reverse direction scores

Note, from Figure 4.14b, that there is pressure on all the erroneous frames (i.e. frames 5, 6 and 10 whose estimates are roughly multiples of the other estimates) to bring them into line. Note also (from Figure 4.14c) that if the changes were made to all frames simultaneously, frame 6 would not be corrected because its resistance threshold is higher than the score for change. This raises the question of the order in which to make the changes. If the changes were made sequentially according to some scheme, as each frame is updated its influence on its neighbours would change. This, in turn, would alter the pressures for change upon the surrounding frames. In the next section such a scheme is sought that can correct erroneous estimates whilst securing correct ones.

4.2.3.2 Estimate Correction

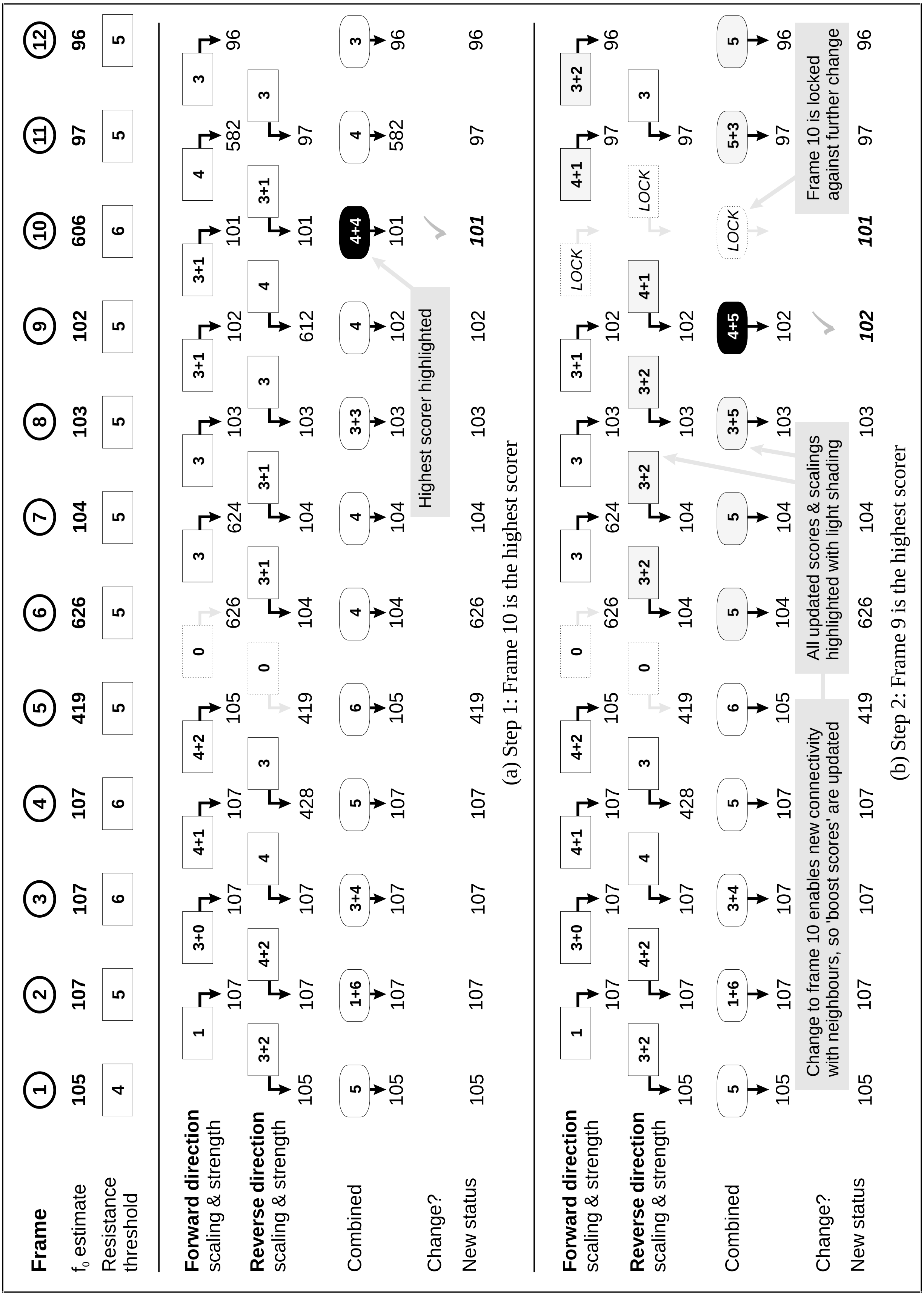
In a situation where two frames are both trying to change each other, say one wants to double the other's estimate, while the other is trying to halve this one's, and both have scores above each others' resistance thresholds, the outcome is dependent on which frame is changed first. Obviously there is no point in changing both frames!

Correction by Highest Scorer

Initially, a method was tried which made changes to the highest scorers first, working downwards. (See Figure 4.15 on the next two pages, which uses the same example data as the previous two figures.) After each change, the changed value would be locked against further change and its influence on neighbours is reassessed given the new f_0 value. In the example of Figure 4.15, frame 10 is changed first (see Figure 4.15a). Following the change (see Figure 4.15b), frame 10 no longer tries to alter frame 11, but instead tries to lock it to its existing (correct) value. Not only does frame 10 affect its immediate neighbours; by virtue of the new connectivity, boost scores are increased in both directions. As a result, there is increased pressure for change on frame 6. When frame 5 is later changed (see Figure 4.15c), this further increases the pressure on frame 6.

The outcome of using a highest scorer priority method (in Figure 4.15e) is the desired one in this example. The method is effective because isolated erroneous frames (which have pressure from both sides) are changed first. Following the change there is increased connectivity which enables propagation of boost scores, thereby increasing the pressure on other erroneous frames.

Unfortunately, if the data includes 'burst errors', where a succession of estimates are at the same multiple/fraction, the scheme can backfire. Erroneous frames toward the centre of the burst are pressurised by both neighbours to secure their f_0 estimates with high combined scores. Therefore when the changes are committed, the incorrect values at the centre of the burst are changed "not at all" before the incorrect elements at the burst boundaries can be altered. See the example in Figure 4.16 (on the two pages following Figure 4.15). This includes a burst of three erroneous frames in frame 6-8 as well as the other errors from the previous example. The outcome (in Figure 4.16d) is that the first six frames are correct and connectable (including the correction of frames 5 and 6), whilst the latter six frames are incorrect and connectable (having wrongly altered the previously correct frames: 9, 11 and 12).



Frame	1	2	3	4	5	6	7	8	9	10	11	12
f_0 estimate	105	107	107	107	419	626	104	103	102	606	97	96
Resistance threshold	4	5	6	6	5	5	5	5	5	6	5	5
Forward direction scaling & strength		1 107	3+0 107	4+1 107	4+2 105	0 626	3 624	3 103	3+1 102	3+1 101	4 582	3 96
Reverse direction scaling & strength	3+2 105	4+2 107	4 107	3 428	0 419	3+1 104	3+1 104	3 103	4 612	3+1 101	3 97	
Combined	5 105	1+6 107	3+4 107	5 107	6 105	4 104	4 104	3+3 103	4 102	4+4 101	4 582	3 96
Change?										✓		
New status	105	107	107	107	419	626	104	103	102	***101***	97	96

Highest scorer highlighted

(a) Step 1: Frame 10 is the highest scorer

Frame	1	2	3	4	5	6	7	8	9	10	11	12
Forward direction scaling & strength		1 107	3+0 107	4+1 107	4+2 105	0 626	3 624	3 103	3+1 102	*LOCK*	4+1 97	3+2 96
Reverse direction scaling & strength	3+2 105	4+2 107	4 107	3 428	0 419	3+2 104	3+2 104	3+2 103	4+1 102	*LOCK*	3 97	
Combined	5 105	1+6 107	3+4 107	5 107	6 105	5 104	5 104	3+5 103	4+5 102	*LOCK*	5+3 97	5 96
Change?									✓			
New status	105	107	107	107	419	626	104	103	***102***	***101***	97	96

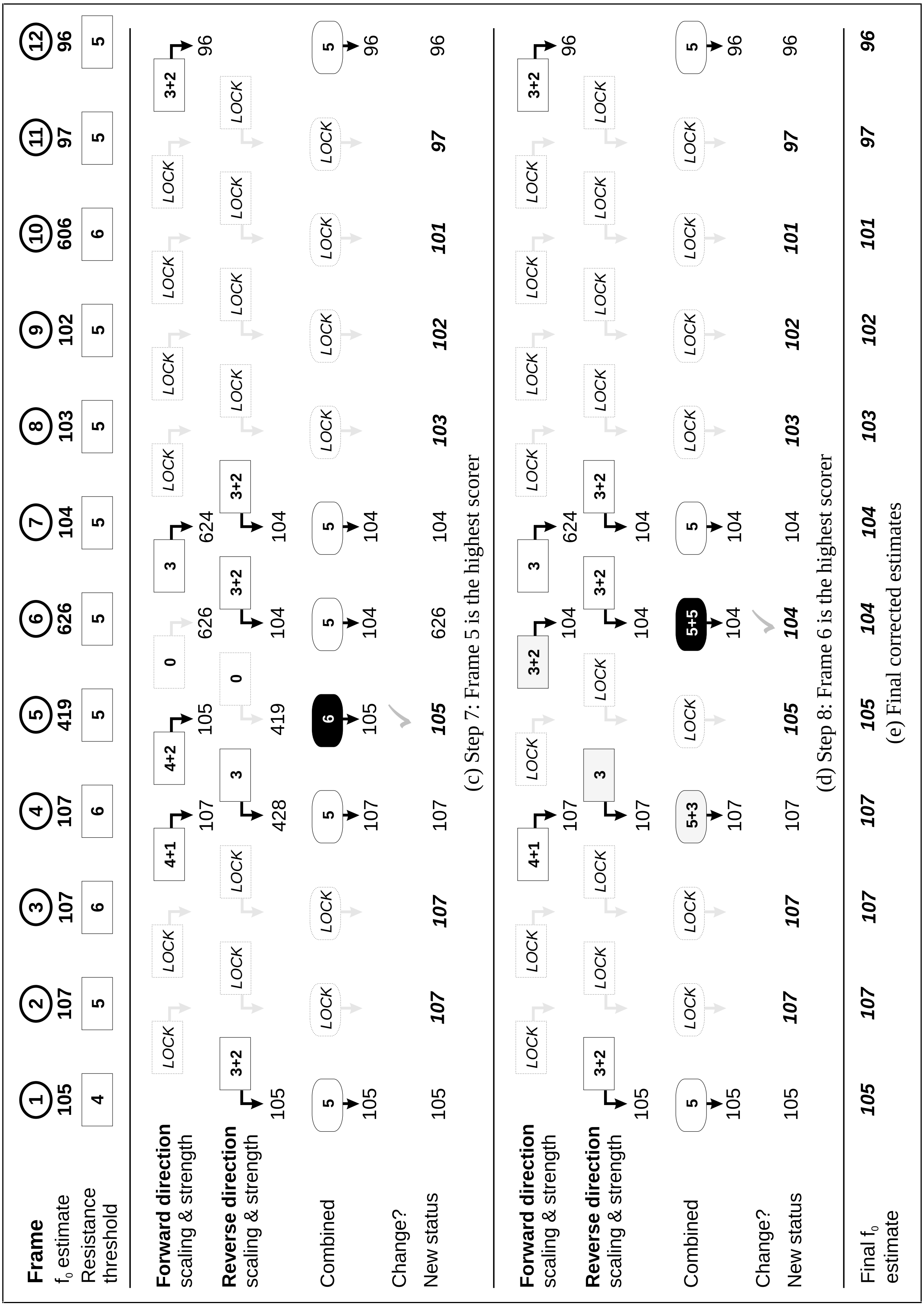
Change to frame 10 enables new connectivity with neighbours, so 'boost scores' are updated

All updated scores & scalings highlighted with light shading

Frame 10 is locked against further change

(b) Step 2: Frame 9 is the highest scorer

Figure 4.15 – The Highest Scorer correction scheme (for previous example data)



Frame	1	2	3	4	5	6	7	8	9	10	11	12
f_0 estimate	105	107	107	107	419	626	104	103	102	606	97	96
Resistance threshold	4	5	6	6	5	5	5	5	5	6	5	5
Forward direction scaling & strength		LOCK	LOCK	4+1 → 107	4+2 → 105	0 → 626	3 → 624	LOCK	LOCK	LOCK	LOCK	3+2 → 96
Reverse direction scaling & strength	3+2 → 105	LOCK	LOCK	3 → 428	0 → 419	3+2 → 104	3+2 → 104	LOCK	LOCK	LOCK	LOCK	
Combined	5 → 105	LOCK	LOCK	5 → 107	6 → 105	5 → 104	5 → 104	LOCK	LOCK	LOCK	LOCK	5 → 96
Change?					✓							
New status	105	*107*	*107*	107	*105*	626	104	*103*	*102*	*101*	*97*	96

(c) Step 7: Frame 5 is the highest scorer

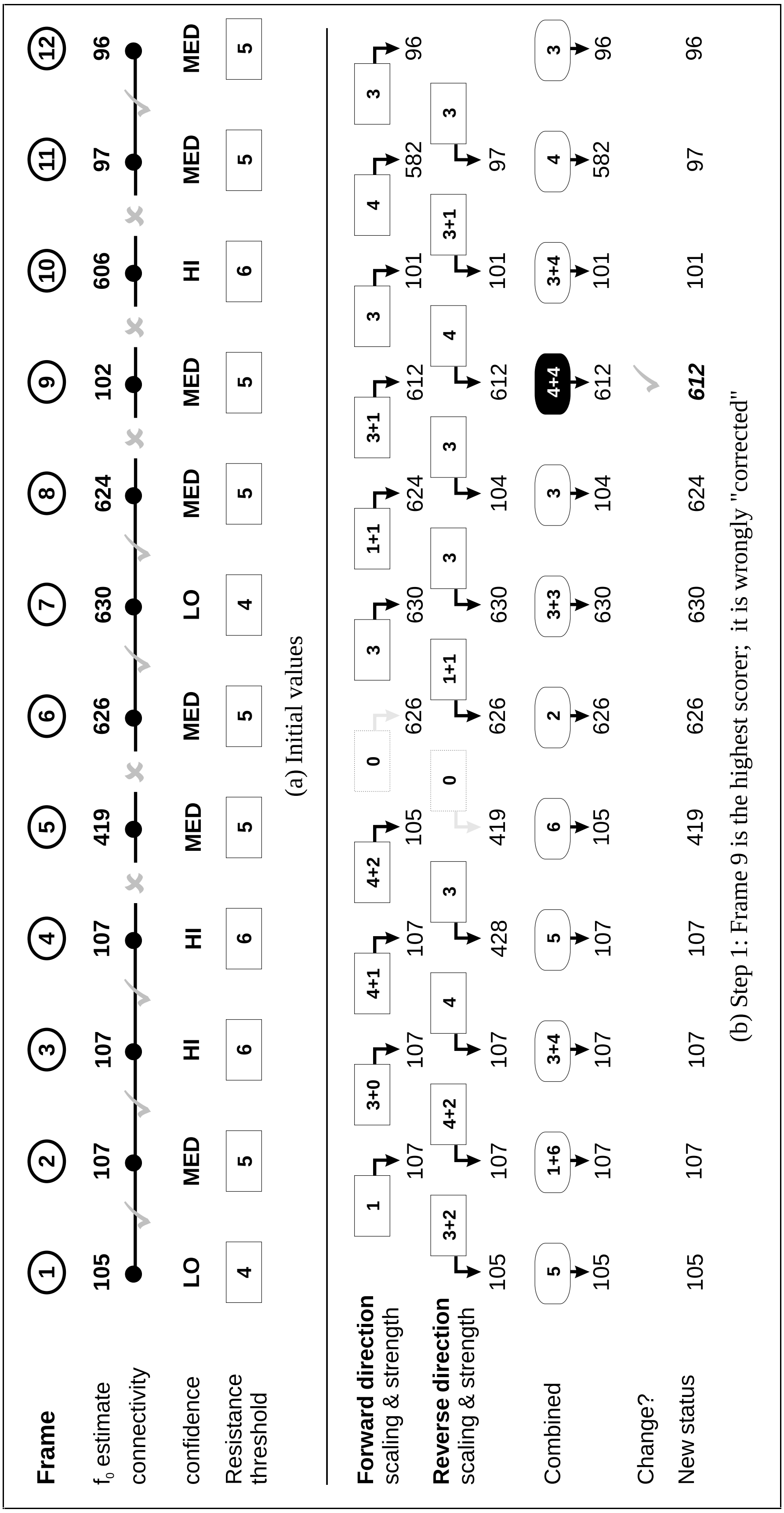
Frame	1	2	3	4	5	6	7	8	9	10	11	12
Forward direction scaling & strength		LOCK	LOCK	4+1 → 107	LOCK	3+2 → 104	3 → 624	LOCK	LOCK	LOCK	LOCK	3+2 → 96
Reverse direction scaling & strength	3+2 → 105	LOCK	LOCK	3 → 107	LOCK	3+2 → 104	3+2 → 104	LOCK	LOCK	LOCK	LOCK	
Combined	5 → 105	LOCK	LOCK	5+3 → 107	LOCK	5+5 → 104	5 → 104	LOCK	LOCK	LOCK	LOCK	5 → 96
Change?						✓						
New status	105	*107*	*107*	107	*105*	*104*	104	*103*	*102*	*101*	*97*	96

(d) Step 8: Frame 6 is the highest scorer

Frame	1	2	3	4	5	6	7	8	9	10	11	12
Final f_0 estimate	*105*	*107*	*107*	*107*	*105*	*104*	*104*	*103*	*102*	*101*	*97*	*96*

(e) Final corrected estimates

Figure 4.15 – The Highest Scorer correction scheme (for previous example data)



Frame	1	2	3	4	5	6	7	8	9	10	11	12
f_0 estimate	105	107	107	107	419	626	630	624	102	606	97	96
connectivity	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	
confidence	LO	MED	HI	HI	MED	MED	LO	MED	MED	HI	MED	MED
Resistance threshold	4	5	6	6	5	5	4	5	5	6	5	5

(a) Initial values

	1	2	3	4	5	6	7	8	9	10	11	12
Forward direction scaling & strength		1 → 107	3+0 → 107	4+1 → 107	4+2 → 105	0 → 626	3 → 630	1+1 → 624	3+1 → 612	3 → 101	4 → 582	3 → 96
Reverse direction scaling & strength	3+2 → 105	4+2 → 107	4 → 107	3 → 428	0 → 419	1+1 → 626	3 → 630	3 → 104	4 → 612	3+1 → 101	3 → 97	
Combined	5 → 105	1+6 → 107	3+4 → 107	5 → 107	6 → 105	2 → 626	3+3 → 630	3 → 104	4+4 → 612	3+4 → 101	4 → 582	3 → 96
Change?									✓			
New status	105	107	107	107	419	626	630	624	***612***	101	97	96

(b) Step 1: Frame 9 is the highest scorer; it is wrongly "corrected"

Figure 4.16 – The Highest Scorer correction scheme (for a burst error example)

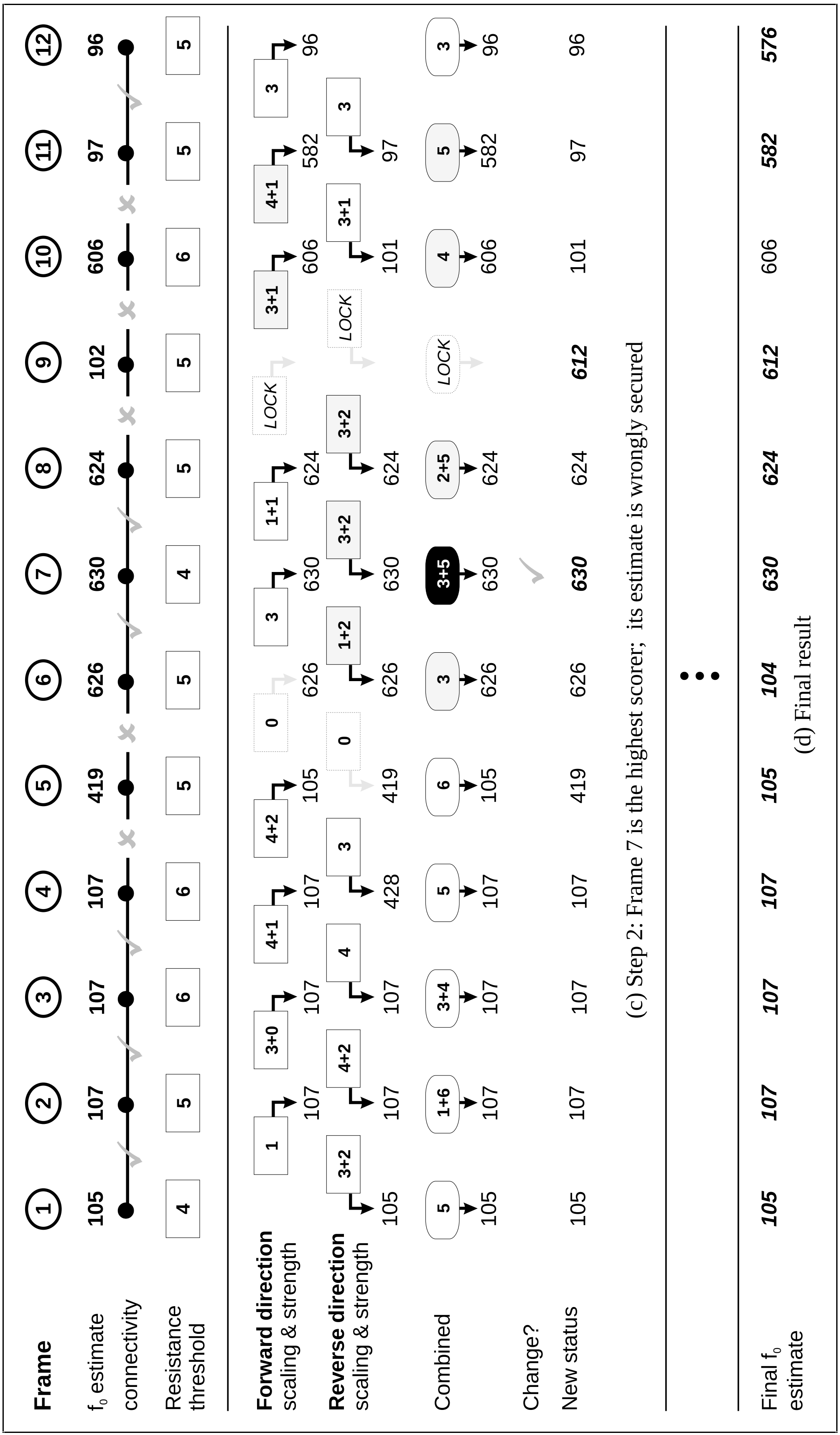


Figure 4.16 – The Highest Scorer correction scheme (for a burst error example)

Correction by Scan-and-Propagate

The problem can be reappraised as the necessity to propagate changes from one point outward, but for this to be successful, it is necessary to initiate changes from a stable base. The final correction method uses the idea of *regions of harmonicity* to achieve this. Observations of the time domain sample or the time-frequency sonogram reveal that the strength of the harmonic structure (i.e. the prominence of periodicity) tends to fade in, become strong for a time and then fade out. This is particularly true for the voiced sections of speech (see Figure 4.17a-d). In addition, speech includes sections where the harmonic strength is low for short periods, although the pitch is continuous. This can happen when there is less strength in the higher harmonics, such as when the mouth cavity is closed mid word. (For example, the “imi” in “imitate”.)

These observations led to a method that scans into a harmonic section, where there is high confidence and good connectivity and propagates the corrections outward from this point. In so doing, the fundamental frequency trajectories are often extended into the lower confidence ‘tails’ of a harmonic section. As part of the new algorithm, ‘harmonic regions’ are defined which not only clarify the extent of each fundamental frequency trajectory, but also delineate the boundaries, so that linking can be avoided in the non harmonic regions.

The algorithm scans forward until a condition for harmonicity is met (see Figure 4.17e). From this point it works in the reverse direction, correcting and locking frames and extending the region until a stopping condition is met. Then it does the same in the forward direction – correcting, locking and extending the region – starting from the same point. This completes the definition of one region. From the end of the region, the scan begins afresh, until all the harmonic sections of the sound have been mapped into regions. The parameters used for this were:

Conditions for Detection of Harmonicity

1. 2 consecutive connectable **HI**’s, or
2. 1 **HI** and 1 **MED** (consecutive connectable, in any order), or
3. 4 consecutive connectable **MED**’s.

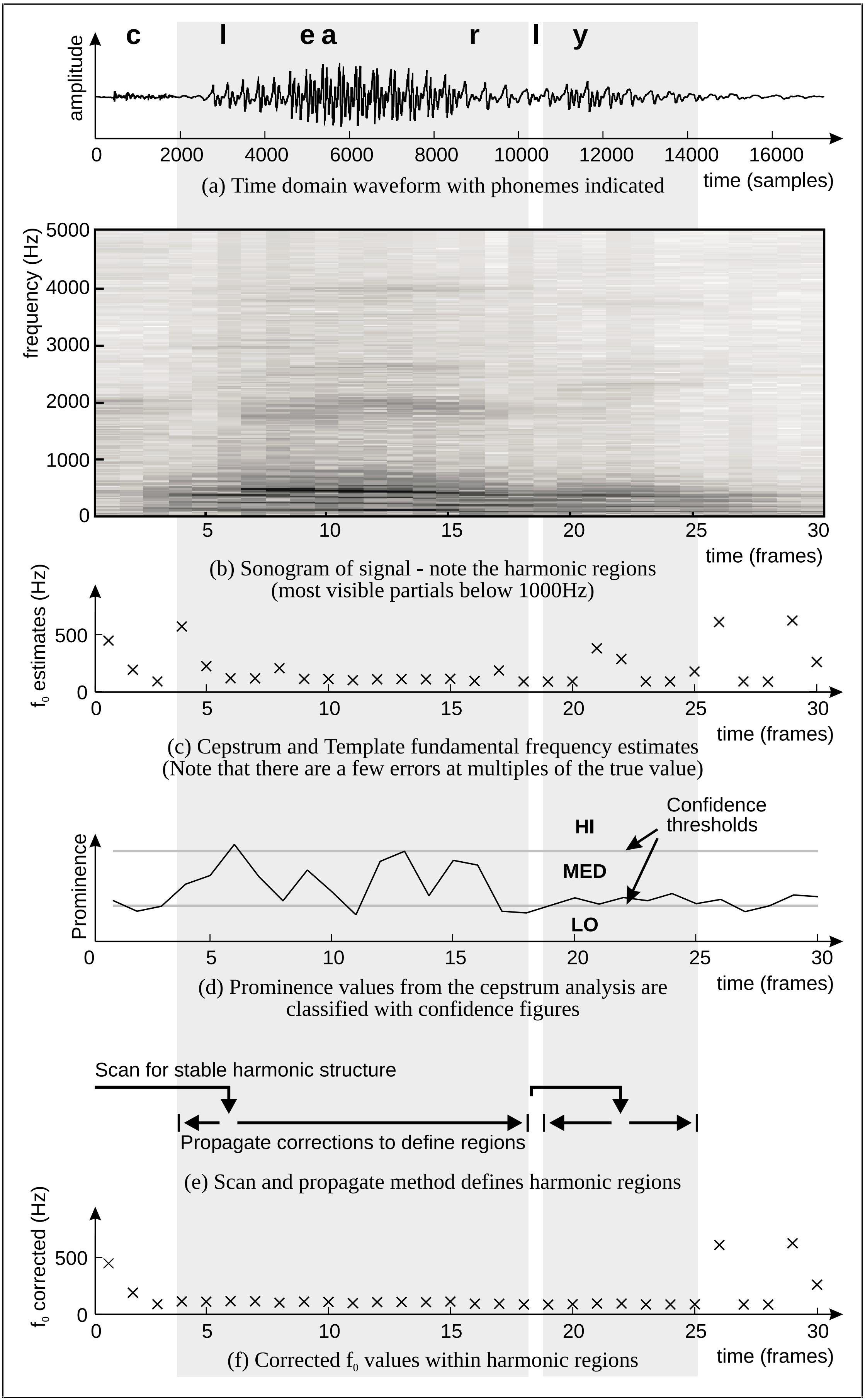
(Connectable means f_0 estimates are closer together than the maximum frequency deviation threshold.)

Conditions for Terminating a Harmonic Region

1. The first occurrence of a non connectable f_0 estimate, or
2. 2 consecutive **LO**’s.

Therefore, even frames which have a confidence of **LO** can be included within a harmonic region (to a consecutive depth of 2), so long as the fundamental frequency estimates are connectable, which improves synthesis into the low confidence region tails.

As a final step, if consecutive regions are touching (as in the example of Figure 4.17) and their corrected fundamental frequencies are connectable then the regions are merged. This situation occurs when there are two to four connectable frames of confidence **LO**, in which the first region is forced to terminate at the second **LO** (see ‘Conditions for Terminating a Harmonic Region’ above).



(a) Time domain waveform with phonemes indicated

(b) Sonogram of signal - note the harmonic regions (most visible partials below 1000Hz)

(c) Cepstrum and Template fundamental frequency estimates (Note that there are a few errors at multiples of the true value)

(d) Prominence values from the cepstrum analysis are classified with confidence figures

(e) Scan and propagate method defines harmonic regions

(f) Corrected f_0 values within harmonic regions

Figure 4.17 – Harmonic analysis of a speech sample (male voice, spoken)

Figure 4.18 (on the next three pages) illustrates the scan-and-propagate scheme for updating frames, step by step for the burst error example of Figure 4.16. In this, a pair of connectable **HI**'s (frames 3 and 4) trigger detection of a stable harmonic region. The first of these is taken as the starting point for propagating change and its estimate is secured. Using strength score values, calculated in the same way as before and updated after each frame change, the changes are first propagated backwards and then forwards. This example is too short to demonstrate conditions for terminating the harmonic region, so the first and last frames become the region boundaries by default. The final result (see Figure 4.18i) shows that all the frames' estimates have been successfully corrected into the most likely connectable sequence.

4.2.4 Matching Peaks to Partials and Linking the Frames

Each harmonic region, by definition, describes a set of consecutive frames over which the fundamental frequency has acceptable connectivity. Therefore the task of linking frames, per se, is already done. To complete the linking process however, the detected spectral peaks must be matched to the harmonic structure (or discarded), within each frame. The process reduces to labelling peaks with a partial index, after which they are automatically linked to peaks with the same index in adjacent frames (as shown earlier in Figure 4.2), or faded in/out if no link is made.

A simple matching process was developed which searches for peaks within a tolerance band around the integer multiples of each fundamental frequency. If more than one peak satisfies the criterion, the one which is closest to the target frequency is designated as the partial.

In the implementation of this algorithm, provisions were made for a more sophisticated matching method, which would take into account not only frequency location, but also amplitude and the closeness to other competing peaks. However, the results from 'closest frequency' matching were good enough not to warrant the extra complexity.

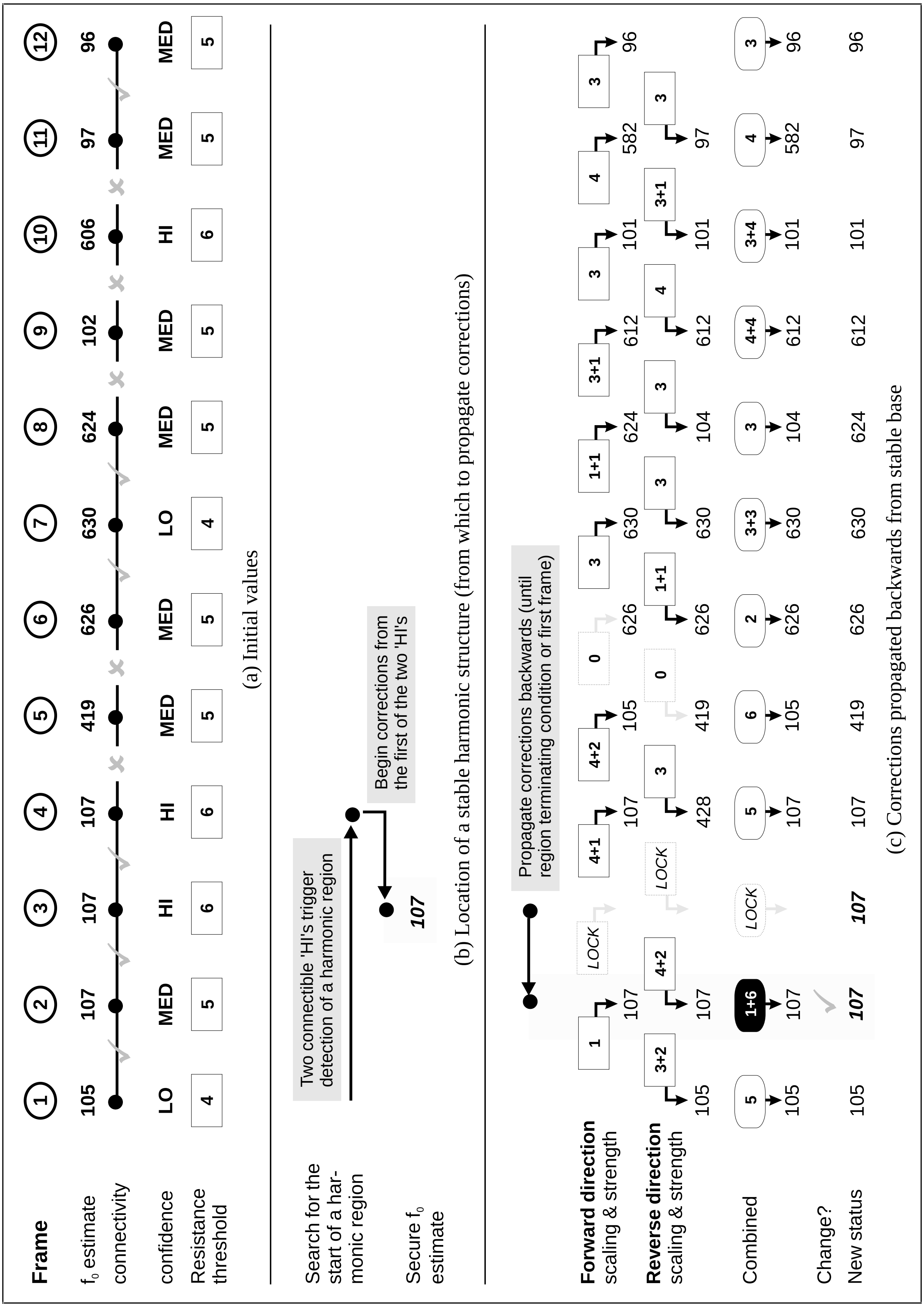
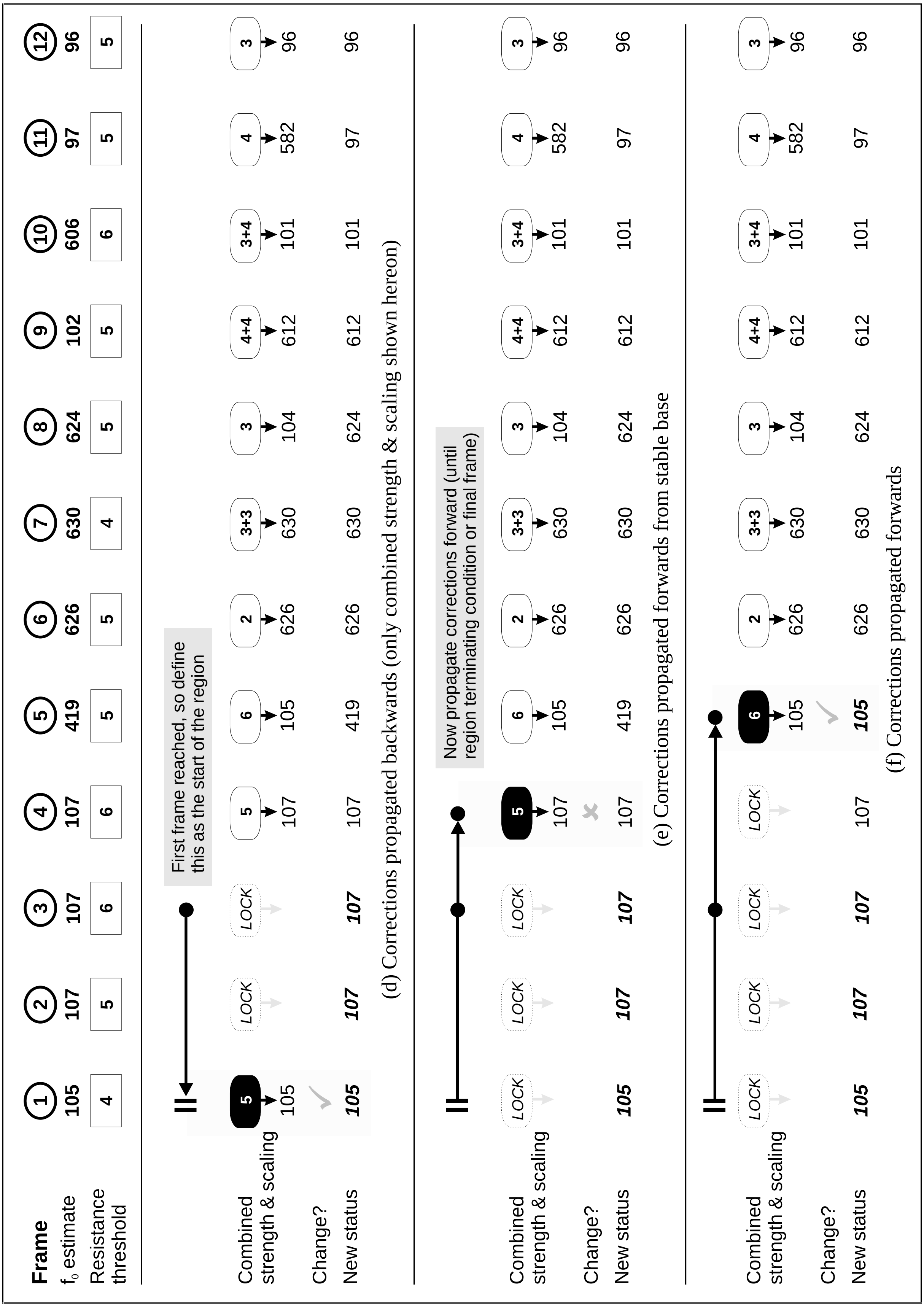

Frame
1 2 3 4 5 6 7 8 9 10 11 12
f₀ estimate
105 107 107 107 419 626 630 624 102 606 97 96
connectivity
confidence
LO MED HI HI MED MED LO MED MED HI MED MED
Resistance threshold
4 5 6 6 5 5 4 5 5 6 5 5
(a) Initial values
Search for the start of a harmonic region
Two connectible 'HI's trigger detection of a harmonic region
Begin corrections from the first of the two 'HI's
Secure f₀ estimate
107
(b) Location of a stable harmonic structure (from which to propagate corrections)
Propagate corrections backwards (until region terminating condition or first frame)
Forward direction scaling & strength
1 LOCK 4+1 4+2 0 3 1+1 3+1 3 4 3
107 107 105 626 630 624 612 101 582 96
Reverse direction scaling & strength
3+2 4+2 LOCK 3 0 1+1 3 3 4 3+1 3
105 107 428 419 626 630 104 612 101 97
Combined
5 1+6 LOCK 5 6 2 3+3 3 4+4 3+4 4 3
105 107 107 105 626 630 104 612 101 582 96
Change?
New status
105 107 107 107 419 626 630 624 612 101 97 96
(c) Corrections propagated backwards from stable base

Figure 4.18 – The Scan-and-Propagate correction scheme (for the burst error example)



Frame	1	2	3	4	5	6	7	8	9	10	11	12
f_0 estimate	105	107	107	107	419	626	630	624	102	606	97	96
Resistance threshold	4	5	6	6	5	5	4	5	5	6	5	5

First frame reached, so define this as the start of the region

	1	2	3	4	5	6	7	8	9	10	11	12
Combined strength & scaling	5 → 105	LOCK	LOCK	5 → 107	6 → 105	2 → 626	3+3 → 630	3 → 104	4+4 → 612	3+4 → 101	4 → 582	3 → 96
Change?	✓											
New status	105	107	107	107	419	626	630	624	612	101	97	96

(d) Corrections propagated backwards (only combined strength & scaling shown hereon)

Now propagate corrections forward (until region terminating condition or final frame)

	1	2	3	4	5	6	7	8	9	10	11	12
Combined strength & scaling	LOCK	LOCK	LOCK	5 → 107	6 → 105	2 → 626	3+3 → 630	3 → 104	4+4 → 612	3+4 → 101	4 → 582	3 → 96
Change?				✗								
New status	105	107	107	107	419	626	630	624	612	101	97	96

(e) Corrections propagated forwards from stable base

	1	2	3	4	5	6	7	8	9	10	11	12
Combined strength & scaling	LOCK	LOCK	LOCK	LOCK	6 → 105	2 → 626	3+3 → 630	3 → 104	4+4 → 612	3+4 → 101	4 → 582	3 → 96
Change?					✓							
New status	105	107	107	107	105	626	630	624	612	101	97	96

(f) Corrections propagated forwards

Figure 4.18 – The Scan-and-Propagate correction scheme (for the burst error example)

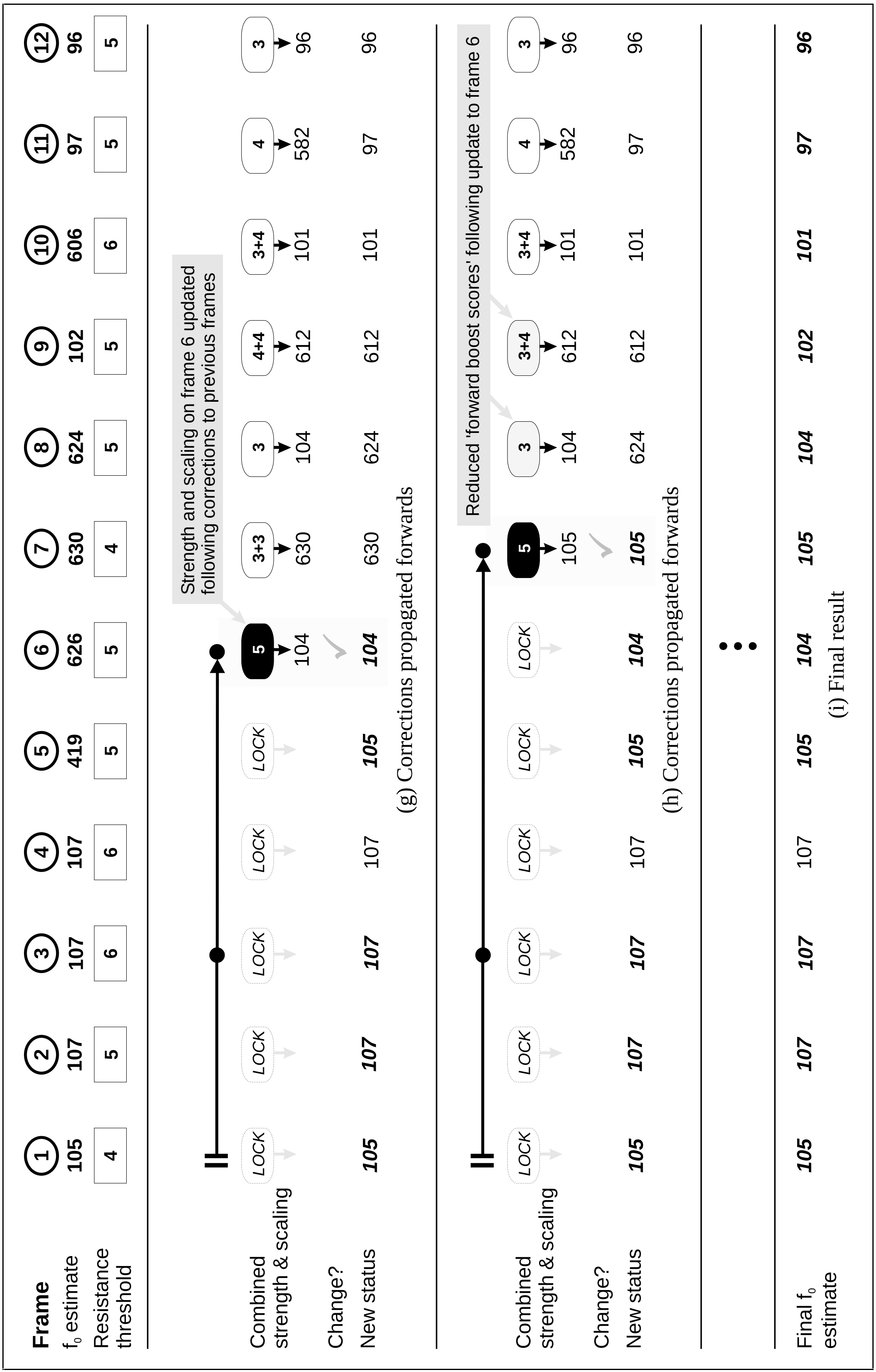


Figure 4.18 – The Scan-and-Propagate correction scheme (for the burst error example)

4.3 Performance of the Technique

Section 4.1 set out to show why the harmonic structure is a necessary component of the peak linking strategy. Section 4.2 has presented an implementation, based around the hybrid of two popular fundamental frequency estimation techniques, by way of justifying this claim. In this section the results of the scheme are given, first in terms of the subjective impact and then objectively from a signal representation standpoint. Note that the cello example of Figure 4.3e was generated by the method presented.

4.3.1 Subjective Results

Comparison with the Nearest Frequency Method

As demonstrated in Figure 4.3c earlier, the Nearest Frequency method of linking peaks leads to a scrambled spectrum at higher frequencies, where the distortion artifacts are more prevalent. If short trajectories are removed as a way of rejecting unstable components, this results in an apparent low-pass filtering effect. With the harmonic linking method there is no such restriction and the synthesised sound retains the ‘openness’ of the original.

The Nearest Frequency method was also found to be prone to susurrus artifacts from the deterministic synthesis, which resulted from erroneous peak links that managed to form trajectories long enough to be accepted, or from legitimate trajectories that occasionally linked to a ‘bad’ peak. In the harmonic linking method, only peaks that are roughly harmonic are linked, so even when there are errors, the results *sound* right.

Overall Quality

The harmonic synthesis, like the deterministic, suffers from a metallic quality, that is probably the result of the smoothly interpolating, over-simplified partial trajectories. This is largely reduced with the addition of the stochastic aspect, but as with the deterministic-stochastic classification, the harmonic-stochastic classification does not fuse perfectly. However, this deficiency is discussed in detail in section 6.5.

Occasionally the method can lock onto a multiple of the true fundamental frequency. During such bursts a fraction of the partials are therefore correctly identified and linked. For voice, the effect is similar to falsetto speech. Fortunately the stochastic aspect synthesises the missed partials (although with random phase) and fills in the spectral holes sufficiently that the final synthesised sound is only minimally degraded.

4.3.2 Objective Performance

The two-stage structure of the analysis: in-frame and inter-frame has resulted in some benefits to the linking process. The algorithm can distinguish between weak harmonic structure and random correlations, by virtue of the scan-and-propagate method as it tracks fundamental frequency continuity into the harmonically weak frames that are usually at the ends of pitched sounds.

4.3.2.1 Matching the Higher Partials to Peaks in the Spectrum

In one sound example, of Tori Amos singing a note with a fundamental frequency of about 350Hz, the sonogram clearly revealed the presence of partials as high as 16kHz. (Indeed the sudden absence of higher partials would suggest that the upper frequency limit was attributable more to the recording and mastering processes, than to her singing voice.) This demonstrates that the algorithm needs to be capable of matching in the order of fifty partials.

In the algorithm, peaks are identified as partials by their closeness (in frequency) to a multiple of the estimated fundamental frequency. Errors in this estimate become similarly scaled with the partial index and even a small error can result in gross mismatches for the higher partials.

The problem in assessing the algorithm's accuracy is that there is no fixed definition for the true value of the fundamental frequency, where the partials are not exactly harmonic. So there is no way of measuring estimate error with any degree of certainty. Any number of mathematical measures could be used. The first partial (the fundamental itself)? Or a least squares fit between partials? Or the mean of frequency differences between partials? Any one of many statistical bases could be used to provide a variety of answers. Possibly the most useful value for assessment purposes is the value that allows for the greatest number of correctly identified partial peaks. Obviously, this requires foreknowledge of which peaks are partials – not always an easy task, even by eye.

Assuming that there will be some small error in the estimate, the step of matching partials to peaks becomes less reliable as the partial order increases. First there is the chance of matching a spurious peak as a partial. For the higher partials it is even possible to match a true partial peak to the wrong partial index.

In light of this, it must be expected that gross errors are being made by the algorithm. However, errors due to mismatches are not readily apparent in the synthesised sound. There are two possible reasons for this: firstly, most of the waveform energy is usually contained in the lowest 5-10 partials which are unlikely to be mislinked, so the differences between source and synthesised sounds should be subtle; secondly, even where there are errors, links are made that follow a roughly harmonic trajectory, so the disparity is likely to be small.

4.3.2.2 Matching Partials to Noise Peaks

Once the algorithm has detected a harmonic structure and decided upon an estimate for the fundamental frequency, any peak lying close (and closest) to a multiple of the fundamental will be matched as a partial. This is very useful where distortion has caused a partial peak to display multiple maxima, because only a single maximum (per peak) will be selected, representing a reduction in artifacts, at the cost of slight inaccuracy.

However, due to resonances in the sound creation mechanism (e.g. formants of the human voice), the partials are usually prevalent above the noise floor for part of the spectrum. So it is quite conceivable that the algorithm matches some partial indices to peaks that result from noise variations. It is not possible to verify or deny this situation in many cases, because the same spectral shape in the STFT could have been created by slowly varying noise or a distorted rapidly varying partial. Until a time-frequency representation is developed that can make the distinction, it is unlikely that this situation can be corrected.

Despite this potential flaw, the results bear close resemblance to the source sound. Possibly, in the cases where noise maxima are mistaken for partial peaks in the deterministic analysis, the stochastic analysis 'fills in the gaps', thus re-randomising those regions of the spectrum.

4.3.2.3 Problems with Reverberation and Extraneous Noises

The cepstrum and template methods were chosen with polyphonic fundamental frequency estimation in mind. Multiple source estimation is a target for the sound model, but this is seen as a future upgrade and has not been implemented in the present system. Therefore the initial fundamental frequency estimate in each frame, the cepstrum estimate, is looking for a *single* prominent peak, from which to identify the period of a signal. When there are extraneous noises (which might for example include wind noise or a more pitched sound, like traffic noise, on an outside voice recording), this can appear as a second conflicting pitch, either momentarily or continuously. Reverberation has a similar effect, where the echo of a previous pitch appears in the spectrum concurrently with the new pitch (as might occur on a concert recording of any orchestral instrument).

If the desired signal is strong, secondary pitches can weaken the prominence calculation, but generally this does not pose a problem. It is when the desired signal is weaker that the conflict arises. Quite often in these situations, the signal sources (both the desired and the secondary sources) have some amplitude variation, so they take it in turn to 'win' in the cepstrum analysis. This causes the frame-by-frame estimates to jump between the two pitches, depending on which is stronger in a particular frame. Naturally, this toggling between the two values destroys the desired continuity of the synthesised sound. Therefore, at present the algorithm is only suitable for clean, single source (monophonic) recordings, such as a voice recording in a studio. The question of upgrading to polyphonic estimation is discussed in chapter six (section 6.2.2.2).

In summary…

The harmonic linking algorithm successfully defines regions which extend into the uncertain 'tails' of voiced sections of speech, thereby extending the range over which pitch can be perceived. It captures partials high in frequency, where the spectral peaks are not well-defined. Occasionally, it locks onto a multiple of the true fundamental, and in these cases there are artifacts caused by stochastic randomising of the missed partial peaks, but even here the sense of pitch is predominant. The greatest asset is the sense of continuity in the synthesised sound.

The implementation is by no means perfect, but compared to the nearest frequency method, it represents a vast improvement. There is a great reduction in audible artifacts, without any low-pass filtering effects. The intention behind this chapter has been to present a methodology that demonstrates the potency of frame linking via the harmonic structure, and in this respect the algorithm has proved a total success.

CHAPTER FIVE

IMPROVED SYNTHESIS OF ATTACK TRANSIENTS

5. IMPROVED SYNTHESIS OF ATTACK TRANSIENTS

5.1 Problem Description and Solution Definition

In section 2.1.4, the amplitude envelope was discussed, in particular the initial transient at the onset of a note, the 'attack' stage. It was recognised that the attack plays a vital role in our perception of timbre. It was also noted that the attack of a percussive sound is short and contains rapidly changing elements, which makes it difficult to study.

It is therefore no surprise that attacks are not well understood and are not well represented within the analysis-resynthesis model. Both deterministic and stochastic aspects of the Initial Model assume a gradually evolving spectrum. There is no method for dealing with rapid spectral changes, so the model fails to capture sufficient detail of short transient events. Drum sounds, clicks and tapping sounds become diffuse or dissolved, or even cause momentary drop-outs.

The problem was recognised by Serra, who proposed a method of storing the original attack and the transient portion immediately following it, as a time domain sample [Serra,1989,pp.128-131]. This would be spliced with the (phase-correct) synthesis of the rest of the sound. The suggestion was an option offered but not explored: no method was proposed for detecting attacks and there was no consideration of parameters that might be required.

Although this provides a solution to the problem of synthesis with a realistic attack, it also raises problems of its own. What should be done upon synthesis with time-stretch? Should the sample be inserted unchanged, and the remainder of the sound over-stretched to compensate? Upon, pitch-shift, should the sample be resampled to a new pitch, with the inevitable alteration in duration? It also signifies a dead-end for potential future improvements, because there has been no capture of features, beyond direct capture of the waveform. Conceptually it is unsatisfactory because it represents a different storage format, outside the model, and makes holistic transformation awkward.

In this chapter, the aim is to find a solution that incorporates attack transients into the model. The problems of diffusion and drop-outs are examined to ascertain the causes. This information is coupled with observations about the signals under analysis, to generate a strategy for avoiding the problem situation.

5.1.1 The Disparity between Attack Transients and their Representation

Observable Properties of Attack Transients

Despite the fact that all signal analysis for the model is based on the time-frequency representation, the most useful observations of attack transients can be gained from the time domain waveform and foreknowledge about the sources of percussive sounds:

1) Percussive sounds have a short attack – the amplitude function rises rapidly – causing an abrupt change to the signal;

2) The attack waveform results from a new sound source, therefore the new sound element is unconnected to the preceding sound signal, and cannot be predicted from it;

3) During and immediately following the attack, the waveform can include chaotic elements, but they often stabilise quickly;

4) Short attacks are generated from plosive or impulsive events, where the greatest concentration of energy is present during the attack and the signal decays thereafter.

Spectral Analysis for the Deterministic Plus Stochastic Model

The first step in the analysis process is the time-frequency representation, which is calculated using the STFT. Although each frame of the STFT is assumed to represent the instantaneous spectrum of the signal under analysis, in reality it is calculated using the FFT from a finite, short-time window on the waveform. The length of this window determines the time and frequency resolutions, which are inversely related. The practical constraint of separating partials in the frequency domain forces a preference for good frequency resolution, at the expense of time resolution. (A more comprehensive description of the analysis process can be found in section 2.3.2.)

The central assumption when using the FFT is that of stationarity – the spectrum of the signal is assumed to be unchanging within the analysis window. Small deviations from this generate a small, but tolerable, amount of distortion. This results from the averaging effect of the FFT, which gives a smoothed magnitude spectrum. The combination of poor time resolution and the requirement for the waveform to approximate stationarity restricts (good quality) analysis to slowly evolving signals.

Analysis of Attack Transients Using the Deterministic Plus Stochastic Model

The first observed property (above – percussive sounds have a short attack) describes an event which is highly localised in time. This is directly in contrast with the FFT's assumption of stationarity. The FFT attempts to describe the sudden change as a stationary, periodically repeating feature, whose fundamental is the length of the analysis window. Hence there appear to be components throughout the spectrum, which correspond to the harmonics of this interpretation. These can obscure the true partial peaks or they can contain additional spectral peaks themselves, that result in drop-outs and artifacts upon deterministic analysis and resynthesis.

Compounding this, the second observed property describes a situation in which the waveform preceding the note onset is largely uncorrelated with the waveform following its initiation[1]. In this situation, the smoothing function of the FFT will produce a con-fused average of both spectra.

During calculation of the STFT, the analysis window advances a fixed hop-distance between frames. The hop-distance is a fraction of the window-length – usually a half or a quarter –

[1] The sound preceding the note onset may continue, so the spectra may not be *totally* uncorrelated.

which guarantees that the above problems *will* occur, if the source signal contains an attack transient. Figure 5.1 shows an example.

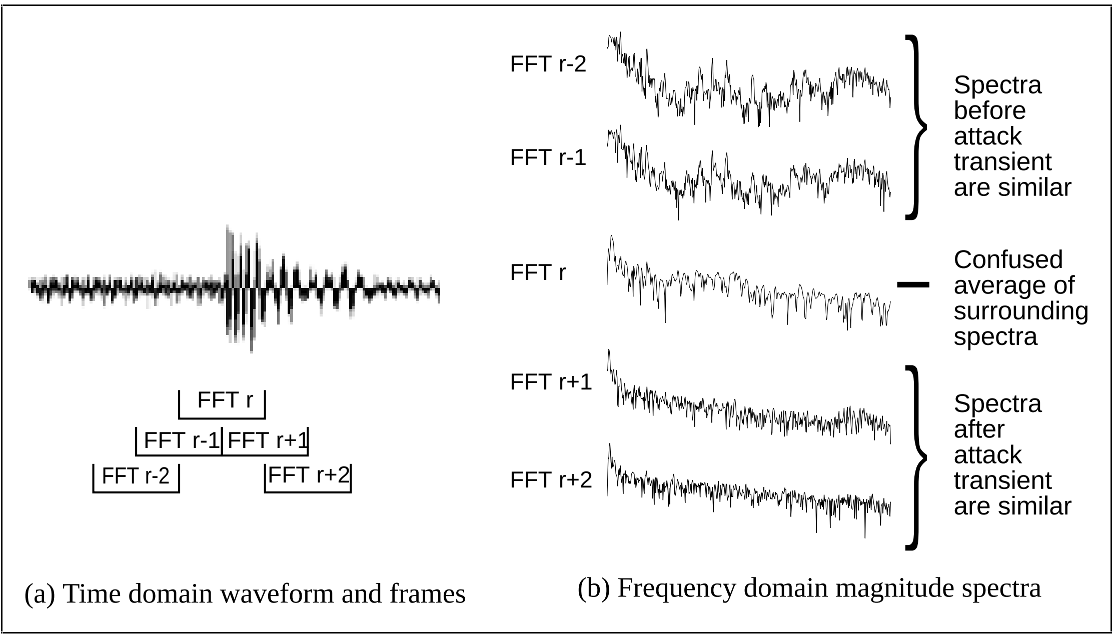


Figure 5.1 – The 'confused' spectra surrounding an attack transient

Even if the deterministic analysis were error free, attack transients would be diffused upon synthesis because of the way that frames are linked; see Figure 5.2. In the deterministic analysis, partial trajectories are constructed by linking peaks between frames. Where links are not possible, during synthesis the partials are slowly faded in/out over a frame's duration. In the stochastic analysis, each spectral envelope represents an averaged snapshot of the spectrum, and upon synthesis, there is smooth crossfading between frames.

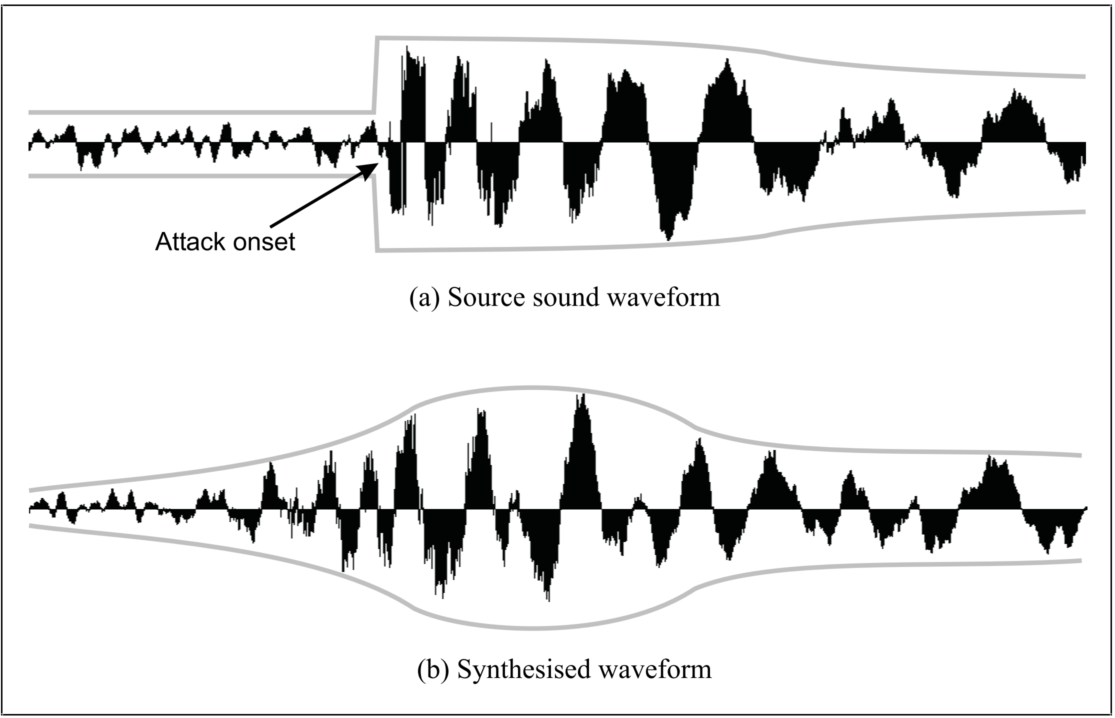


Figure 5.2 – Attack transients sound 'diffused' upon synthesis

5.1.2 Modifications to the Model to Incorporate Attack Transients

The solution, presented in the following sections, synchronises the analysis and synthesis processes to the percussive note onsets. This prevents the analysis window from straddling any events (which avoids 'confused' spectra) and forces an abrupt change at these locations during synthesis (to overcome the diffusion from slow fades). The method involves:

1) Detecting the presence of a transient event and locating its onset – the *event boundary*;
2) Synchronising the 'hopping' analysis window to each event boundary, such that the spectra either side are never simultaneously captured;
3) Disallowing any peak linking (deterministic) or spectral interpolation (stochastic) across the event boundary;
4) Performing a fast crossfade at the event boundary upon synthesis, to retain the abruptness of the original sound.

Implementation of the proposed solution requires four changes to the model, corresponding to the four method steps above

1) Addition of a pre-analysis scan, to identify the event boundaries and compile them into a 'transient event region (TER) list';
2) Bind both analysis processes to the region list, such that each region is considered independently;
3) Extrapolate the first and last frames of each region toward the region boundaries – this 'fills in the gap' so that peaks/spectral envelopes, that would previously have been slowly faded out/in, are extended up to the region boundaries;
4) Manage the fast crossfade at the region boundaries.

These are presented graphically in Figure 5.13 on page 149. Figure 5.14 on page 153 shows the impact upon the model structure.

Sections 5.2 and 5.3 following describe the implementation of the above steps. Section 5.2 describes and compares three methods for determining the location of transient event boundaries (step 1). Section 5.3 discusses implementation details for synchronising the analysis and synthesis processes to those boundaries (steps 2-4).

5.2 Detection and Location of Transient Events

Initially, a frequency domain detection method was developed, which observed short term shifts in energy localisation; this is the algorithm published in [Masri & Bateman,1996]. The frequency domain was chosen because of its ability to show changes in the spectrum even if they are not accompanied by significant changes in the time domain envelope. It was also convenient to employ a technique that utilised existing algorithms within the model (i.e. STFT) because it would be implemented quickly.[2]

Through work on a separate project by the author, two alternative detection methods were created. The first, a time domain envelope-based method, was developed for its simplicity and speed of programming and operation. The second was developed as a method for gauging temporal complexity, using the spectral dissimilarity between frames as a measure of how rapidly the spectrum was changing. In itself, this is also a viable method for attack transient detection.

All three methods (energy distribution, attack envelope and spectral dissimilarity) are presented in the next subsections, followed by a discussion and comparison of their relative merits. In all cases the detection scheme would be implemented as a pre-analysis scan that generates a list of attack detections and their sample locations. This information would then be used to synchronise the analysis and synthesis (to be discussed later, in section 5.3).

5.2.1 Detection by Energy Distribution[3]

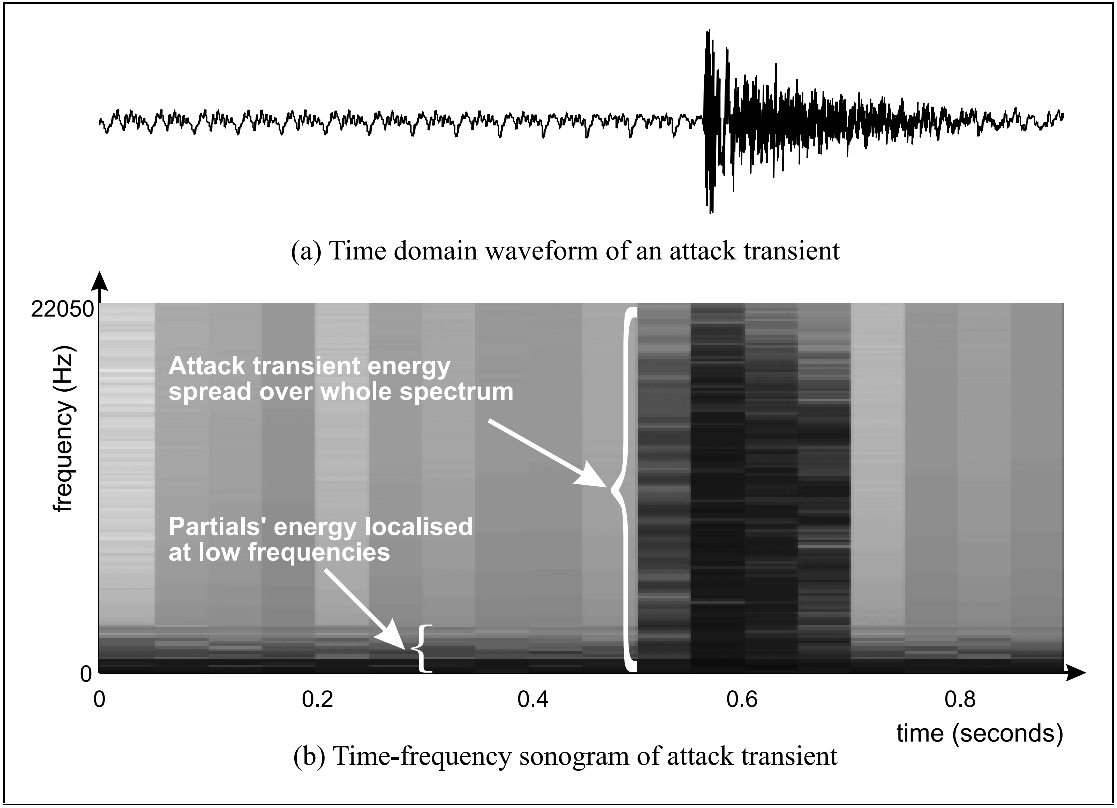
The Energy Distribution method bases its detection function on two of the observed properties considered in section 5.1.1: the suddenness of the onset (the first observation) and the increase in energy (the third observation). A ‘hit’ (a positive detection) should be registered by the detection function when there is a significant rise in energy, coupled with an increased bias towards the higher frequencies.

The test for a rise in energy is justified by the third observation. The latter test is based on the expectation that a sudden change to the waveform will introduce a phase discontinuity (or at least discontinuities in the derivatives of phase) at the point where the new sound source impacts the waveform. This discontinuity would be represented in the FFT with increased energy throughout the spectrum. The lower frequencies tend to contain the most dominant partials, so the increase in energy would be more prominent at higher frequencies. The end result is therefore a sudden overall shift in energy toward the higher frequencies. (An alternative explanation is that the discontinuities would cause increased spectral leakage,

[2] It should be noted that some sophisticated methods for note onset detection have been developed using the wavelet transform [Grossman & Holschieder et al.,1987; Solbach & Wöhrmann,1996; Tait & Findlay,1996]. These were not considered here because of the computational expense of calculation and the increase in system complexity that would arise from implementing an additional time-frequency representation structure.

[3] In order to establish the efficacy of the complete method (outlined above in 5.1.2), the Energy Distribution detector was created, as a simple and effective solution. It was published in [Masri & Bateman,1996] and has been included here for the sake of completeness, in comparison with alternative methods devised since publication.

similar to an FFT of a signal which has a discontinuity between the ends of the window, that would inevitably result in an increase in amplitude at the higher frequency bins.) See Figure 5.3.


(a) Time domain waveform of an attack transient

(b) Time-frequency sonogram of attack transient

Figure 5.3 – The sonogram reveals a large increase in energy throughout the spectrum for an attack transient (bass drum)

5.2.1.1 Implementation

The pre-analysis scan is performed using the STFT, although with a considerably shorter time window, for increased time resolution. Naturally, this compromises the frequency resolution, but a detailed frequency breakdown is not required, merely an indication of the trends. It was found through experimentation that there was a natural lower limit to the window-length; settings below this would result in false hits, due to the increased significance of spectral leakage (particularly where the window-length was much smaller than a single period of the signal). The hop-distance was fixed at half the window-length, so that every transient event would be captured clearly (i.e. toward the centre of the window) at least once.

Since frequency resolution was not an issue, zero padding was not necessary. Because phase was also unimportant, the window-length was set to a power of two, for easy calculation of the FFT. In summary, useful values for a sound sampled at 44.1kHz were found to be:

Window-length[4] = 512 samples (≈12ms),
Hop-distance = 256 samples,
Window function = Hamming,
No zero padding.

In each frame, the Energy and High Frequency Content (HFC) functions are calculated. The detection function uses values of these from consecutive frames to generate a Measure of Transience (MoT). This is compared with a threshold, a value above which registers a 'hit'.

Construction of the Detection Function

The energy function is calculated as the sum of the amplitude squared for the specified frequency range:

$$E = \sum_{k} \left\{ \left| X(k) \right|^{2} \right\} \tag{5.1}$$

where *E* is the energy function for the current frame,
k is the frequency bin index (for the FFT array),
X(*k*) is the *k*th bin of the FFT.

The function to measure high frequency content was set to a weighted energy function, linearly biased toward the higher frequencies:

$$HFC = \sum_{k} \left\{ \left| X(k) \right|^{2} \cdot k \right\} \tag{5.2}$$

where *HFC* is the High Frequency Content function for the current frame,
(all other symbols defined as above).

Both equations include a summation for all *k*; in practice the lowest two bins were discarded to avoid unwanted bias from the DC component (hence the lower limit is actually *k*=2). Since the analysis window was sometimes shorter than a single period of the lowest frequency waveform, there was a real chance of a significant DC component.

The detection function combines the results from each pair of consecutive frames thus:

$$MoT_{r} = \frac{HFC_{r}}{HFC_{r-1}} \cdot \frac{HFC_{r}}{E_{r}} \tag{5.3}$$

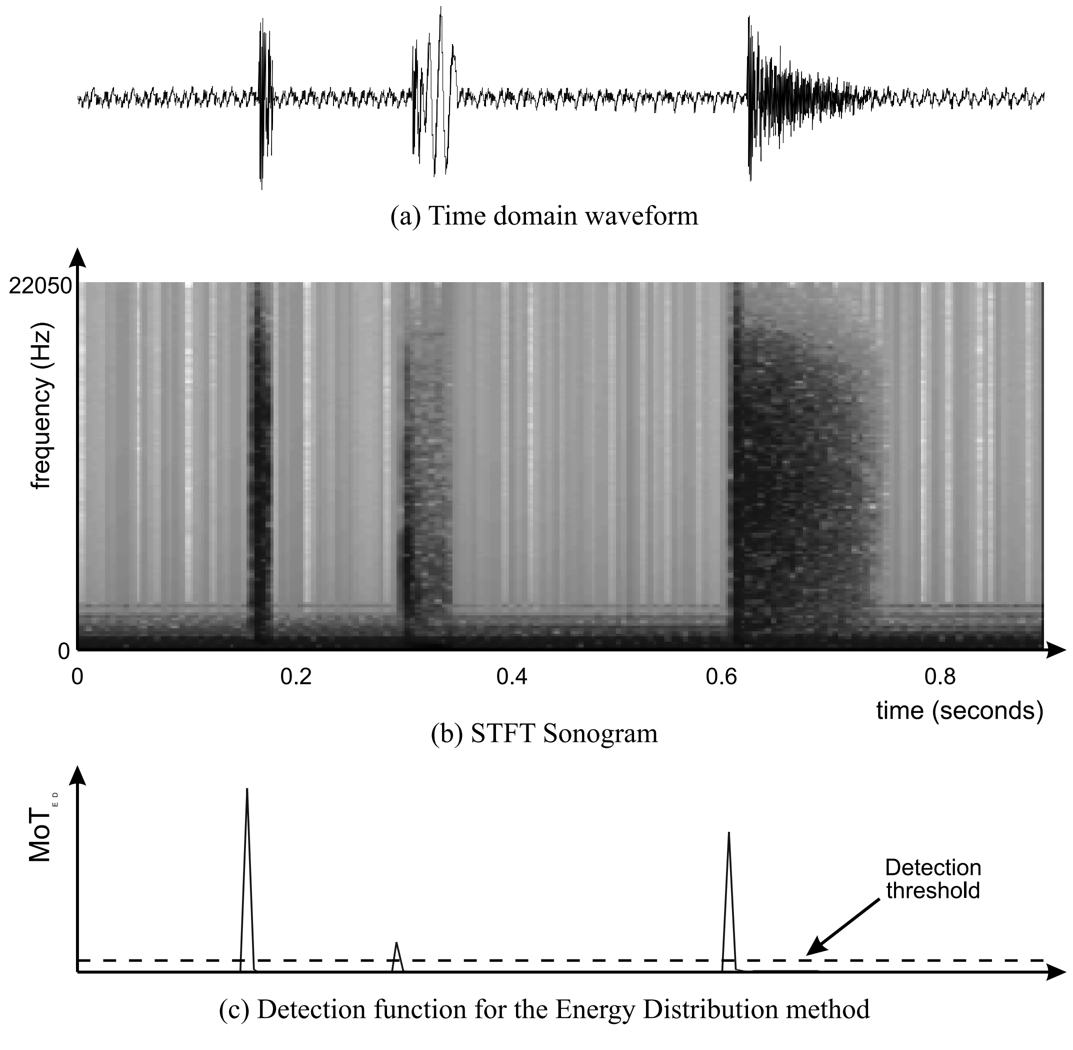
$$\text{Condition for Detection: } MoT_{r} > T_{D,ED} \tag{5.4}$$

where subscript *r* denotes current frame (equals latter of two in detection function),
subscript *r*-1 denotes the previous frame,
MoT_r is the Measure of Transience at the current frame,
$T_{D,ED}$ is the Detection Threshold (for the Energy Distribution method), above which a

[4] In the published paper [Masri & Bateman,1996], the window-length of 128 was used. After further testing it was found that the threshold for the detection function was not stable enough, due to the bias introduced for sounds with low frequency components. For greater stability, the window-length was revised upwards (with the inevitable loss in resolution).

hit is registered,
$HFC_{r\text{-}1}$ and E_r are constrained to have a minimum value of one, to avoid the potential 'Divide by zero' computation error.

The detection function is the product of the rise in high frequency energy between the two frames and the normalised high frequency content for the current frame. The first part of the product can easily be interpreted as the detector for sudden increases in high frequency energy. The second part of the product ensures that a hit is only registered when there is significant high frequency energy, thus preventing accidental triggering on large *relative* increases for low level *absolute* values. See Figure 5.4 for an example showing a sequence of three drum hits.



(a) Time domain waveform

(b) STFT Sonogram

(c) Detection function for the Energy Distribution method

Figure 5.4 – Energy Distribution detector (for a sequence of three drums)

For attacks whose growth is slightly slower, but whose onset is nevertheless sudden, it is possible for the detection function to be triggered on more than one frame. To avoid multiple detections, the algorithm was given a parameter for the minimum closeness of two hits. In practice, setting this to 2 frames was adequate for the majority of sounds (i.e. only disallowing consecutive hits).

5.2.1.2 Time Resolution

The accuracy, in time, of the detection process is equal to the hop-distance. The above values of STFT parameters give a resolution of 5.8ms, which is comparable to the accepted resolution of the ear, of 2-10ms.

The *transient event boundary* – the point of onset of the transient – is stored as the start of the second analysis window of the detection pair. It is the second window that contains the event, so this placement ensures that the whole of the attack is located after that point. In this way, any errors in the deterministic analysis, caused by inaccuracy within this process, are confined to the frame containing the percussive onset, where the suddenness of the attack should dominate perceptually.

5.2.2 Detection by Attack Envelope

The Attack Envelope method is conceptually the simplest approach. It operates in the time domain, calculating the envelope of the waveform and detecting when there is a sudden rise. It has the inherent advantage over frequency domain methods that its resolution can (theoretically) be a single sample.

The envelope is calculated using a peak follower algorithm that mimics the behaviour of an ideal capacitive voltage peak follower circuit. A detection is registered when the present output of the peak follower is a certain ratio greater than the preceding output. This method has the additional option of estimating the length of the chaotic region following the attack, before resetting the detector, so that the minimum distance between hits can be adaptive. Figure 5.5 shows the peak follower output for the example in Figure 5.3a.

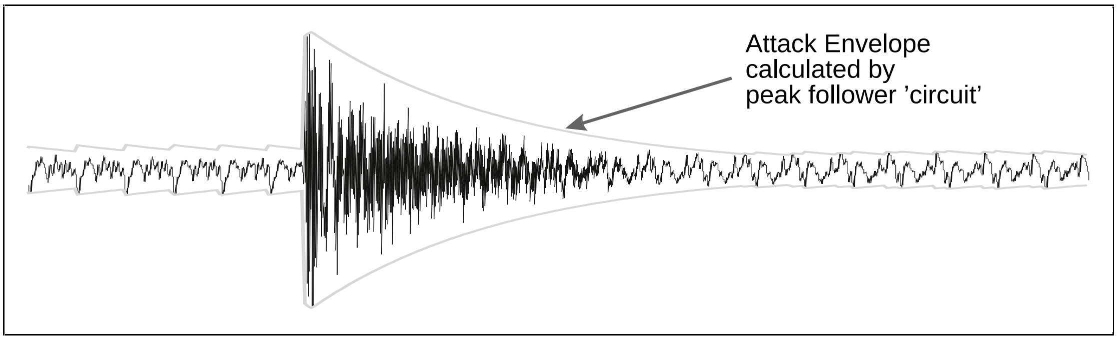


Figure 5.5 – An attack transient with the time domain envelope overlaid

5.2.2.1 Implementation

Block-based Scan

For practical reasons of computational speed, the peak follower is set to consider samples in blocks of 50 (≈1.1ms @ 44.1kHz sample rate). Prior to the scan, the waveform is divided into these (consecutive, non overlapping) blocks and the maximum absolute value is passed to the detection algorithm:

$$y(n) = \max_{m=0}^{m=M-1}\left\{\left|x(nt+m)\right|\right\} \tag{5.5}$$

where $x(t)$ is the sample sequence (starting with t=0),
$y(n)$ is the sample-block sequence that is passed to the detection algorithm,
m indexes samples within each block,
M is the block size (50, in this case)

The value of 50 was chosen as the block size because it enables a significant computational saving, whilst retaining a more-than-sufficient degree of accuracy.

Detection

At each sample-block, the peak follower is updated to the new sample-block value if that is higher, or a fixed proportion (close to 1) of its old value, so as to simulate exponential decay:

$$P(n) = \max\left\{ y(n), P(n-1) \times k_{DECAY} \right\} \tag{5.6}$$

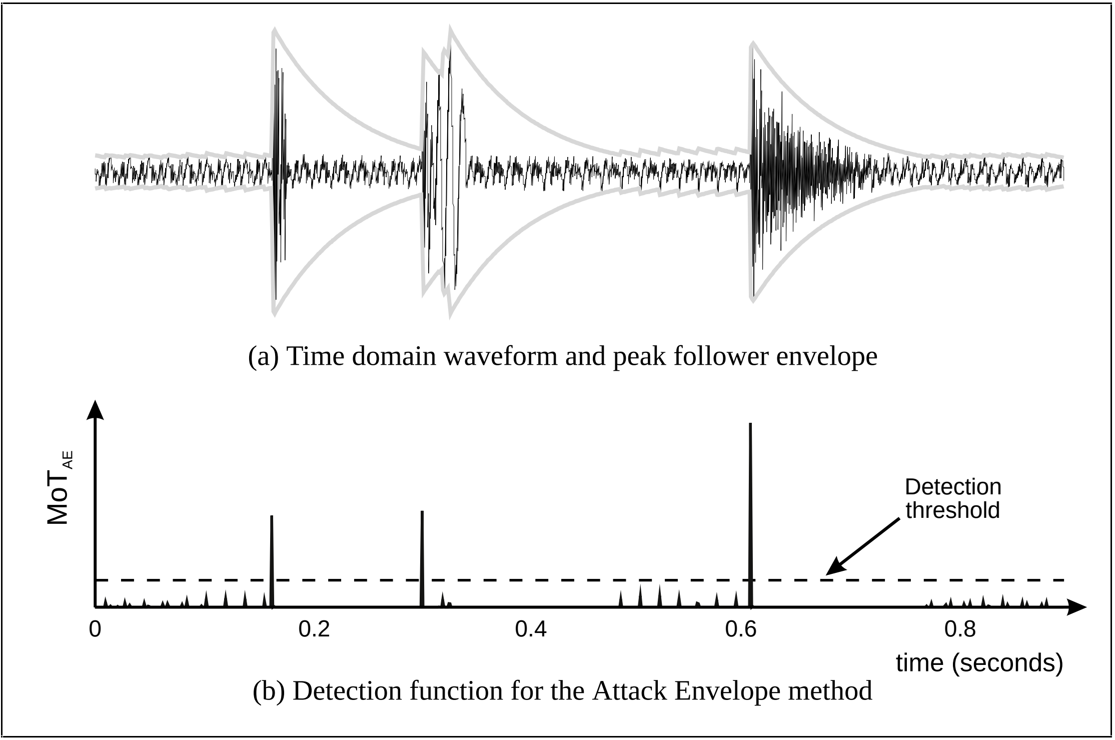
where $P(n)$ is the peak follower output for sample-block n,
k_{DECAY} is the decay factor,
(all other symbols defined as above).

A detection is registered when:

$$\frac{P(n)}{P(n-1)} > T_{D,AE} \tag{5.7}$$

where $T_{D,AE}$ is the Detection Threshold (for the Attack Envelope method), above which a hit is registered,
(all other symbols defined as above).

See the example in Figure 5.6 (which uses the same sound example as Figure 5.4).



(a) Time domain waveform and peak follower envelope

(b) Detection function for the Attack Envelope method

Figure 5.6 – Attack Envelope detector (for a sequence of three drums)

Following a detection or at the start of the scan, it is possible for the detector to falsely trigger on the first few samples, simply because the peak follower has been reset and may initially lie in an insignificant envelope trough. To avoid this, *P* is initialised to the maximum of a train of sample-blocks, starting with its initialisation location. By experimentation a value of 10 blocks (equal to 500 samples) was found reasonable – larger values can cause the detector to miss a true attack, whilst smaller values are insufficient to get around the problem.

Estimation of Transient Duration

The end of the transient is estimated as the point at which the envelope drops below a predefined fraction of its peak value or the maximum duration, whichever is shorter. The former condition is based on the expectation of a decay immediately following the attack, which is characteristic of struck or plucked instruments. The latter condition caters for 'lead' instruments that produce a sustained note through a continuous injection of energy (e.g. by bowing or blowing).

The peak value is initialised with the peak follower value upon detection of an attack transient. Thereafter it is updated at each sample-block:

$$P_{SUSTAIN} = \max\left\{P_{SUSTAIN} \;, y(n)\right\} \tag{5.8}$$

where $P_{SUSTAIN}$ is the peak value following a detection,
(all other symbols defined as above).

The end of a transient is detected when:

$$\frac{P(n)}{P_{SUSTAIN}} < T_{S,AE} \quad \text{or} \quad n \geq n_{DETECT} + T_{DUR} \tag{5.9}$$

where $T_{S,AE}$ is the Sustain Threshold (for the Attack Envelope method), below which a transient is terminated,
n_{DETECT} is the sample-block index at which the detection was made,
T_{DUR} is the maximum duration for a transient (in sample-blocks),
(all other symbols defined as above),

whichever occurs first.

5.2.2.2 Time Resolution

The Attack Envelope method has the potential for accuracy to the sample, but this is both unnecessary and computationally expensive. As explained above, the samples are grouped into blocks of 50 (for a sample rate of 44.1kHz), which gives a more than adequate resolution of ≈1.1ms.

The transient event boundary for each attack transient is the location of the first sample in the sample-block that triggered the detector, less a predefined offset. Although the block that caused the detection contains the attack, there is the possibility for the attack to have begun to build (whilst remaining below the threshold) at the end of the previous block. The previous block could therefore contain the new spectrum. For a block size of 50, an offset of 25 samples was found to be good, to guarantee that the start of an attack would lie after the output location. This has effectively reduced the accuracy of location, so that the maximum error is 75 samples (≈1.7ms). Once again, this is sufficient.

5.2.3 Detection by Spectral Dissimilarity

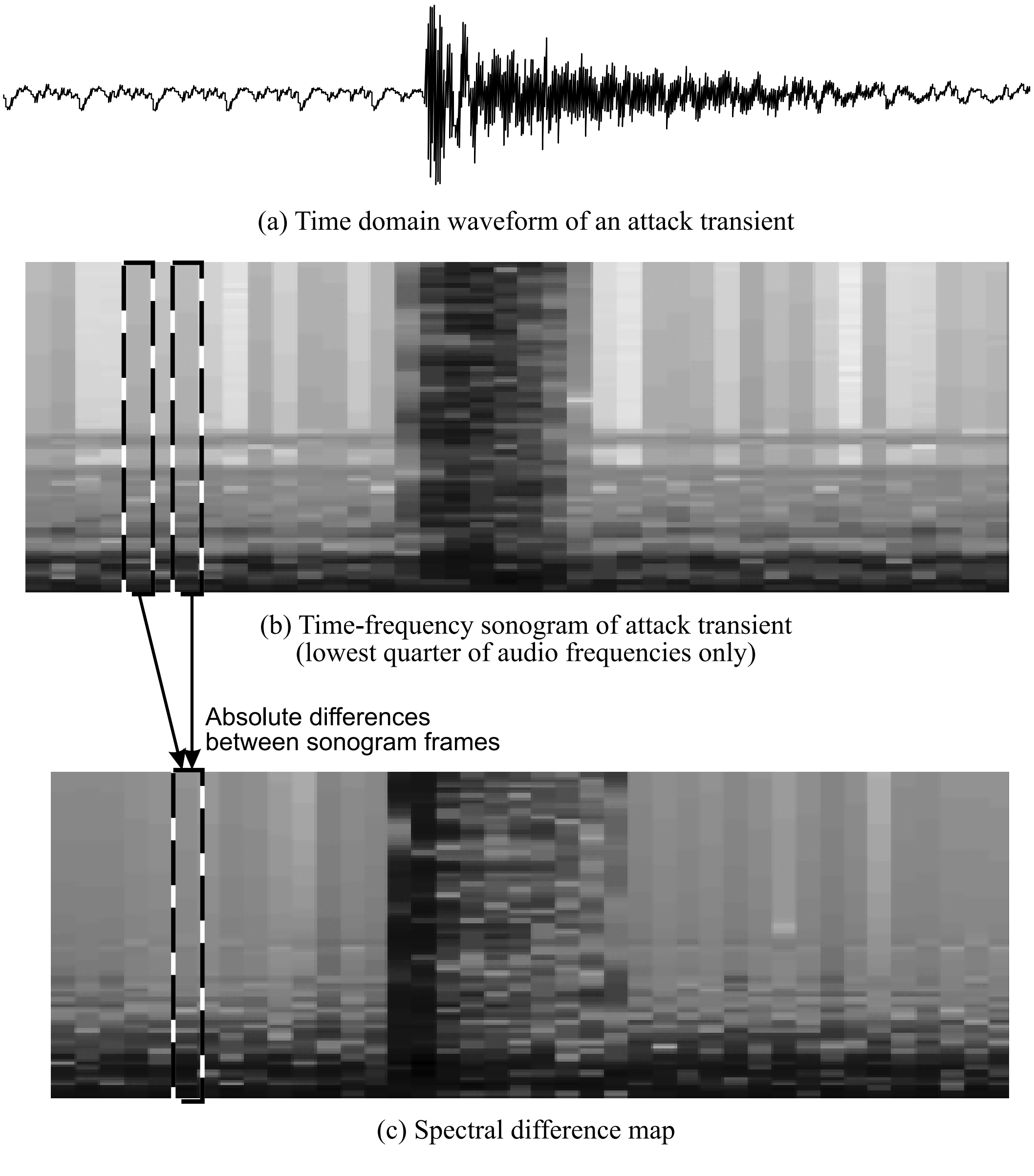
The Spectral Dissimilarity takes an alternative frequency domain approach to that of the Energy Distribution method. Instead of looking for the distortion that is generated by a sudden change in the waveform, this method looks for differences in the spectra between frames, wherever they may appear in frequency; see Figure 5.7. A sudden change in energy (whether rising or falling) would be detected, but so would a sudden change in the spectral content, even if it was not accompanied by localisation in the higher range of frequencies. Of course, an attack transient will most likely produce both effects, resulting in a reinforced chance of detection.

Therefore the Spectral Dissimilarity method aims to react to the first, second and fourth observed properties, listed in section 5.1.1.

5.2.3.1 Implementation

The test is a measure of dissimilarity between corresponding FFT bins of overlapping portions of the waveform. As with the Energy Distribution method, there is a lower limit to the FFT window-length, below which the method will be affected by the rise and fall of the waveform within each period. So once again the FFT length is set to 512 samples, only a quarter of the main analysis STFT.

The dissimilarity figure is measured for frames that are half overlapping. The fact that frames are overlapping means that they share common data and this obviously reduces the dissimilarity figure, but experiments show that there is sufficient difference between the frame


(a) Time domain waveform of an attack transient

(b) Time-frequency sonogram of attack transient (lowest quarter of audio frequencies only)

(c) Spectral difference map

Figure 5.7 – Generation of the dissimilarity matrix from the sonogram

content for the method to still work. (Experiments also showed that greater overlap does reduce sensitivity, whilst less or no overlap reduces the time accuracy of the detector.)

The maximum hop-distance that enables detection of all attack transients is the frame separation (i.e. the distance between frames that are compared, equal to half the window-length). Shorter hop-distances reveal more detail in the detection function but at the cost of extra calculation. The extra detail can however make the difference between a peak rising sufficiently to be detected or not. From experiments, a good value was found at half the maximum hop-distance (equal to half the frame separation or a quarter of the window-length). In summary, useful values for a sound sampled at 44.1kHz were found to be:

Window-length	=	512 samples (≈12ms),
Hop-distance	=	128 samples,
Window function	=	Hamming,
No zero padding.		

The dissimilarity function is calculated:

$$D_r = \frac{\sum_k \left| \left| X_r(k) \right| - \left| X_{r-2}(k) \right| \right|}{E_{r-2}} \tag{5.10}$$

where D_r is the dissimilarity function for frame r,
$X_r(k)$ is the kth bin of FFT frame r,
E_r is the energy function (as defined in Equation 5.1).

The dissimilarity function is therefore the normalised, cumulative magnitude difference between spectra, where the energy function normalises the values to the earlier of the two frames. It could be alternatively written:

$$D_r = \sum_k \frac{\left| \left| X_r(k) \right|^2 - \left| X_{r-2}(k) \right|^2 \right|}{\left| X_{r-2}(k) \right|^2} \tag{5.11}$$

This would have identical results to equation 5.10 (with a change to the detection threshold below), but might make the computation faster, because it would not require a separate summation.

The condition for detection is simply:

$$D_r > T_{D,SD} \tag{5.12}$$

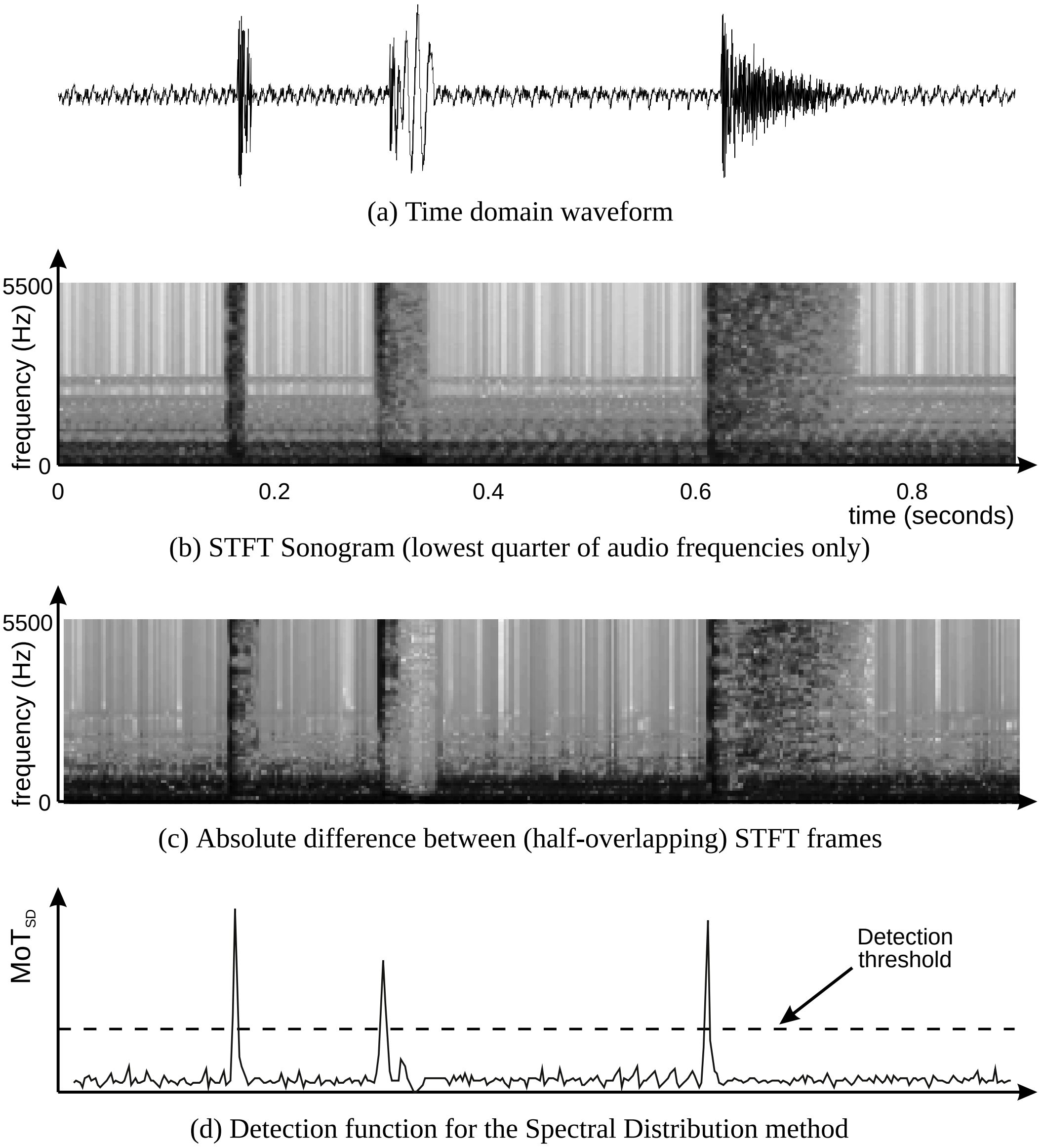
where $T_{D,SD}$ is the Detection Threshold (for the Spectral Dissimilarity method), above which a hit is registered,
D_r is defined as above.

Although equation 5.11 shows that the dissimilarity function is a summation over all FFT bins, it was found by experiment that only the lowest quarter of the spectrum needs to be used. (Similarly the energy equation only needs to be calculated for the lowest quarter of the spectrum.) See Figure 5.8 (same sound example as Figure 5.4). If the function uses less than the lowest quarter of the audio range, then the reliability of the method suffer, particularly for higher pitched sounds (such as the ride cymbal).

The computational cost of this method is quite high because of the FFT length, but if the zoom FFT is used (because only the lowest quarter of the positive spectrum is required), then the speed becomes comparable to that of a 128-length FFT.

5.2.3.2 Time Resolution

The time resolution is largely determined by the frame separation, the distance between FFT frames that are compared, which is equal to 256 samples (≈5.8ms) using the above values.


(a) Time domain waveform

(b) STFT Sonogram (lowest quarter of audio frequencies only)

(c) Absolute difference between (half-overlapping) STFT frames

(d) Detection function for the Spectral Distribution method

Figure 5.8 – Spectral Dissimilarity detector (for a sequence of three drums)

This is marginally affected by the hop-distance, if it is smaller. However, much smaller hop-distances reveal more of the detection function curve, but have a diminishing impact on the accuracy of locating the transient event boundary. With the hop-distance of 128 samples, the accuracy varies in the range 2.9-5.8ms, which is on the borderline with the threshold of the ear.

5.2.4 Comparison of Detection Methods

5.2.4.1 Reliability

The three detection methods have been compared for a variety of sound types, some of which are presented in the figures below. The detection thresholds are marked on the graphs at approximately 500 (for Energy Distribution method), 1.7 (for Attack Envelope) and 5 (for Spectral Dissimilarity). These values were determined through a number of tests comparing detector outputs with human scoring to maximise correct hit detection whilst keeping the threshold as high as possible above the 'background noise' of the detector. (All the comparative sound example figures: Figure 5.9 to Figure 5.12, can be found at the end of this section, pp.144-147.)

Sound Example 1: Simple drum sequence with tonal background

In the case of drum beats with no background sound, all the methods perform well. As the level of background sound is increased (whether tonal or not), their capabilities diminish. The first method to suffer is the Energy Distribution method. Figure 5.4, Figure 5.6 and Figure 5.8 show a sequence of three drums with a low pitched sustained note (similar to the timbre of a decaying piano note) and the results of the three detection methods. All methods succeed in detecting the percussive onsets, but already a large degree of variability is noticeable for the Energy Distribution method.

Sound Example 2: 'Tricky' drum sequence with various background sounds

The second example also shows a variety of percussive hits, this time with much greater background sound. See Figure 5.9. The background is composed of a decaying low pitched tonal sound and a growing higher pitched female voice singing 'oo'. Also, between the final two beats a downward sweeping noise source is (smoothly) introduced. This example was composed deliberately to be a tricky situation in which all the drums should nevertheless be detected.

The Energy Distribution method fails to detect the first beat which is preceded by the tail end of another higher pitched drum and the last beat which occurs during the noise sweep. In the former situation, the overall pitch drops at the point of onset, so the increase in higher frequency energy is diminished. A typical example that occurs frequency in rock and pop music is the cymbal followed by the bass drum, which has a similar effect on the spectrum. In the latter situation, the whole spectrum is flooded with noise (see the sonogram in Figure 5.9b), almost completely obscuring the drum beat. Ultimately these examples determined that the Energy Distribution technique should be abandoned. However, for completeness the remaining sound examples include its results.

The Attack Envelope and Spectral Dissimilarity methods perform well, although the Attack Envelope method is noticeably affected by the ‘spikiness’ of the time domain envelope, unlike the Spectral Dissimilarity method.

Sound Example 3: Pitched melody with smooth transitions (legato clarinet)

The next two examples include tonal instrument melodies with contrasting attack profiles. The first is a solo clarinet (Figure 5.10). This example includes onsets which are relatively smooth, so much so that some cannot be detected by eye from the waveform. Although there are definite points of onset of the notes, it is debatable whether an attack transient detection method should pick these out. In other words, the onsets are not percussive and could be adequately captured by the standard analysis-resynthesis model. Nevertheless, it is interesting to see how the detection schemes cope; such information could be useful to future developments of the model, even if not for the purpose of percussive onset detection.

The Attack Envelope method only detects the onset of the first new note that follows a largely decayed previous note. The other note onsets do not register above the detection threshold, although peak clusters can be observed in the detection function at their locations. The Spectral Dissimilarity method also registers a detection for the first note only, although the onsets of the other notes are clearly visible in the detection function, more so than with the Attack Envelope detector.

Sound Example 4: Pitched melody with sharp transitions (staccato piano)

The second tonal example is a synthesised piano playing single notes and chords (Figure 5.11). The piano was played staccato (short, unconnected notes) so that the onsets would be both percussive and clearly defined, as indeed they are in the time domain waveform. The Attack Envelope method picks these out easily. The Spectral Dissimilarity detector is not as effective, displaying peaks at the note onsets but none sufficient to trigger a ‘hit’. It would appear that the commonality of harmonics from one note to the next hinders the process.

Sound Example 5: The spoken voice

The final example is a male spoken voice saying “Please talk clearly” (Figure 5.12). From the time domain waveform (Figure 5.12a), the unvoiced plosives that begin each word (‘p’, ‘t’, ‘c’) are visible as noise bursts prior to and separate from the much louder voiced sections. The plosive onsets obviously qualify as attack transients and they are visible in either the time or time-frequency representations (‘p’ is more visible in the sonogram, while ‘c’ is more visible in the time domain), yet the focus of the detectors is mainly the onset of the voiced sections which contain the greater amount of energy.

There are two interesting exceptions. The Attack Envelope method picks out the onsets of both the plosive and the voiced sections of “clearly”, whilst the Energy Distribution method picks out the plosive of “talk”. The best performer in this speech example is the Attack Envelope method, which picks out all the voiced onsets. Once again the Spectral Dissimilarity method displays peaks (at the voiced onsets) which are insufficient to trigger detections.

5.2.4.2 Temporal Accuracy

The above figures confirm the expectations of the previous sections, that the Attack Envelope method has a far superior time resolution, by virtue of not being limited to FFT frames. The accuracy is in the order of 2-4 times better.

5.2.4.3 Computation Speed

Once again, by virtue of not using the FFT, the Attack Envelope method performs the fastest. Both of the other methods use the same length FFT, but the Spectral Dissimilarity method calculates double the number of frames, because of its smaller hop-distance. However, implementation with the zoom FFT, gives the Spectral Dissimilarity method a clear advantage over the Energy Distribution method.

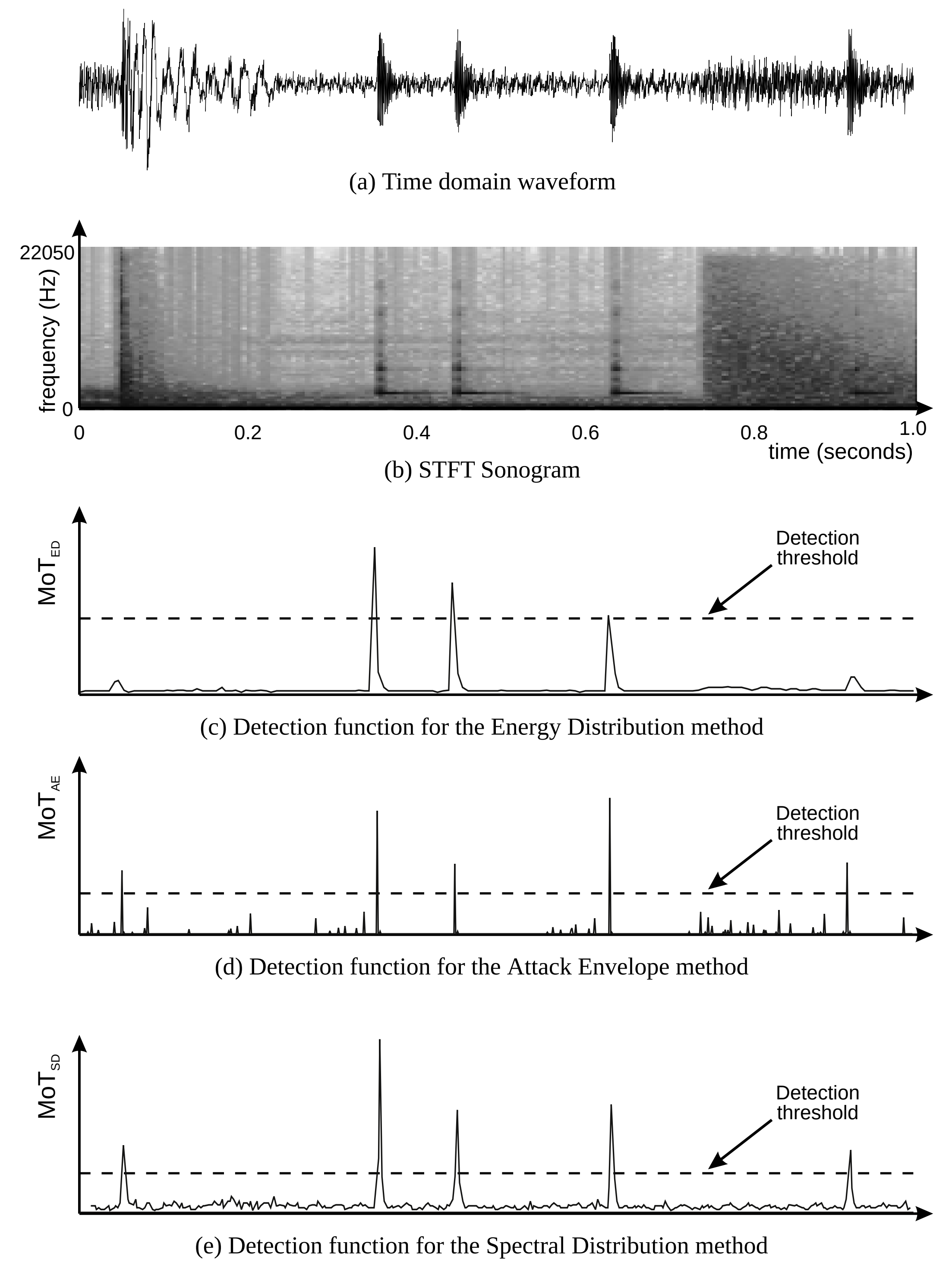
5.2.4.4 Summary

The most appealing method conceptually is Spectral Dissimilarity, because it has the potential to pick out spectral changes that are not accompanied by dramatic envelope changes. In practice though, attack transients are the targeted feature and these tend to begin with a large burst of energy (see observations 1 and 4 in section 5.1.1). Therefore the Attack Envelope method performs best.

The Attack Envelope method is prone to peaks due to amplitude modulation and the rise and fall within single periods of low pitched waveforms (as seen in the examples), but these peaks always seem to remain well below the detection threshold. The Spectral Dissimilarity method is able to display peaks at spectral changes not noticeable by the other methods, but these never seem to be significant enough to register a detection. Also the method can be fooled by note onsets where the new notes share the same harmonics as the previous note, thus diminishing its ability to detect the change.

In summary, each of the methods has situations under which it can fail. For the Energy Distribution method, it is the transition from higher pitches to lower pitches. For the Attack Envelope method, it is sudden changes that do not exhibit large amplitude changes. For the Spectral Dissimilarity method, it is the transition between notes with shared harmonics. Of all these, the shortfall of the Attack Envelope method is least significant, because it is very unlikely to occur, given that the aim is to detect attack transients.

The reliability of the Attack Envelope method, coupled with its fast execution and good time resolution promotes it as the best detection scheme for attack transients.



(a) Time domain waveform

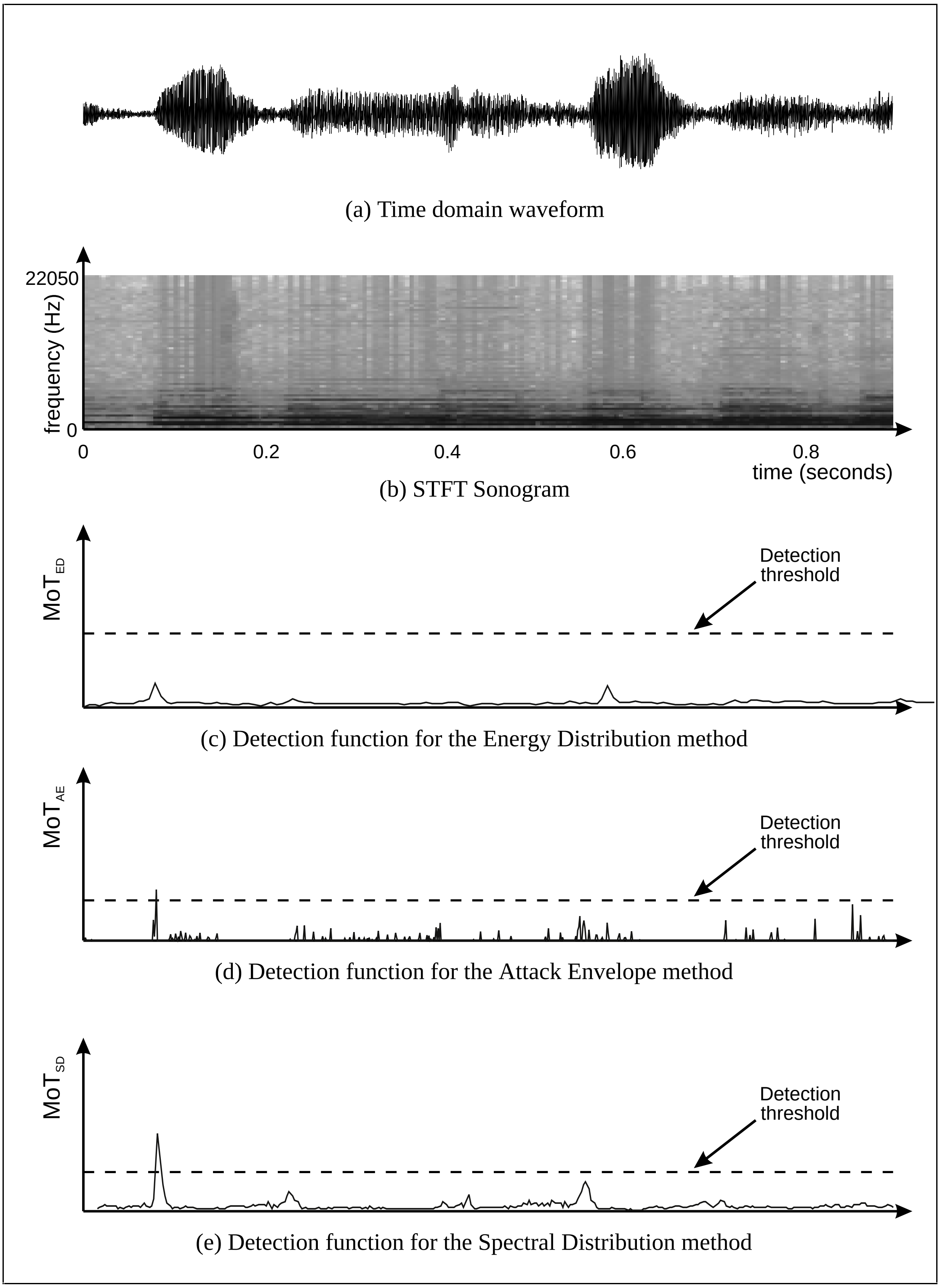
(b) STFT Sonogram

(c) Detection function for the Energy Distribution method

(d) Detection function for the Attack Envelope method

(e) Detection function for the Spectral Distribution method

Figure 5.9 – Comparison for 'Tricky' drum sequence with various background sounds



(a) Time domain waveform

(b) STFT Sonogram

(c) Detection function for the Energy Distribution method

(d) Detection function for the Attack Envelope method

(e) Detection function for the Spectral Distribution method

Figure 5.10 – Comparison for a legato clarinet melody

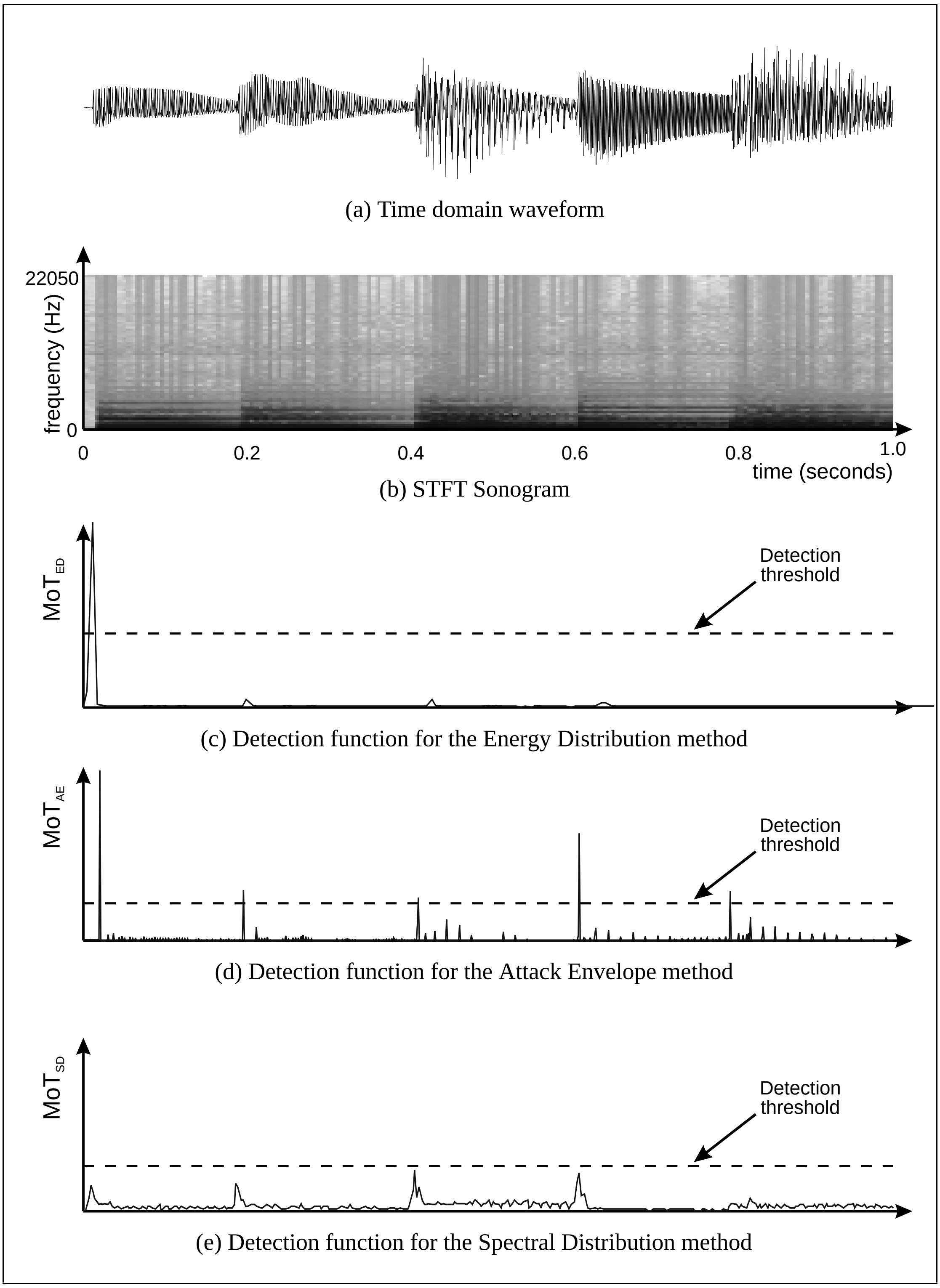


Figure 5.11 – Comparison for a staccato piano melody

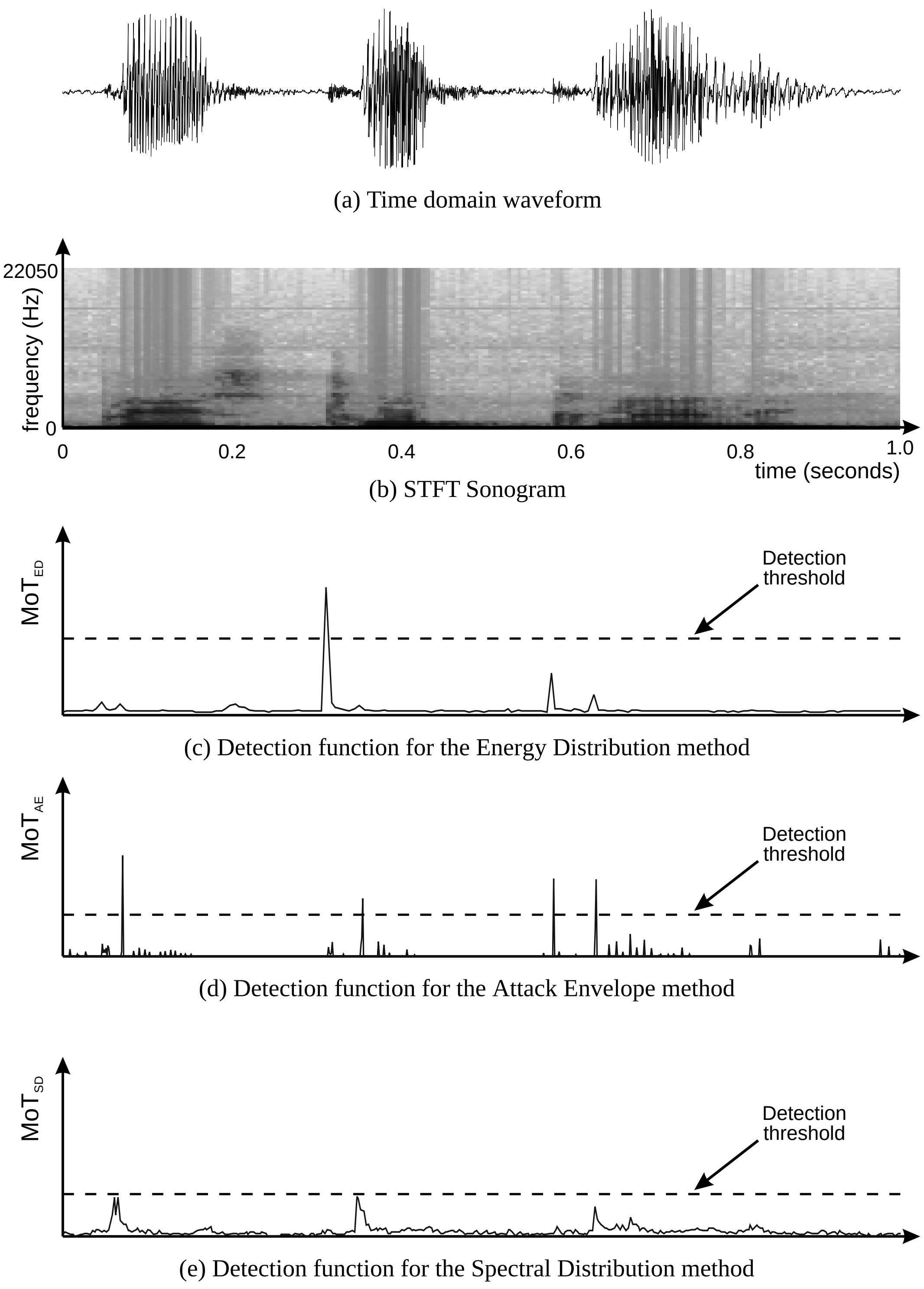


Figure 5.12 – Comparison for 'Please talk clearly', a male spoken voice

5.3 Synchronisation of Analysis and Synthesis

5.3.1.1 Region Definition

The detection process is carried out as a pre-analysis scan and its results are compiled as a *region list,* where the region boundaries correspond to the detected transient event boundaries, as shown in Figure 5.13a. The regions are then used to synchronise the analysis and synthesis. In essence, analysis and synthesis are performed for each region independently, thus ensuring that spectral 'confusion' is avoided and sudden changes are possible at the region boundaries. Implementation details are presented in the following subsections.

5.3.1.2 Synchronised Analysis

See Figure 5.13b. At the start of each region, the analysis window is positioned with its trailing end at the first sample, the region's start boundary. Thereafter analysis proceeds, as normal, the window advancing by the hop-distance for each new frame.

The first frame whose window touches or crosses the region's end boundary is 'snapped' so that its leading edge coincides with the last sample of the region. Naturally, this means that the final hop-distance is reduced (unless the last sample of the window already coincides with the region boundary).

For strict accuracy, during deterministic analysis, the parameters governing the peak linking process should be adjusted to account for the reduced hop-distance. In practice however, the reduced interval means that the frame's spectra are more alike, and therefore have less linking conflicts. So for simplicity, no further modification is made.

Extrapolation towards Region Boundaries

The data for each frame notionally corresponds to the instantaneous situation at the centre of the frame. Frames that are linked generate approximations to the actual spectral evolution through peak / spectral envelope interpolation. However the bounding frames of each region are not linked to one another, therefore the first half of the first frame and the last half of the last frame in each region are undefined.

To provide continuity, the data in these frames are extrapolated outward to the respective region boundaries, by fixing the values of the features. See Figure 5.13c. (i.e. The frequencies and amplitudes of partials and the stochastic spectral envelopes remain constant from the centre of the last frame to the region boundary. Similarly, the values on the other side of the boundary remain constant at the settings of the first frame, from the region boundary to the centre of the frame.)

This makes the implicit assumption that the waveform is slow changing in these zones. Whereas this may be accurate at the end of a region, we already know that there are rapid changes at the start of a region. Despite this, the STFT provides no way of gaining extra detail, if the frequency resolution is to be maintained.

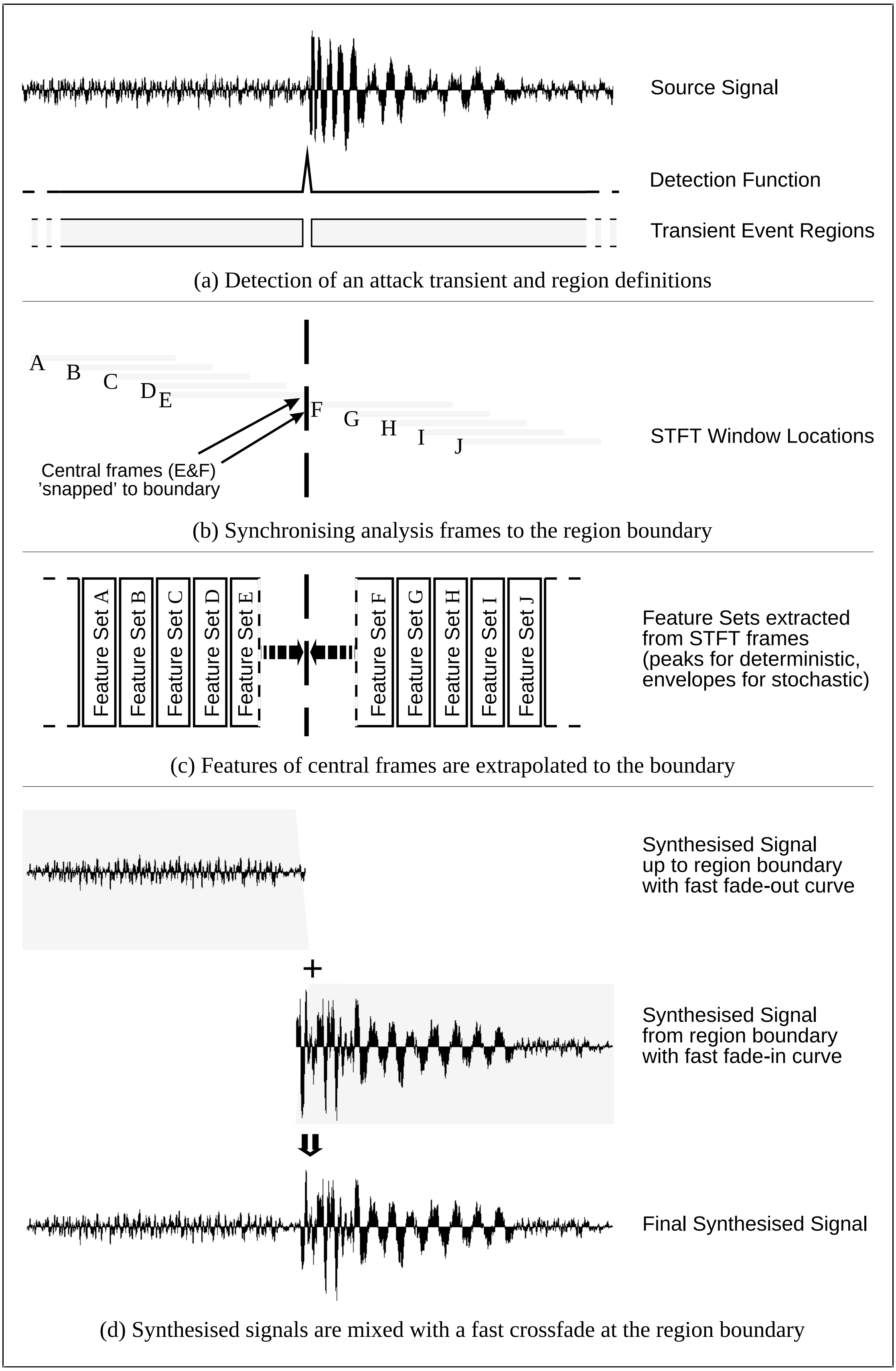

Source Signal
Detection Function
Transient Event Regions
(a) Detection of an attack transient and region definitions
A
B
C
D
E
F
G
H
I
J
STFT Window Locations
Central frames (E&F) 'snapped' to boundary
(b) Synchronising analysis frames to the region boundary
Feature Set A
Feature Set B
Feature Set C
Feature Set D
Feature Set E
Feature Set F
Feature Set G
Feature Set H
Feature Set I
Feature Set J
Feature Sets extracted from STFT frames (peaks for deterministic, envelopes for stochastic)
(c) Features of central frames are extrapolated to the boundary
Synthesised Signal up to region boundary with fast fade-out curve
+
Synthesised Signal from region boundary with fast fade-in curve
Final Synthesised Signal
(d) Synthesised signals are mixed with a fast crossfade at the region boundary

Figure 5.13 – The four steps to incorporating the feature of attack transients

5.3.1.3 Synchronised Synthesis

Upon synthesis, both deterministic and stochastic, the extrapolated spectra are synthesised beyond the region ends by a short amount. This provides sufficient excess waveform to allow a crossfade between regions. See Figure 5.13d.

The crossfade length has been set to ≈5.8ms (256 samples @ 44.1kHz sample rate), where it is short enough to reproduce the suddenness of the original attack, but not so short that an audible click (termed a 'glitch') is produced.

Automatic Synchronisation for Sound Transformations

Extrapolation is built into the data during analysis and the crossfade method is built into the synthesis, therefore transformations like time-stretch require no further modifications to the synthesis process. As the frame interval is scaled, the region boundaries are automatically scaled.

It is recommended that the crossfade length should remain constant under time-stretch, in order to preserve the timbral properties of the attack transients. This deviates from truly proportional time scaling, but the question arises: Is the aim to create a transformation that is mathematically perfect or one that sounds right? With a fixed crossfade, the attacks sound sharp even when a sound is slowed down, which is more realistic to music played at a slower tempo or speech spoken more slowly.

5.4 Results

5.4.1 Performance of the Technique

Figure 5.13a&d (above) demonstrates that the proposed method is able to retain the suddenness of an attack transient. The success of the method is further confirmed through listening tests. Sounds synthesised by the new method retain the crispness of the original sound. The importance of the modification is further emphasised when time-stretch is applied to slow down a sound. The sharpness of the attacks is preserved, which gives the perceptual impression of a sound played more slowly, rather than a recording slowed down. This is particularly true for drum patterns, where the rhythmical impact persists after stretching, so that even for large stretches (say to 200%, to halve the tempo), the output is still usable in a drum sequence.

The performance is good even where the spectral evolution has been simplified. (Some spectral detail is inevitably lost following a transient event boundary, as a consequence of using the STFT for analysis.) It would appear that the abrupt rise in volume and the sudden change in spectral content are the most prevalent perceptual features. In other words, the reproduction of observed properties 1, 2 and 4 (from section 5.1.1) somewhat mask the inability to track the chaotic behaviour of observed property 3.

Tests were made for sounds that contained percussive elements with steady sounds continuing in the background. These examples included both tonal sounds (such as in the examples of Figure 5.4 {& Figure 5.6 & Figure 5.8} and Figure 5.11) and noisy sounds (such as in the example of Figure 5.9), which would be tracked in the deterministic and stochastic aspects of the model respectively. In all cases, the background sound is perceived as continuous. Therefore the errors in its analysis (which are inevitable during rapid changes just after the percussive onset) are masked by the perceptually more dominant attack transient.

Performance During Misdetections

No detection scheme is likely to be foolproof to every situation presented, although the Attack Envelope method has shown itself to be very reliable. So the potential situations of false hits and false misses are examined here.

A false miss occurs when the detector fails to detect an attack transient. In this situation, the analysis and synthesis would proceed as for the Initial Model and the results would be the problems described at the start of this chapter. The situation of drop-outs (in which the deterministic analysis fails to capture any significant energy during the attack transient) only tends to occur when there is a very dramatic change in the waveform and the spectrum, so it is extremely unlikely that this type of failure could happen. More likely are additional artifacts and the diffusion effects, due to small scale erroneous peak selection and the slow-fade method of updating the spectrum during synthesis, respectively. In the end, failure due to a false miss is no worse than if the detector had not been implemented, so it is merely a waste of computation power.

A false hit occurs when the detector registers a 'hit' when it shouldn't. This is a more likely scenario for the Attack Envelope detector, since large-scale, rapid amplitude modulation could force the detection function to rise above the threshold. In this case the analysis and synthesis would be erroneously synchronised to the detection location. Detail would be lost during analysis, because the STFT window would be prevented from 'hopping' through the detection point. At synthesis, the spectra either side would be extrapolated. For steady state signals, this presents no problem, because little change is likely to have happened during that period, but the situation is more likely to occur for sounds that are changing in some way, so some variation must be expected at the boundary of the regions. In such cases there is likely to be a small but noticeable artifact (not as severe as a glitch).

Furthermore, the continuity will be affected by the phase of the synthesised signal, therefore to safeguard against false hits, the synthesis would need to be phase correct (i.e. synthesising phase as well as frequency and amplitude, or somehow aligning phases either side of the misdetection location).

Room For Improvement

Some degradation is noticeable for sounds with greater 'depth', such as a booming bass drum, where some of its power is lost. This is probably due to the rapidity of change following the transient onset, the necessary detail of which, the analysis is still unable to extract.

One area where the modification is unable to help is the rapid succession of transient events, such as from a drum roll, or drums with gated reverb. In these cases, the pre-analysis scan is often still able to detect the event boundaries, but the events are too close to one another to apply the synchronisation. This is because the resultant regions are shorter than a single window-length for the main analysis FFTs.

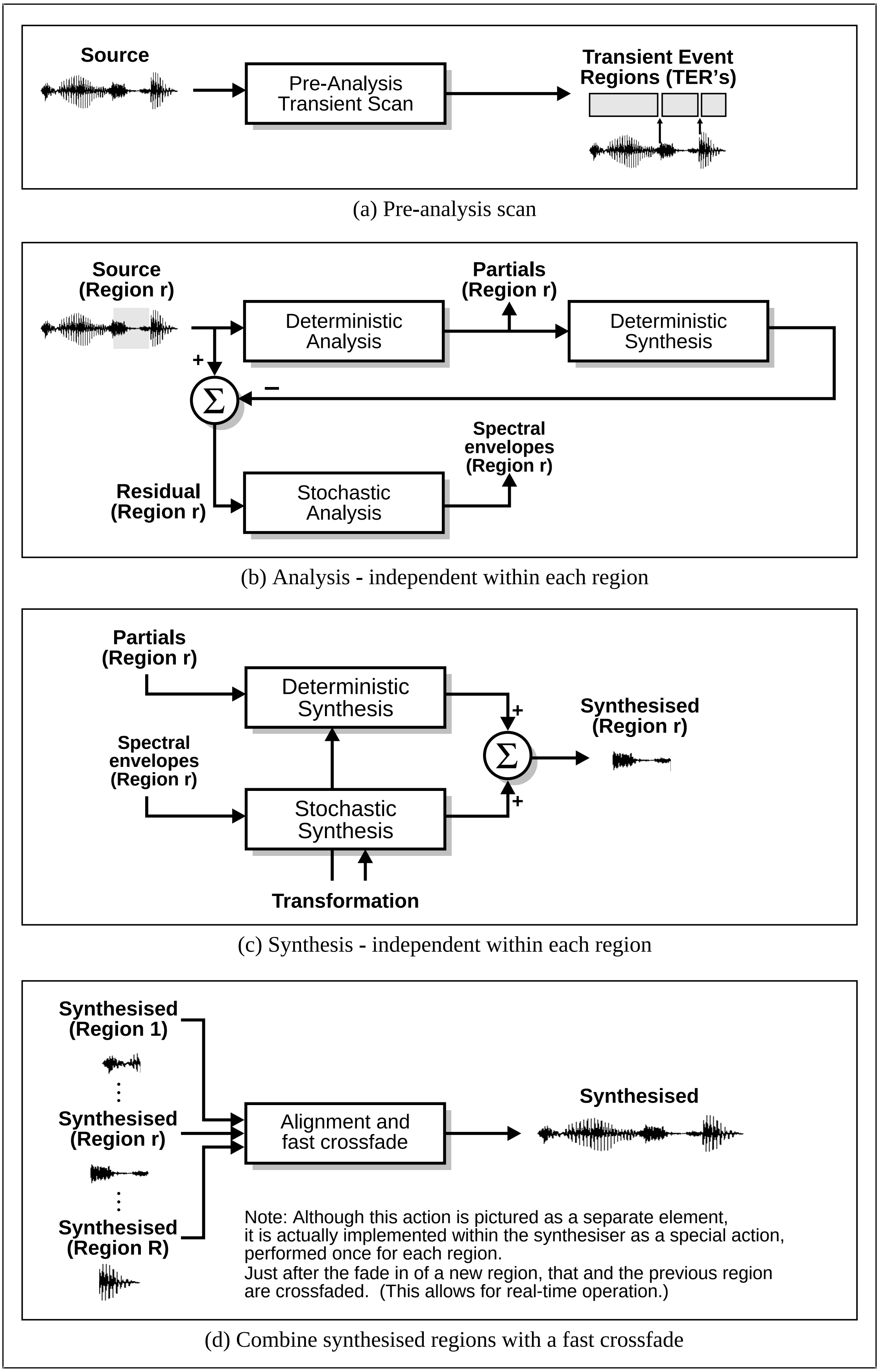
5.4.2 Impact upon the Model

The modifications presented in this chapter include:

- a pre-analysis scan, to detect attack transients and define regions;
- modification to the analysis to work within regions;
- modification to the hop-distance at the region boundaries;
- extrapolation of the feature sets (for both deterministic and stochastic);
- modification to the synthesis to work within regions;
- inclusion of a scheme to perform (or simulate) the crossfade.

Structural Impact

The pre-analysis scan sits outside the existing analysis-resynthesis system structure, so it is simple to implement. The inclusion of regions to both the analysis and synthesis requires some modification, but in practice this can be implemented quite simply as a shell around each of the processes.



(a) Pre-analysis scan

(b) Analysis - independent within each region

(c) Synthesis - independent within each region

Note: Although this action is pictured as a separate element, it is actually implemented within the synthesiser as a special action, performed once for each region.
Just after the fade in of a new region, that and the previous region are crossfaded. (This allows for real-time operation.)

(d) Combine synthesised regions with a fast crossfade

Figure 5.14 – The modified model, incorporating attack transients

The region defines which parts of the waveform are available to the analysis and then launches the process. The only change required within the analysis is detection of the last frame, so that the hop distance can be modified. Frame extrapolation is implemented notionally. That is, there is common understanding between the analysis region shell and the synthesis region shell that the bounding frames of each region must be extrapolated, so no further intervention is necessary.

For synthesis, there needs to be an overlap between regions to effect the crossfade, so the modification is a little more involved. Once again a region shell launches the synthesis process, providing it with the region's data and the transformation parameters. Within the region, samples are synthesised from half the crossfade duration *before* the region boundary up to the centre of the first frame (the point at which interpolation begins). If this is not the first region, at this point the crossfade is implemented: imposing a fade-out on the previous region (which will have been synthesised *beyond* the region boundary by half a crossfade duration), imposing a fade-in on the current region, and mixing the signals. Thereafter synthesis of the region proceeds as normal. (This process may be easier to visualise with reference to Figure 5.13d.)

By incorporating the crossfades into the synthesis process, the algorithm can be implemented on a real-time synthesiser, without significant time lag (although such a system would have to synthesise faster than real-time, so that the signals surrounding each region boundary can be prepared and crossfaded, before they are needed for outputting).

Figure 5.14 shows the modified analysis-synthesis structure. From this it can be seen that the modifications have included new elements, but have not required a drastic change to the existing structure to do so.

Computational Impact

The pre-analysis scan imposes a computational burden on the system, but this is very small compared to the STFT of the main analysis. The region management in the 'shells' described above takes minimal computation time. In fact the additional computational expense can largely be traded with the reduction in FFT calculation, since the number of FFTs is reduced by the synchronisation at region boundaries. Analysis is therefore virtually unaffected, in terms of the processing load.

Synthesis acquires some extra tasks (shell management and region crossfade), but these are small compared to preparing the samples of each frame, even if a Fast Additive Synthesis method is used (see sections 2.3.3.2 and 2.3.6.1).

In summary...

It was observed that rapid attack transients were not well represented within the Initial Model. The result upon resynthesis was a degradation of sound quality at these points, in the form of diffusion or dissolved features. The method presented in this chapter includes a robust pre-analysis detection scheme that identifies and locates the percussive onsets to a good accuracy. The onset data are then used to synchronise the analysis and synthesis processes, to minimise erroneous analysis and retain the suddenness of the attacks upon synthesis.

Furthermore, it has been shown that it is possible to incorporate this new feature of ‘attack transients’ into the model at very little computational expense and with minimal disruption to the existing system.

The results, both visually in terms of the synthesised waveform and audibly through listening tests, have confirmed the effectiveness of the presented technique. These results are further proved under time-stretch transformation, where the timbral properties of the attacks are preserved to give the impression of a more natural slowing of the sound.

CHAPTER SIX

OUTCOMES AND FUTURE DIRECTIONS

6. OUTCOMES AND FUTURE DIRECTIONS

This chapter looks at the outcomes of the presented work and proposes new directions for improving the sound model and its implementation. The first three sections correspond to the advances of chapters three to five respectively. The fourth section looks at how they work together in a single system and examines what synergies can be gained. In all these sections, the results are placed in context, to evaluate how they have improved upon the benchmark Initial Model and how they have impacted the model structure (in terms of system complexity, computational cost, implementation issues, and so forth). The results are analysed in a similar manner to the 'Critique and Modification' sections of chapter two, with the intention of discovering which aspects need improving. On these grounds, suggestions for future work are proposed. Where there have been recent advances by other authors that are relevant to the studies presented here, the methods are briefly presented, with a cursory evaluation where possible.

The fifth and final section presents the basis for a new model of sound. The deterministic-stochastic classification solved the problem of system inflexibility, with respect to musical transformation, primarily by freeing the system from phase dependency. However, this classification has its own shortcomings. By examining where and how this classification fails, the basis for a new model is formed. The principle of 'the Noisy Partial' is presented, with evidence to support its premises. Potential routes for investigating and developing the technique are also considered.

Providing a reference for comparison, Figure 6.1 and Figure 6.2 display a schematic of the steps for analysis and synthesis in the Initial Model (described in full in section 2.3). Note that all the modified or new elements of the following sections are highlighted in the figures with a thick border.

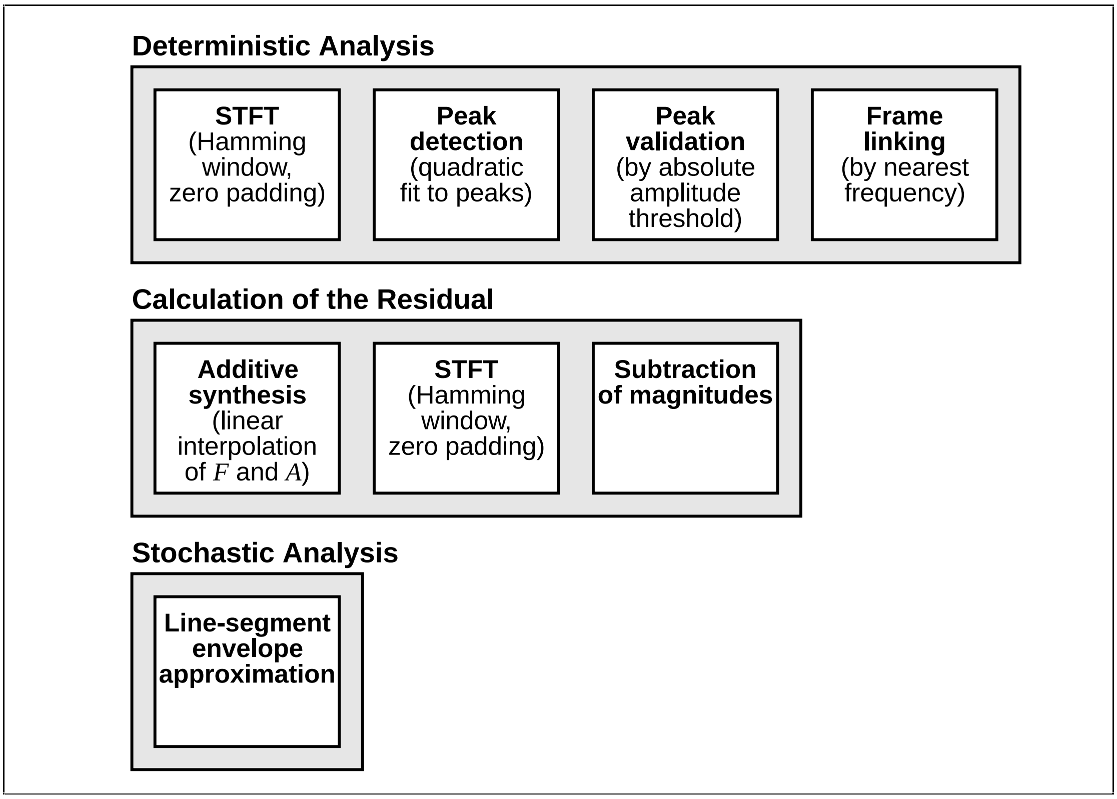

Deterministic Analysis
STFT (Hamming window, zero padding)
Peak detection (quadratic fit to peaks)
Peak validation (by absolute amplitude threshold)
Frame linking (by nearest frequency)
Calculation of the Residual
Additive synthesis (linear interpolation of *F* and *A*)
STFT (Hamming window, zero padding)
Subtraction of magnitudes
Stochastic Analysis
Line-segment envelope approximation

Figure 6.1 – Analysis in the Initial Model

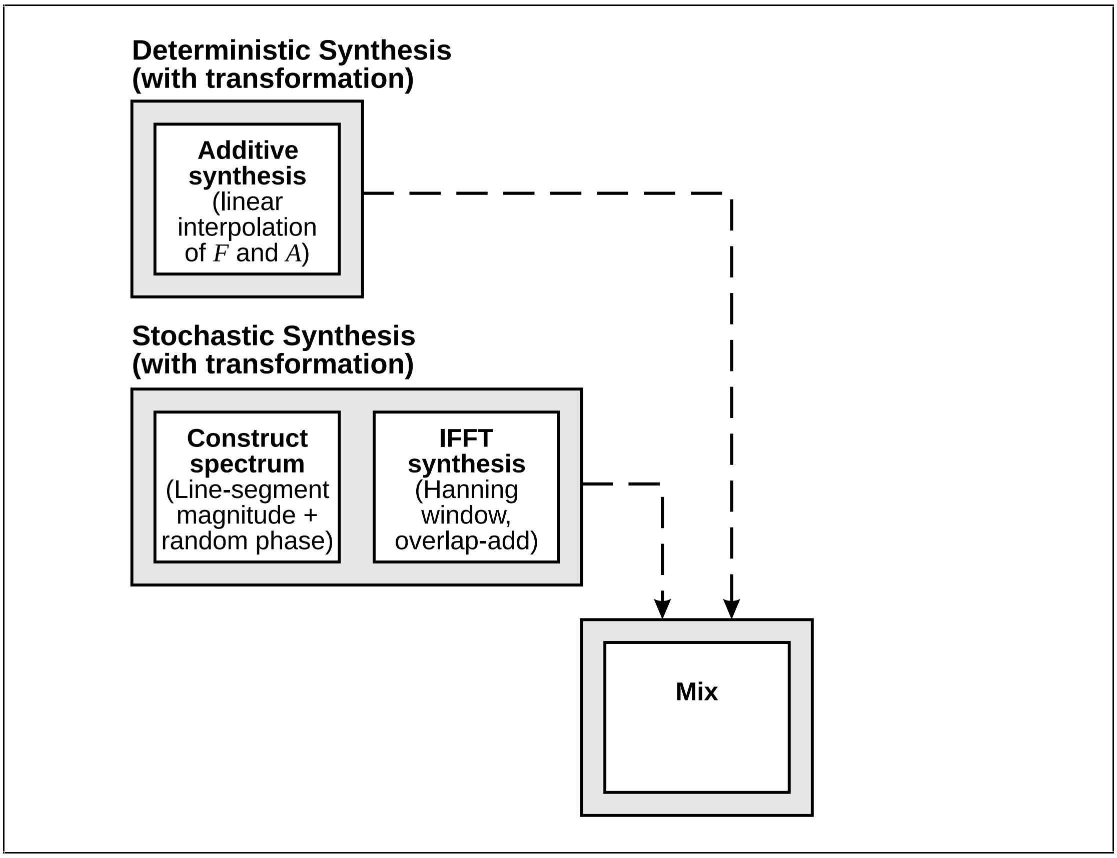

Deterministic Synthesis (with transformation)
Additive synthesis (linear interpolation of *F* and *A*)
Stochastic Synthesis (with transformation)
Construct spectrum (Line-segment magnitude + random phase)
IFFT synthesis (Hanning window, overlap-add)
Mix

Figure 6.2 – Synthesis in the Initial Model

6.1 Phase Distortion Analysis (PDA)

6.1.1 Impact upon the Model

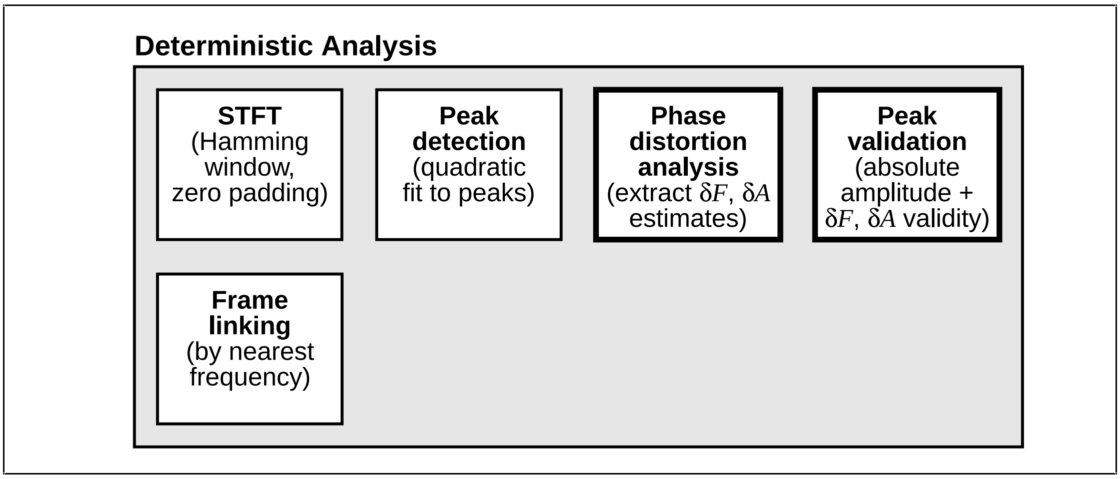

Deterministic Analysis
STFT (Hamming window, zero padding)
Peak detection (quadratic fit to peaks)
Phase distortion analysis (extract δF, δA estimates)
Peak validation (absolute amplitude + δF, δA validity)
Frame linking (by nearest frequency)

Figure 6.3 – Changes to the Initial Model to incorporate Phase Distortion Analysis

The phase distortion analysis method described in section 3.3 adds a new step to the deterministic analysis, as shown in Figure 6.3. This step does not perform an action to the data; instead it measures properties of the data. The technique is attractive from a system viewpoint, because it slots into the existing framework, without disturbing it, and yet its results can be made use of in a number of the system's processes.

There is the potential for the method to impact peak validation, where it can help to identify which peaks belong to actual sinusoids and which are noise or representation artifacts. There is also the potential for assisting in frame linking, where the extracted rates of change of frequency and amplitude could aid in the choice of which are the most likely peak-to-peak links between frames. Within deterministic synthesis it could also aid in the detail of the frequency and amplitude trajectories. (See section 3.4.2.)

Computational Impact

PDA introduces some additional computation to the analysis, but no more than that for locating the maximum of a detected peak. For each peak, the phase offsets are read $\frac{1}{8}$th bin from the maximum, their sum and difference calculated and used to index tables that convert them into LFM and EAM values. Compared to the calculation of the STFT, this is minuscule.

6.1.2 Critique and Future Directions

Reliability and Simplicity

PDA is a novel technique that appears to gain information at almost no cost. The method for extracting the trajectory data is perhaps the simplest method: using the minimum of two measurements to evaluate the two parameters (LFM and EAM), from just the phase offsets at points close to and equidistant from the detected peak maximum. Unfortunately this degree of simplicity means that the method is not very robust – for rapidly changing portions of sound only the most prominent partials give reliable readings. The suspected reason for this is that as the partial peaks distort, becoming wider in frequency, the degree of additive interference increases and the phase profile across the less prominent partial peaks is affected significantly.

Several potential solutions exist:

1) Iteratively estimate the partial trajectories from the most prominent peak downwards, removing each peak from the spectrum after it has been analysed;

2) Take phase measurements at two or more offsets either side of each partial peak and use a more sophisticated decision procedure for determining the trajectories;

Recent advances in IFFT synthesis now make it possible to generate the spectral profile of a linear chirp (LFM sinusoid) [Goodwin & Kogon,1995]. Therefore it should be possible to subtract the profile of a peak, once its trajectories have been estimated. In this way the additive interference to neighbouring partial peaks would be reduced and reliable estimates should be possible for a large number of the partials. This process should work in principle and it would be interesting to see how well it performs. If successful, it could lead the way to iterative techniques that allow the time-frequency domain to be 'picked apart' little by little, even for polyphonic sounds with complex trajectories. The disadvantage of the technique is that it would probably be prohibitively slow to use in anything but an off-line system (say a re-mastering algorithm).

If more points were read around each main-lobe there would be the opportunity to reduce or remove the effects of the additive interference from its neighbours (assuming the interference varies smoothly across the measured points). However the larger the number of points, the more complex the decision method is likely to become. A solution to this that would retain much of the speed and simplicity would be to replace the decision method with a neural network [Haykin,1994; Hertz et al.,1991]. The neural network would be trained to associate phase profiles (based on an array of readings across the main-lobe) and possibly amplitude profiles, with trajectories. The aptitude of neural networks for pattern recognition, especially in situations where the data is not ideal, makes this an attractive solution.

Higher Order Data Extraction

Since it has been shown that a) the FFT contains all information about a signal, stationary and nonstationary, b) the 'distortion' is localised around the main-lobe, and c) second order information can be extracted, there is every likelihood that higher order information can also be extracted from the main-lobe. However it is also anticipated that with every increase in order, the data will probably be less easy to extract (due to more complex associations between the modulation law and the peak's distortion).

A method is proposed that could automatically test this hypothesis. Just as in the above example, where a neural network is suggested to associate a phase profile pattern with the LFM and EAM trajectory parameters, a neural network could be trained to associate phase and amplitude profiles with a number of higher order parameters. The added complexity in distortion analysis would enable extraction of more complex frequency and amplitude trajectories within each FFT frame, thereby relaxing the window-length restrictions further.

This would enable the use of longer window-lengths, with the associated improvement in frequency resolution and better separation (hence isolation) of the partial peaks.

6.1.2.2 Alternative Techniques and Recent Advances

A New Kernel for the Bilinear Distribution

William Pielemeier and Gregory Wakefield have proposed a new kernel for the Cohen class of bilinear distributions [Pielemeier & Wakefield,1996]. In the ambiguity domain (whose axes are time lag and Doppler; i.e. time and frequency shifts respectively) the time and frequency closeness of interacting components is mapped. Components that are closely spaced in frequency have a low Doppler value. Therefore auto-terms appear close to zero whilst cross-terms have higher values.

Pielemeier-Wakefield's kernel is effectively a low-pass filter in the ambiguity domain that filters out (c.f. suppresses) the cross-terms. The filter's cut-off location has an upper limit so that it suppresses cross-terms of adjacent partials (for a low fundamental of around 100Hz) and a lower limit so that auto-terms with a large bandwidth are not compromised. This lower limit on the filter places an upper limit on the bandwidth of signal components. Therefore the cross-term suppression is at the expense of temporal resolution. The authors claim an overall improvement in time-frequency resolution (with respect to the STFT), but with the restriction that rapidly changing elements suffer poor representation[1]. A new analysis-resynthesis method has been introduced using this TFR, which is specifically targeted at modelling piano sounds. The method is described and initial results presented in [Guevara & Wakefield,1996].

A Quadratic Phase Parametric Signal Model

The computer music team within Texas Instrument's DSP research department claim to have achieved "greatly enhanced" time-frequency resolution over the STFT. They have applied a quadratic phase (linear FM) approach to parametric modelling of sinusoids for the deterministic aspect of the Deterministic Plus Stochastic model. Further details are not available at present (and hence an assessment is not possible) because technical papers are awaiting publication [Ding & Qian,1997; Qian & Ding,1997].

In summary…

Phase Distortion Analysis has been proven capable of extracting second order information about the frequency and amplitude trajectories of sinusoids. However the two-point method is not sufficiently robust for practical use in the multicomponent environment of harmonic sounds. Two routes have been proposed as future directions: one based on iterative data extraction and the other using more data points and a neural network association method.

The principle behind PDA – that data may be extracted from the apparent distortion of a TFR if a change is made to the assumptions and expectations associated with it – remains valid for

[1] Rapidly changing elements have wide bandwidths. In the ambiguity domain, auto-terms and cross-terms have a wider Doppler spread. Therefore the auto-terms will be partially filtered and the cross-terms will not be completely filtered by the Pielemeier-Wakefield kernel.

any invertible transform, and the technique could also be applied with the higher order TFRs that are currently being developed to squeeze extra detail from the representations.

6.2 Harmonic Frame Linking

6.2.1 Impact upon the Model

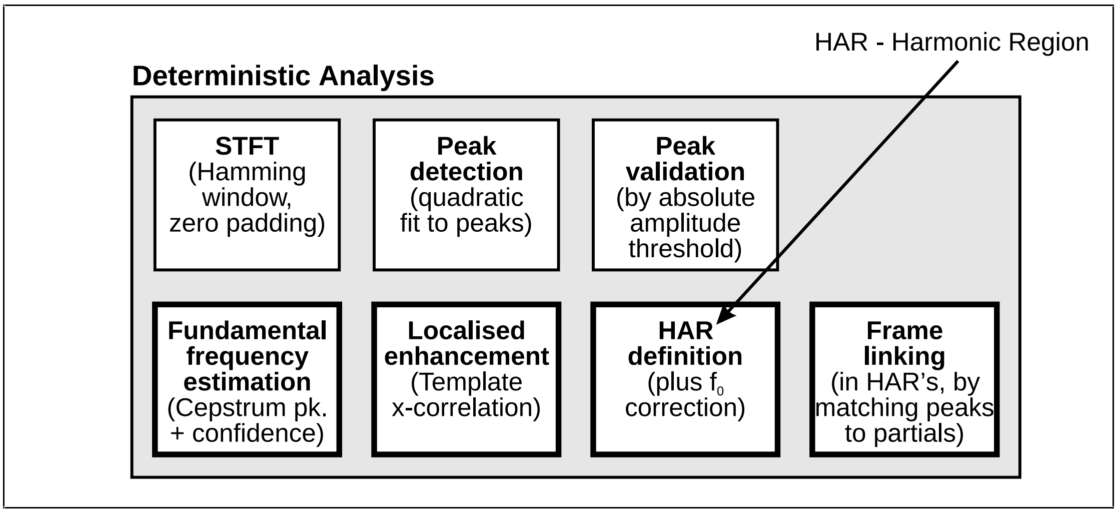

HAR - Harmonic Region
Deterministic Analysis
STFT (Hamming window, zero padding)
Peak detection (quadratic fit to peaks)
Peak validation (by absolute amplitude threshold)
Fundamental frequency estimation (Cepstrum pk. + confidence)
Localised enhancement (Template x-correlation)
HAR definition (plus f_0 correction)
Frame linking (in HAR's, by matching peaks to partials)

Figure 6.4 – Changes to the Initial Model to incorporate Harmonic Frame Linking

The switch from nearest frequency frame linking to harmonic linking is essentially the replacement of an existing block of the deterministic analysis with a new method. See Figure 6.4. As with Phase Distortion Analysis, the new algorithm slots into the existing framework without disturbing it.

The impact upon the model feature set and the method of extraction is that the sinusoidal trajectories more closely represent the partials of the source sound. Advantage is taken of the fact that the source partials are roughly harmonically spaced, and this enables extraction of meaningful trajectories (i.e. without artifacts) even for the higher partials. In terms of the ideal that the model should be musically meaningful, the harmonic linking method brings the system closer to that ideal.

Computational Impact

There is a significant computational load associated with detection of the fundamental frequency and matching partials to peaks. The Nearest Frequency method alone is simple and computationally inexpensive, but it has been shown to be inadequate. Serra introduced heuristics to improve the performance at the expense of further computation (see section 2.3.2.5). Unfortunately there seems to be no alternative to introducing extra computation if the frame linking is to be improved. Nevertheless, the use of two fundamental frequency estimators is particularly expensive (even for a development system), so this is listed as one of the weaknesses of the algorithm, in the following section.

6.2.2 Critique and Future Directions

This section looks at the strengths and weaknesses of the particular implementation of harmonic linking described in chapter four, not in comparison with the Nearest Frequency method, but in its own right with a view to making improvements.

Summary of Method Strengths

The singular strength of chapter four's algorithm actually arises from the need to cope with deficiencies in the fundamental frequency estimators. The need to correct individual frame estimates led to the inter-frame Scan-and-Propagate method of defining harmonic regions. This gave the algorithm the capability of discriminating between harmonically weak frames (usually at the tails of harmonic regions) and random correlations in the spectrum.

Summary of Method Weaknesses

The weaknesses include:

- The chosen fundamental frequency estimators, the cepstrum and template methods, are prone to suggesting multiples or fractions of the true fundamental, as just mentioned.
- The hybrid fundamental frequency estimator is processing intensive because it has two estimation stages (cepstrum and template) as well as the correction, propagation and partial matching stages;
- There is an upper limit to the range of detectable fundamental frequencies, which results from a limitation of the cepstrum method (see the last paragraph of section 4.2.2.1);
- There is an upper limit to the number of partials that can be identified (see section 4.3.2.1);
- Once a structure is found, the algorithm assumes that partials are present throughout the spectrum, and attempts to match any peak (close to a harmonic frequency), even if that peak is actually localised noise (see section 4.3.2.2);
- Although the algorithm works well for quiet, solo recordings, it is not robust in the presence of noise or reverberation, which make the sound appear polyphonic (see section 4.3.2.3).

6.2.2.2 Localisation in Time and Frequency

Localisation is the Key

A new method is sketched out here for future investigation. The method arises from the observations that there are inevitable mismatches of the higher partials and that spectral variations of noise are matched as partials. The present method works globally, to ascertain whether the spectrum as a whole has a harmonic structure and if so, to match partials to the whole spectrum. However observations of real spectra show that harmonicity can be quite localised. For example, some waveforms are only composed of three or four partials, in which case global harmonicity is weak and detection becomes difficult. Also, in some sounds like the spoken voice, there are multiple resonances (the formants) where harmonicity is locally strong and other parts of the spectrum which appear to contain no harmonic structure (to the best resolution of the FFT).

Localisation, therefore, is the key to the new method proposed. First of all a fundamental frequency detection would be carried out focusing on the lower partials. Once the harmonic spacing is ascertained, the algorithm would scan through all frequencies for localised matches. These would be rated according to spectral shape (more peak-like → higher score). The propagation would then be extended in time, to search for similar matches in frequency, according to variations tracked in the fundamental. Where there are scores above a threshold, the time-frequency regions are confirmed harmonic, highest scorers first.

A Potential Method

The template method would be a good candidate for the initial estimate, using a short template with priority toward the lower partials (e.g. the fifth template of Figure 4.8 on page 100). For localised frequency scanning, a template of fixed scale could be generated at the spacing of the most prominent fundamental frequencies, for cross-correlation in the *linear frequency-scale* spectrum. A Scan-and-Propagate method similar to that described in chapter four could be used to link the fundamental frequency between frames. By normalising the frequency scales of each frame to each fundamental frequency trajectory, the partials would line up horizontally. Therefore a statistical maximum likelihood method could be used to determine which time-frequency regions are truly part of a harmonic structure. Potential methods include the Hough transform (commonly used for detection of lines or shapes in image segmentation [Vuillemin,1994]), or a Hidden Markov Model (HMM) method similar to Depalle, Garcia & Rodet's partial tracking algorithm [Depalle & Garcia & Rodet,1993].

Problem Solutions and Expansion to Polyphonic Modelling

The cepstrum method was useful in providing a confidence figure. However, such a figure would not be necessary for the localised search, because each part of the spectrum would be evaluated separately. Removal of the cepstrum method would reduce the computational complexity a little and would also solve the problem of an upper limit on the fundamental frequency. With a refinement to the template method (e.g. to reduce the cross-correlation output where there is energy in-between template peaks[2]), the problem of matching multiples would be reduced. These two alterations may solve many of fundamental frequency estimation limitations of the combined cepstrum-template method.

The problem of increased significance of fundamental frequency error would be solved by the frequency scan, since this would focus on partial spacing rather than absolute multiples. Therefore the error would be a fixed value throughout the spectrum. This has the added bonus of being able to capture stretched partials (see section 2.1.2.2).

The localisation in frequency would also provide a solution to classifying noise as partials, since partial peaks would only be sought in time-frequency regions confirmed as harmonic. Finally, the problem of extraneous noise would only interfere with partials that are in the same locality, so this would only be significant where it obstructs the initial fundamental frequency estimation.

[2] For example, if the minima of the template had values below zero, they would reduce the cross-correlation output according to the signal strengths at frequencies energy in-between the harmonics.

Polyphonic sounds or monophonic sounds with reverberation have the potential of being solved by this method if it is allowed to select more than one fundamental frequency, say iteratively from the most prominent downwards. If successful, this could provide the route to incorporating polyphony into the sound model.

6.2.2.3 Alternative Techniques and Recent Advances

As stated at the beginning of chapter four, general approaches to fundamental frequency estimation were considered rather than specific algorithms, to avoid the need for a time-consuming evaluation of the vast number of publications. This section describes some of the alternative techniques that have been developed or used for computer music applications in the last few years.

One approach that is gaining popularity is the neural network estimator. This takes frequency domain magnitudes as its inputs and is trained with examples that are either synthetic or that have been hand-scored. The outputs of the neural network are a set of probabilities, where each output corresponds to a particular frequency. Because frequency is a continuous variable, accurate estimation would require thousands of output nodes. Instead neural network estimators are usually applied to note detection systems for pitch-to-MIDI conversion or automatic score notation, where only the frequencies corresponding to pitches of the musical scale are required. One such example is [Petroni & Degrassi & Pasquariello,1994].

Closer to music analysis-resynthesis, a Harmonic Plus Noise model was developed for flexible transformation of speech [Laroche,1993; Laroche & Stylianou & Moulines,1993]. Like the Deterministic Plus Stochastic model (and in fact based on this decomposition), the harmonic aspect models partials and the noise aspect models the residual, although the analysis and synthesis methods differ. Within the harmonic aspect, the first step is fundamental frequency detection (as with the method of chapter four), but this is done in the time domain and therefore provides no upgrade path to polyphony.

Perhaps the most promising fundamental frequency estimation technique for music analysis-resynthesis is the maximum likelihood estimator developed by Boris Doval and Xavier Rodet [Doval & Rodet,1991]. The authors state the aims of providing a robust algorithm “that would give a correct value each time a human observer is not confused”. Using a probabilistic model, the estimate is first coarsely estimated and then refined within the most likely frequency range. Although the method is different from the work of chapter four, this latter aspect has the same basis as the template method (where the harmonics contribute to the estimate with a finite frequency distribution, scaled according to the frequency domain magnitude). In a later development inter-frame tracking was introduced using a Hidden Markov Model method [Doval & Rodet,1993]. The authors claim good results for both aspects. As with the above methods, this section contains only summaries of alternative techniques without evaluation.

6.3 Attack Transients

6.3.1 Impact upon the Model

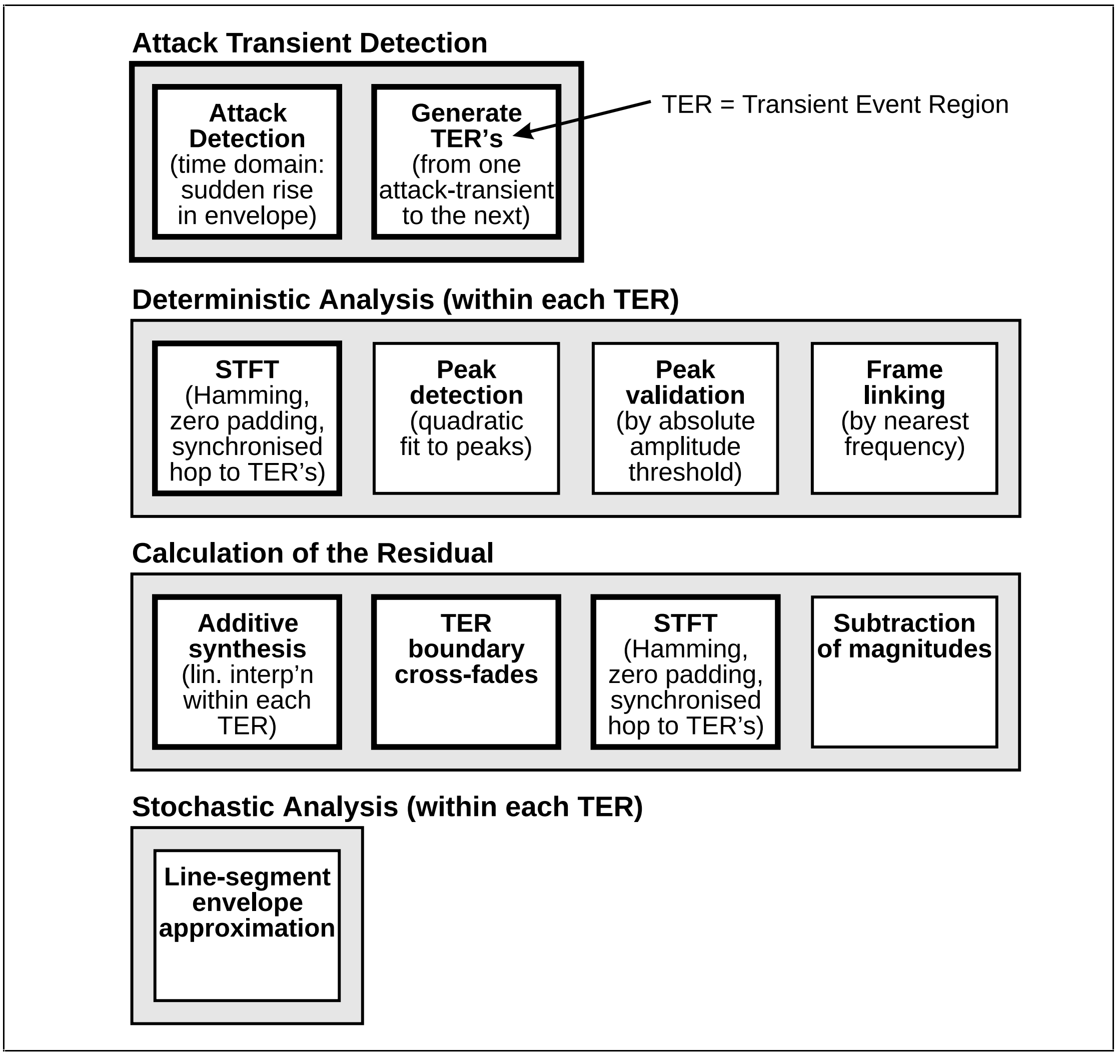

Attack Transient Detection
Attack Detection (time domain: sudden rise in envelope)
Generate TER's (from one attack-transient to the next)
TER = Transient Event Region
Deterministic Analysis (within each TER)
STFT (Hamming, zero padding, synchronised hop to TER's)
Peak detection (quadratic fit to peaks)
Peak validation (by absolute amplitude threshold)
Frame linking (by nearest frequency)
Calculation of the Residual
Additive synthesis (lin. interp'n within each TER)
TER boundary cross-fades
STFT (Hamming, zero padding, synchronised hop to TER's)
Subtraction of magnitudes
Stochastic Analysis (within each TER)
Line-segment envelope approximation

Figure 6.5 – Changes to the Initial Model's Analysis to incorporate Attack Transients

The Initial Model has no means for modelling rapidly varying spectra, so the inclusion of attack transients requires a greater modification to the model structure. Nevertheless, it has been achieved without much disruption to the existing structure. There is a new section: the pre-analysis, which generates the Transient Event Regions (TERs). There is also an alteration to the control structure of the analysis and synthesis, to work within regions. See Figure 6.5 (above) and Figure 6.6 (overleaf). Figure 5.14 in section 5.4 (page 153) shows this in a more signal-oriented form. Section 5.4 also discusses the computational impact of introducing attack transients into the analysis-resynthesis system.

In terms of performance, the inclusion of attack transients as a model feature has a significant and immediately perceptible effect upon the model. Time-frequency resolution remains a problem in sound modelling, but the device of synchronising analysis and synthesis to percussive note onsets has been a significant step towards overcoming these limitations. It is

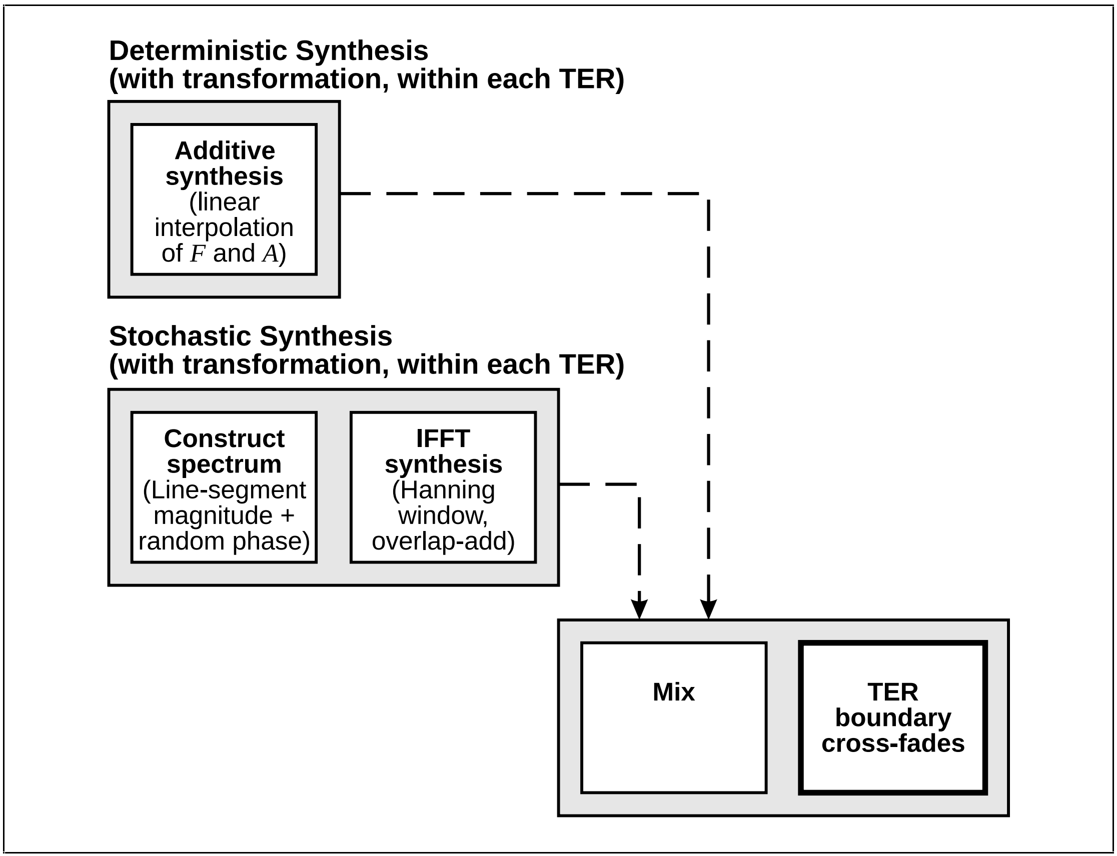

Deterministic Synthesis
(with transformation, within each TER)
Additive synthesis (linear interpolation of F and A)
Stochastic Synthesis
(with transformation, within each TER)
Construct spectrum (Line-segment magnitude + random phase)
IFFT synthesis (Hanning window, overlap-add)
Mix
TER boundary cross-fades

Figure 6.6 – Changes to the Initial Model's Synthesis to incorporate Attack Transients

recognised that ultimately the TFR needs upgrading and that this should yield a more elegant and complete solution.

6.3.2 Critique and Future Directions

The reliability of the time domain attack transient detection method has been demonstrated. Synchronisation has also been effective in retaining the crispness of the attack – the sharp rise in the envelope and the sudden change in the spectrum account for this. However there is (perceptual) detail missing from the synthesised signal immediately following the attack. The first half frame includes extrapolated data, so the spectrum is static and so is the time domain envelope.

A short term improvement might be afforded for some sounds by imposing the time domain envelope immediately following the attack. In this way, sounds with a fast decay will retain that decay. For further improvement, a better understanding is needed about the rapidly changing spectrum during and immediately following the attack. A clue to this may be found by examining the ongoing advances in physical modelling.

Ultimately, improvements are dependent on the ability to see more in the time-frequency domain. Therefore transient modelling relies on improvements to the TFR. Such improvements will probably also yield the resolution to be able to individually capture closely spaced attacks, such as from a drum roll.

6.3.2.1 Alternative Techniques and Recent Advances

Since 1987, the Wavelet Transform has been recognised as a powerful tool for detecting sudden changes in the spectrum [Grossman & Holschieder et al.,1987] of musical sounds, and new methods continue to be developed (e.g. [Solbach & Wöhrmann,1996; Tait & Findlay,1996]) on this basis. Like the FFT, the WT provides both a magnitude and a phase spectrum, but because it contains different time-scales at different frequencies, sudden changes in the spectrum can be located with very high accuracy. The WT magnitude spectrum naturally displays a sudden rise in energy at an attack transient, but it is the phase spectrum that has been found most effective for pinpointing the location of sudden changes.

6.4 The Results Combined in a Single System

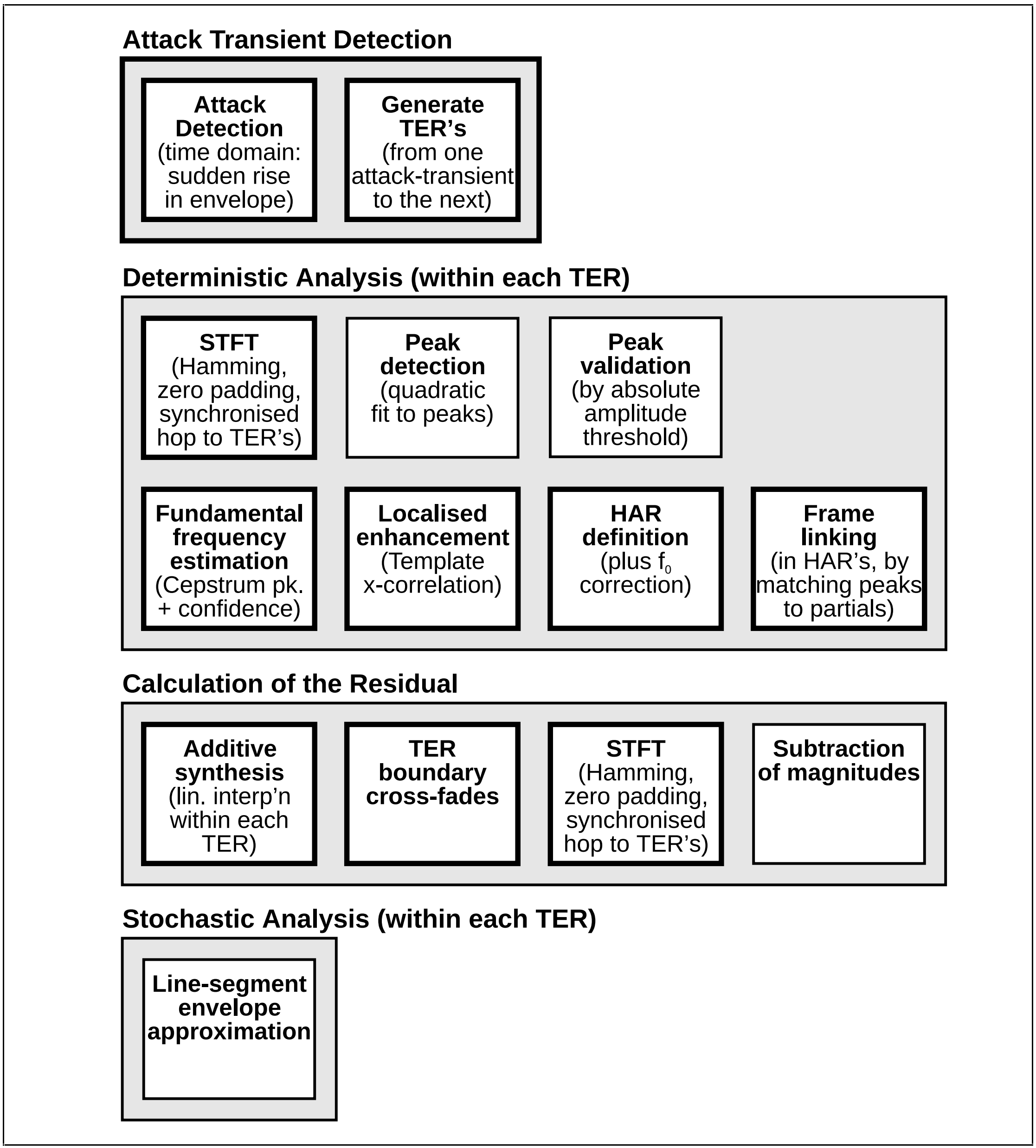

Attack Transient Detection
Attack Detection (time domain: sudden rise in envelope)
Generate TER's (from one attack-transient to the next)
Deterministic Analysis (within each TER)
STFT (Hamming, zero padding, synchronised hop to TER's)
Peak detection (quadratic fit to peaks)
Peak validation (by absolute amplitude threshold)
Fundamental frequency estimation (Cepstrum pk. + confidence)
Localised enhancement (Template x-correlation)
HAR definition (plus f_0 correction)
Frame linking (in HAR's, by matching peaks to partials)
Calculation of the Residual
Additive synthesis (lin. interp'n within each TER)
TER boundary cross-fades
STFT (Hamming, zero padding, synchronised hop to TER's)
Subtraction of magnitudes
Stochastic Analysis (within each TER)
Line-segment envelope approximation

Figure 6.7 –Harmonic Linking and Attack Transients combined within one system (Analysis only shown)

The Final Model of the thesis, in contrast with the Initial Model, includes the modifications for harmonic frame linking and attack transients. Phase Distortion Analysis with its potential for many attractive enhancements is not included, because the technique in its present form is not sufficiently robust for making measurements from practical multicomponent spectra. The reasons for this and future work solutions were discussed in section 6.1.2.

Since the presence of an attack transient indicates a sudden change in the spectrum, it would be undesirable for a harmonic region to be defined that encloses it. Therefore the implementation of Figure 6.7, in which transient event regions are defined first and take precedence over the definition of harmonic regions, ensures compatibility between the two modifications. In fact they work well together, particularly for speech, where transient event regions locate plosives and harmonic regions locate voiced phonemes. (Harmonic linking is implemented without changes to the synthesis processes, so Figure 6.6 shows the synthesis system for the Final Model.)

If Phase Distortion Analysis were included...

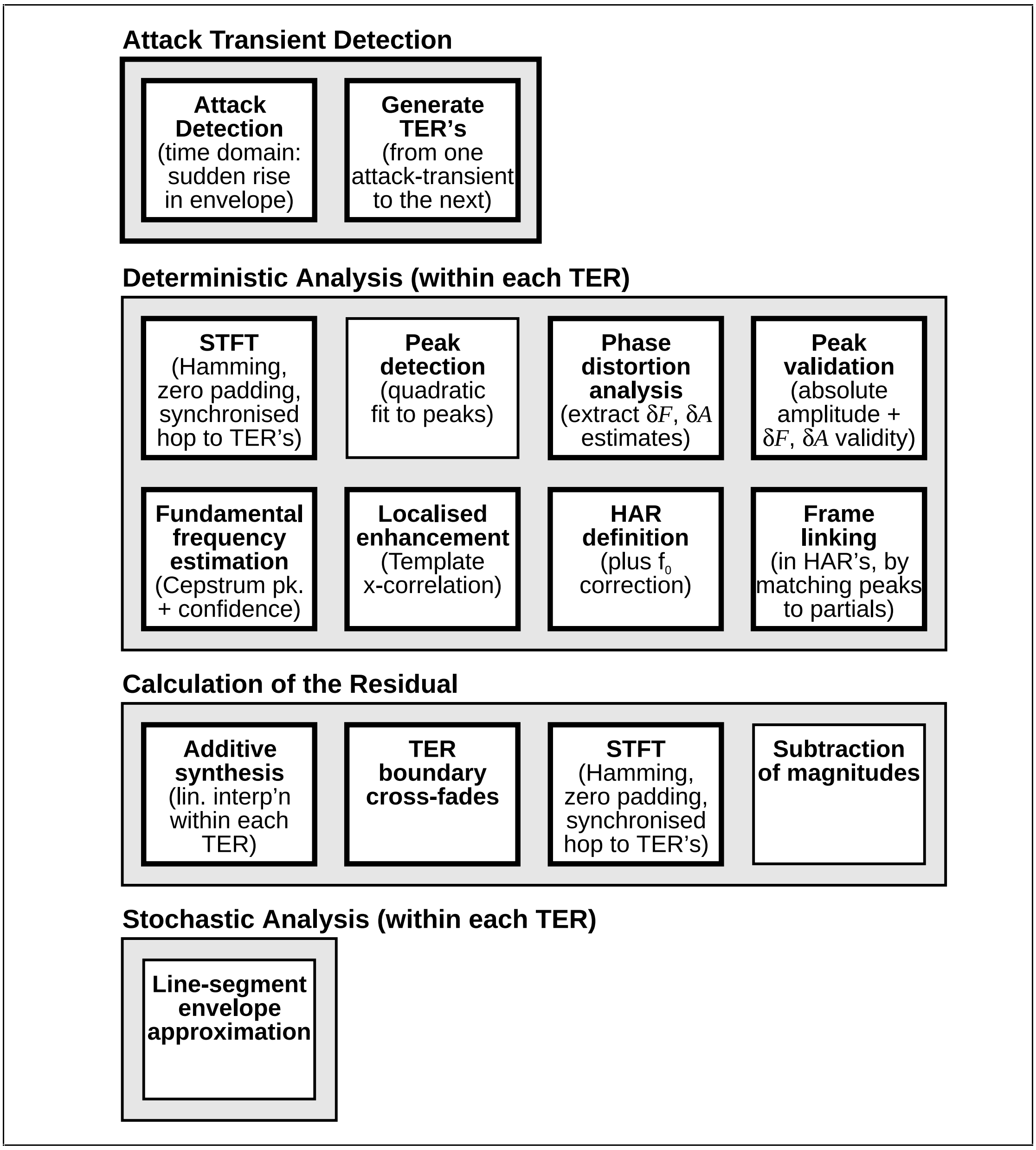

Attack Transient Detection
Attack Detection (time domain: sudden rise in envelope)
Generate TER's (from one attack-transient to the next)
Deterministic Analysis (within each TER)
STFT (Hamming, zero padding, synchronised hop to TER's)
Peak detection (quadratic fit to peaks)
Phase distortion analysis (extract δF, δA estimates)
Peak validation (absolute amplitude + δF, δA validity)
Fundamental frequency estimation (Cepstrum pk. + confidence)
Localised enhancement (Template x-correlation)
HAR definition (plus f0 correction)
Frame linking (in HAR's, by matching peaks to partials)
Calculation of the Residual
Additive synthesis (lin. interp'n within each TER)
TER boundary cross-fades
STFT (Hamming, zero padding, synchronised hop to TER's)
Subtraction of magnitudes
Stochastic Analysis (within each TER)
Line-segment envelope approximation

Figure 6.8 – All three innovations within one system (Analysis only shown)

With the expectation that PDA will be reworked for greater reliability, we can conjecture the benefits of the three presented advances in association with one other. See (Figure 6.8.) Attack transient analysis would improve representation and synthesis of percussive note onsets. Harmonic linking would improve deterministic synthesis of harmonic sound signals. Phase distortion analysis would validate peaks prior to frame linking and would give extra detail to the deterministic synthesis trajectories.

There would also be synergies, where roles overlap. For example, PDA measurement applied to harmonic linking could aid in validating peaks for matching to partials. Not only could side-lobes be rejected prior to matching; it could also assist in resolving conflicts, by providing the additional information of frequency trajectories.

6.5 Noisy Partials – The Basis for a New Model of Sound

> "According to the philosopher Ly Tin Wheedle, chaos is found in greatest abundance wherever order is being sought. It always defeats order, because it is better organised."
>
> ——— *Terry Pratchett (in Interesting Times [Pratchett,1995])*

In this section, the foundations for a new model of sound are presented. Although this sounds like a radical reinvention of the sound model, it is actually an evolutionary step. Nevertheless, it will require significant reshaping of the model and in particular the analysis and synthesis methods. The idea has come from several observations of different aspects of the current model, including its conceptual basis, properties of the signals it aims to model and the strengths and weaknesses of its results. These are discussed in section 6.5.1 following. In section 6.5.2 the basis for the new model is introduced, which has been termed the 'Noisy Partial'. A preliminary sketch describes how the model could be implemented as the feature set and the analysis and synthesis processes, and how transformations would be effected. This sketch is intended as a suggested research direction for further investigations.

6.5.1 Observations about the Deterministic Plus Stochastic Model

6.5.1.1 Observations Regarding Capture of Feature Detail

Subjective

The noisiness in synthesised sounds does not always fuse with the tonal aspect – they appear to come from slightly different sources – an effect which is particularly noticeable for speech. (Inaccuracies in *speech synthesis* are probably the most apparent because we are so attuned to the human voice for communication.) The deterministic aspect alone has a 'metallic' quality, that probably derives from the smooth interpolations of the partial trajectories. Although the stochastic aspect was designed to add back the missing roughness, it is not completely successful at masking the metallic quality and this may account for the perceived lack of fusion.

System Inaccuracy

In the critique of the calculation of the residual (section 2.3.4.1) it was noted that even when the deterministic aspect captures frequency and phase accurately, there is a difference in the width of the main lobe peaks between the source sound and the deterministic synthesis, the source sound's being wider. From the observations of chapter three (section 3.3) we know that deviation from stationarity causes widening of the main lobe. Yet the gradual frame-to-frame variations *are* captured by the deterministic. This suggests that there are additional nonstationarities which must occur on a time-scale much smaller than a frame.

Furthermore, the PDA analysis can successfully extract the frequency and amplitude trajectories (of the higher amplitude partials), which implies that the signal is a good approximation to LFM and EAM in each frame. Therefore any undetected detail in the frequency and amplitude trajectories must be small relative to the average frequency and amplitude. In other words, the analysis is failing to capture small-scale rapid variations of the partial trajectories.

Sound Generation

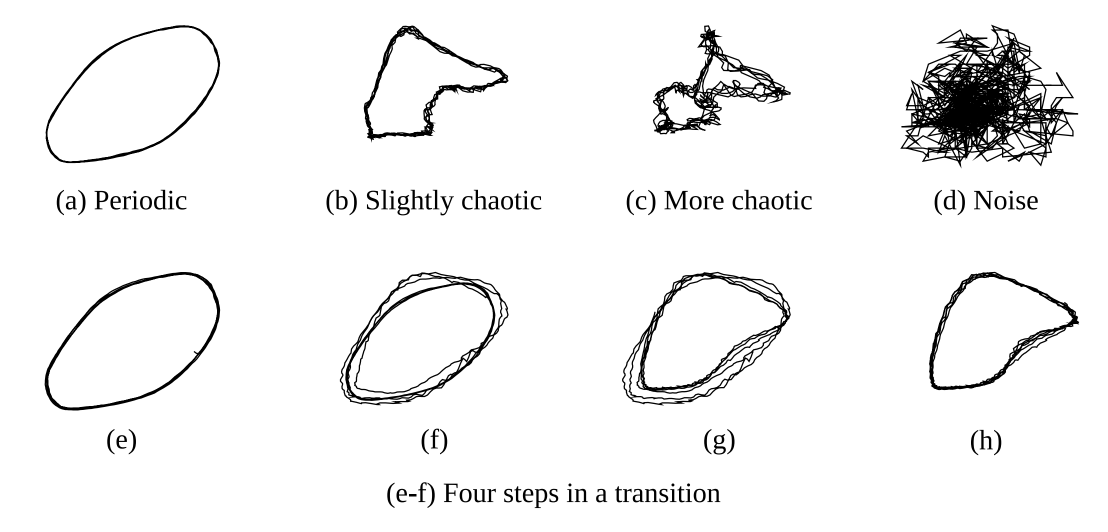
Chapter two's description of what is a musical signal revealed that not all noise is additive (section 2.1.5.1). This concurs with the above observations, in suggesting that there is more detail to the modulation, but it also suggests that the undetected modulation is noisy or chaotic. Intuitively this sounds correct, since the sense of missing detail in the deterministic synthesis is a lack of *roughness*, which in textural terms implies random or quasi-random mid-to-high frequency modulation.

Evidence from the field of physical modelling of instruments also points to the need for small-scale instabilities (see section 2.2.1.3). In the case of these models, they are introduced through nonlinear feedback, such that the partial trajectories are perturbed about their average value in a small-scale chaotic sense. (There is sufficient instability to cause the system not to stabilise to a steady-state condition, but not so much as to destroy the resonant modes which give rise to the partials).

Signal Representation

Neither the time domain waveform nor the time-frequency sonogram provide the necessary evidence of the postulated small-scale chaotic modulation. However the phase space plots the time domain sample against a delayed version of itself (for the two-dimensional case), so by comparison from one period to the next, it is capable of revealing micro-variations that occur within fractions of the period.

Every waveform shape presents a different shape in phase-space, just as it does in a time domain plot. However at the same point in each period, the x and y co-ordinates of the phase space will be more or less identical (by definition of periodicity). See Figure 6.9.



(a) Periodic (b) Slightly chaotic (c) More chaotic (d) Noise

(e) (f) (g) (h)

(e-f) Four steps in a transition

Figure 6.9 – Two-dimensional phase space representation of sound signals

If the signal is truly periodic over the measurement interval, then the phase space will display a closed figure, since each period will exactly overlay the previous period (Figure 6.9a). Note that the complexity of the shape is a function of the waveform and has **no** bearing on complexity of the partial trajectories.

If there are small variations between periods, the shape will still appear closed, but instead of a single curve there will be multiple strands that *approximately* trace the same path. For signals with greater and greater period-to-period variation, the variation in the strand paths will increase until, in the limit, the signal is totally aperiodic or random and its phase space representation will no longer be a closed shape (Figure 6.9a-d). Often, where the signal is gradually evolving (as assumed by the Deterministic Plus Stochastic analysis-resynthesis method), the shape will appear to morph as strands progressively deviate in an ordered manner from one position to another (Figure 6.9e-h). In contrast, small-scale chaotic modulation can be observed as sudden, temporary deviations of a strand path from the average path (Figure 6.9b-c).

6.5.1.2 Observations Regarding The Deterministic-Stochastic Classification

The Binary Classification is Unnatural

The assertion here is that nature is rarely black or white, or in the language of Boolean logic: TRUE or FALSE, A or NOT A; this is an underlying principle borrowed from fuzzy logic [Kosko,1994]. Put another way, it says that classification is something imposed by an observer and is not an intrinsic property of an object. In nature, properties of objects are often amorphous – they differ only slightly from one object to another, or one moment to the next – whereas bivalent classification imposes hard thresholds that state "This side of the boundary is property A; that side is NOT A". This is the case with the Deterministic-Stochastic classification, which is intended to have the meaning 'DETERMINISTIC or NOT DETERMINISTIC'. (In actual fact the classification as implemented is not as complete as this suggests, as discussed under the heading *The Deterministic-Stochastic Classification is Incomplete* to follow.)

Classifications are undoubtedly useful devices for organisation and they can be helpful when learning new concepts. However, when applied to a model that acts on real data, the hard thresholds between properties cause problems.

The deterministic classification models sound elements as time-varying sinusoids, whilst the stochastic classification models sound elements as dynamic broadband noise. Wherever the threshold may be placed between these two, a problem will arise. More dynamic sinusoids will have significant bandwidth and although best synthesised as sinusoids, may be rejected from the deterministic classification. Similarly, fairly stable, fairly narrowband noises may be classified as deterministic and erroneously synthesised as sinusoids.

Fuzzy logic philosophy summarises this by saying that problems arise close to the threshold where the actual property is not 100% A, nor is it 100% NOT A. Therefore a binary system that considers only the classifications A or NOT A must, by definition, actively introduce errors.

The Deterministic-Stochastic Classification is Incomplete

In addition to the problems of the bivalent classification between deterministic and stochastic, the descriptions are actually incomplete in their description of the timbre space of all musical sounds. See Figure 6.10. 'Deterministic' is designed to model stable sinusoids. This encompasses signal components that are very narrowband, gradually varying and not short-lived. 'Stochastic' is supposed to model the remainder but, as implemented, it is only suitable for wideband, gradually varying components (with no restriction on longevity).

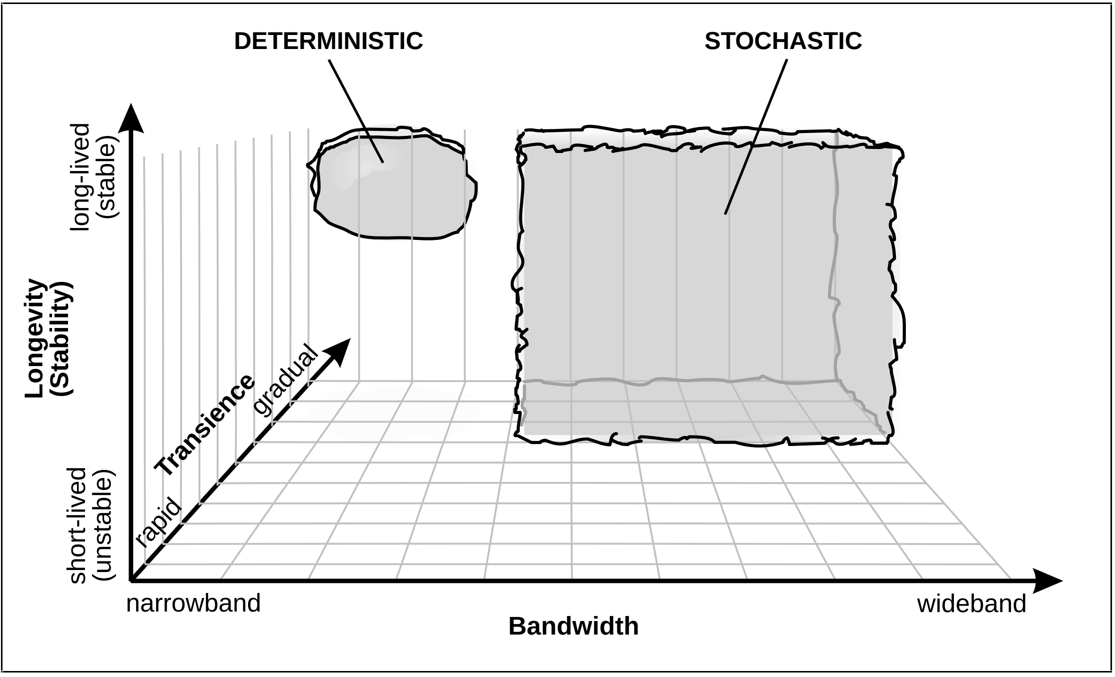


Figure 6.10 – Sound component capture by the Deterministic-Stochastic classification

Signals of a small-to-medium bandwidth may be rejected from the deterministic classification, but the stochastic aspect (whose features are line-segment spectral envelopes) does not have the resolution to capture more than the broad envelope profile. It is debatable, though, whether a deficiency in frequency detail is actually problematic, because the ear is less pitch sensitive to noisy sounds (see section 2.1.5). Certainly, when a sound with both deterministic and stochastic components is synthesised, the lack of frequency detail in the stochastic synthesis is not noticeable. Sounds that do not have a deterministic component, though, sound too rough when synthesised from the line-segment spectral envelope. (Good capture is possible by using the residual magnitude spectrum directly, without making a line-segment approximation. However this requires many times the data storage space.)

Similarly, rapidly varying components, whether narrowband or wideband, are rejected from the deterministic classification, but stochastic modelling is also frame-based at the same resolution, so these elements are not represented here at all. Figure 6.10 shows the unmodelled space between narrowband and wideband, the short-lived sinusoids "under" the deterministic capture region and all rapidly varying components. (Note that the deterministic-stochastic as described here refers to the definitions and implementation given by Serra. Work in this thesis has altered the classifications somewhat; for example, the inclusion of attack transients enables capture of the most rapidly changing elements at the front of the graph in Figure 6.10.)

The Feature Set Requires Two Incompatible Data Sets

The deterministic data set is composed of sinusoidal trajectories described by lists of frequencies and amplitudes. The stochastic data set is a list of spectral envelopes. This is undesirable and limiting from a number of perspectives:

- The analysis requires two passes. In this model, they cannot be done in parallel and they require a (deterministic) synthesis in-between. This contributes largely to the inefficiency of the current model. The sequential analysis also means that errors in one aspect of the model are propagated to the subsequent stages (where they can have an increased effect). i.e. errors in deterministic analysis appear in the stochastic analysis where they can be duplicated.
- Two data sets are required, which inevitably have an increased storage requirement. It also impacts transformation since two modification functions are required for each transformation.
- The data types are incompatible with one another. This reduces the advantages of system unification that could be facilitated by an acoustic waveform model (discussed in the next chapter on model applications). It also impacts applications that need to use the model as a basis for more complex functions, since the data types cannot be reconciled with one another. For example, in auditory scene analysis[3], partials can be grouped together on the basis of their frequency spacing and the common trajectory modulation, but it is impossible to say which part of the stochastic aspect comes from which source or how it associates with the partial trajectories.
- Both data types overlap in the time-frequency domain. From a modelling perspective this is undesirable because some parts of the time-frequency domain are described twice.
- Two synthesis architectures are required, which doubles the system complexity, computation time and the number of synthesis parameters, all of which are hindrances to real-time synthesis.

6.5.2 The Noisy Partial

Having observed a deficiency in the model's sound quality and having hypothesised the source of the problem, it remains to find a solution in terms of the model feature set and the analysis-resynthesis processes. Since it appears that detail is missing from the partial trajectories and that this could account for the lack of roughness of the deterministic aspect, a solution is proposed that adds the noise to the partial trajectory itself, thereby creating a noisy partial.

A 'Noisy Partial' is a partial that is chaotic *to some extent*. If it is 0% chaotic, it is a stationary sinewave. If it is 100% chaotic, it is white noise. The former is the limit of narrowband stability, whilst the latter is the limit of wideband chaos. Close to these extremes

[3] Comparable to visual scene analysis, in which an image can be segmented into its component objects (e.g. cars, trees, road, sky from a landscape photograph), auditory scene analysis aims to segment sound by the sound sources (e.g. vocals, strings, oboes, drums, …). This is attempted by analysing the spectrum and then grouping together features which are most likely to have arisen from a single source.

lie the old definitions of deterministic and stochastic. Further towards the centre lie more ambiguous definitions: fairly narrowband noise, fairly chaotic sinusoid.

Model Feature Set

'Noisy Partial' has a more fluid definition than the traditional sinusoidal partial, so it cannot be described as easily in terms of 'a frequency', 'an amplitude' or 'a phase'. Instead it is replaced by the concept of a distribution, which has certain properties to describe its frequency location, its shape and size and its dynamic. An example might be centre frequency, amplitude, bandwidth and an index/indices for the degree of chaos $\left(F_C, A, B, \xi\right)$.

Of these parameters, the first three describe placement and scaling within the time-frequency plane. The chaos factor is less obvious; it describes how an element's distribution is filled in terms of the phase function, varying from ordered to random.

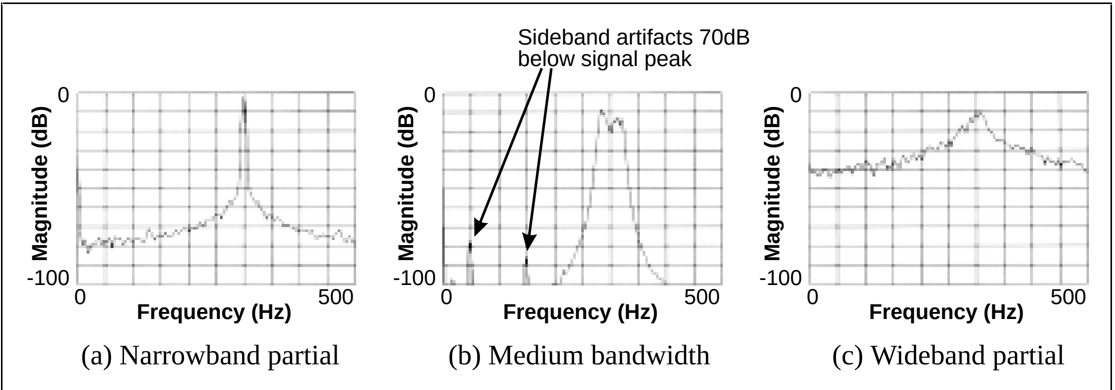
Analysis

The Noisy Partial is a distribution and the form of a partial in the STFT is also a distribution. Although the shapes do not have a direct correspondence, there is some correlation in that the more chaotic a partial, the wider its peak in the STFT representation. As a first approximation, the STFT partial peak measurements could be directly substituted into the model parameter set.

For a more thorough investigation the true correspondence would have to be ascertained. This could be achieved using the bispectrum, recently shown to demonstrate the degree of small-scale frequency variations of partials [Dubnov & Tishby & Cohen,1996]. Alternatively, a PDA-style approach (see section 3.3) could be used, in which the FFT distribution (or whatever supersedes it) is directly observed and characterised for different degrees of chaos.

Synthesis

The concept of the Noisy Partial is a sinusoid, frequency-modulated by noise, so this is also the obvious basis for the synthesiser architecture. For a sinusoid partial (0% chaos) the modulation would be zero, and for white noise (100% chaos) it would be maximal (in theory infinite). In valedictory experiments a sinusoid's frequency was perturbed at different rates and by different amounts according to a flat random distribution and achieved outputs that were controllable between these extremes. (In effect, the experiments used a single square wave oscillator of fixed frequency and random amplitude, where the frequency and amplitude range were under programmer's control.) Examples are presented in Figure 6.11.



(a) Narrowband partial (b) Medium bandwidth (c) Wideband partial

A Noisy Partial with centre frequency at 300Hz and three different modulations
Each diagram shows the mean of magnitude FFTs of eight recordings

Figure 6.11 – Noisy Partials generated by noise modulation of a sinusoid

Further investigations are required to ascertain a good level of control. Fitz and Haken [Fitz & Haken,1995] recently introduced a technique which they term Bandwidth Enhanced Sinusoidal Modelling, to include some noise into the partials (although their system does not go as far as to obviate the need for a separate stochastic aspect). (In terms of the Noisy Partial approach their system remains relatively close to the sinusoidal end: say 0-10% chaos factor.) They claim good results using narrowband noise to frequency modulate the partials. In light of this, a promising approach would be to experiment with noise whose frequency profile is a normal distribution, where amplitude, centre frequency and bandwidth can be controlled. The suggested synthesis scheme is summarised in Figure 6.12.

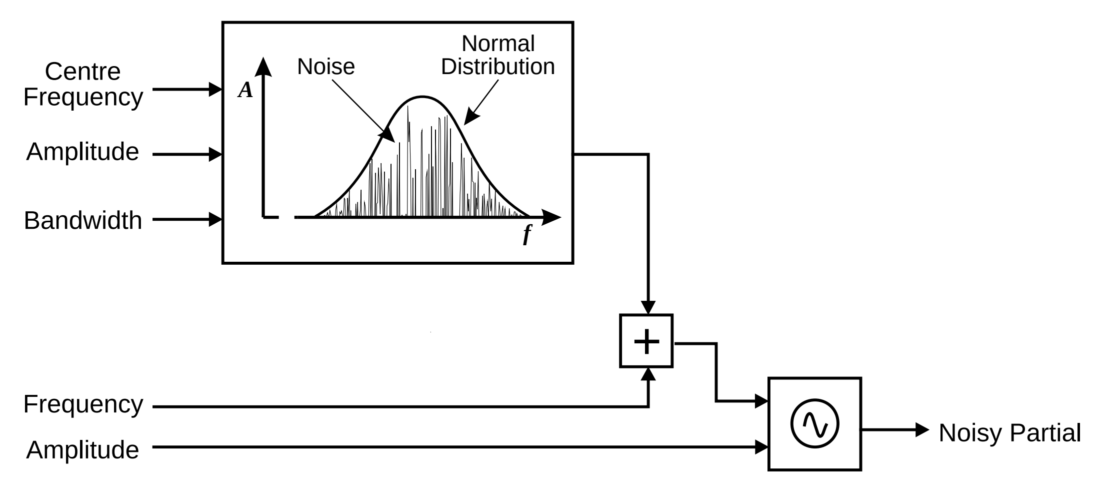


Figure 6.12 – Synthesis of a Noisy Partial

Transformation

From the first page of the thesis, the necessity for musically relevant parameters has been stated and restated, because they provide flexibility for musical transformation. The Noisy Partial model provides a means for describing time-varying partials and shaped noise (whose relevance was established in section 2.1) with a single description.

In addition to the now commonplace transformation of time-stretch, pitch-shift and cross-synthesis, the Noisy Partial enables a whole host of new transformations. Through variations in the index of chaos, spectral elements can be made smoother or rougher, more tonal or more noisy; through variations in bandwidth, they can become localised or dissolved into the mass of spectral energy.

In summary…

Observations from different sources combine to indicate that the Deterministic Plus Stochastic model does not represent noise accurately. It seems that the partial trajectories of the deterministic aspect are too smooth and therefore a new scheme has been proposed that introduces noise into the partials through chaotic frequency modulation. The presented Noisy Partial model actually proposes that all noise for the sound model could be encapsulated, manipulated and synthesised from this format.

Although this may appear to be a radical re-invention of the sound model, it actually follows as a multivalent description in which ‘deterministic’ and ‘stochastic’ are the bivalent limits. By providing a smooth transition between these extremes, it is anticipated that real signals will be better represented. In particular it has a better chance of modelling the noisiness of sounds because its description of noise (including additive and higher order modulation noise) has its foundation in observed phenomena. By exchanging exact partial trajectories for distributions, the Noisy Partial model unifies all sound components within a single format. By virtue of this and its new parametric controls, it also facilitates much greater power for transformation, and at the same time reduces the computational burden for analysis and synthesis.

Initial investigations demonstrate that the idea has promise and suggestions are put forward for its development. In terms of the observed deficiencies of the Deterministic Plus Stochastic model, all the points raised in section 6.5.1 are solved… with the exception of the subjective result – a point that remains to be tested.

CHAPTER SEVEN

APPLICATIONS OF THE SOUND MODEL

7. APPLICATIONS OF THE SOUND MODEL

7.1 Applications in Musical Systems

Music synthesis can be broken down into two classes: generative and transformative. The former refers to synthesis schemes that generate sound by virtue of their architecture. This includes all the classic schemes employed in commercial music synthesisers, such as analogue, FM, digital waveshaping and now physical modelling (digital waveguide). A common feature to these schemes is that highly complex sounds can be generated and controlled with a handful of synthesis parameters.

Transformative synthesis refers to sound generation schemes that rely on a source sample which is transformed (or directly reproduced). These include samplers and now analysis-resynthesis systems. Unlike generative synthesisers, the architecture is general – it is not specific to a sound type or instrument – so it can accommodate a wide and potentially infinite range of sounds. Also unlike generative synthesisers, the samples or parameters have large memory requirements so that they, rather than the architecture, describe the sound type.

The sound model of this thesis naturally takes the form of an analyser-resynthesiser, which is a transformative technique. Its applications in this role are discussed in section 7.1.1. Through customisation with a suitable interface, the sound model can also be used as a generative synthesiser. Though little explored, this potential is discussed in 7.1.2.

7.1.1 Analysis-Resynthesis for Advanced Post-Production Editing

7.1.1.1 Post-Production Editing Suites

Traditionally the reserve of wealthy music studios, the post-production audio editing suite is a tool whose role lies in-between the recording and mastering stages of music production. It has all the functionality of a traditional ‘multi-track tape recorder plus mixing desk’ set-up but with much more flexibility and new tools. With a mixing desk, it is possible to set the ‘mix levels’ so that each audio track appears at the desired level relative to the other tracks. In addition, pan and EQ settings may be adjusted and, using the Send and Return ports, tracks may be transformed through external effects units (e.g. reverb units, compressors).

The post-production editing suite allows all of these operations to be performed within a single system, with the extra functionality that tracks can be non-linearly edited (i.e. cut, splice, duplicate). The ‘post-production’ tag describes the fact that the volume, pan etc. requirements for each track can be set and changed individually, and at the leisure of the operator. (The mixing desk approach may provide much freedom, but the tracks are being recorded in real-time, so there is a limit to how much can be done.)

The last few years have seen the emergence of post-production audio editing suites using custom hardware but controllable through a personal computer. The affordability, quality and flexibility of such systems has guaranteed their use from Hollywood film sound tracks to semi-professional music studios. (Recently, 'audio sequencers' have also become available which add limited track-based audio editing capabilities to PC-based MIDI sequencers, but without the advantages of additional hardware[1].)

Using hard disks as the storage media, these editing suites can store in excess of three track-hours of CD-quality audio on a single disk[2]. The access speeds of hard disks are now sufficiently fast to facilitate real-time recording and playback of several tracks simultaneously, so they are transparent, appearing to the user as accessible as onboard memory.

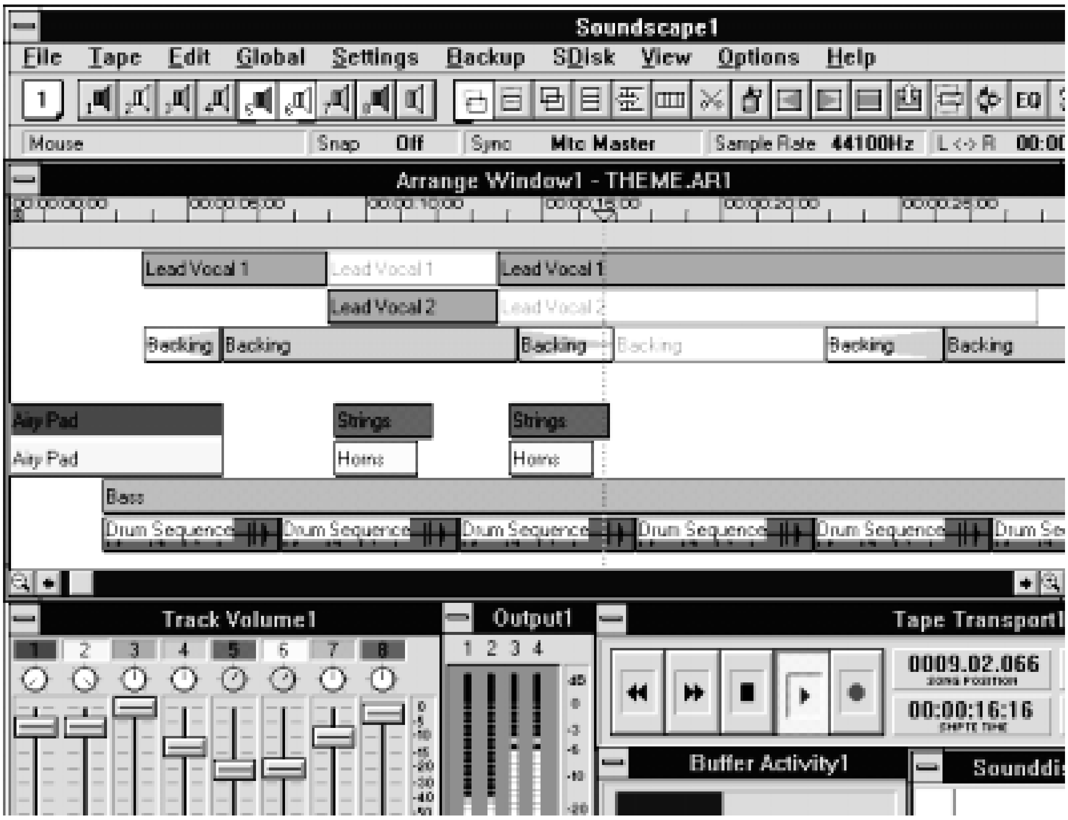


Figure 7.1 – Arrangement of Audio Tracks in a Post-Production Editing Suite (example screen-shot courtesy of Soundscape Digital Technology Ltd.)

Within the graphical user interface (GUI) of the editor, the user places the recordings as coloured blocks into an 'arrangement window', where the height is one track and the width is the duration of the recording. See Figure 7.1. The power of the editor is that multiple non-linear edits can be made without modifying or destroying the recorded material. Current systems also have the processing capacity to apply time varying volume, pan and EQ settings. In addition, more complex transformations can be applied in non real-time, although this requires extra storage space so that both the original recording and the modified version can be

[1] namely speed (which translates to track capacity) and signal cleanness (PC soundcards are notoriously noisy, because they pick up the digital 'chatter' from other components in the system, such as the hard disk or the video card).

[2] A 1Gb hard disk can store 3.38 track hours of 16 bit audio recorded at 44.1kHz.

retained. Upon playback, a cursor advances along the time-line of the arrange window and the output sound is a mix (the additive combination) of all tracks that co-exist at the location of the cursor.

There are three primary applications for post-production editing suites:

1) Audio tracks and voice-overs/voice-dubbing for synchronisation with film and video sequences;

2) Audio sequences and voice-overs for radio commercials and jingles;

3) Music composition and arrangement in a style similar to MIDI sequencing, but in which every sound element is actually a recording.

Audio for video and audio-only applications require tight control of timing and duration, so time-stretch transformation is an essential tool. For subtle retuning, pitch-shift can also prove useful. Arrangement of music tracks opens up more creative opportunities for transformation tools, that can play an active role in composition.

7.1.1.2 Enhancement Through a Sound Model

The sound model, employed as an analysis-resynthesis tool, is the perfect component to satisfy these requirements for transformation. Unfortunately, whereas the musical applications may forgive modelling imperfections if flexibility is forthcoming, the more commercially important audio and audio-video applications require resynthesis that is indistinguishable from the original sound. This is especially true for speech recordings.

Advances in sound modelling (especially the move to using harmonic regions – see chapter four) have improved the accuracy of the resynthesised sound (with a particular emphasis on speech). However, shortfalls still exist that have to date prevented the employment of analysis-resynthesis in these editors. The problems include, for example, the non-fusion of deterministic and stochastic aspects which lends a metallic quality to speech, and the pre-/post-echo of the stochastic aspect, as discussed previously.

The model weaknesses appear to be largely problems of capturing detail, rather than an outright inability to detect features. The synthesised sound is generated directly from the model's feature set, so analysis failings translate to compromised synthesis quality. As a route to finding immediate practical applications of sound modelling, the analysis results could be used to enhance proven techniques of sound manipulation, where feature detection alone is sufficient. To test this premise, a simplified analyser was used to enhance a standard time-stretch algorithm. In this experiment, emphasis was placed on sound quality for speech.

The results demonstrated improvements in sound quality and rhythmic accuracy for time-stretched speech that were better than a widely accepted commercial time-stretch algorithm. The same was true for the singing voice. In fact the sounds could be stretched up to about 150% before any degradation was noticeable. In addition, poor quality speech recordings and solo instruments were time-stretched to a higher quality than the sounds resynthesised from the model alone. For these latter sound examples, the comparison with the commercial time-stretch showed that certain aspects were improved upon and others were found not to be as robust. (Improvements to these aspects are ongoing.) The algorithm was even capable of reasonable time-stretch quality for complex recordings containing many instruments, the human voice and percussion, scenarios that are considered very challenging for time-stretch and too challenging for the sound model at present.

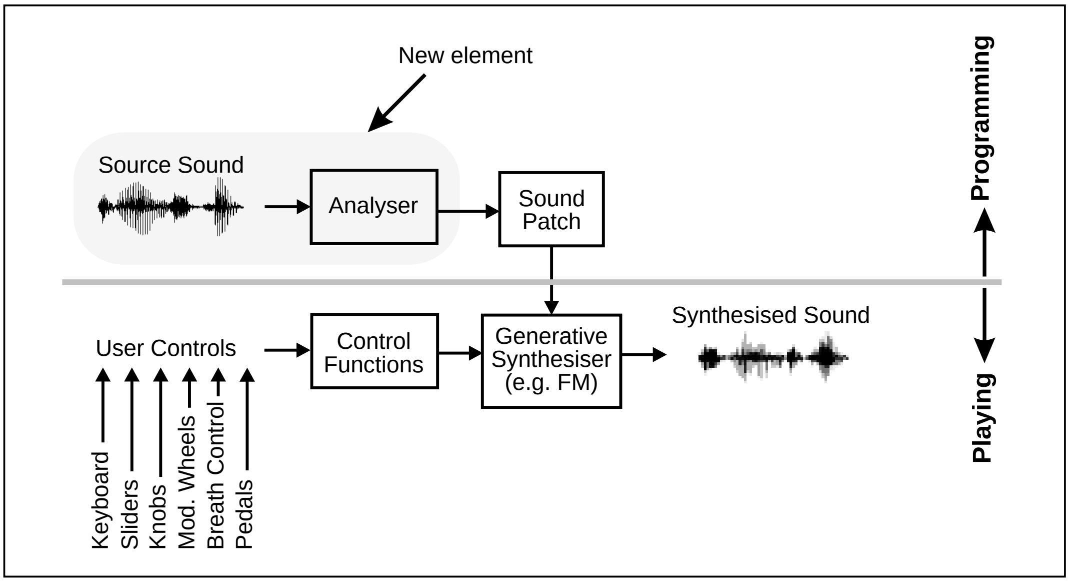
7.1.2 Generative Synthesis of Sound Model Features

Synthesis architectures naturally tend towards being either generative or transformative. However, with adaptation, either can be made to behave like the other. In section 2.2.2.3, methods were mentioned that place an analysis engine in front of FM and wavetable synthesis schemes, thereby providing them with the capability for direct reproduction and transformation. Similarly, by replacing the analyser of an analysis-resynthesis system with a parametric control structure, it becomes possible to impose generative-like synthesis upon the sound model. See Figure 7.2 (overleaf).

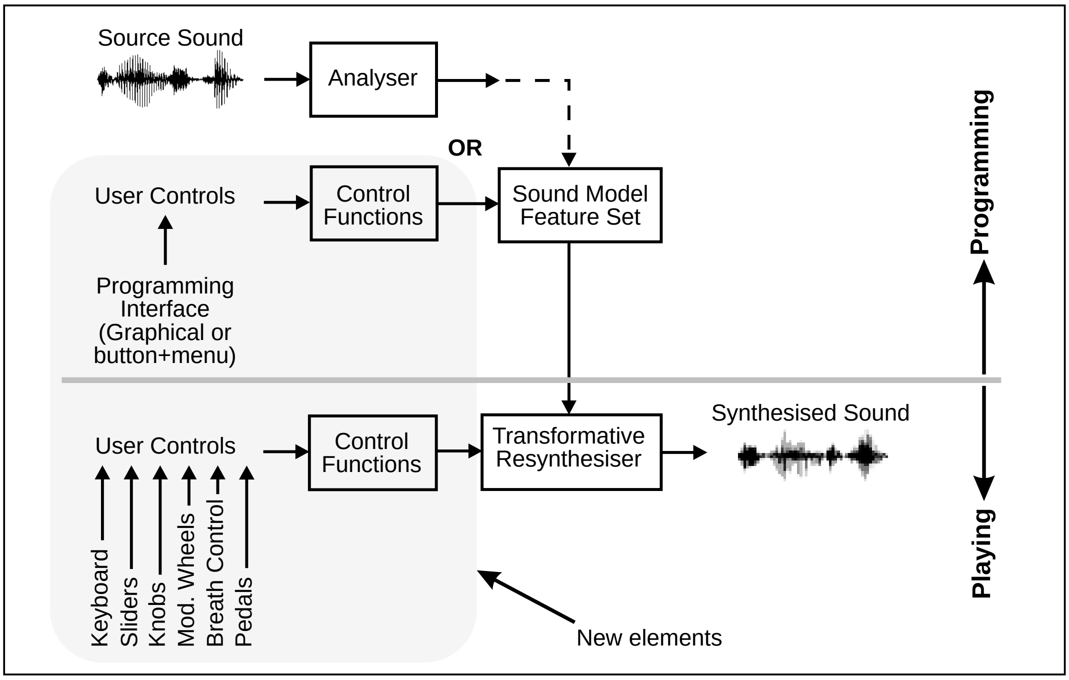
Why might we want to alter the ‘natural’ role of these synthesis schemes? In the case of making a generative architecture transformative, the answer is that even though the system only has a few parameters, it becomes capable of the realism normally reserved for sample-based systems. With an automated parameter-matching process, such as a genetic algorithm, it becomes possible to render real instrument sounds (or a much closer approximation); see Figure 7.2a.

Conversely, the sound model is inherently capable of realism. It also has the desired flexibility because of its non-specific architecture and parameterised synthesiser. However, the large number of parameters make it unwieldy. Also, it has no reference to the instrument that created the sound, so transformations are limited to global effects and cannot mimic the natural timbre variations of the source instrument. With a suitable interface, the vast number of parameters could be manipulated through a small handful of controls; see Figure 7.2b.

In the latter case, each control would have a function that can influence many synthesis parameters simultaneously, perhaps with interdependence on other controls. In this way the control functions could be used to simulate a particular instrument, discussed in section 7.1.2.1 following, or to emulate a particular synthesis architecture, discussed in section 7.1.2.2.



(a) Generative to Transformative - Adding realism to control



(b) Transformative to Generative - Adding control to realism

Figure 7.2 – Swapping the roles of generative and transformative synthesisers

7.1.2.1 Instrument Simulation

Synthesisers have always faced the problem of creating a realistic sound – one that evokes the qualities of a real musical instrument. The synthesis architectures have been developed to give control over spectrally rich timbres with few parameters. Yet, on the one hand the parameters have a complex correspondence to timbral properties that make them notoriously difficult to program, and on the other hand whilst the synthesis architecture enables generation of complex spectra, it can still only generate sounds that fall into a subset of the complete timbral space.

In terms of timbral space, the sound model aims to provide total freedom and every design iteration or improvement moves closer to that ideal. Moreover, through the analysis process, the location of a source sound is closely approximated in that space, automatically.

By replacing global transformations with a small set of programmable controls, the sound model can function like a conventional synthesiser. For example, one set of controls could be used for programming so that a change in pitch yields a change in timbre correspondent to the source instrument, and so forth. Another set of controls could be assigned to real-time modulators, such as after-touch and modulation wheels. The user interface would then be similar to the programming controls on a conventional synthesiser.

In addition to the improved accuracy of the synthesised sound, the model-based synthesiser offers the opportunity for a new range of much more musician-friendly controls. The flexibility of the model's internal parameters for modification to the sound spectrum makes it possible to introduce 'natural' controls that correspond to descriptions used in normal language. For example, controls could be made to vary brightness, sharpness, fizz, rasp or some other descriptive (i.e. non-technical) qualities. Moreover, because the controls would result from mathematical functions, any number of new controls could be custom programmed in software.

7.1.2.2 Architecture Simulation

The controls of the previous section were created so as to vary the sound model features according to variations in playing parameters. In fact, a control could be made to vary the features according to any law, since the controls are simply functions. As a progression of this line of thought, sound analysis could be *replaced* by the control law so that all features are artificially generated. The combination of a sound model and a reconfigurable control interface would become the basis for software simulation of any synthesis architecture. Popular architectures like analogue or FM could be emulated and new architectures with new parameter sets could also be designed.

With a suitable graphical interface, custom sound design could be greatly simplified so that a musician could, say, draw waveforms, spectral curves or control curves to describe aspects of the sound patch. Indeed, this would become a perfect prototyping environment enabling and encouraging third-party designers to easily and cheaply create new synthesisers.

Many of the synthesisers that have been released over the past 15 years have been variants on a theme. Musicians have bought new synthesisers that largely duplicate their existing hardware because they want the new sounds and because programming has been so challenging that it has become its own deterrent. The combination of sound model and control interface could rectify that situation, as the basis for a general-purpose synthesiser. Just as software packages are bought for personal computers to enable different applications, including the ability to program the system, so too sound packages, new architectures and synthesis architecture design applications could become the norm for use on software-configurable synthesisers.

7.1.2.3 System Unification

The ability to bring any synthesis scheme under the same umbrella as a sample-based system unifies what have been, to date, incompatible formats. Sample-based systems have the capacity to record and reproduce for non-linear editing (i.e. cut, copy, paste) but have remained inflexible for control of individual sound features. Synthesisers, in contrast, have always had limitations for realistic mimicry but through the parametric set, can enable sophisticated control.

By unifying both within a common framework and with a universal data format, synthesised sounds can be manipulated like samples and sampled sounds can be brought under flexible real-time control. Indeed, the previous section suggests that with the growing processing capabilities of DSPs, it should be possible to create a system that can emulate, through software, any sound or instrument-based musical architecture.

In addition to the advantages of modular integration, new sound generation possibilities emerge where, for example, sampled sounds can play an active role within a synthesis architecture (i.e. more than an additive role for attack transients or breath noise loops). A sampled sound feature could modulate a synthesis parameter to weave together sounds in a more elaborate manner than cross-synthesis or timbral morphing.

The cost of the advances indicated in this section is the need for greater data storage and the need for greater processing power. The former would not be a cost problem if the system were hard disk based and the latter may not be much of an obstacle because the cost of processing power is forever falling.

7.2 Applications in Ongoing Research

The shortfalls of the present analysis-resynthesis algorithms and the inaccuracies of the sound model features have been discussed in some detail: first in chapter six with a view to identifying solutions, and then in the first half of this chapter relating to quality issues of applying the sound model to musical systems for general use. This may give the impression that the sound model is not very useful in practice. Actually the current status of the sound model is sufficiently advanced for certain applications.

Two research projects, ongoing at the University of Bristol's *Digital Music Research (DMR) Group*, are described in this section. They apply the model to musical instrument synthesis (section 7.2.1) and speech synthesis (section 7.2.2).

7.2.1 Instrument Timbre Modelling (ITM) Synthesis

The sound model features, when applied to a single note from a musical instrument, describe the timbre of that instrument for the pitch, loudness and playing style of that note. The analysis is waveform-based, with no reference to the instrument or how it was played, so the feature set is too restrictive to be called an instrument model. However, if there is a means by which the timbral features can be associated with the playing parameters, then the sound model could be used for timbre modelling of musical instruments.

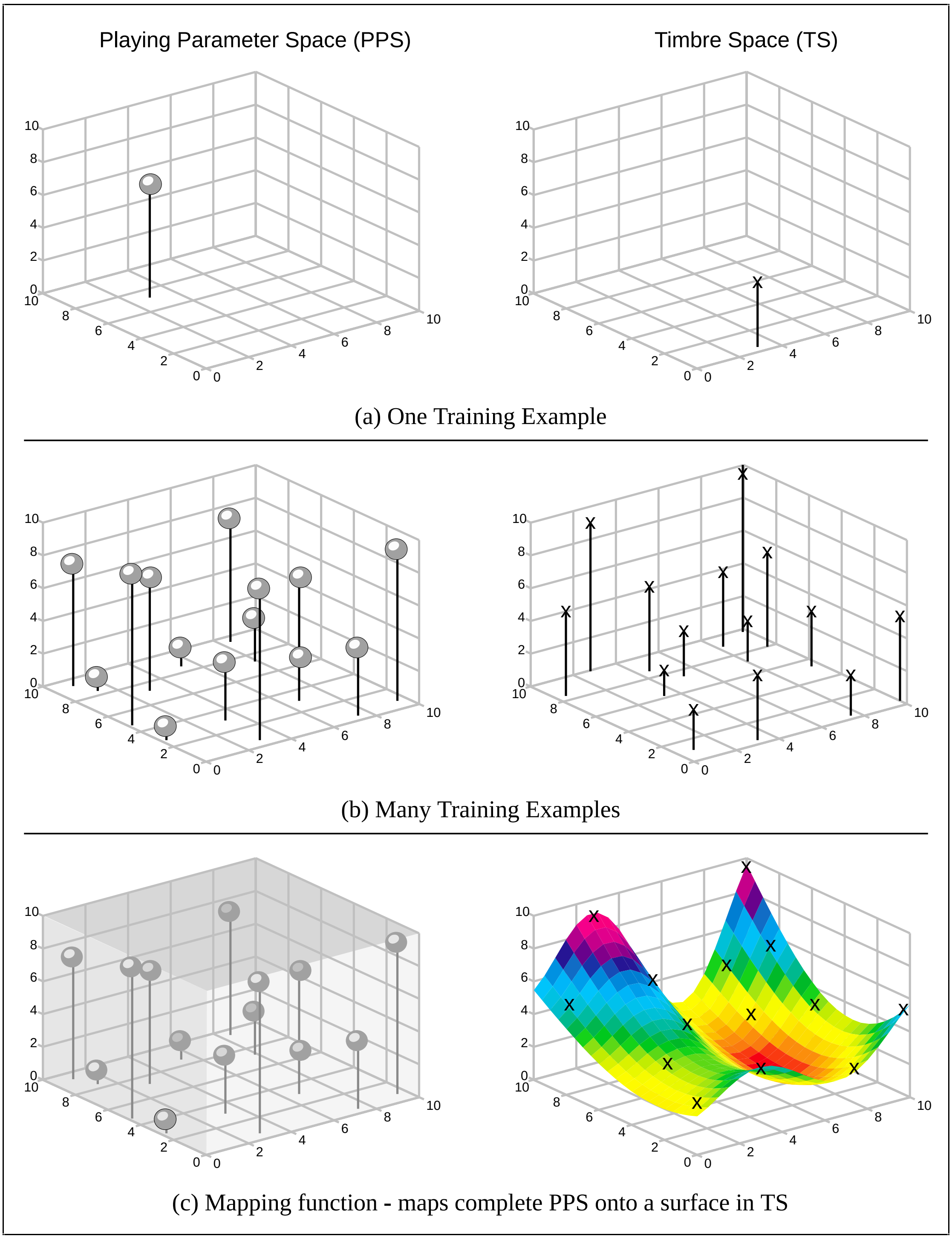
The forte of neural networks (NN) is to associate data sets that have complex relationships, and to approximate the general situation from a set of specific training examples (a property termed 'generalisation'). A three year programme is underway within the *DMR Group*[3] to use NNs to make the association that will link the recorded sounds back to the source instrument and its playing parameters.

The Musical Instrument Character Map (MICM)

The association between timbre space (TS – the domain of all sounds, as per Figure 2.1) and the playing parameter space (PPS – the domain of all playing gestures and instrument states) is made by a mapping function, termed hereon the Musical Instrument Character Map (MICM). The mapping is unique for every instrument but the map format is universal, so any analysed instrument can be synthesised from the same architecture by using the relevant MICM. The MICM is a set of functions that are encoded in the inter-nodal weightings of the neural network.

The inter-nodal weights are determined by training, using real data examples. Each training example associates one point in PPS with a point in TS. See Figure 7.3a (overleaf). When the NN is presented with a number of examples distributed throughout the PPS, a surface is generated in the TS. See Figure 7.3b-c. The surface represents the estimated domain of the instrument in TS. It is smooth and continuous, passing through the training example points

[3] The work is funded by a ROPA award granted by the Engineering and Physical Sciences Research Council (EPSRC) of the UK. Equipment for this project was provided by Texas Instruments under their *Elite* programme of European university research support.



(a) One Training Example

(b) Many Training Examples

(c) Mapping function - maps complete PPS onto a surface in TS

Figure 7.3 – Using a neural network to determine the mapping function

and effectively interpolating between them. Therefore it becomes possible to estimate the timbre of the instrument for a set of playing parameters that was not used in training.

The process of NN design begins with selection of a suitable architecture and training scheme. During the training phase, the network is periodically tested in order to find the optimal number of training examples, as well as to assess the effectiveness of the design. The tests are made by presenting the network with a set of examples (whose PPS-TS correspondence is known to the designer, but which were not used in the training set) and comparing the network's estimated mapping with the known association.

Synthesis and Recognition Roles

Neural networks are unidirectional in their mapping function. One data set must be designated the input and the other the output. If the function is required to be bi-directional, the mapping function must either be decoded and inverted or another NN must be trained in the reverse direction.

With the PPS as input and the TS as output (implied in the training description preceding), the NN has the role of synthesiser. It takes a set of playing parameters (which could be MIDI codes for example) and outputs a set of features that describe the timbre space. These features correspond to the features of a sound model, so the model's synthesis architecture can then generate the audible sound sample.

In the reverse direction (by training a separate NN with TS as input and PPS as output), the NN becomes a recognition system, capable of estimating properties of how a sound was played from the sound features alone. As in the synthesis case, the TS is a model feature set, so the NN would be preceded by a waveform feature analyser.

Research Goals

The role of synthesis has an obvious immediate musical application as a creative tool for musicians. The capability of recognition is seen as a first step toward a number of systems. Firstly, with the ability to assess how a note was played on an instrument, simply from a recording, it becomes possible to create an interactive expert tutor system, that can teach people how to play a musical instrument and provide constructive feedback. Taking this a step further to the more complex sound production mechanisms of speech, the tool could become the basis for speech therapy.

Secondly, if the playing parameters are replaced by general instrument features, then the recognition system becomes capable of recognising and identifying specific instruments. This capability would provide vital information to auditory scene analysis, both in pattern recognition to identify the existence and location of individual sound sources and in pattern classification to identify which instruments are which. The technique promises true automatic score notation including instrument labelling and expression markings. Just as the expert tutor could be adapted to a speech applications, so the auditory scene analyser could be applied to solve a problem that persists in the design of intelligent hearing aids, namely the cocktail party effect. These hearing aids try to pick out and enhance a single voice, but when presented with a cacophony of voices find such discrimination difficult. If the system were able to identify individual voices from the hullabaloo, then it could enhance just the nearest voice and suppress the others.

7.2.2 Expressive Speech Synthesiser

The *Speech Processing for Unconstrained English (SPRUCE) Group* at the University of Bristol is developing a system for realistic automatic speech synthesis. Modern speech synthesisers used in telecommunications are based on the Linear Predictive Coding model (see section 3.2.4.3) or more specifically Codebook-Excited Linear Prediction (CELP), in which a dictionary of codes is used for excitation instead of simply pulses or white noise. In this application, there is a source speech signal, analysis of which provides the parameters for synthesis. If there is no speech analysis stage, as in *automatic* speech synthesisers, such as text-to-speech converters (used, for example, in automatic book readers for the blind), the system must already know how each phoneme sounds. So sentence construction becomes a process of concatenating phonemes from a phone database or ‘dictionary’.

Two problems dominate in current automatic speech synthesisers: the sound quality is poor and expression is limited. Joint research between the *SPRUCE Group* and the *DMR Group* (both at the University of Bristol) is applying sound modelling techniques to overcome these problems. The sound model synthesis quality is sufficient for good quality speech reproduction and the model’s versatility in transformation enables smooth transitions between phonemes and expressive intonation.

The basic sounds used in speech will be compiled into a dictionary of phones (individual utterances) or diphones (pairs of utterances which capture the transitions), in terms of a sound model feature set. At synthesis, the elements will be selected and concatenated. The inter-element transitions will be smooth, because the database will include ‘normalised’ sounds, whose pitch, loudness, etc. are matched, enabling splicing between partial trajectories or noise spectra. Initial experiments have already verified this. Furthermore, because voiced speech approximates closely to a truly harmonic structure and because the database will contain normalised elements, the model feature set can be a simplified version of the general sound model of this thesis.

A sentence, constructed as described, would sound very artificial, although smooth and intelligible, because all the natural variations have been normalised out of the phonemes. Instead, the intonation will be imposed on the features during synthesis, with time-varying values for pitch, volume and speed variations. The fact that there is no intonation in the constructed sentence actually provides a neutral starting point, making it easier to impose whatever variations are required. The intonation will be implemented through formant-correct pitch-shift, amplitude scaling and time-scaling.

CHAPTER EIGHT

CONCLUSIONS

8. CONCLUSIONS

8.1 The Global Aims of the Thesis

The *Introduction* chapter set out the goal of the thesis: to develop a tool that combines good sound quality with the flexibility for musical transformation. This was shown to require a sound model whose feature set can fully describe all sounds, and an analysis-resynthesis system to convert between the time domain and the model's feature domain.

Sound models and analysis-resynthesis systems have been in existence for about twenty years and have gradually evolved from general signal capture toward sound-specific feature representation. In doing so, the systems have improved in their flexibility for transformation, becoming viable tools for musical use. However, the range of features is still limited and the systems require significant manual intervention for good quality synthesis.

The aim of this thesis has been to create an automatic model – one whose parameters are global (requiring no setting up by the user) and sound-independent. This means that the user need have no knowledge about sound or its signal properties, a vital attribute for a system if it is to be useful to musicians at large. This can be achieved by keeping the model parameters internal to the system, available only to the system designer for pre-setting during development.

In addition to automation, it was desired that the model should retain the flexibility of the best systems to date, whilst improving sound quality. The means for achieving this latter goal have been to improve the scope of the sound model and the accuracy with which features are captured. (The scope of the model can be thought of as the coverage of the domain of all possible sounds and is a function of the choice of feature set; the wider the coverage, the more sounds that can be accurately represented.) The feature set and the analysis processes were targeted rather than the synthesis process, because the synthesis method is obvious from the parameters of the features and easy to implement (at least for an unoptimised system).

8.2 Summary of Achievements

8.2.1 The Specific Goals of the Thesis

As its starting point, the 'Initial Model' of the thesis was based on a particular variant of the sinusoidal model, in which sound features are either classed as deterministic or stochastic. The deterministic aspect is based on the Fourier Series principle of describing a periodic signal with a set of harmonically related sinusoids, like the sinusoidal model. However, it tries to be more intelligent in its feature extraction, so as to capture only those features of a sound signal that are truly members of a harmonic structure (i.e. partials). The stochastic aspect then classifies the remaining signal components as time-varying noise.

In chapter two's description and critique of this model, several shortcomings became apparent. Whereas, in its conception, the deterministic-stochastic classification enables a wide scope for sound representation, in reality the analysis processes were unable to capture the features with sufficient accuracy. The initial observation was that errors in the deterministic analysis were being propagated to the stochastic analysis, where they would be captured in degraded form. The intention for model development was to focus on improving the deterministic analysis, thereby improving overall sound quality.

With further examination, this intention was clarified into three studies:

- to examine ways of improving the time-frequency resolution of the initial spectral estimator, on which the whole system relied for its information about the sound signal under analysis;
- to improve the frame linking process – the analysis step that achieves feature continuity, linking the instantaneous features extracted from the frame-based time-frequency representation into continuous evolving trajectories;
- to find a way of incorporating percussive sound features into the model; these are abrupt changes in the signal and were the worst represented feature in the system.

8.2.2 Improved Time-Frequency Resolution

The sound model bases the whole of its analysis on the time-frequency representation (TFR) of the signal. Therefore any deficiencies in the TFR will limit the accuracy (or cause gross errors) in the feature estimators. The TFR of the Initial Model is the Short-Time Fourier Transform (STFT). The STFT has many desirable properties in the way that it represents the spectrum (and in particular, the simplicity with which partials are represented as peaks in the magnitude spectrum). However it has limited time-frequency resolution, which makes tracking of rapid sound events and rapidly changing components difficult (e.g. higher partials).

Chapter three sought to find a solution, on the one hand looking at ways of gaining more resolution from the STFT, and on the other hand looking at alternative mathematical distributions that would satisfy the criteria for time-frequency representation. There has been much research into 'higher order' spectral distributions that are able to gain better time-frequency resolution, by virtue of being a closer approximation to nonlinear reality than linear estimators (like the STFT). Unfortunately, the time-frequency resolution trade-off is swapped for a resolution-interference trade-off that affects multicomponent signals like sound; in comparisons, the STFT was found to give a better representation for such signals.

Given the limited advancement of alternative TFRs, a study reported in chapter three tried to find ways of gaining extra detail from the STFT. The FFT (which makes up each frame of the STFT) is an invertible transform that does not create or destroy information, yet it is always applied with the assumption that the analysed signal is locally stationary. By examining this further, it was discovered that nonstationarities of a signal *are* represented, but that they appear as distortion. Treating this distortion as an information source, the Phase Distortion Analysis (PDA) method was created, with which it was demonstrated that the nonstationarities of signals can be measured. In the presented algorithm, frequency and amplitude trajectories were estimated in addition to the normal absolute frequency, amplitude and phase values. Unfortunately, the algorithm is not sophisticated enough as yet, to cope with the additive interference of close and larger neighbouring signal components. Future work suggestions were given for how this might be overcome and how the technique could be extended to the extraction of more complex nonstationarities.

The extraction of linear FM and exponential AM or higher order modulation laws not only proves that more information can be obtained from the STFT; it also represents an effective improvement in time-frequency resolution.

8.2.3 Improved Frame Linking

As a by-product of the limited time-frequency resolution of the STFT, there have been difficulties in validating which peaks in the spectrum represent partials and which result from ‘side-lobes’ or other representation artifacts. This is especially true for more rapidly varying components, like the higher frequency partials. Within each frame, it is difficult to determine which peaks are which (although the PDA method above, promises a solution to this). So much of the decision making results from the ability to link the peaks between frames, creating continuous partial trajectories.

The standard ‘Nearest Frequency’ method tries to link peaks on the basis that partial peaks will not vary much between frames. Unfortunately, with the limited resolution of the STFT, the time spacing between frames can allow significant inter-frame changes. As a result the Nearest Frequency method causes scrambling of the higher frequencies due to incorrect linking, leading to the familiar susurrus artifacts upon synthesis.

Chapter four promoted the use of the harmonic structure as a basis for linking frames. Although this is not, of itself, a novel concept, a method was designed which aimed to be more intelligent than existing approaches. Key to this was the observation of inter-frame trends in the fundamental frequency estimation. First of all, the fundamental frequency was estimated in each frame individually. The estimation techniques are known to sometimes lock onto a multiple or a fraction of the true fundamental frequency, so the inter-frame comparisons facilitated the correction of erroneous estimates. An unforeseen advantage is that the method can distinguish between frames that possess weak harmonicity and those with random correlations. This enables linking of partials into the weak frames at the start and end of harmonic regions, thereby improving feature capture and synthesis quality.

As a result, the artifacts of the Nearest Frequency method were eliminated and in their place higher frequency detail sounded more natural. From the experience of developing this technique, a future work direction was suggested for improving fundamental frequency estimation. This would simultaneously estimate the fundamental frequency using localised time and frequency information, instead of sequentially using global frequency information and then localised time information. The expected advantage is that localised evidence of a harmonic structure would not be corrupted other features in the global frequency representation.

8.2.4 Improved Modelling of Percussives

It has been established that the limited time-frequency resolution of the STFT prevents tracking of rapid changes. The worst-affected sound features are percussive note onsets, which can be near instantaneous, causing a abrupt rise in energy and a sudden change to the spectral content. During analysis, the STFT becomes a confused mixture of the energy localisations preceding and following a percussive onset, which causes degraded analysis – sometimes very little information is captured during deterministic analysis, resulting in drop-outs upon synthesis. The synthesis method was also not designed for recreating rapid changes – the gradual interpolation and slow fades between frames diffused what information had been captured by the analysis.

No alternative TFR has been found that can yield the required time resolution (without compromising the frequency resolution to extract partials), so a solution was sought in chapter five that would avoid the problem situation altogether. Since the full detail of an attack transient could not be estimated from a TFR, a method was designed that could detect an attack transient and then force the analysis and synthesis to synchronise around its location.

Three detection methods were designed and tested, based on the rise in energy, the rise in the time domain envelope and the suddenness of spectral change. Of these the time domain method proved the most robust and the most accurate. For synchronisation, the analysis was allowed up to and starting from each attack transient, but not across the percussive onset itself. At synthesis, the waveforms either side of each attack transient were separately generated and then joined with a fast cross-fade that retained the suddenness of the change.

The outcome, at very little computational expense and with high reliability, was the elimination of diffusion and drop-outs. Instead, the crispness of attack transients in the source sound was preserved, as demonstrated both objectively (through signal observations) and subjectively (by listening to the results).

8.3 The Future of Sound Modelling

8.3.1 Development of a New Model

During the investigations of this thesis, it became apparent that the model was not capturing the full detail of the partial trajectories. The cause of the model's inaccuracy was identified as the hard classification between deterministic and stochastic. In this classification, signals are only accurately represented if they are stable slowly-varying sinusoids or broadband slowly-varying noise.

Chapter six (section 6.5) proposed a new direction for computer modelling of sound, by replacing the deterministic-stochastic classification with the concept of the 'Noisy Partial'. The Noisy Partial is a sound component with a chaos index. When it is set to 0% chaos, the Noisy Partial corresponds to the 'deterministic' classification and when it is set to 100% chaos, it corresponds to the 'stochastic' classification. In-between the extremes, signal components can be sinusoids with varying amounts of chaotic modulation or noise components with varying degrees of localisation.

Initial experiments and the results from related work of other authors suggest that this is a more effective way of encapsulating the noisiness of sounds and that it is more true to the actual physical mechanisms. In addition, the simplification from the dualistic representation to a single, more flexible feature type yields simplifications to the analysis-resynthesis structure that should enable some improvement in computational efficiency.

8.3.2 Applications of the Model

A number of musical applications were identified in chapter seven, where the model could be a catalyst for an improved range of studio-based and live performance tools, that have more musician-friendly controls. Unfortunately, it was also noted that many of these applications would require further maturity of the sound model for wide acceptance by professional musicians.

As an indirect route to achieving that quality increment, a new type of 'Model-Enhanced' transformation technique was created, in which the methods of sound modelling are applied to standard sample manipulation algorithms to make them more intelligent. Often, these standard techniques have little or no information about the signal content, and so they apply the manipulation in the same manner to all parts of a sound irrespective of content. With a little additional information, these methods might be made more adaptive. Although the Model-Enhanced approach is not advocated as a replacement to the sound model, it is anticipated that these techniques could provide an interim solution by enhancing the quality and capabilities of standard algorithms.

Two applications of the sound model in new research projects were also described in chapter seven. In one, the sound model's capacity is increased to become an instrument model. Not only can the model capture the features of an instrument's sound; it can also capture the way in which the instrument was played. By using neural networks to generate a Musical Instrument

Character Map, the playing parameters can be associated with the sound features to create an expressive synthesiser that inherently possesses the quality of realism, or to create a recognition system that can identify how an instrument was played from the sound alone.

In the second application, the sound model is being used at the heart of an automatic speech synthesiser, where its sound transformation capabilities are enabling more natural intonation of the artificially constructed sentences and phrases. (Both of these projects are ongoing at the University of Bristol's *Digital Music Research Group*.)

APPENDICES

APPENDIX A

PARTIAL DOMAIN SYNTHESIS OF MUSIC

co-authored with Andrew Bateman

Originally published:

1994. Proceedings of Digital Signal Processing UK (DSP-UK-94).
Vol.1. (pages unnumbered)
Organised by New Electronics.

Partial Domain Synthesis of Music

Paul Masri, Prof. Andrew Bateman
Digital Music Research Group, University of Bristol
1.4 Queens Building, University Walk, Bristol BS8 1TR, United Kingdom
Tel: +44 117 928-7740, Fax: +44 117 925-5265, email: paulm@ccr.bris.ac.uk

Abstract

'Partial' is the collective term for the fundamental and overtones of musical sound, which define the pitch, loudness and timbre (tonal quality). Partials of a sampled sound can be identified, by post-processing FFT's of short waveform segments. The data for Partial Domain Synthesis is in the form of an array of oscillators, whose frequency, amplitude and phase correspond to the instantaneous state of the partials. The sound can then be created from the partials by Additive Synthesis. This paper presents the methods for analysis and synthesis, and shows that each can be achieved in real time, using a single standard DSP device.

Furthermore, because the synthesis parameters relate directly to the tonal properties of sound, there is a unique level of control when transforming sounds. Examples are given of two of the latest (and most sought after) sound effects: time-scale and pitch-shift. The musical power of partial domain synthesis combined with the DSP power of real time synthesis create the potential for a new generation of music sythesisers.

Introduction

A fundamental aim of any form of music synthesis is the ability to transform sound in interesting ways. However the first step in creating a new method of synthesis is to prove the method by direct reproduction of familiar sounds. This is particularly true for analysis-based synthesis where familiar sounds can be captured as source material, through digital sampling. This paper presents a method of (analysis-based) music synthesis, assesses its performance in direct reproduction, and with the aid of some demonstrations, explores the potential for new and interesting sound effects.

As a production tool, off-line synthesis may be adequate, but for performance, real time synthesis is essential. A method is presented for achieving this with a DSP device. Calculations also show that real time analysis is possible using a DSP. These are significant qualifications for a new generation of music tools, in today's music technology market.

The Current Market in Music Technology

It is a universal goal of all music synthesis techniques that they enable the musician to use sound in ways that were not previously possible. In pursuit of this, music technology has followed two paths: generation and transformation. In the former category the sound wave is generated from a synthesis architecture composed of oscillators, filters, modulators and envelope controllers. In the latter, 'real' sounds are recorded and transformed.

Synthesisers offer the programmer control over the sound generation process and are therefore very flexible. However there is often no intuitive link between the values of a parameter and the timbre (tonal quality) of the sound. (This is a property of the synthesis scheme, and is not connected with the issue of the human-machine interface.) This is particularly true of FM synthesis, arguably the most popular synthesis scheme of the last fifteen years. As a result, many musicians make do with the factory presets of a particular synthesiser and this has in turn encouraged manufacturers to develop a vast range of sound modules with near identical synthesis engines, but sporting different sounds in their presets. It is not uncommon for professional and semi-pro musicians to have racks-full of synth modules from which they pick and choose their sounds. The more successful producers, keen to have a unique sound for their bands, now employ a sound programmer - an expert who translates the musician's description of a sound into a set of parameter values.

By virtue of a parameterised synthesis architecture, there is flexibility, but each architecture has a finite range of possible sounds. There are a few different synthesis architectures [J.Chowning 1973, M.LeBrun 1979] which broaden the overall range, but each new architecture also brings with it a new set of obscure parameters.

Unlike *generation* which originates sounds, *transformation* manipulates existing sounds. Until recently, sounds were recorded on analogue tape and transformed in real-time with analogue effects. The modern equivalent uses digital sampling technology where effects are achieved with digital signal processing. Commercial units called samplers (and more recently PC desktop packages also) provide the means for recording, playing and editing sounds at the waveform level. However the range of tools is still quite restricted: software cut and splice, filtering and altering the rate of playback. Recently two new effects have been introduced to the market and are quickly becoming essential: time-scale and pitch-shift.

Time-scaling alters the duration of a sound, stretching or compressing it without affecting the pitch. This is equivalent to changing the rate of speaking for the human voice, or the tempo for rhythmical music. Conversely, pitch-shift provides a change in pitch whilst retaining the original temporal properties, and can be compared with transposition. Both effects overcome the limitations of simply changing the playback rate by making the tempo and pitch independent.

In summary, synthesisers have the potential for flexibility, whilst samplers have a means of accessing sounds of the real-world (including the sounds of traditional instruments). What is desired are instruments which can combine these two properties.

Flexibility and Realism

The solution adopted by manufacturers has been to integrate samples into synthesisers, to add realism to an already flexible framework. The samples are typically very short. They are used to create a realistic attack at the onset of notes (such as the piano hammer or the noise burst at the start of a trumpet tone). Alternatively samples are looped (repeated over and over) within a sound to add complex textures, such as the breathiness of a wind instrument. Listening to a single note it is very successful, and this has been proved by its success in the marketplace. However the samples restrict any variation in the sound, a natural property of traditional instruments. Each note sounds the same, so very quickly the sound of a Sample-plus-Synthesis (S+S) synthesiser can be recognised. There is flexibility in the creation of a new sound which is realistic, but very little of either property in using the sound expressively.

Research has taken different approaches to solving the problem of combining flexibility with realism. Physical modelling makes the synthesis architecture more specific, by tailoring it to model a specific traditional instrument. The synthesis parameters bear direct relevance to the physical dimensions and playing technique of a particular type of instrument. Conversely, analysis-based synthesis makes the synthesis architecture more general, catering for a wider range of sounds. Its parameters relate to specific features of sound, which are extracted from samples by analysis. Partial Domain Synthesis is a method of analysis-based synthesis, whose parameters are the partials, the basic elements of musical sound. There is therefore an intuitive link between parameters and tonal quality.

Fourier Theory, Partials and the principle of Partial Domain Synthesis

The perception of pitch results from the periodicity of musical sound waves. The period is determined by the fundamental frequency, and the timbre (tonal quality) is affected by the strengths of the overtones. Collectively the fundamental and overtones are termed 'partials'. It is the aim of Partial Domain Synthesis (PDS) to represent sounds in terms of their partials, and to recreate the sounds (or transformed versions) from the partials.

Fourier theory states that any periodic waveform can be generated by the superposition (addition) of harmonically related sine waves. However musical sound waves are not eternally periodic, as required by pure Fourier theory. The waveforms, though repetitive, change gradually over time, hence the description quasi-periodic. In order to apply Fourier theory, an assumption is made that short segments of the sound can be considered to be periodic.

After applying the Fourier transform to each segment, the instantaneous frequency, amplitude and phase of each partial is found. Each segment gives rise to one frame of PDS data. The synthesis architecture consists of multiple sinusoidal oscillators, whose parameters are the partial data. During each frame, the outputs of the oscillators are summed to generate the audio waveform, a technique known as Additive Synthesis. At the frame boundaries, the oscillator parameters are updated with the next frame of partial data. This is achieved such that the output waveform suffers no discontinuities at the frame boundaries.

Analysis

A method for extracting partial information from musical sound waves is given in [Serra 1990], a simplified summary of which is presented next. The use of multiple frequency bands is introduced in this paper to make the analysis more generalised and automatic - the "Sinusoidal Model" scheme proposed in [Serra 1990] is not well-suited to complex sounds containing chords or multiple instruments, and the system parameters need manual setting, dependent on properties of the sound.

Partial Extraction

When a segment of sound is selected, there is no guarantee that it will contain an exact number of periods of the waveform, so the first step is to apply a window function. (Serra uses the Kaiser window function in which the value of β is manually chosen, but the authors have found the fixed Hamming function adequate.) Since multiplication in the time domain is

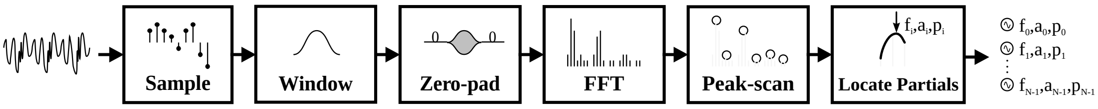


Fig. 1 - Block Diagram of the Analysis Process

equivalent to convolution in the frequency domain, a copy of the window's frequency spectrum appears centred about each partial frequency. The frequency spectrum of the Hamming window has the peak of its main lobe at the centre, so this peak lies at the actual frequency of each partial in the windowed segment's spectrum. The partials are located by taking the FFT of the windowed segment and picking out the peaks. The resolution of the FFT is inferior to that of the ear[1], so interpolation between the bins is required. This can be achieved to sufficient accuracy through a combination of zero-padding the windowed segment (prior to FFT) and fitting quadratic curves to the FFT bins whose maxima are then equated. This is summarised in figure 1.

The final stage of analysis is to link partials between frames. During synthesis, the frequency and amplitude of partials are linearly interpolated between frames. Where new partials are created or old partials are 'killed', they are faded in/out over the period of one frame.

Multiple Frequency Bands

The approach of splitting the frequency spectrum into bands, which are each analysed separately, has been introduced to overcome sound-dependence in the choice of segment length, and to cater for more complex sounds. Previously the segment length was tailored to the sound; now there are many fixed segment lengths, each tailored to a range of frequencies. Each band is analysed as before, but partials are only selected within its frequency range.

Two factors influence the match between frequency range and segment length. For a given frequency: the segment must include sufficient signal to reliably locate partial peaks; and the segment should be as short as possible to preserve information about transients. If there is insufficient signal of a particular frequency, some frames fail to identify the peak and the synthesised sound becomes degraded. Theoretical tests show that more than 1.5 cycles of a given frequency must be included. In practice a safety factor of about 10 is used, because the shape of actual partial peaks is rarely as ideal as the theory[2]. At the other extreme, if the segment is too long, transient information becomes blurred. This is particularly evident in percussive sounds, where the transient becomes spread over a whole frame.

An ideal band allocation would require tens, possibly hundreds of bands. Since the aim of this work includes making it practical for cheap real-time implementation, there has been a compromise. Segment lengths have been chosen to be powers of 2 for the sake of the FFTs. Frequency bands have been allocated to those segment lengths such that between 16 and 32 periods of any frequency are included. The lowest band is from 0-700Hz, so there are 6 bands in total (spanning the audio region).

Synthesis

Additive Synthesis

Each partial is simulated by an oscillator whose frequency, amplitude and phase parameters are assigned from that partial. The audio waveform is generated by summing the outputs of all the active oscillators. The equation for this is :-

$$x[n] = \sum_{i=0}^{N-1} a_i[n] \cos\left(\theta_i[n]\right) \qquad \textbf{[1]}$$

where n is the sample number
x[n] is the sample value (for output)
N is the number of oscillators
i (suffix) is the index to each oscillator
$a_i[n]$ is the amplitude function
$\theta_i[n]$ is defined as:-

$$\theta_i[n] = 2\pi f_i[n]t + p_i \qquad \textbf{[2]}$$

where $f_i[n]$ is the frequency function (in Hz)
p_i is the phase offset (p_i=phase at t=0)
t is the time (in seconds) from the start of the frame, calculated as:-

$$t = \frac{n}{F_s} \qquad \textbf{[3]}$$

where F_s is the sampling frequency (in Hz)

[1] A typical FFT length is 1024 points. For sounds sampled at 44.1kHz (the standard used by Compact Disks), the resolution is ≈43 Hz. At its most sensitive, a discriminating ear can detect a difference between tones 0.03 Hz to 0.08 Hz apart [Rakowski 1971].

[2] Partial peaks are distorted by waveform segments which are not very periodic, close partials with overlapping spectra, and the presence of noise.

For test purposes this was implemented as a non real-time program whose output was stored as a sample file. Oscillators were created in software for each of the partials in a frame; the frequency and amplitude were linearly interpolated between frames, according to the links identified during analysis. For multi-band synthesis, the oscillators were organised in sets, where each set corresponded to an analysis band and whose frames were updated at the required rate for that band.

The first use of this algorithm was to prove the analysis process by directly reproducing the original sounds with no transformation. Thereafter, the traditionally complex transformations of time-scale and pitch-shift were implemented to demonstrate the power of this synthesis scheme. Time-scale is achieved by altering the frame rate, whilst pitch-shift involves scaling the frequencies of the partials, both trivial calculations. Many commercial algorithms limit the effects to ±10% (about 2 semitones pitch-shift); PDS imposes no such restrictions.

Single band analysis yields good quality on sounds with a high tonal content, low noise, and little pitch variation. Multi-band analysis improves on this by handling melodies which span several octaves. A high tonal content is still preferred, since this is where partials are present. However non-tonal sounds like the hammer at the start of a piano note are picked up. Although these do not strictly contain partials, they do produce peaks in the FFT which can be picked out and treated like partials. The reproduction of these is good for high frequency transients, where PDS has good temporal resolution. Orchestral instruments such as strings, wind, brass and piano are synthesised to high quality. Where PDS requires greater refinement is in capturing low frequency transients (where the PDS segments are long) and bands of noise. These occur in bass percussion and breathiness, particularly the spoken voice. These lose some resolution and may sound 'metallic', as noise assumes a tonal quality.

DSP Implementation of Analysis and Synthesis

Feasibility of Real-time Analysis

Real-time analysis was not originally a goal of the project, but it was found to be possible on standard DSPs such as the TMS320C30 or M56000. In calculations, the assumption was made that almost all the processing time is spent on FFT calculation (which is not unreasonable). The Cooley-Tukey, radix-2, 3-butterfly algorithm is used, for which the required processing is catalogued in [Burrus & Parks 1985]. The calculations also assumed that almost all of the FFT processing time is taken with multiplications and additions. Assuming that each of these operations takes a single cycle, and allowing an overhead for other operations such as windowing, peak location and linking, this indicates that real-time analysis can be achieved by processors with clock speeds of 25MHz (for Multiplication and Addition on separate cycles) or 16MHz (where a Multiply and an Add can be done in parallel in a single cycle).

Real-time Synthesis Algorithm and Architecture

In implementing Additive Synthesis as an algorithm, the oscillators are updated at each frame and there is frequency and amplitude interpolation between frames. The interpolation is represented as a fixed increment per sample, which operates over the whole frame. A sinusoidal look-up table simplifies evaluation of the cosine function to an indexed read instruction. The phase, frequency and frequency interpolation values are accordingly translated to table indices and offsets.

For a single band, the entire synthesis algorithm can be written :-

```
For each frame {
    For each oscillator i {
        Initialise $\phi_i$, $F_i$, $A_i$, $\delta F_{i,}$ $\delta A_i$
    } [Next oscillator]
For each sample in frame {
        Initialise $x_{sum} = 0$
        For each oscillator {
            $x_{sum} = x_{sum} + c[\phi_i] * A_i$
            $\phi_i = < \phi_i + F_i >_L$
            $F_i = F_i + \delta F_i$
            $A_i = A_i + \delta A_i$
        } [Next oscillator]
        Output $x_{sum}$
    } [Next sample]
} [Next frame]
```

where ϕ_i is the instantaneous phase index
F_i is the instantaneous frequency index
A_i is the instantaneous amplitude scalar
δF_i is the frequency interpolation increment
δA_i is the amplitude interpolation increment
x_{sum} is the synthesised sample output
c[] is the cosine function, as a table read
L is the length of the cosine wavetable
$<>_L$ means modulo L - to keep the phase in the range 0-360°

The synthesis algorithm implicitly requires that all the data for the next frame is available locally by the next frame boundary. For this synthesiser to be useful within an interactive, integrated architecture, it must be capable of receiving its data in real time. Figure 2 shows an architecture which enables simultaneous data transfer and synthesis.

While the current frame is being synthesised, the data for the next frame is downloaded. At the simplest level this uses two banks of data. While one is being used for synthesis, the other is loaded with the next frame's data. At the frame boundary, the roles of each bank are switched.

The chosen output architecture is one in which samples are calculated and stored in a circular buffer, from which the DAC can take its data when required. An interrupt routine would handle the data output at the request of the DAC.

In the case of multiple frequency bands, each band has its own Bank A and Bank B. At synthesis, the additive synthesiser uses oscillator data from the current bank of each band. The code for switching banks at frame boundaries is necessarily more complicated, so the circular output buffer is a necessity, and redundancy must be built into the synthesis cycle to allow the buffer to be filled slightly faster than it is emptied by the DAC.

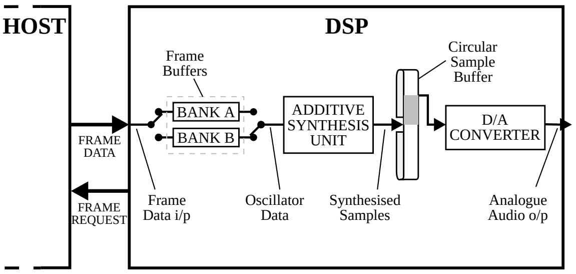


Fig. 2 - Architecture for real-time synthesis using a DSP

Conclusions and Future Work

The multiple frequency band approach has fulfilled its purpose in overcoming the problems of dependence on the source sound. It presents a generalised form of the analysis method and is totally automatic. As a bonus, the sound quality is improved for more complex sounds; those containing chords, multiple instruments, or where the melody traverses several octaves.

There is still room for improvement, particularly where the original sound contains noise elements or low frequency transients. The final chapter of [Serra 1990] presents a method for handling these stochastic sound elements, by modelling them as dynamic band-limited noise. This and other techniques are currently being investigated.

Real-time synthesis hardware is being built using a single TMS320C30. This will facilitate triggering from a MIDI keyboard, where pitch-shift is linked to the key pressed. In addition, it will make time-scale and pitch-shift alterable in real-time, via MIDI controllers such as a pitch bend wheel, a step beyond the current commercial boundaries.

Time-scale and pitch-shift demonstrate some of the power of partial domain synthesis, but its real strength lies in its ability to access the timbre directly. One new effect, for which PDS is well suited, is sound morphing, where the properties of a sound can change between one instrument and another - analogous to video morphing much used in advertising and films like T2 and The Mask. Applying it to more than sound effects, PDS could form the basis of a new generation of truly flexible, realistic synthesisers.

Author Profiles

Paul Masri is a PhD student at the University of Bristol's Digital Music Research Group. His background is in Electronic Engineering, for which he has a BEng honours degree (University of Bristol) and two year's industrial experience. A long-term interest in the interface between technology and art/music provided motivation for the current research.

Andrew Bateman is professor of signal processing at the University of Bristol, specialising in applications to sound synthesis, telecommunications, medical electronics and image analysis. This research team comprises more than 80 full-time staff.

Acknowledgements

The Digital Music Research Group has grown out of research into digital signal processing for speech and communications. This research programme owes much to the Centre for Communications Research and its director, Prof. J. P. McGeehan. The author also wishes to express personal thanks to Andrew Bateman for his help in making the project possible.

References

C.S. Burrus & T.W. Parks. 1985. *DFT/FFT and Convolution Algorithms*. Pub. by John Wiley & Sons. ISBN 0-471-81932-8.

J. Chowning. 1973. *The synthesis of complex audio spectra by means of frequency modulation.* Journal of the Audio Engineering Society, 21(7): pp526-534. Reprinted in C. Roads and J. Strawn (eds). 1985. *Foundations of Computer Music.* Cambridge, Massachusetts, MIT press: pp6-29.

M. LeBrun. 1979. *Digital waveshaping synthesis.* Journal of the Audio Engineering Society, 27(4): pp250-266.

A. Rakowski. 1971. *Pitch discrimination at the threshold of hearing.* Proceedings of the Seventh International Congress on Acoustics, vol.3 (Budapest).

X. Serra. 1990. *A system for sound analysis / transformation / synthesis based on a deterministic plus stochastic decomposition.* Ph.D. diss., Stanford University.

APPENDIX B

IDENTIFICATION OF NONSTATIONARY AUDIO SIGNALS USING THE FFT, WITH APPLICATION TO ANALYSIS-BASED SYNTHESIS OF SOUND

co-authored with Andrew Bateman

Originally published:

1995. IEE Colloquium on Audio Engineering.
Digest No. 1995/089.
pp. 11/1 – 11/6.
Published by the Institute of Electrical Engineers (IEE), Savoy Place, London WC2R 0BL.

Identification of Nonstationary Audio Signals Using the FFT, with Application to Analysis-based Synthesis of Sound

Paul Masri, Prof. Andrew Bateman
Digital Music Research Group, University of Bristol
1.4 Queens Building, University Walk, Bristol BS8 1TR, United Kingdom
Tel: +44 117 928-7740, Fax: +44 117 925-5265, email: paulm@ccr.bris.ac.uk

Abstract

In the analysis of sound (for synthesis), digitally sampled audio is processed to extract certain features. The resulting data can be synthesised to reproduce the original sound, or modified before synthesis to musically transform the sound. The analysis primarily uses a harmonic model, which considers a sound to be composed of multiple nonstationary sinusoids. The first stage of analysis is often the Fast Fourier Transform(FFT), where they appear as peaks in the amplitude spectrum.

A fundamental assumption when using the FFT is that the signals under investigation are Wide Sense Stationary(WSS); in terms of sinusoids, they are assumed to have constant frequency and amplitude throughout the FFT window. Since musical audio signals are in fact quasi-periodic, this assumption is only a good approximation for short time windows. However the requirement for good frequency resolution necessitates long time windows. Consequently the FFT's contain artifacts which are due to the nonstationarities of the audio signals. This results in temporal distortion or total mis-detection of sinusoids during analysis, hence reducing synthesised sound quality.

This paper presents a technique for extracting nonstationary elements from the FFT, by making use of the artifacts they produce. In particular, linear frequency modulation and exponential amplitude modulation can be determined from the phase distortion that occurs around the spectral peaks in the FFT. Results are presented for simulated data and real audio examples.

1. Introduction

In the analysis-based synthesis of sound, the harmonic model plays a primary role. Sounds that possess pitch have waveforms that are quasi-periodic. That is, they display periodicity, but in the short term only. In the harmonic model of sound, the waveform is a multi-component signal, additively composed of sinusoids whose frequencies are harmonically related. Traditionally the harmonic analysis has been performed using the Short Term Fourier Transform (STFT), a time-frequency representation whose time-frames are each calculated using the Fast Fourier Transform (FFT) algorithm [4].

One of the fundamental assumptions of the FFT is that the signal under analysis is stationary. Where this is true, each spectral component within the signal appears as a narrow peak, whose frequency, amplitude and phase can be estimated from the maximum of the peak. The assumption of the harmonic model is that sound waveforms change slowly enough to approximate stationarity, over short time segments. However, this constraint for short FFT windows is in conflict with the constraint for good frequency resolution, where a long window is desirable. In practice, the latter condition is favoured and the system is made tolerant to some distortion in the FFT representation.

For a spectral component that is significantly modulated within the analysis window, its peak in the FFT is smeared, becoming wider and suffering phase distortion. However, if the modulation is not severe, the instantaneous frequency, amplitude and phase at the centre of the time-window can still be estimated from the maximum of the peak. The conventional approach to estimating parameters for the harmonic model has therefore been to scan the FFT for peaks, and to determine the frequency, amplitude and phase at their maxima, ignoring distortion to the shapes of the peaks.

On the whole this has been successful, but there are two major drawbacks. Firstly, a peak is only considered if the amplitude ratio of its maximum to the adjacent minima is greater than a certain threshold. This aims to reject peaks arising from spectral leakage - the 'side lobes' - which are normally much smaller than the important peaks - the 'main lobes'. Where there is distortion due to nonstationarity, some main lobes are rejected and some side lobes accepted, resulting in audible distortion upon synthesis. Secondly, the constraint for long windows forces the loss of information about the dynamics of the sound. Upon synthesis, certain sounds audibly lose the sharpness of their transients.

In this paper, the authors present evidence that information about nonstationarities can be obtained from the distortions themselves. The method is explained and results are displayed for simulated and real data. Finally, the merits of an FFT *with* nonstationary information is compared to the abilities of alternative nonstationary (higher order) spectral estimators.

Throughout the paper, the symbols F, A, Φ, t are used to denote frequency, amplitude, phase and time respectively.

2. Detection and Measurement of Non-stationarities using the FFT

It is well known that the FFT contains a complete description of the time domain signal, because:

$$\mathrm{IFFT}\big(\mathrm{FFT}\{x\}\big) = x$$

where IFFT is the Inverse FFT function

Therefore spectral components within a signal that are nonstationary *are* represented by the FFT. It is simply that the nonstationarities are represented as distortions.

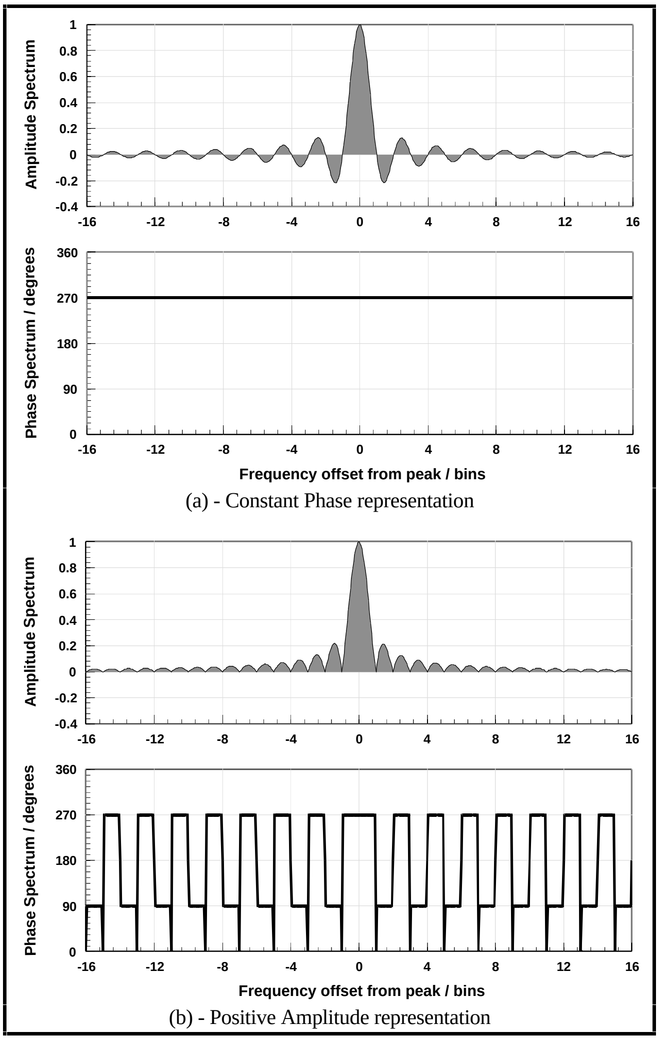
The FFT of a windowed, stationary sinusoid is the Fourier transform of the window function, centred about the frequency of the sinusoid, and sampled at frequencies corresponding to the FFT bins. It is also scaled according to the amplitude of the sinusoid, and rotated to the instantaneous phase of the sinusoid at the centre of the time-window. Modulation of the frequency and/or amplitude of the sinusoid results in a widening of the spectral shape, distortion to its form (particularly around the main lobe), and phase distortion. However the frequency location, amplitude and phase at the maximum of the main lobe are minimally affected, unless the distortion is severe.

The discussion in this paper concentrates on the phase distortion that occurs in the main lobe (also referred to as 'the peak'). Also, information about nonstationarities is limited to detection and measurement of linear FM 'chirps' (quadratic phase law) and exponential AM. In all cases, the measurements were found to be invariant of the frequency and amplitude of the modulated sinusoid. Also, the modulation is described in absolute terms; i.e. not relative to the modulated signal.

The distortion is dependent on the window function but experiments on the rectangular, triangular, Hamming and Hanning windows suggest that the *form* of the distortion is identical, and it is the actual values that differ. Hence the presented technique could be applied to any window function, but the measurements would need re-calibration. Results detailed in this paper are primarily for the Hamming window function, which the authors use in their sound analysis process.

2.1 Phase of an Unmodulated Sinusoid

For an unmodulated sinusoid, the phase is constant across the main lobe and all the side lobes as shown in figure 1(a). However its amplitude oscillates about zero, so for an FFT whose amplitudes are all represented as positive, the phase will appear to be shifted by 180° at certain points (see figure 1(b)).



(a) - Constant Phase representation

(b) - Positive Amplitude representation

Fig. 1 - Fourier transform of a sinusoid (rectangular window)

2.2 Linear Frequency Modulation

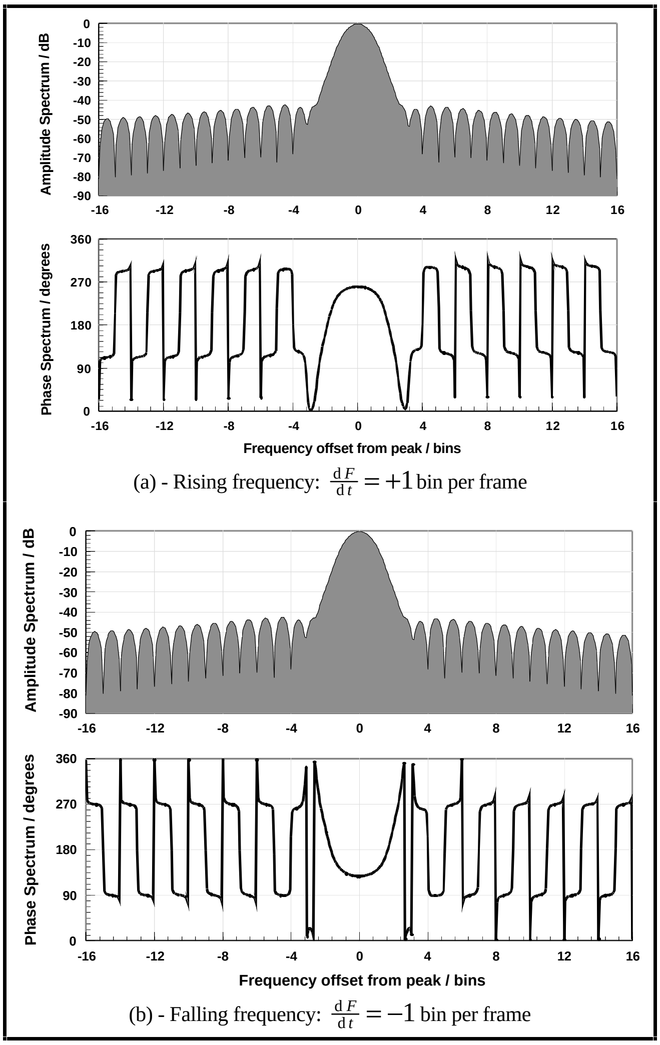
For sinusoids of linearly increasing frequency, the phase either side of the maximum is reduced, as shown in figure 2(a). For a given $\left|\frac{dF}{dt}\right|$, the amplitude spectrum is the same regardless of whether the frequency is rising or falling. Conversely, for a given $\left|\frac{dF}{dt}\right|$, the *degree* of phase distortion is identical, but the orientation depends on the sign of $\frac{dF}{dt}$; these effects can be observed by comparing figures 2(a) and 2(b).

The measurements at fractions of an FFT bin in all figures were made by zero-padding the time domain signal by a factor of 16 prior to the FFT.[1]

The degree of phase distortion is dependent on the rate of change of frequency, according to the curves shown in figure 3. The curves measure the phase distortion at different frequency offsets from the maximum. The similarity of the curves indicates that measurements can be taken at any offset within the main lobe, if $\frac{dF}{dt}$ is to be

[1] Zero-padding provides greater spectral detail of the Fourier Transform (FT), even though it does not increase the spectral resolution of the actual signal. i.e. it samples extra points along the FT curve of the unpadded FFT.

determined from the phase distortion. Note that there is not a unique mapping between $\frac{\mathrm{d}F}{\mathrm{d}t}$ and $\mathrm{d}\Phi$. In determining $\frac{\mathrm{d}F}{\mathrm{d}t}$ from $\mathrm{d}\Phi$, this restricts the usage to $\frac{\mathrm{d}F}{\mathrm{d}t} \in [0,4]$.



(a) - Rising frequency: $\frac{\mathrm{d}F}{\mathrm{d}t} = +1$ bin per frame

(b) - Falling frequency: $\frac{\mathrm{d}F}{\mathrm{d}t} = -1$ bin per frame

Fig. 2 - Linear frequency modulation (Hamming window)

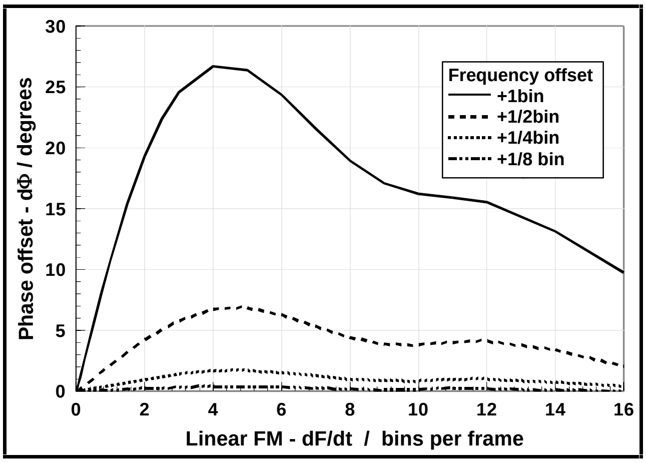
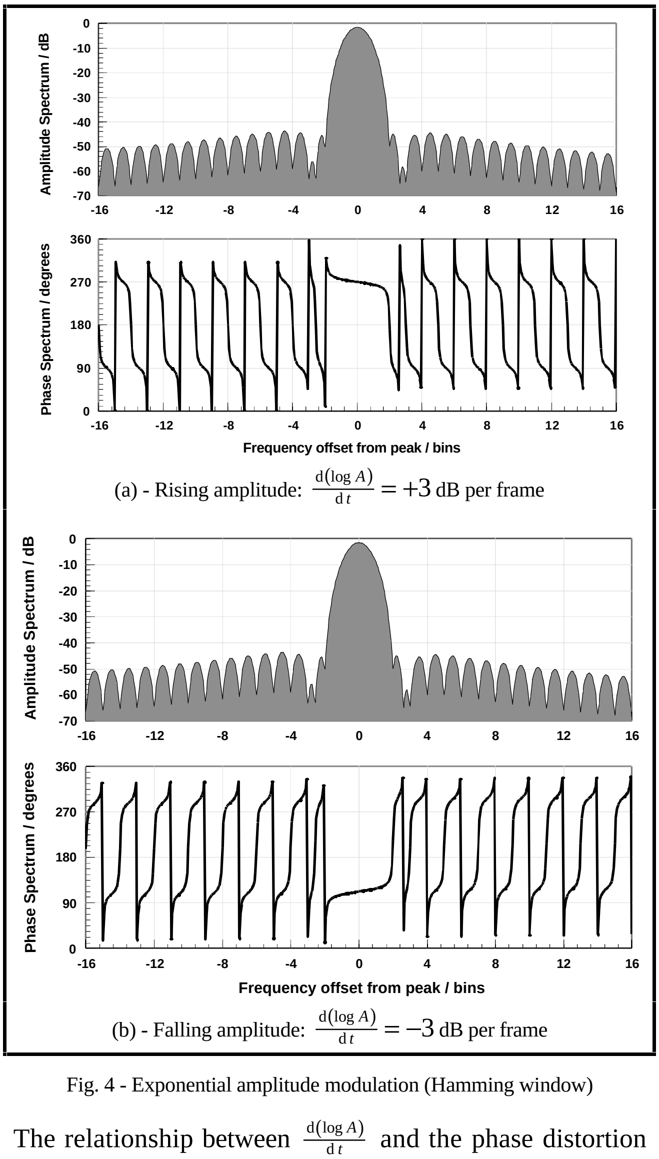


Fig. 3 - Linear FM phase distortion at various frequency offsets from the maximum (Hamming window)

2.3 Exponential Amplitude Modulation

Whereas the phase distortion for linear FM is equal either side of the maximum, in the case of exponential AM, the phase distortion is of equal magnitude but opposite sign. For exponentially increasing amplitude, the phase at a positive frequency offset from the maximum is negative, whilst at a negative frequency offset, it is positive. See figure 4(a). The amplitude spectrum is identical for a given $\left|\frac{\mathrm{d}(\log A)}{\mathrm{d}t}\right|$ regardless of whether the amplitude is rising or falling. Also, although the degree of phase distortion is identical for a given $\left|\frac{\mathrm{d}(\log A)}{\mathrm{d}t}\right|$, its orientation depends on the sign of $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$. Compare figures 4(a) and 4(b).



(a) - Rising amplitude: $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = +3$ dB per frame

(b) - Falling amplitude: $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = -3$ dB per frame

Fig. 4 - Exponential amplitude modulation (Hamming window)

The relationship between $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$ and the phase distortion at a given offset from the maximum appears to be linear, as displayed in figure 5. This linear relationship appears to exist for all the curves, suggesting that $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$ can be determined from $\mathrm{d}\Phi$ at any frequency offset within the main lobe. Unlike the linear FM case however, there *is* a unique mapping between $\frac{\mathrm{d}(\log A)}{\mathrm{d}t}$ and $\mathrm{d}\Phi$ within the range measured, thus placing no further restriction on the range of usage.

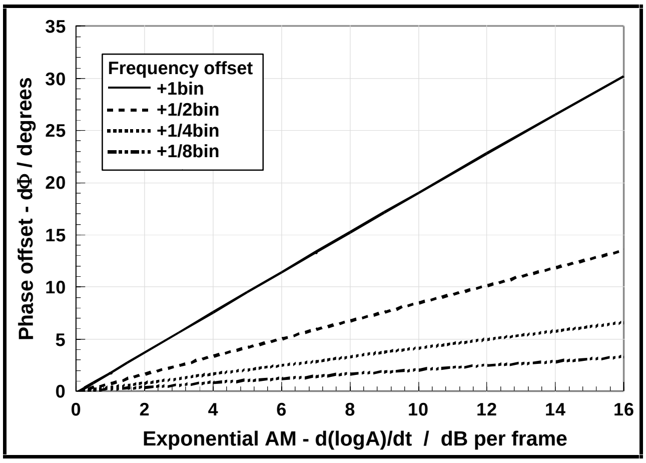
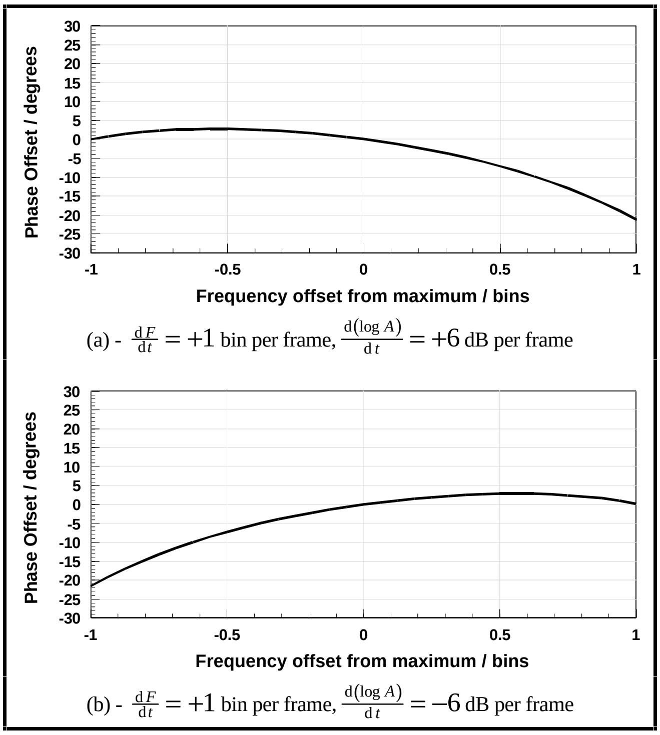


Fig. 5 - Exponential AM phase distortion at various frequency offsets from the maximum (Hamming window)

(Note that if exponential AM is displayed with amplitude in dB, it will appear as a linear modulation.)

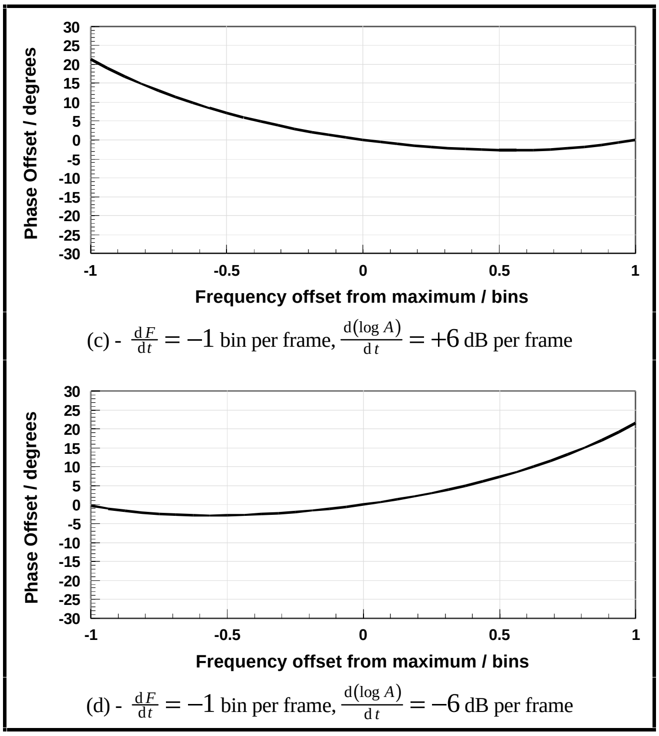
2.4 Concurrent FM and AM

Perhaps surprisingly, the phase distortion of linear FM and exponential AM are additive. At any offset from the maximum, in the range -1 to +1 bin, the total phase distortion is the sum of the distortion due to the linear FM and the distortion due to the exponential AM. The four graphs of figure 6 display combinations of rising and falling FM and AM.



(a) - $\frac{\mathrm{d}F}{\mathrm{d}t} = +1$ bin per frame, $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = +6$ dB per frame

(b) - $\frac{\mathrm{d}F}{\mathrm{d}t} = +1$ bin per frame, $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = -6$ dB per frame



(c) - $\frac{\mathrm{d}F}{\mathrm{d}t} = -1$ bin per frame, $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = +6$ dB per frame

(d) - $\frac{\mathrm{d}F}{\mathrm{d}t} = -1$ bin per frame, $\frac{\mathrm{d}(\log A)}{\mathrm{d}t} = -6$ dB per frame

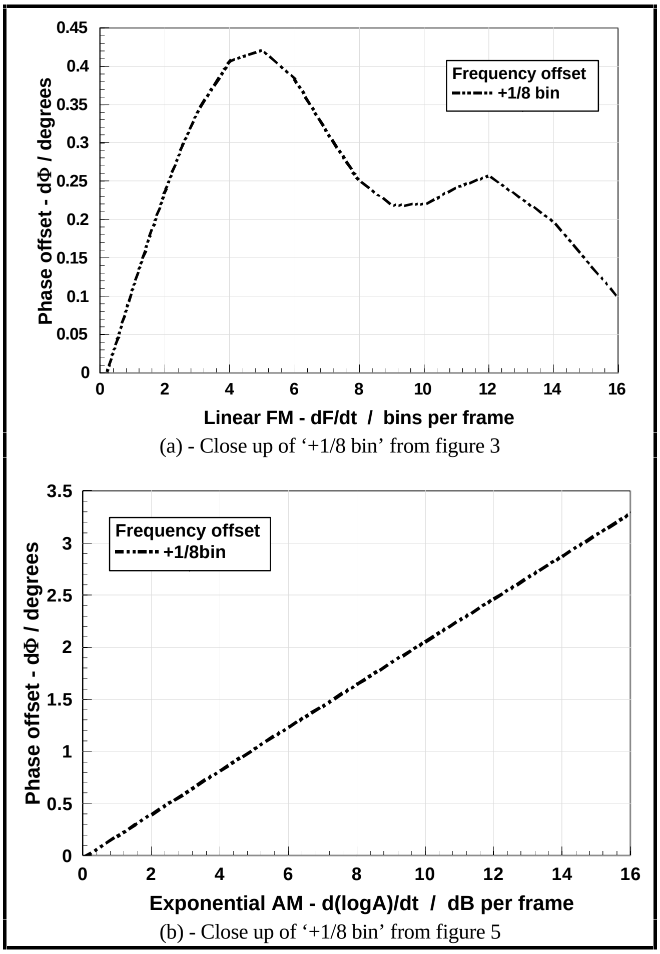
Fig. 6 - Combinations of rising and falling linear FM and exponential AM (Hamming Window)

If two measurements are taken, then the amount of distortion due to each can be separated. For example, if they are taken one either side of (and equidistant from) the maximum, then the amount of distortion due frequency and amplitude are, respectively, the sum÷2 and the difference÷2.

3. Application of Theory

In sound analysis, spectral components which are close in frequency additively interfere, affecting eachothers' amplitude and phase spectra. It is therefore desirable to make all phase distortion measurements close to the maximum of a peak, so as to maximise the influence from that peak and minimise the influence from adjacent peaks. In the following examples the measurements were made at $\frac{1}{8}$th bin from the maxima (see figure 7, based on figures 3 and 5). (Measurements were not taken closer to the maxima, because the phase distortion becomes small enough that the numerical resolution of the processor becomes significant.)

In a practical situation, such as application to audio signals, the frequency and amplitude modulation will not follow such idealised trajectories as linear FM and exponential AM. However the methodology can be used successfully if its estimations are largely accurate, when there is a presence of higher order modulation.



(a) - Close up of '+1/8 bin' from figure 3

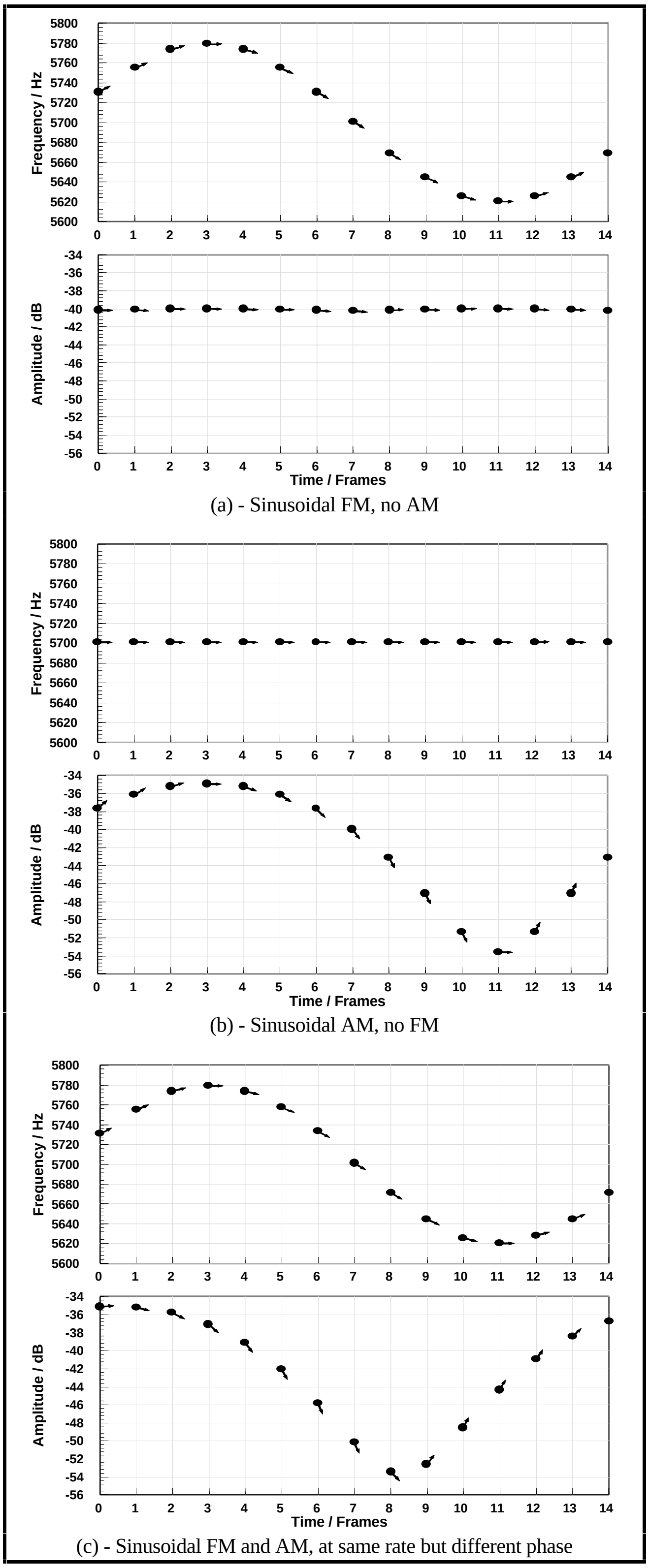
(b) - Close up of '+1/8 bin' from figure 5

Fig. 7 - Graphs used to 'decode' phase distortion of real audio data (Hamming Window)

3.1 Performance for Simulated Data

Figure 8 shows three examples of simulated audio signals. The points indicate the frequency/amplitude measured at the maximum of the peak, and the arrows indicate the frequency/amplitude trajectories measured from phase distortion.

The examples display sinusoidal FM and AM where the FFT window is short enough to approximate line segments of the frequency/amplitude curve. Consequently, the arrows approximate tangents to the curves. Figure 8(a) is the analysis of sinusoidal FM (with parameters comparable to vibrato of a musical instrument), where the amplitude is constant. Figure 8(b) is the analysis of sinusoidal AM (comparable to realistic tremolo), where the frequency is constant. Figure 8(c) shows a combination of FM and AM. The rate of modulation of each is the same, but the phase has been offset by 60° to demonstrate that the technique is not reliant on correlation between frequency and amplitude. Note that the amplitude modulation does not *appear* to be sinusoidal because a logarithmic (dB) scale is used.



(a) - Sinusoidal FM, no AM

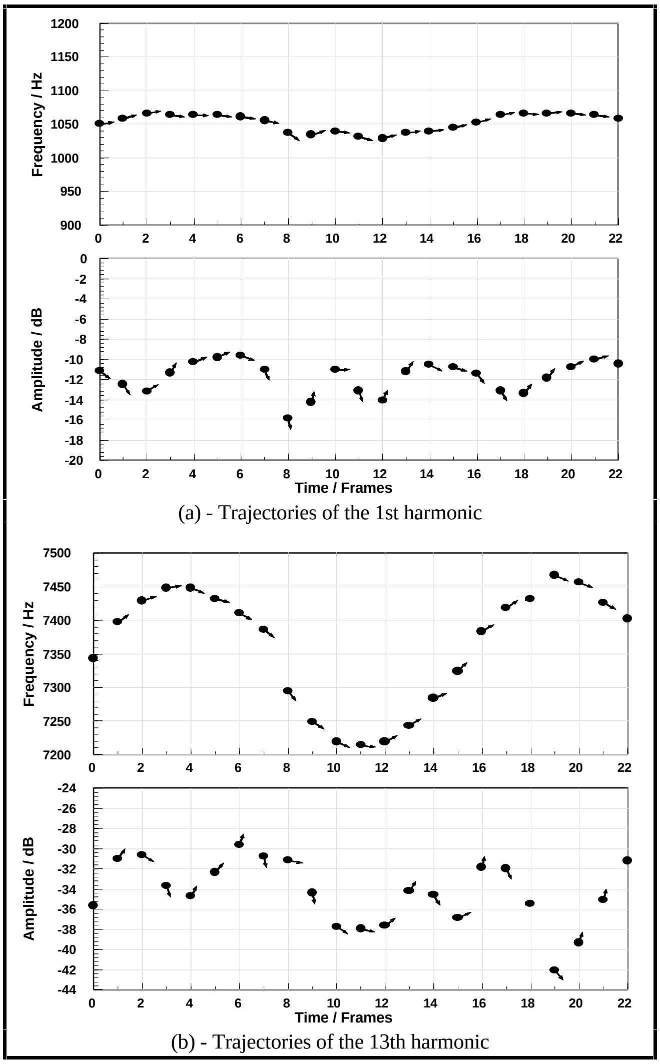
(b) - Sinusoidal AM, no FM

(c) - Sinusoidal FM and AM, at same rate but different phase

Fig. 8 - Measurements of trajectory for simulated data

3.2 Performance for Real Audio Data

Finally, the two graphs of figure 9 show the technique applied to harmonics of real audio: a cello note with a large amount of vibrato. Figure 9(a) tracks the 1st harmonic centred about 1050Hz, where the frequency modulation is slight, and figure 9(b) tracks the 13th harmonic centred about 7350Hz, where the modulation is more pronounced.



(a) - Trajectories of the 1st harmonic

(b) - Trajectories of the 13th harmonic

Fig. 9 - Frequency and amplitude trajectories of a cello note with vibrato (Graphs display 290ms segment)

3.3 Application as a Sound Analysis Technique

In order to preserve the continuity of sounds upon synthesis, the harmonics are tracked from one frame to the next. To date, this is achieved by scanning a frame of spectral data and identifying which peak (if any) is closest *in frequency* to each peak in the previous frame. In this respect, information from the phase distortion can improve the success rate, by searching for peaks lying closest to a *frequency trajectory*. As can be observed from figure 9, the amplitude changes more erratically than the frequency, but since tracking is solely conducted on frequency-based data this will cause no problems.

The current synthesis method uses linear interpolation of frequency and amplitude between frames, based on the absolute values at the start and end of each frame. With the inclusion of $\frac{dF}{dt}$ and $\frac{d(\log A)}{dt}$ data, synthesis can be achieved with cubic interpolation. Hence some of the dynamic information that was lost by using long FFT windows can now be regained.

4. FFT with Phase Distortion Analysis as an Alternative to Higher Order Spectra

The FFT has been viewed traditionally as incapable of yielding more than a linear phase representation. As a result higher order phase representations, which can describe nonstationarities of frequency, have been (and continue to be) developed. These are largely based on the Wigner-Ville transform, which achieves quadratic phase (linear FM) representation.

For signals that are mono-component, nonstationary, these higher order spectra (HOS) have proved very useful. However for multi-component signals such as sounds, the spectra display peaks not only for each component (the auto terms), but also for the correlation between components (the cross terms). The cross terms are often large enough to be indistinguishable from the auto terms, and can even mask them at times [2]. Current research is attempting to overcome this problem by developing techniques that suppress the cross terms; e.g. [1,3].

The technique presented here is capable of yielding second order phase information, without the complications associated with the Wigner-Ville distribution and its descendants. In addition, it yields information about amplitude nonstationarity.

The compromise for these abilities is that: 1) there must be sufficient frequency separation between concurrent components; 2) the information can only be gained for a limited range of linear FM as indicated by figure 3. The first restriction is one already present in sound analysis, and the second is largely unrestrictive for sound analysis.

The simplicity of the method presented indicates potential for extending this technique to higher orders of modulation. This is especially promising, since distortion from modulation (of whatever order) appears to be concentrated around the maximum of the associated spectral peak.

References

[1] S. Barbarossa, A,Zanalda. 1992. *A Combined Wigner-Ville and Hough Transform for Cross-terms Suppression and Optimal Detection and Parameter Estimation.* Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-92; Vol V)

[2] B.Boashash, B.Ristich. 1992. *Polynomial Wigner-Ville Distributions and Time-Varying Higher Order Spectra.* Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Victoria,BC,Canada)

[3] R.S.Orr, J.M.Morris, S.-E.Qian. 1992. *Use of the Gabor Representation for Wigner Distribution Crossterm Suppression..* Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-92; Vol V)

[4] X. Serra. 1990. *A system for sound analysis / transformation / synthesis based on a deterministic plus stochastic decomposition.* Ph.D. diss., Stanford University.

APPENDIX C

IMPROVED MODELLING OF ATTACK TRANSIENTS IN MUSIC ANALYSIS-RESYNTHESIS

co-authored with Andrew Bateman

Originally published:

1996. Proceedings of the International Computer Music Conference (ICMC96).
pp.100-103.
Organised by the International Computer Music Association (ICMA).

Improved Modelling of Attack Transients in Music Analysis-Resynthesis

Paul Masri, Andrew Bateman
Digital Music Research Group, University of Bristol
5.11 Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, U.K.
Tel: +44 117 954-5203, Fax: +44 117 925-5265, email: Paul.Masri@bristol.ac.uk

Abstract

Current music analysis-resynthesis models represent sounds through a set of features, which are extracted from a time-frequency representation. So that each time-frame can present a good approximation to the instantaneous spectrum, it is necessary to analyse the waveform in short segments. This is achieved with a window function whose position is advanced by a fixed amount between frames. When the window encompasses a transient event, such as the percussive onset of a note, it contains information both before and after the event. These partially-correlated spectra often become confused during analysis and cause audible 'diffusion' upon resynthesis.

This paper presents a simple, novel technique to avoid the problem, by synchronising the analysis window to transient events. Event locations are identified by observing short-term changes in the spectrum. Thereafter the position of the analysis window is constrained, to prevent it capturing the signal both sides of an event simultaneously. This method, which has been automated, yields an improvement that is clearly audible, particularly for percussive sounds which retain their 'crispness'.

1. Introduction

It has long been known that the onset of a note, the *attack*, plays an important role in our perception of timbre [pp.9-12, Grey, 1975]. In traditional instruments, it is the phase during which resonances are building up, but before the steady state condition of standing waves has been established. Where the attack is short, such as for the trumpet, there are many rapid changes, so that it can sound like a noise burst. For this reason, the attack transient is difficult to study.

It is not surprising therefore, that attacks are not well understood and are not well represented within analysis-resynthesis models. This paper focuses on finding a solution in the context of the popular Deterministic Plus Stochastic model. This model was developed by Xavier Serra[1990], based upon the Sinusoidal model of McAulay and Quatieri[1986]. The presented work was implemented within the authors' own system, from which other model developments have also been forthcoming [Masri & Bateman, 1994, 1995].

1.1 Traditional Spectral Analysis and The Problem Caused by Attack Transients

The first step in the analysis process is the time-frequency representation, which is calculated using the Short Time Fourier Transform (STFT). For each frame, a small portion of the time domain waveform is isolated, by application of an analysis window, and spectral estimation is computed using the Fast Fourier Transform (FFT). Between frames, the analysis window is advanced by a fixed amount, called the *hop-distance*.

During the deterministic analysis, the primary goal is to detect and locate the *partials*, the instantaneously sinusoidal elements that compose the harmonic structure of a pitched sound. For good frequency resolution, it is necessary to have a long analysis window. In opposition to this, for good time resolution, the window must be short. The practical constraint of separating partials in the frequency domain necessarily favours good frequency resolution.

The central assumption when using the FFT is that of stationarity - the waveform is assumed to be truly periodic within the analysis window. Small deviations from this generate a small, but tolerable, amount of distortion.

When the analysis window contains a percussive note onset, there is a dramatic change in the waveform. The introduction of the new sound element is unconnected to the preceding sound signal and cannot in any way be predicted from it. Therefore the waveform preceding the note onset is largely uncorrelated with the waveform following its initiation.[1]

This far-from-stationary situation causes much distortion in the FFT spectrum and often affects the subsequent feature extraction processes adversely. This is demonstrated in Figure 1 where the frames prior to an attack transient are similar; also the frames following its onset are similar to each other; the central frame spanning both regions, though, is a 'confused' average of both spectra.

Within the deterministic analysis, the partials are identified from peaks in each spectrum and continuous trajectories are formed by linking peaks between frames. A frame which contains a percussive attack or other transient event often contains a large burst of energy throughout the spectrum. In such a case, the peak detection algorithm may fail to detect peaks, resulting in a

[1] The sound preceding the note onset may continue, so the spectra may not be *totally* uncorrelated.

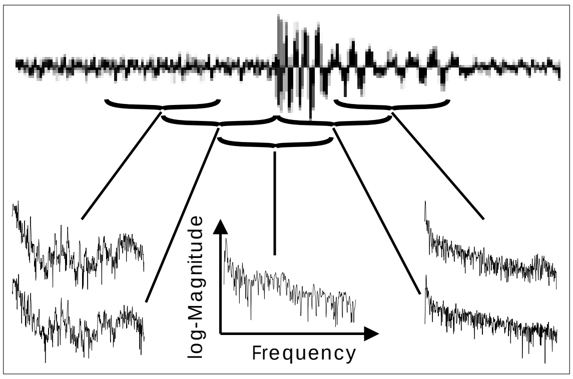
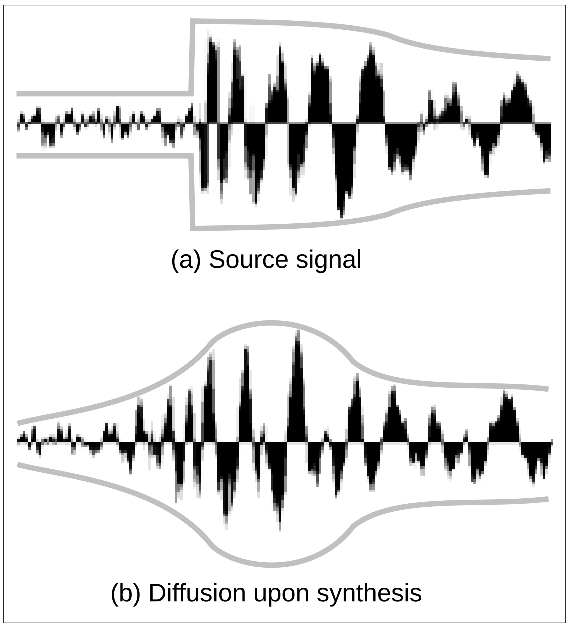


Figure 1 - Spectra surrounding an attack

momentary drop-out upon synthesis. Alternatively, the rapid changes may introduce multiple spurious peaks. Some of these will be discarded at the peak linking stage and some will be erroneously linked, leading to artifacts upon synthesis.

In some cases, where the transient event does not dominate the signal, peak linking may be largely successful. However, upon synthesis, the partial trajectories are smoothly interpolated between frames and the transient events, which were highly localised in time, become diffused or even dissolved completely. The diffusion effect is reinforced by the stochastic aspect of the model: during analysis each spectral envelope represents an averaged snapshot of the spectrum; upon synthesis, there is smooth cross-fading between frames. See Figure 2.


(a) Source signal

(b) Diffusion upon synthesis

Figure 2 - Smoothed attack causing ‘diffusion’

The solution presented in this paper detects transient events and avoids the above problems, by synchronising the analysis and synthesis processes to those event locations.

2. Detection of Transient Events

The detection method was designed to recognise two signal properties associated with a sharp attack: the suddenness of the signal change and the increase in energy. A frequency domain method was chosen because of its ability to reveal both changes in overall energy and the energy concentration in frequency. The frequency location of energy is important because the sudden change to the signal will cause phase discontinuities; in the frequency spectrum this appears as high frequency energy.

Naturally the time resolution for detecting transient events must be smaller than that of the main analysis STFT, if any advantage is to be gained. This necessitates a reduction in frequency resolution, but fine frequency resolution is not an issue here; only the broad spectral envelope is required. The following parameters were found to be useful:-

Window-length =2.9ms (128 samples @ 44.1kHz),
Hop-distance =1.5ms (64 samples @ 44.1kHz),
Hamming window function,
No zero padding.

The hop-distance is set to half the window-length, the maximum value that ensures each transient event will appear toward the centre of the window in at least one frame.

2.1 Detection Function

The energy function is calculated as the sum of the magnitude squared of each frequency bin (in the specified range):

$$E = \sum_{k=2}^{N/2+1} \left\{ \left| X(k) \right|^2 \right\} \qquad (1)$$

where E is the energy function for the current frame,
N is the FFT array length
(so $N/2 + 1$ corresponds to the frequency $F_s/2$,
F_s is the sample rate),
$X(k)$ is the kth bin of the FFT.

The function to measure high frequency content was arbitrarily set to a weighted energy function, linearly biased toward the higher frequencies:

$$HFC = \sum_{k=2}^{N/2+1} \left\{ \left| X(k) \right|^2 \cdot k \right\} \qquad (2)$$

where HFC is the High Frequency Content function for the current frame,
other symbols as defined above.

In both cases the lowest two bins are discarded, to avoid unwanted bias from DC or low frequency components.

The condition for detection combines the results from each pair of consecutive frames thus:

$$\frac{HFC_r}{HFC_{r-1}} \cdot \frac{HFC_r}{E_r} > T_D \qquad (3)$$

where subscript r denotes current frame (equals latter of two in detection function),
subscript r-1 denotes the previous frame,
T_D is the threshold, above which a hit is detected.

(Note that HFC_{r-1} and E_r are constrained to have a minimum value of one, to avoid the potential 'Divide by zero' computation error.)

The detection function is the product of the rise in high frequency energy between the two frames and the normalised high frequency content for the current frame.

For attacks whose growth is slightly slower, but whose onset is nevertheless sudden, the detection function could be triggered on more than one frame. To avoid multiple detections, the algorithm is given a parameter for the minimum closeness of two hits. In practice, setting this to 2 frames is adequate for the majority of sounds (i.e. only disallowing consecutive hits).

2.2 Temporal Resolution and Region Definition

The accuracy, in time, of the detection process is equal to the hop-distance. The above values of STFT parameters give a resolution of 1.5ms, which compares favourably to the accepted resolution of the ear, which is 2-10ms.

The *transient event boundary* - the point of onset of the transient - is stored as the start of the second analysis window of the detection pair. It is the second window that contains the event, so this placement ensures that the whole of the attack is located after that point. In this way, any errors in the deterministic analysis, caused by inaccuracy within this process, are confined to the frame containing the percussive onset, where the suddenness of the attack should dominate perceptually.

The detection process is carried out as a pre-analysis scan, and its results are compiled as a *region list*, where the region boundaries correspond to the detected transient event boundaries, as shown in Figure 3.

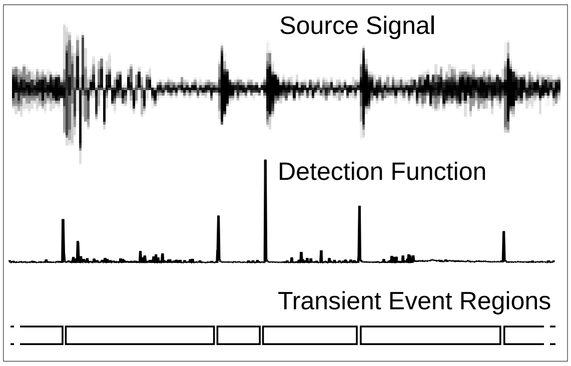


Figure 3 - Event detection and region generation

3. Synchronised Analysis

At the start of each region, the analysis window is positioned with its trailing end at the first sample. Thereafter analysis proceeds, as normal, the window advancing by the hop-distance for each new frame.

The first frame whose window touches or crosses the region boundary is 'snapped' so that its leading edge coincides with the last sample of the region. Naturally, this means that the final hop-distance is reduced.

3.1 Extrapolation Towards Region Boundaries

The data for each frame notionally corresponds to the instantaneous situation at the centre of the frame. Therefore the first half of the first frame and the last half of the last frame in each region are undefined. For simplicity, the data in these frames are extrapolated outward to the respective region boundaries.

This makes the implicit assumption that the waveform is slow changing in these zones. Whereas this may be accurate at the end of a region, we already know that there are rapid changes at the start of a region. Despite this, the STFT provides no way of gaining extra detail.

4. Synchronised Synthesis

Upon synthesis, both deterministic and stochastic, the extrapolated spectra are synthesised beyond the region ends by a short amount. This provides sufficient excess waveform to allow a crossfade between regions. See Figure 4 below.

The crossfade length has been set to 5.8ms (256 samples @ 44.1kHz sample rate), where it is short enough to reproduce the suddenness of the original attack, but not so short that an audible click (termed a 'glitch') is produced.

5. Results

5.1 Performance

Figure 4 shows that the proposed method is able to retain the suddenness of an attack transient. The success of the method is further confirmed through

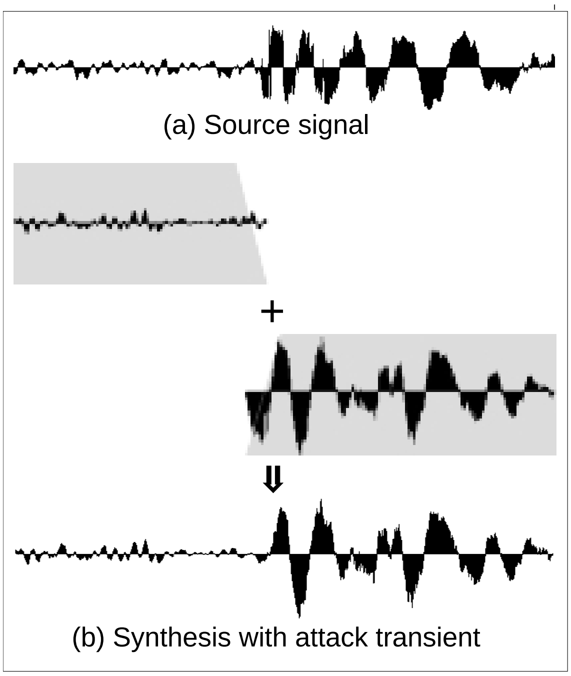


Figure 4 - Crossfade at a region boundary

listening tests. Sounds synthesised by the new method retain the crispness of the original sound.

The performance is good even where the spectral evolution has been simplified. (Some spectral detail is inevitably lost following a transient event boundary, as an inevitable consequence of using the STFT for analysis.) It would appear that the abrupt rise in volume and the sudden change in spectral content are the most prevalent perceptual features.

Some degradation is noticeable for sounds with greater 'depth', such as a booming bass drum or a snare with reverberation, where some of that power is lost. This is probably due to the rapidity of change following the transient onset, the necessary detail of which, the analysis is still unable to extract.

One area where the modification is unable to help is the rapid succession of transient events, such as from a drum roll, or drums with gated reverb. In these cases, the pre-analysis scan is often still able to detect the event boundaries, but the events are too close to one another to apply the synchronisation. That is, the resultant regions are shorter than a single window-length for the main analysis FFT's.

5.2 Cost

The computational cost of the pre-analysis scan is low, when compared to the analysis itself. The FFT's are much shorter - and thereby disproportionately faster[2] - and the hop-distance of half a frame is a reduction in the overlap of analysis windows between frames.

In addition to the pre-analysis scan, some changes have been made to the model structure, but these have an impact only once per transient event, and the impact is minimal.

6. Conclusions

The popular Deterministic Plus Stochastic model makes the assumption that all waveform changes are gradual, an unavoidable consequence of the STFT's limited time-frequency resolution. Consequently, the model fails to capture transient events adequately and audible diffusion or complete drop-outs result.

In this paper a method has been presented that extends the model to incorporate percussive attacks and other transient events. The principle behind the method is that the spectra before and after an attack transient should be treated as different, the change happening instantaneously at the transient onset.

The results have proven successful in preserving the abrupt amplitude change in the waveform and the crispness, perceptually, of attack transients. This is however the first implementation of this technique, and improvements are possible. Two areas of future work are proposed.

The time domain envelope at the start of each region could be captured and imposed on the synthesised output, to improve the 'sense of depth' for certain sounds.

Analysis of closely-spaced attacks may require a change to the time-frequency representation. Higher order spectra (e.g. Wigner-Ville, bispectrum, etc.) do not suffer the same time-frequency resolution restrictions of the linear transforms (e.g. Fourier, Wavelet) [Boashash, 1990][Cohen, 1989]. However they possess their own quirks, and a detailed investigation is needed before they can be applied to music analysis-resynthesis. Further discussion on this subject is soon to be published in [Masri, *to be published*].

Acknowledgements

The authors wish to express their gratitude to **Soundscape Digital Technology** for funding this research and providing essential equipment. Thanks are also due to the **Centre for Communications Research** (University of Bristol) and its director, Prof. Joe McGeehan, for supporting the research with generous access to facilities and the expertise of its staff.

References

[Boashash, 1990] B. Boashash. "Time frequency Signal Analysis" (ch.9) in *Advances in Spectral Analysis.* Ed. S.Haykin. 1990. Vol.1; pp.418-517. Publ. Prentice Hall (Englewood Cliffs, NJ, USA).

[Cohen, 1989] L. Cohen. "Time-frequency Distributions - A Review" in *Proceedings of the IEEE.* 1989. Vol.77:7, pp.941-981.

[Grey, 1975] J.M. Grey. *An exploration of musical timbre using computer-based techniques for analysis, synthesis and perceptual scaling.* Ph.D. dissertation, 1975. Stanford University.

[Masri & Bateman, 1994] P. Masri, A. Bateman. "Partial Domain Synthesis of Music" in *Digital Signal Processing Conference Proceedings (DSP UK 94).* 1994. Vol.1.

[Masri & Bateman, 1995] P. Masri, A. Bateman. "Identification of Nonstationary Audio Signals Using the FFT, with Application to Analysis-based Synthesis of Sound" in *IEE Audio Engineering Colloquium Digest.* 1995. pp.11/1-11/6.

[Masri, *to be published*] P. Masri. *Computer modelling of Sound for Transformation and Synthesis of Musical Signals.* Ph.D. dissertation, due for submission in Summer 1996. University of Bristol.

[McAulay & Quatieri, 1986] R.J. McAulay, T.F. Quatieri. "Speech Analysis/Synthesis based on a Sinusoidal Representation" in *IEEE Transactions on Acoustics, Speech and Signal Processing.* 1986. Vol.34:4, pp.744-754.

[Serra,1990] X. Serra. *A system for sound analysis/ transformation/synthesis based on a deterministic plus stochastic decomposition.* Ph.D. dissertation, 1990. Stanford University.

[2] The processing required for an FFT is related to N, the length of the FFT array, by the factor NlogN. Thus a reduction in array length yields more than a proportional advantage in computation speed.

BIBLIOGRAPHY

BIBLIOGRAPHY

R. Bargar; B. Holloway; X. Rodet; C. Hartman. 1995. "Defining spectral surfaces" in *Proc. International Computer Music Conference (ICMC)*. pp. 373-376.

B. Boashash. 1990. "Time frequency signal analysis" (ch.9) in *Advances in spectral analysis*. Ed. S. Haykin. Vol.1. pp. 418-517. Publ. Prentice Hall (Englewood Cliffs,New Jersey,USA)

B. Boashash; G. Frazer. 1992. "Time-varying higher-order spectra, generalised Wigner-Ville distribution and the analysis of underwater acoustic data" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.5, pp. 193-196.

B. Boashash; P. O'Shea. 1991. "Time-varying higher order spectra" in *International Conference on Digital Signal Processing*. pp. 759-762.

C.S. Burrus; T.W. Parks. 1985. *DFT/FFT and Convolution Algorithms - Theory and Implementation.* #1 of *Topics in Signal Processing*. Publ. John Wiley & Sons (New York)

C. Cadoz; A. Luciani; J. Florens. 1984. "Responsive input devices and sound synthesis by simulation of instrumental mechanisms: The Cordis system" in *Computer Music Journal*. Vol.8:3, pp. 60-73.

C. Chafe. 1995. "Adding vortex noise to wind instrument physical models" in *Proc. International Computer Music Conference (ICMC)*. pp. 57-60.

C. Chafe; D. Jaffe; K. Kashima; B. Mont-Reynaud; J.O. Smith. 1985. "Techniques for note identification in polyphonic music" in *Proc. International Computer Music Conference (ICMC)*. pp. 399-405.

S.-K. Chan; J. Yuen; A. Horner. 1996. "Discrete summation synthesis and hybrid sampling-wavetable synthesis of acoustic instruments with genetic algorithms" in *Proc. International Computer Music Conference (ICMC)*. pp. 49-51.

T. Chang; C.-C.J. Kuo. 1993. "Texture analysis and classification with tree-structured wavelet transform" in *IEEE Transactions on Image Processing*. Vol.2:4, p. 429-441.

H.L. Choi; W.J. Williams. 1989. "Improved time-frequency representation of multicomponent signals using exponential kernels" in *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol.37:6, pp. 862-871.

J.M. Chowning. 1973. "The synthesis of complex audio spectra by means of frequency modulation" in *Journal of the Audio Engineering Society (AES).* Vol.21:7. (Reprinted in Foundations of Computer Music. 1985. (Eds. C.Roads & J.Strawn), pp.6-29.)

M. Clarke. 1996. "TIM(br)E: Compositional approaches to FOG synthesis" in *Proc. International Computer Music Conference (ICMC).* pp. 375-377.

T.A.C.M. Classen; W.F.G. Mecklenbrauker. 1984. "On the time-frequency discrimination of energy distributions : Can they look sharper than Heisenberg?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* pp. 41.B.7.1 - 41.B.7.4.

L. Cohen. 1966. "Generalized phase-space distribution functions" in *J.Math.Phys.* Vol.7, pp. 781-786.

L. Cohen. 1989. "Time-frequency distributions - A review" in *Proceedings of the IEEE.* Vol.77:7, pp. 941-981.

P.R. Cook. 1991. "TBone: An interactive WaveGuide brass instrument synthesis workbench for the NeXT machine" in *Proc. International Computer Music Conference (ICMC).* pp. 297-300.

P. Depalle; G. Garcia; X. Rodet. 1993. "Tracking of partials for additive sound synthesis using hidden markov models" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Vol.1, pp. 225-228.

P. Depalle; L. Tromp. 1996. "An improved additive analysis method using parametric modelling of the Short-Time Fourier Transform" in *Proc. International Computer Music Conference (ICMC).* pp. 297-300.

S. Deutsch; A. Deutsch. 1993. *Understanding the Nervous System : an engineering perspective.* Publ. Institute of Electrical and Electronics Engineers (IEEE) (New York)

Y. Ding; X. Qian. 1997. "Estimation of sinusoidal parameters of musical tones based on global waveform fitting" in *IEEE Workshop on Multimedia Signal Processing.* (Submitted; awaiting publication)

C. Dodge; T. Jerse. 1985. *Computer Music.* Publ. Schirmer Books (New York)

M.C. Dogan; J.M. Mendel. 1992. "Real-time robust pitch detector" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Vol.1, pp. 129-132.

B. Doval; X. Rodet. 1991. "Estimation of fundamental frequency of musical sound signals" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Vol.5, pp. 3657-3660 (Secn. A2.11).

B. Doval; X. Rodet. 1993. "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMM's" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.1, pp. 221-224.

S. Dubnov; N. Tishby; D. Cohen. 1996. "Influence of frequency modulating jitter on higher order moments of sound residual with applications to synthesis and classification" in *Proc. International Computer Music Conference (ICMC)*. pp. 378-385.

S.A. Van Duyne; J.O. Smith. 1993. "The 2-D digital waveguide mesh" in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. (*Final Program and Paper Summaries*); pp. 177-180.

S.A. Van Duyne; J.O. Smith. 1996. "The 3D tetrahedral digital waveguide mesh with musical applications" in *Proc. International Computer Music Conference (ICMC)*. pp. 9-16.

M. Evans. 1993. Future Music (music technology magazine). Issue 5, March 1993. Quote sourced from p.40.

A.B. Fineberg; R.J. Mammone. 1992. "A method for instantaneous frequency tracking of multiple narrowband signals" in *Signal Processing*. Vol.29, pp. 29-44.

K. Fitz; L. Haken. 1995. "Bandwidth enhanced sinusoidal modeling in Lemur" in *Proc. International Computer Music Conference (ICMC)*. pp. 154-157.

K. Fitz; L. Haken; B. Holloway. 1995. "Lemur - A tool for timbre manipulation" in *Proc. International Computer Music Conference (ICMC)*. pp. 158-161.

J.R. Fonollosa; C.L. Nikias. 1992. "Analysis of transient signals using higher-order time-frequency distributions" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.5, pp. 197-200.

A. Freed; X. Rodet; P. Depalle. 1993. "Synthesis and control of hundreds of sinusoidal partials on a desktop compter without custom hardware" in *Proc. International Conference on Signal Processing Applications and Technology*. pp. 1024-1031.

D. Gabor. 1947. "Acoustical quanta and the theory of hearing" in *Nature*. Vol.159:1044, pp. 591-594.

N.L. Gerr. 1988. "Introducing a third order Wigner distribution" in *Proceedings of the IEEE*. Vol.76, pp. 290-292.

J.L. Goldstein. 1973. "An optimum processor theory for the central formation of the pitch of complex tones" in *Journal of the Acoustical Society of America*. Vol.54:6, pp. 1496-1516.

M. Goodwin; A. Kogon. 1995. "Overlap-add synthesis of nonstationary sinusoids" in *Proc. International Computer Music Conference (ICMC)*. pp. 355-356.

M. Goodwin; X. Rodet. 1994. "Efficient Fourier synthesis of nonstationary sinusoids" in *Proc. International Computer Music Conference (ICMC)*. pp. 333-334.

J.M. Grey. 1975. *An exploration of musical timbre using computer-based techniques for analysis, synthesis and perceptual scaling.* Ph.D. thesis.(Stanford University)

R. Gribonval; P. Depalle; X. Rodet; E. Bacry; S. Mallat. 1996. "Sound signals decomposition using a high resolution matching pursuit" in *Proc. International Computer Music Conference (ICMC)*. pp. 293-296.

A. Grossman; M. Holschieder; R. Kronland-Martinet; J. Morlett. 1987. "Detection of abrupt changes in sound signals with the help of the wavelet transforms" in *Advances in Electronics and Electron Physics*. Ed. P.C. Sabatier. (*Suppl. 19 - Inverse Problems: An Interdisciplinary Study*); pp. 289-306.

R.C.L. Guevara; G.H. Wakefield. 1996. "A modal distribution approach to piano analysis and synthesis" in *Proc. International Computer Music Conference (ICMC)*. Vol.99, pp. 350-351.

L. Haken. 1995. "Real-time timbre modifications using sinusoidal parameter streams" in *Proc. International Computer Music Conference (ICMC)*. pp. 162-163.

F.J. Harris. 1978. "On the use of windows for harmonic analysis with the Discrete Fourier Transform" in *Proceedings of the IEEE*. Vol.66:1, pp. 51-83.

S. Haykin. 1994. *Neural Networks: A comprehensive foundation.* Publ. Prentice-Hall (London)

H.L.F. von Helmholtz. 1954. *On the sensations of tone as a physiological basis for the theory of music.* Publ. Dover (New York) (Translation by A.J. Ellis; Originally published 1877)

D.J. Hermes. 1988. "Measurement of pitch by subharmonic summation" in *Journal of the Acoustical Society of America*. Vol.83:1, pp. 257-264.

H. Hertz; A. Krogh; R.G. Palmer. 1991. *Introduction to the theory of neural computing.* Publ. Addison-Wesley (Reading,Massachusetts)

A. Horner. 1996. "Double-modulator FM matching of instrument tones" in *Computer Music Journal*. Vol.20:2, pp. 57-71.

D. Jaffe; J.O. Smith. 1983. "Extensions of the Karplus-Strong Plucked-String Algorithm" in *Computer Music Journal*. Vol.7:2, pp. 56-69.

D.A. Jaffe; J.O. Smith. 1995. "Performance expression in commuted waveguide synthesis of bowed strings" in *Proc. International Computer Music Conference (ICMC)*. pp. 343-346.

J. Jeong; W.J. Williams. 1992. "Kernel design for reduced interference distributions" in *IEEE Transactions on Signal Processing*. Vol.40:2, pp. 402-412.

D.L. Jones; T.W. Parks. 1992. "A resolution comparison of several time-frequency representations" in *IEEE Transactions on Signal Processing*. Vol.40:2, pp. 413-420.

K. Karplus; A. Strong. 1983. "Digital synthesis of plucked-string and drum timbres" in *Computer Music Journal*. Vol.7:2, pp. 43-55.

S.M. Kay. 1988. *Modern spectral estimation: Theory & Application.* Publ. Prentice-Hall (Englewood Cliffs,New Jersey,USA)

P. Kleczkowski. 1989. "Group additive synthesis" in *Computer Music Journal*. Vol.13:1, pp. 12-20.

B. Kosko. 1994. *Fuzzy Thinking - The new science of fuzzy logic.* Publ. Flamingo, HarperCollins (London)

R. Kronland-Martinet. 1988. "The wavelet transform for analysis, synthesis and processing of speech and music sounds" in *Computer Music Journal*. Vol.12:4, pp. 11-20.

J. Laroche. 1993. "HNM: A simple, efficient harmonic+noise model for speech" in *Proc. IEEE-SP International Symposium Time-Frequency and Time-Scale Analysis*. (17-20 Oct 1993, New Paltz,NY,USA) (*Final Program and Paper Summaries*); pp. 169-172.

J. Laroche; Y. Stylianou; E. Moulines. 1993. "HNS: Speech modification based on a harmonic+noise model" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.2, pp. 550-553.

P.J. Loughlin; J.W. Pitton; L.E. Atlas. 1992. "An information-theoretic approach to positive time-frequency distributions" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 125-128.

J.P. Mackenzie. 1995. "Chaotic predictive modelling of sound" in *Proc. International Computer Music Conference (ICMC)*. pp. 49-56.

J. Makhoul. 1975. "Linear prediction: a tutorial review" in *Proceedings of the IEEE*. Vol.63:4, pp. 561-580.

S.G. Mallat; Z. Zhang. 1993. "Matching pursuits with time-frequency dictionaries" in *IEEE Transactions on Signal Processing*. Vol.41:12, pp. 3397-3415.

J.D. Markel; A.H. Gray. 1976. *Linear prediction of speech.* Publ. Springer-Verlag (New York)

S.L. Marple. 1987. *Digital spectral analysis with applications.* Publ. Prentice-Hall (Englewood Cliffs,NJ,USA)

P. Martin. 1982. "Comparison of pitch detection by cepstrum and spectral comb analysis" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 180-183.

P. Masri; A. Bateman. 1994. "Partial domain synthesis of music" in *Digital Signal Processing (UK)*. Vol.1, (Pages unnumbered). (Organised by New Electronics) **See Appendix A.**

P. Masri; A. Bateman. 1995. "Identification of nonstationary audio signals using the FFT, with application to analysis-based synthesis of sound" in *IEE Colloquium on Audio Engineering*. pp. 11.1-6. **See Appendix B.**

P. Masri; A. Bateman. 1996. "Improved modelling of attack transients in music analysis-resynthesis" in *Proc. International Computer Music Conference (ICMC)*. pp. 100-103. **See Appendix C.**

R.J. McAulay; T.F. Quatieri. 1986. "Speech analysis/synthesis based on a sinusoidal representation" in *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol.34:4, pp. 744-754.

M.L. Meade; C.R. Dillon. 1986. *Signals and Systems: models and behaviour.* #8 of *Tutorial Guides in Electronic Engineering* (series eds. G.G. Bloodworth, A.P. Dorey, J.K. Fidler). Publ. Van Nostrand Reinhold (UK) (Wokingham,Berkshire,England)

A.M. Noll. 1964. "Short-time spectrum and "Cepstrum" techniques for vocal-pitch detection" in *Journal of the Audio Engineering Society (AES)*. Vol.36:2, pp. 296-302.

A.H. Nuttall. 1981. "Some windows with very good sidelobe behavior" in *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol.29:1, pp. 84-91.

K. Ohya. 1995. "A sound synthesis by recurrent neural network" in *Proc. International Computer Music Conference (ICMC)*. pp. 420-423.

A.V. Oppenheim; J.S. Lim. 1981. "The importance of phase in signals" in *Proceedings of the IEEE*. Vol.69:5, pp. 529-541.

P. Pabon. 1994. "Real-time spectrum/cepstrum games" in *Proc. International Computer Music Conference (ICMC)*. p. 361. (Proceedings only include abstract)

E.R.S. Pearson; R.G. Wilson. 1990. "Musical event detection from audio signals within a multiresolution framework" in *Proc. International Computer Music Conference (ICMC)*. pp. 156-158.

N.C. Petroni; F. Degrassi; G. Pasquariello. 1994. "Detection of pitch in random acoustic signals by neural networks" in *Journal of New Music Research*. Vol.23:4, pp. 369-399.

D. Phillips; A. Purvis; S. Johnson. 1994. "A multirate optimisation for real-time additive synthesis" in *Proc. International Computer Music Conference (ICMC)*. pp. 364-367.

D. Phillips; A. Purvis; S. Johnson. 1996. "Multirate additive synthesis" in *Proc. International Computer Music Conference (ICMC)*. pp. 496-499.

W.J. Pielemeier; G.H. Wakefield. 1996. "A high resolution time-frequency representation for musical instrument signals" in *Journal of the Acoustical Society of America*. Vol.99:4, pp. 2382-2396.

J.R. Pierce. 1989. "Introduction" (ch.1) in *Current Directions in Computer Music Research*. Eds. M.V. Mathews, J.R. Pierce. #2 of *System Development Foundation Benchmark Series*. pp. 1-4. Publ. The MIT Press (Cambridge,Massachusetts; London,England)

J.R. Pierce. 1992. *The Science of Musical Sound. (2nd ed.)* Publ. W.H. Freeman and Co. (New York)

R. Porter; N. Canagarajah. 1996. "A robust automatic clustering scheme for image segmentation using wavelets" in *IEEE Transactions on Image Processing*. Vol.5:4, pp. 662-665.

T. Pratchett. 1995. *Interesting Times. (2nd ed.)* Publ. Corgi Books, Transworld Publishers Ltd.,London,UK (London/Australia/Aukland)

X. Qian; Y. Ding. 1997. "A novel approach to estimating sinusoidal parameters of musical tones" in *IEEE Signal Processing Letters*. (Submitted; awaiting publication)

A. Radunskaya. 1996. "Chaos and non-linear models" in *Proc. International Computer Music Conference (ICMC)*. pp. 440-443.

O. Rioul; M. Vetterli. 1991. "Wavelets and signal processing" in *IEEE Signal Processing Magazine*. (October 1991) pp. 14-38.

J.-C. Risset. 1991. "Timbre analysis by synthesis: Representations, imitations and variants for musical composition" in *Representations of musical signals*. Eds. G. De Poli, A. Piccialli, C. Roads. pp. 7-43. Publ. The MIT PRess (Cambridge,Massachusetts; London,England)

J.-C. Risset; M.V. Mathews. 1969. "Analysis of musical-instrument tones" in *Physics Today*. Vol.22, pp. 23-30.

C. Roads. 1985. "A tutorial on nonlinear distortion or waveshaping synthesis" in *Foundations of Computer Music*. Eds. C. Roads, J. Strawn. pp. 83-94. Publ. The MIT Press (Cambridge,Massachsetts/ London,England)

C. Roads. 1988. "Introduction to Granular Synthesis" in *Computer Music Journal*. Vol.12:2, pp. 11-13.

X. Rodet. 1994. "Stability/instability of periodic solutions and chaos in physical models of musical instruments" in *Proc. International Computer Music Conference (ICMC)*. pp. 386-393.

X. Rodet; P. Depalle. 1992a. "A new additive synthesis method using inverse Fourier transform and spectral envelopes" in *Proc. International Computer Music Conference (ICMC)*. pp. 410-411.

X. Rodet; P. Depalle. 1992b. "A physical model of lips and trumpet" in *Proc. International Computer Music Conference (ICMC)*.

X. Rodet; Y. Potard; J.-B. Barrière. 1984. "The CHANT project: From the synthesis of the singing voice to synthesis in general" in *Computer Music Journal*. Vol.8:3, pp. 15-31.

X. Rodet; C. Vergez. 1996. "Physical models of trumpet-like instruments: Detailed behaviour and model improvements" in *Proc. International Computer Music Conference (ICMC)*. pp. 448-453.

G.P. Scavone. 1996. "Modeling and control of performance expression in digital waveguide models of woodwind instruments" in *Proc. International Computer Music Conference (ICMC)*. pp. 224-227.

X. Serra. 1989. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.* Ph.D. thesis.(Stanford University)

X. Serra; J.O. Smith. 1989. "Spectral Modeling Synthesis" in *Proc. International Computer Music Conference (ICMC)*. pp. 281-284.

Z. Settel; C. Lippe. 1994. "Real-time musical applications using FFT-based resynthesis" in *Proc. International Computer Music Conference (ICMC)*. pp. 338-343.

D.W.N. Sharp; R.L. While. 1993. "Determining the pitch period of speech using no multiplications" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.2, pp. 527-529.

J.O. Smith. 1987. "Music applications of digital waveguides" in *Department of Music Technical Report (CCRMA,Stanford)*.

J.O. Smith. 1991. "Waveguide simulation of non-cylindrical acoustic tubes" in *Proc. International Computer Music Conference (ICMC)*. pp. 304-307.

J.O. Smith. 1992. "Physical modeling using digital waveguides" in *Computer Music Journal*. Vol.16:4, pp. 74-91.

J.O. Smith. 1993. "Efficient synthesis of stringed musical instruments" in *Proc. International Computer Music Conference (ICMC)*.

L. Solbach; R. Wöhrmann. 1996. "Sound onset localization and partial tracking in Gaussian white noise" in *Proc. International Computer Music Conference (ICMC)*. pp. 324-327.

T. Stainsby. 1996. "A system for the separation of simultaneous musical audio signals" in *Proc. International Computer Music Conference (ICMC)*. pp. 75-78.

C. Tait; W. Findlay. 1996. "Wavelet analysis for onset detection" in *Proc. International Computer Music Conference (ICMC)*. pp. 500-503.

T. Todoroff. 1996. "A real-time analysis and resynthesis instrument for transformation of sounds in the frequency domain" in *Proc. International Computer Music Conference (ICMC)*. pp. 432-435.

M. Tohyama; R.H. Lyon; T. Koike. 1993. "Source waveform recovery in a reverberant space by cepstrum dereverberation" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol.1, pp. 157-160.

B. Truax. 1990. "Composing with real-time granular sound" in *Perspectives of New Music*. Vol.28:2, pp. 120-134.

B. Truax. 1994. "Discovering inner complexity: time-shifting and transposition with a real-time granulation technique" in *Computer Music Journal*. Vol.18:2, pp. 38-48.

S. Venkateshananda. 1984. *Yoga Vasishtha.* Publ. Albany: State University of New York Press (Original text by Sage Vasistha in Sanskrit; translated by Swami Venkateshananda.)

J. Ville. 1948. "Théorie et applications de la notion de signal analytique" in *Cables Transmission*. Vol.2A, pp. 61-74.

J.E. Vuillemin. 1994. "Fast Linear Hough Transform" in *IEEE International Conference on Application Specific Array Processors*. pp. 1-9.

E.P. Wigner. 1932. "On the quantum correction for thermodynamic equilibrium" in *Physics Review*. Vol.40, pp. 749-759.

E. Zwicker; U.T. Zwicker. 1991. "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system" in *Journal of the Audio Engineering Society (AES)*. Vol.39:3, pp. 115-126.