

Modular and Adaptive Control of Sound Processing

Douglas Van Nort



Music Technology Area
Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

January 2010

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

© 2010 Douglas Van Nort

Abstract

This dissertation presents research into the creation of systems for the control of sound synthesis and processing. The focus differs from much of the work related to digital musical instrument design, which has rightly concentrated on the physicality of the instrument and interface: sensor design, choice of controller, feedback to performer and so on. Often times a particular choice of sound processing is made, and the resultant parameters from the physical interface are conditioned and mapped to the available sound parameters in an exploratory fashion. The main goal of the work presented here is to demonstrate the importance of the space that lies between physical interface design and the choice of sound manipulation algorithm, and to present a new framework for instrument design that strongly considers this essential part of the design process. In particular, this research takes the viewpoint that instrument designs should be considered in a musical control context, and that both control and sound *dynamics* must be considered in tandem.

In order to achieve this holistic approach, the work presented in this dissertation assumes complementary points of view. Instrument design is first seen as a function of musical context, focusing on electroacoustic music and leading to a view on gesture that relates perceived musical intent to the dynamics of an instrumental system. The important design concept of mapping is then discussed from a theoretical and conceptual point of view, relating perceptual, systems and mathematically-oriented ways of examining the subject. This theoretical framework gives rise to a mapping design space, functional analysis of pertinent existing literature, implementations of mapping tools, instrumental control designs and several perceptual studies that explore the influence of mapping structure. Each of these reflect a high-level approach in which control structures are imposed on top of a high-dimensional space of control and sound synthesis parameters. In this view, desired gestural dynamics and sonic response are achieved through modular construction of mapping layers that are themselves subject to parametric control. Complementing this view of the design process, the work concludes with an approach in which the creation of gestural control/sound dynamics are considered in the low-level of the underlying sound model. The result is an adaptive system that is specialized to noise-based transformations that are particularly relevant in an electroacoustic music context.

Taken together, these different approaches to design and evaluation result in a unified framework for creation of an instrumental system. The key point is that this framework

addresses the influence that mapping structure and control dynamics have on the perceived feel of the instrument. Each of the results illustrate this using either top-down or bottom-up approaches that consider musical control context, thereby pointing to the greater potential for refined sonic articulation that can be had by combining them in the design process.

Abrégé

La présente dissertation expose une recherche sur la création de systèmes pour le contrôle de la synthèse et du traitement des sons. Les travaux portant sur la conception d'instruments de musique numérique se concentrent le plus souvent, à juste titre, sur la dimension physique de l'instrument et de l'interface: la conception de capteurs, le choix de contrôleurs, l'interaction avec le musicien, etc. Le plus souvent, on fait d'abord un choix particulier de traitement des sons pour ensuite prendre, en fonction d'expérimentations les décisions quant à quels paramètres de l'interface physique contrôleront quels paramètres sonores. Prenant le contre-pied de cette tendance, le but principal du présent travail est de démontrer l'importance de l'espace qui se situe entre la conception d'interface physique et le choix d'algorithmes de manipulation sonore, et de présenter un nouveau cadre pour la conception d'instruments qui prend mieux en considération cette part essentielle du processus de conception. Plus précisément, cette recherche adopte le point de vue selon lequel la conception d'instruments devrait partir d'un contexte de contrôle musical, et que le contrôle et les dynamiques sonores devraient être considérées en tandem.

Pour réaliser cette approche holistique, les différents travaux présentés dans cette dissertation adoptent des points de vue complémentaires. À partir du cas de la musique électroacoustique, la conception d'instrument est d'abord vue comme fonction du contexte musical. Cette idée permet une compréhension des gestes qui lient l'intention musicale perçue aux dynamiques d'un système instrumental. Le concept important de "mapping" est ensuite examiné d'un point de vue théorique et conceptuel, en mettant en relation des approches perceptives, systémiques et mathématiques de ce sujet. Seront élaborés à partir de ce cadre un espace de conception de mapping, une analyse fonctionnelle de la littérature pertinente, l'implémentation d'outils de mapping, des designs de contrôle d'instruments, et plusieurs études perceptives qui explorent l'influence de la structure de mapping. Chacune de ces études reflète une perspective de haut niveau dans laquelle les structures de contrôle sont superposées à un espace de contrôle et de paramètres de synthèse sonore de grande dimension. Selon cette perspective, les dynamiques gestuelles et les réponses sonores souhaitées sont obtenues à partir d'une construction modulaire de niveaux de mapping qui sont elles-mêmes sujettes à un contrôle paramétrique. En complément de cette conception du processus d'élaboration, ce travail se conclut avec une approche selon laquelle la création de contrôles gestuels et de dynamiques sonores sont considérés au niveau le plus bas du

modèle sonore sous-jacent. En résulte un système adaptatif spécialisé pour les transformations fondées sur le bruit, particulièrement pertinentes dans le contexte de la musique électroacoustique.

Ensemble, ces différentes approches de conception et d'évaluation résultent en un cadre unifié pour la création d'un système instrumental. Le point essentiel est que ce cadre répond à l'influence des structures de mapping et des dynamiques de contrôle sur la sensation perçue de l'instrument. Chacun des résultats illustrent cela par une approche descendante ou ascendante qui considère le contexte de contrôle musical, illustrant ainsi le plus grande potentiel d'articulation sonore qui peut être obtenu en les combinant dans le processus de conception.

Acknowledgments

There are a number of individuals who inspired, supported or otherwise helped to shape this work in various ways. This certainly includes my advisors Philippe Depalle and Marcelo Wanderley, who helped guide me into new avenues of music technology research, and without whom this project couldn't possibly be as interdisciplinary as it has turned out to be. I would also like to thank the exceptionally bright and creative colleagues and staff from the Music Technology Area at McGill, particularly from the IDMIL and SPCL labs, for inspiration and assistance that has helped shape my musical views as well as technical skill-set. Many thanks to Dr. Brandon Gavett for very helpful feedback on my user studies. Thanks as well to the Centre for Interdisciplinary Research in Music Media and Technology for facilitating this research through its excellent staff, labs and through project and travel funding. Similarly I owe a great deal of gratitude to Richard H. Tomlinson and the Tomlinson fellowship program for supporting this research at its outset, and the McGill Majors program for continuing this support. Thank you to my family for being supportive in spite of the fact that you couldn't care less about music technology research. Finally – at the risk of ending with a cliché – far and away the deepest appreciation and gratitude is reserved to Stacy Denton, who has inspired and supported me not only in this project, but in every facet of existence. This work certainly could not have been completed without her.

Contents

1	Introduction	1
1.1	Mapping in Digital Music	4
1.1.1	Articulating Boundaries	6
1.1.2	Modularity and Separability	6
1.1.3	Timbre and Geometric Space	7
1.2	Control and Articulation in Sound Synthesis	8
1.2.1	Keyboard Paradigm	9
1.2.2	Early Electric Sound	9
1.2.3	Modular Synthesis	10
1.2.4	Classic Vocoder	10
1.2.5	Digital Synthesis	11
1.2.6	Post-Digital-Synthesis	12
1.3	Originality and Contributions	18
2	From Control to Sound Gestures	21
2.1	Control Gesture	23
2.2	Sonic Gesture	25
2.2.1	From Sound Objects to Sonic Gestures	27
2.2.2	Perceptual Criteria of Sonic Gesture	30
2.2.3	Form/Matter and Gesture/Texture: Mapping Considerations and Structure-Bearing Principles	32
2.2.4	Towards Sonic Gesture Features	34
2.2.5	From Sonic Gesture Back to Control Gesture	79
2.3	Chapter 2 Summary	84

3	Mapping Theory	87
3.1	What is Mapping?	88
3.1.1	Musical Control Context	89
3.1.2	Perspectives on Mapping	91
3.2	Boundary of Signal Conditioning, Mapping and Control Gesture Design . .	96
3.3	Functional View on Mapping	97
3.3.1	Geometric and Topological Framework	99
3.3.2	Multi-Layered Approach	108
3.3.3	Spatial vs. Temporal View of Control	110
3.4	Towards a Toolbox of Mapping Functions	112
3.4.1	Parallel Developments	113
3.4.2	Previous Work and My Extensions	115
3.4.3	Mapping Design Space	119
3.5	LoM Toolbox	121
3.6	Chapter 3 Summary	126
4	Mapping: Perception, Practice and Gestural Dynamics	128
4.1	Influence of Mapping Trajectory on Perception of Control	129
4.1.1	Preliminaries	131
4.1.2	Qualitative Assessment: Expressivity and Ease-of-Use	132
4.1.3	Quantitative Analysis of the Effect of Visual Feedback	133
4.1.4	Preliminary Study Conclusions	135
4.2.3	Navigation and Sonic Gesture Target Acquisition	159
4.3	Modular and Parametric Control Structures	178
4.3.1	Transformation of Resonant Models	179
4.3.2	Multi-Layered Approach	182
4.3.3	Granular Scrubbing	184
4.4	Signal Conditional, Gesture Conditioning	193
4.5	Case Studies: Fabric-based Interaction Design	195
4.5.1	The Tapestry	196
4.5.2	The Blanket	201
4.5.3	Tapestry and Blanket Projects: Reflection	206
4.6	Chapter 4 Summary	208

5	Sound Modeling and Control	210
5.1	Defining Control/Sound Gesture Dynamics, Considering Sonic Context . .	210
5.2	State-Space Analysis/Synthesis	212
5.2.1	The Phase Vocoder	212
5.2.2	State-Space Phase Vocoder	213
5.2.3	Example Effects	217
5.2.4	Beyond SSSPV: From Effect to Transformation	221
5.3	Introduction of Recursive, Infinite Length Windows	221
5.4	Kalman Filter-Based Phase Vocoder	223
5.5	Chandrasekhar Implementation	227
5.6	Additive Layer and Higher-Level Architecture	229
5.7	Sound Transformations	233
5.7.1	Textural Effects	234
5.7.2	Spectral Effects	241
5.7.3	Cross-Synthesis	242
5.8	Adaptive Control of Sound Transformations	244
5.8.1	Application 1: Control of Modulated Source-Filter Model	247
5.8.2	Application 2: Control Dynamics for Partial/Residual Modification	249
5.9	Chapter 5 Summary	259
6	Conclusion	261
6.1	Mapping, Gestural Dynamics and Musical Control Context	261
6.2	Future Work	263

List of Figures

2.1	Spectrogram for test sound x_{89}	36
2.2	Plot of Spectral Centroid for example sound objects.	40
2.3	Plot of TFS measure for example sound objects	42
2.4	TFS for dust noise example ($N=8192$, $\text{hop}=4096$, $F_s = 44100$)	44
2.5	TFS for all MUMS acoustic examples ($N=8192$, $\text{hop}=4096$, $F_s = 44100$)	46
2.6	Spectrogram of IMF3 from x_{89} (0-1800 Hz), FFT window size = 4096, 50 trials, with (a) Max iterations = 10 and (b) Max iterations = 20	51
2.7	Spectrogram of IMF3 from x_{89} for $f=0-6\text{kHz}$, FFT window size = 4096, 50 trials, Max iterations = 10	52
2.8	Spectrogram of IMF4 from x_{89} for $f=0-4\text{kHz}$, FFT window size = 4096, 50 trials, Max iterations = 10	53
2.9	Spectrogram of signal x_{89} from 100 to 500Hz.	53
2.10	Magnitude of bin located near 150Hz over time as extracted from spectrogram of signal x_{89} . STFT magnitude is non-normalized, with $N=4096$	54
2.11	Magnitude of bin located at 161 Hz over time for IMF3 (top) IMF4 (bottom) of signal x_{89} . STFT magnitude is non-normalized, with $N=4096$	55
2.12	Magnitude of bin located near 150Hz over time for signal x_{89} (solid) and extracted IMF4 (dotted). STFT magnitude is non-normalized, with $N=4096$	57
2.13	Roughness for seven lowest IMFs for signal x_{92} . Note the saliency of the spectral grain signal IMF3.	57
2.14	Spectrogram of x_{92}	58
2.15	Extracted spectral and iteration grains for signal x_{92} , captured respectively by IMF3 (top) and IMF4 (bottom) of the EEMD analysis with max sifting = 10	60

2.16 Spectrogram of transient grain for x_{92} ($f = 0$ to 12.5 kHz, $N = 4096$) . . .	61
2.17 TFS measure for identified transient, spectral and iteration grains for signal x_{92} . The transient grain TFS is much higher, while the iteration grain TFS is nearly zero.	62
2.18 Relative Grain Spectral Spread for all perceptually relevant grain signals. .	63
2.19 Relative Grain Spectral Spread: Expanded view of spectral (top) and iteration (bottom) grains.	64
2.20 Relative Grain Spectral Placement for transient (top), spectral (middle) and iteration (bottom) grain signals	64
2.21 Relative Grain Spectral Placement: Expanded view of spectral (top) and iteration (bottom) grains	65
2.22 Relative grain weight for transient grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.	66
2.23 Relative grain weight for spectral grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.	67
2.24 Relative grain weight for iteration grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.	68
2.25 RMS envelope for signal x_{89} (top) the EEMD-extracted trend (middle) and residual (bottom).	70
2.26 Amplitude Envelope for all example sounds related to the the example Solfege du L'Object Sonore ($N = 8192$, Hop = 4096, $F_s = 44.1\text{kHz}$)	72
2.27 Ratio of Maximum Amplitude to Total Length, Extracted from RMS Envelope	74
2.28 Ratio of Temporal Centroid to Total Length, Extracted from RMS Envelope	75
2.29 Temporal Flatness for all Sound Examples, Extracted from RMS Envelope	76
2.30 Temporal Increase (a) and Decrease (b) for all example sound objects . . .	78
2.31 Transient grain and next-largest IMFS for a double bass playing C1 bowed (top) muted (middle) and martelé (bottom). Temporal Fine Structure (TFS) window size was 8192, hop 4096 and f_s of 44.1k.	80
2.32 Possible control structures for different sonic gestures types (attack, attack/decay and graduated continuant 1,2,3), including control type for all portions of a sound's life. Also included are the relevant form/matter features that are perceptually relevant and so may be affected (dynamic profile, instantaneous matter, dynamic matter, motion, grain and dynamic grain).	85

3.1	A generalized systems-oriented view of mapping from control to sound parameters.	93
3.2	A generalized functional view of mapping from control to sound parameters.	94
3.3	(a) Direct embedding of 2D control surface into 3D sound synthesis space, including pointwise mapping. (b) Interrelation of parameters due to direct embedding.	104
3.4	Various views of a mixture embedding between two dimensional control space to three dimensional sound output space. The continuous input trajectory, sampled every 20 ms, travels around the borders of the rectangular input space (solid, black) and then across both diagonals (dashed, blue) to provide the given output trajectory. The output is continuous, but the space is split and passes through itself, thus the mapping is not a homeomorphism. Dimensions are indexed along the axes as a point of reference.	106
3.5	A projection from regions of two dimensional control space onto a one dimensional intermediate space, which determines the ultimate topology when mapped into a three dimensional sound synthesis parameter space (with grey “shadow” on X-Y plane).	110
3.6	Table for editing text values or inputting new output sound states to lom.si object.	123
3.7	In this example a two dimensional control space (Wacom position) is mapped to a nine dimensional space of granular synthesis parameters. The control data is sent to lom.si, which interpolates this input vector and outputs a list of sound parameters. A second output from lom.si sends triangle information to the lom.jit.si abstraction which then draws the parameter space using the native OpenGL rendering tools (bottom window). As with the implementation from [1], nodes of the triangular complex refer to stored parameter sets, and the input point lets the user know where they are in parameter space. This visual feedback is most useful for initial learning of an instrument.	124
4.1	Tablet controller as surface embedded in a three dimensional FM synthesis space.	131

4.2	Visualization of trajectory across a mapping surface (a) Multilinear 3D trajectory and (c) 2D Projection. (b) RST 3D trajectory and (d) 2D Projection.	134
4.5	Mean Normalized Ratings for Sound Preference.	146
4.6	Mean Normalized Ratings for Controllability.	147
4.7	Mean Normalized Ratings for Transparency.	148
4.8	Sound preference rating mean for Composers of EA Music (top) EA Music Listeners (middle) and Subjects with Musical Training (bottom).	151
4.9	Controllability rating mean for Composers of EA Music (top) EA Music Listeners (middle) and Subjects with Musical Training (bottom).	152
4.10	Sound preference rating mean for the low EA-experience / musical training group (top) and the high EA-experience / musical training group (bottom).	153
4.11	Controllability rating mean for the low EA-experience / high musical training group (top) and the high EA-experience / low musical training group (bottom).	155
4.12	Controllability rating variance for the low EA-experience / high musical training group (top) and the high EA-experience / low musical training group (bottom).	156
4.13	Spectrotemporal Profile of GC3 Gesture for MI Granular Instrument. The overall envelope (top) increases in amplitude while the Spectral Smoothness (middle) decreases as the sound's spectrum becomes continually more jagged. Meanwhile the Relative Grain Weight (bottom) decreases at a similar rate as more widespread spectral energy overrides the fine-grain fluctuations.	163
4.14	Average distance to target for all subjects in regards to Amplitude Envelope (top) and Spectral Smoothness (bottom).	167
4.15	Average distance to target for all subjects in regards to (a) Spectral Centroid and (b) Spectral Flatness.	168
4.16	Average distance to target for all subjects in regards to Relative Grain Spectral Flatness (top) and Relative Grain Spectral Smoothness (bottom). MI mapping structure leads to better performance for both instrument sets.	170
4.17	Average distance to target for all subjects in regards to TFS (top) and RGW (bottom). SI mapping structure leads to better performance for both instrument sets.	171

4.18	Average distance to target for all subjects in regards to Relative Grain Spectral Centroid.	172
4.19	Average distance to target for Spectral Centroid (top) Spectral Flatness (middle) and Spectral Smoothness (bottom) for the coherent subgroup formed across AE, SC, SS, SF and RGW contour tracing.	175
4.20	Average distance to target for Amplitude Envelope (top) RGW (middle) and TFS (bottom) for the coherent subgroup formed across AE, SC, SS, SF and RGW contour tracing.	176
4.21	Different spatial layout of sound model presets on tablet surface, and their respective approaches to model interpolation for example 4.3.1: movement or insertion of new sound models, creating new localized regions (left) vs. varying weightings for each fixed sound model, creating a different geometry for the model interpolation (right). The former allows for more defined sonic gestures while the latter makes it easier to define a global sound quality within a given region.	181
4.22	Control structure for example 4.3.2: two concurrent mappings control the possible sound combinations (control-to-sound) as well as the control/sound temporal response (control response self-map). Combined they determine the resultant sonic gestures.	182
4.23	Control structure for example 4.3.2: the control response self-mapping acts as meta-control, affecting the responsiveness in parallel with the mapping from control to high-dimensional sound parameter space.	185
4.24	Overlap-and-min of parallel control spaces, in order to construct one-dimensional topology in sound space.	187
4.25	Control structure is dynamically altered as input velocity influences the intermediate topology space as well as shape of mapping-to-sound space. . .	189
4.26	Example mapping structure in which scrubbing is based on temporality of gesture. Speed of input action determines granular processing through SI mapping, and response and dynamics of control gesture are conditioned by another filtering layer that is itself gesturally-controlled by position or orientation.	192
4.27	Velocity-to-response mapping. The actual path in this space is determined by control of RST mapping's smoothness and tension parameters.	192

4.28	Feedback adaptive mapping structure which changes mapping layer based on sonic gestural output.	194
4.29	Sensing bird from the Tapestry, woven with conductive thread.	197
4.30	Users interacting with the tapestry instrument.	197
4.31	The Blanket instrument (a) sans human interaction (b) collective play along the interior. The colored lights projected from the top are for theatrical effect.	203
4.32	Different Approaches to the Blanket topology include (a) the rectangular interaction between columns and (b) the circular interaction between boundary and center.	203
4.33	Sensor data from Blanket (left) is “tapped” as time series (center) and autocorrelation (right top, function of lag time) and cross correlation (right bottom, function of spatial lag and time) series are extracted in order to provide information about regularity and phase relationships, respectively. Above data results from flapping the Blanket edge closest to the chosen column.	205
5.1	Log FFT plot for stochastic state-space processed sinusoid. Gaussian noise added to analysis dynamics matrix in outer triangular portion (a) and to diagonal (b).	218
5.2	Spectrogram of sinusoid affected by noise propagated through analysis state matrix at one column alone (a) and coupled between a column/row pair (b). Sampling frequency = 11.025 kHz.	220
5.3	Model Architecture for Kalman-based Additive and Recursive Exponential STFT Models, shown here in parallel. Amplitude or phase values may be extracted for processing during the respective transformation stages.	231
5.4	Example piano note for KF-based analysis ($f_s = 22050$ Hz)	236
5.5	Observation noise after analysis	237
5.6	Process noise related to first several partials after analysis	237
5.7	Spectrogram of input piano attack	239
5.8	Spectrogram of piano attack after SOLAF time-stretching of each state noise process, with stretch factor proportional to temporal support. Window size is 256 with overlap of 128.	240

5.9	Spectrogram of piano attack after SOLAF time-stretching of each state noise process, with stretch factor randomized with mean factor as well as window size proportional to frequency of related partial. Overlap is 50% in each case.	240
5.10	Spectrogram showing different partial evolution for sound example from previous section, after altering state matrix window decay parameter.	243
5.11	Spectrogram showing a frequency-dependent delay and time smearing after altering state matrix frequency values.	243
5.12	The mapping structure for the given instrument design. X-Y tablet values feed the physically-inspired dynamics of the leaky integrator and mass-spring-damper systems, respectively. As with the similar example depicted in figure 4.26, time (of contact) is an implicit input that controls the “scrubbing”. This control is outside the model and so is not discussed in this chapter, but note that it may control either sound input u or the model parameters by way of some time-scaling algorithm such as a SOLA technique. Finally, the analyzed sound u is an input to the analysis process, which is then affected by the control dynamics before being resynthesized into output \hat{u}	254
5.13	Nonlinear state-space system for controlling time/frequency model. Analyzed input sound and control values are observed, while linear dynamical systems for control/sound are combined into nonlinear control-sound dynamics model, augmented by an EKF. This dynamics model predicts the control/sound state value, and this is re-synthesized to produce new sound output.	258

List of Tables

2.1	Morphological Qualities and their Corresponding Sound Features	71
3.1	Mapping strategies and certain relevant properties	121

List of Acronyms

ACF Autocorrelation Function

AD Attack-Decay

ADSR Attack, Decay, Sustain, Release

AE Amplitude Envelope

AM Amplitude Modulation

ARMA Autoregressive Moving Average

CB Colorblobs

CEM Continuously Excited Middle

DFT Discrete Fourier Transform

DSP Digital Signal Processing

EA Electroacoustic

EKF Extended Kalman Filter

EMD Empirical Mode Decomposition

EEMD Ensemble Empirical Mode Decomposition

FFT Fast Fourier Transform

FM Frequency Modulation

GC	Graduated Continuant
GIS	Geographic Information System
GM	Gaussian Mixture
HB	Hyperbolic
HCI	Human-Computer Interaction
HT	Harmonic Timbre
IMF	Intrinsic Mode Function
KF	Kalman Filter
LFO	Low-frequency Oscillator
LoM	Library of Maps (toolbox)
LPC	Linear Predictive Coding
MA	Moving Average
MI	Multilinear Interpolation
MnM	Music is not Mapping (library)
NC	(general) Nonlinear Curve
NEM	Non-Excited Middle
NN	Natural Neighbor
PL	Piecewise Linear
RESTFT	Recursive Exponential Short-Time Fourier Transform
RGW	Relative Grain Weight
RGsc	Relative Grain Spectral Centroid
RGsf	Relative Grain Spectral Flatness

RGss Relative Grain Spectral Spread

RMS Root Mean Square

RST Regularized Spline with Tension

SC Spectral Centroid

SF Spectral Flatness

SI Simplicial Interpolation

SS Spectral Spread

SSSPV Stochastic State-Space Phase Vocoder

STFT Short-Time Fourier Transform

TFS Temporal Fine Structure

TPS Thin Plate Spline

WIMP Windows, Icons, Menus, Pointers

List of Publications

The following are relevant publications whose subject and writing appear in this dissertation, with page numbers noted as appropriate:

1. Doug Van Nort. Instrumental Listening: sonic gesture as design principle. Organised Sound 14(2):177-187, August 2009. Appears within pages 21-71.
2. Doug Van Nort, David Gauthier, Sha Xin Wei and Marcelo Wanderley, Extraction of Gestural Meaning from a Fabric-Based Controller, in Proc. of the International Computer Music Conference 2007 (ICMC-07), Copenhagen, Denmark, August 2007. Appears within pages 201-207.
3. Doug Van Nort and Marcelo Wanderley, Control Strategies for Navigation of Complex Sonic Spaces, in Proc. of the International Conference on New Interfaces for Musical Expression 2007 (NIME-07), New York, NY, June 2007. Appears within pages 179-185.
4. Doug Van Nort and Marcelo Wanderley, The LoM Mapping Toolbox for Max/Msp/Jitter, in Proc. of the 2006 International Computer Music Conference (ICMC 06), New Orleans, LA, November, 2006. Appears within pages 121-126.
5. Doug Van Nort and Philippe Depalle, A Stochastic State-Space Phase Vocoder for Synthesis of Roughness, in Proc. of the 2006 International Conference on Digital Audio Effects (DAFx 06), Montreal, QC, September, 2006. Appears within pages 212-220.
6. Doug Van Nort and Marcelo Wanderley, Exploring the Effect of Mapping Trajectories on Musical Performance, in Proc. of the International Conference of Sound and Music Computing (SMC 06), Marseille, France, May 18-20, 2006. Appears within pages 129-136.

7. Doug Van Nort, Marcelo M. Wanderley and Philippe Depalle. On the Choice of Mappings based on Geometric Properties. in Proc. of the 2004 International Conference on New Interfaces for Musical Expression (NIME 04), Hamamatsu, Japan, June 3-5, 2004. Appears within pages 98-108.

Chapter 1

Introduction

The scope of this research is to examine, assess and create strategies for controlling perceptually relevant timbral and textural aspects of digital sound synthesis in a way that is both modular and adaptive. The techniques implemented and created in this dissertation were chosen because of their ability to be combined, to complement one another and to be tuned to changing performance contexts. The term “performance” is a bit of an extended usage here, as the application of interest is equally geared towards live performance with digital musical instruments and studio-based composition work requiring real-time digital sound processing. Therefore the issue of communicating expressivity or gestural information from performer to audience is not of concern, but rather the focus is squarely on the sonic expressive potential of an instrumental system and what has often been called its degree of “intimacy”.

At this point, there is already a rich history of artistic experimentation as well as formal evaluation of the use of electrical and electronic technology in the direct production and transformation of music, dating back to the beginning of the previous century at least. In a more scientific and technological vein, a great deal of work has gone into the development of devices and techniques for sound synthesis, analysis and transformation, primarily relying on results from telecommunications (resulting in various signal models) or applying research in acoustics (resulting in physical models). As a result of this work many high-quality, real-time tools have become available to musicians and sound artists working with technology. Thus, while there is still much work to be done here, I have decided not to focus on new sound methods per se, but

rather the confluence of sound transformation/processing techniques and the means by which these are controlled. *It is a major stance of this work that re-considering control in the context of underlying sound models employed is a direct path to subtle and expressive performance control.* I would consider this a burgeoning and new area of computer music and music technology, as the ever-increasing power of the personal computer has only (relatively) recently allowed for the exploration of complex control methods to be applied to the more advanced techniques for sound processing. In the same way that the ability to analyze and re-synthesize musical sounds naturally led to discussions on the definition and meaning of musical timbre (and related elements), this ability for real-time control forces one to re-examine what precisely is meant by “musical instrument” in the context of computer-based music. Essential questions are raised when one considers that the coupling between a performer’s physical actions and the sonic result – an inherent feature of any acoustic sound-making object – is no longer an intrinsic quality of a musical instrument in the digital realm. As a result, a second major issue that I address in this work, and that I find to be of central importance to the design of new instruments, is the way in which we associate human input with sound output. This act of assigning control to sound parameters is referred to in the literature as *mapping*, and throughout this dissertation I preserve this language while extending the definition to other parallel meanings that I feel mapping subsumes, including one’s perception of intentionality. I will further explore this definition in regards to the role of mapping in defining instrumental dynamics.

Throughout the course of this dissertation I will cover three main areas that deal with control aspects of digital musical instrument systems. This first will be to consider perceptual aspects of controlling sound – both in the sense of action-to-sound as well as the perception of causality and control within sound. I will situate the musical context for this work in an electroacoustic tradition by focusing on Schaefferian theory to consider these perceptual elements. This part of the work, found in chapter 2, serves both as the underlying philosophical approach to music and sound throughout the dissertation as well as a framework for deriving sound analysis techniques that will be applied in the perceptual studies of chapter 4. The second main body of work deals with the issue of mapping, its role and definition. The focus of chapter 3 is to apply theoretical rigor to the subject, providing a functional analysis of different mapping

approaches using a geometric interpretation. These properties give rise to a set of design criteria that one may consider when creating a system to drive multi-parametric sound synthesis and processing¹ techniques in a complex fashion. Along this line, real-time implementations of certain mapping techniques have been written, with this work discussed at the end of chapter 3. Following this, chapter 4 presents a set of perceptual user studies aimed at establishing an initial formal examination of the interplay between mapping structure and perceptual control structure. Rounding out this chapter will be several musical control structures that build on the instruments of the user study, and finally two installation-based interactive projects that cast mapping in a different light: as an issue of gestural conditioning, extending the definition and contrasting with the view taken in the design of multi-parametric instrumental control. The third axis of this research, discussed in chapter 5, is entirely complementary to the geometric and spatial view of chapters 3 and 4 in that it focuses entirely on the temporal aspect of control. The interest will be in expressing the confluence of sound and control dynamics, illustrating the advantage of including a consideration of control gestures in the process of creating a sound model at a low level. The techniques applied to this end are borrowed from the control theory literature as applied in this signal processing context, framing the musical instrument – defined in terms of control/sound parameters and their dynamical evolution in state-space – as a “plant” that is steered by the exogenous disturbance of human action. The system can adapt to the disturbance based on a prediction of instrumental dynamics. The particular application of choice for this work will be time-frequency models, developed into several different control structures and resulting in a novel approach to instrument design.

The sum total of this work presents a unified approach to designing an instrumental system in a digital music context. The emphasis has been placed on methods that may be used modularly – in the sense of combining control structures, controllers and sound processing – and adaptively in the sense of being tunable for different performance contexts or different “disturbances” (i.e. control/sound signal types). Modularity has a long history in electronic music, and I seek to build upon this through extensions to real-time control; adaptation within the varying “noise” of performance context (of a

¹ From this point forward, I will refer to sound effects processing as well as analysis, synthesis and transformations collectively as sound processing, both for compactness and to highlight that these all refer to signal processing techniques applied to sound signals.

room, of the instrument itself, etc.) is done intuitively by an acoustic performer, yet is something that needs to be modeled explicitly in a digital world.

The sound-focused view of chapter 2 provides the scope of the work and definition of “perceptual relevance”. Within this context, two methodologies are outlined, where the geometric-based mapping approach and the state-space dynamical systems approach can be seen to stand on opposite ends of the spectrum of musical control. One method seeks to impose a control “surface” as a mediating layer between musician and sound production technique – a top down approach, while the other seeks to create a representation that internalizes a description of control over time – a bottom up approach. The former places no presupposition on the sound that is being controlled, but rather we seek to find appropriate mappings (i.e., those that “sound good”) by exploring within the totality of control-sound space, constraining this through mapping geometries and an understanding of perceptual relevance. In the latter approach, the concern is to build a model in which the control dynamics are partially or completely described and the systems state is updated by virtue of constraints that one might consider as the mapping in this context. These two approaches are complementary in the sense that part of the problem lies in choosing the proper sound parameters that we would like to control and constraining them to certain behaviors (addressed by the former) while we also need to consider the chosen model for sound production, its potential states and dynamics as well as how we might effectively control them (the latter approach). Together they serve to enhance ones ability to translate the proper physical gesture into a sonic one. More generally speaking, the idea here is to strike a balance between an approach determined by theory and an empirical one, allowing one to construct physical dynamics yet be able to tune this through experimentation. Before describing the work of these three core areas, I will present some of the history of ideas that have set the stage for this research.

1.1 Mapping in Digital Music

Generally speaking, the notion of associating action to sound can be considered as old as music itself: deciding where one focuses an action to produce musically-relevant sound can be considered as either a compositional (if planning a musical sequence of

actions) or parametric (if constructing a sound-making device) type of mapping decision. In this sense mapping arises in all instrument design: deciding that a non-dominant hand selection would map to pitch choice while bow angle/pressure/speed would control loudness and timbral attributes is a parametric mapping decision for bowed string instruments. However, the fact remains that exciting a string and pressing down at differing lengths will always change the resultant pitch that is perceived: it is a physical constant. Moving into the realm of electronic music, the theremin has a very simply action-to-sound association: proximity to antennas controls fundamental frequency and amplitude of an analog oscillator. This decision too can be seen as a mapping, and yet the principle behind the system – capacitance sensing that drives the rate and amplitude of said oscillator – is constrained by the physics of the circuitry so that with such a circuit design one must always affect these parameters in an up/down fashion. At the same time, in a digital context all of the action-sound decisions are arbitrary for a given system, from manner of conditioning gestural action to parameter mapping, synthesis type, and so on. Therefore, while these other examples can be seen to have some sort of mapping, I find it most useful to reserve this term for the digital case in which all elements of action-to-sound transduction must be imagined and constructed. As this is at the heart of what makes mapping interesting as a topic of discussion, it is where I will focus for the remainder of this work.

In the digital realm it is only a relatively recent phenomenon that high-quality sound processing and synthesis can be produced in real-time. Therefore while the literature on digital audio for music is extensive, discussions of how to map out real-time control of digital sound is still very limited by comparison. In the case of instrument-like controllers [2] on the input side and physical models on the sound synthesis side, mapping is much more strongly constrained. Even still, Rován et al. [3] found significant effects by varying a parameter mapping from the instrument-like Yamaha WX-7 to additive synthesis. They found that, in order to achieve a sufficient level of musical expressivity, the mapping between controller (embouchure, breath pressure, key) and additive synthesis variables needed to be interrelated in a complex fashion. The class of mappings that were found to be most expressive were dubbed *convergent* mappings, which map several input variables to a single sound synthesis variable. The

authors state that *divergent* mappings – when one input variable controls several sound variables – may be immediately expressive but over time become limited in that they do not allow for “internal features” of sound output to be controlled in a subtle or direct manner. These two types are contrasted with one-to-one parameter mappings, found to be the least expressive.

1.1.1 Articulating Boundaries

Similar views on mapping are discussed in [4], where Hunt and Kirk adopt the language of one-to-one, one-to-many and many-to-one to talk about mapping parameter complexity. They reported on a user study, conducted by the first author, which presented participants with different mappings of varying degree of interdependence. They found that the mapping alone exerted influence on the perception of control, with the non one-to-one mappings that required user’s energy for sound output and more than one limb were more engaging to users. A follow-up study that normalized the type of control device was presented in [5], and similar results were reported. As a follow-up to these results as well as the discussion of mapping from [6], Hunt et al. [7] present one of the first attempts at a formal study of mapping. One of their primary distinctions lies between parameter mappings that are explicitly defined and those that are created implicitly, through some sort of “internal adaptations of the system”. The latter type has primarily focused on the use of neural networks to learn input/output mappings [8][9] [10]; at least one other study has applied genetic algorithms to this end [11], while Bevilacqua et al. [12] treat the mapping as a multi-dimensional linear input/output problem, and use singular value decomposition to find the “best-fit” mapping in a linear regression sense.

1.1.2 Modularity and Separability

Another issue that was implicitly addressed by [3] and which led into the work presented in [13] is the notion of having multiple *layers* of mappings, in order to address specificities of the control device and sound processing algorithm separately. In the former the intermediate layer was a two-dimensional grid-based layout of additive instrument parameters, organized according to pitch and dynamics. Thus the first

mapping is from control parameters into this space of (dynamics, pitch) perceptual parameters, while the second is from this “perceptual space” into the underlying additive model parameters, extracted from analysis of actual instruments. Therefore this perceptual space is tied to the underlying sound parameters in that it is defined by the nature of the sound analysis. The work of [13] seeks to promote modularity between the control and sound synthesis sides by introducing the concept of an intermediate “abstract parameter space” wherein the control-side or sound-side mapping layers could be changed independently of one another if respective device or algorithm were replaced. They present this as an approach to “composed instruments”, an idea that is further explored in this dissertation. Later work that articulated this same idea with a conceptual discussion of “gestural perceptual” and “sound perceptual” spaces can be found in [14].

1.1.3 Timbre and Geometric Space

This notion of “spaces” arises in the consideration of mapping in two distinct ways:

1. When one considers each (control or sound) parameter from a set as being ordered (e.g. less or more amplitude) and being independent from one another in regards to controllability.
2. When a notion of distance arises between perceived events, such as when a given sound is perceived to be close to or far from another.

This second interpretation forms the basis for the notion of a *timbre space*, wherein the idea is to represent sounds in regards to this multidimensional attribute by placing them in a geometric space relative to one another, based on subjects’ pairwise similarity judgements. Key early studies to this end can be found in [15] and [16]. In the same timeframe, researchers were considering the perceptual influence of *acting on* a timbre space, such as the influence that modifying a sound’s spectra would have on its position in timbre space [17]. In [18] Wessel directly articulated the idea of timbre space itself as a *control structure*, wherein musicians could navigate within this. Later studies then began to provide a link between perceptual timbral correlates and underlying signal attributes, notably in [19], providing more insight into those characteristics of sound that one might attempt to influence in order to effectively

move about a given timbre space, moving this idea closer to the parametric notion of space presented in item 1 above. This inversion on the timbre space-as-representation - approaching it as a control structure - gives rise to mapping questions such as how best to continuously move within a multidimensional space, and how to interpolate between known sonic entities or states. This was approached in the aforementioned works of [3] and [13], the former by interpolating between additive model parameters and in the latter through interpolating user-defined parameters for steady-state sounds, laid out within an N-dimensional grid. Other works have taken this spatial, geometric approach to continuous and multi-parametric control, notably [20], [21], [22], [1], [23] and [24], which will all be discussed in more detail later in this dissertation. While studies such as [19] seek to move from a perceptual space back towards controllable sound features, one can say that these works all begin with an underlying parameter set and work towards perceptual relevancy through defining personalized control space. A large part of chapters 3 and 4 will be devoted to building on this trajectory of research by unifying and contextualizing the approach, extending this by way of new mapping techniques and suggesting a new methodology for testing the perceptual relevance of a space itself as well as its control.

1.2 Control and Articulation in Sound Synthesis

The history of sound synthesis is quite rich, with preludes such as Hermann von Helmholtz's experimental oscillators in the mid-to-late 1800's and Thaddeus Cahill's Telharmonium from the turn of the 20th century leading into the RCA Mark II, the analog synthesis from the Cologne studio, the modular synthesizers of Moog and Buchla, Chowning's FM synthesis, the Karplus-Strong plucked string algorithm and waveguide synthesis from Stanford, digital synthesis from Mathews at Bell labs and Risset's analysis/synthesis, the musical use of the phase vocoder, the extensions to analysis/synthesis and the phase vocoder at IRCAM in Paris, and so on. Rather than cover the many facets of these results and the many more that coincided, the reader is directed to sources such as [25], [26] and [27]. Instead, my brief overview will focus on the manner in which *control* was conceived of and embedded in different sound synthesis and processing contexts, as this notion is crucial to my work presented here.

1.2.1 Keyboard Paradigm

The keyboard is one of the most basic and essential tools for musical learning, as it provides a visual representation of such concepts of high vs. low notes, scales, chords, and it makes learning temporal sequences much easier through visual feedback. As such it is a gold standard amongst musical instruments, and it is no surprise that this method of input has dominated in design of music synthesizers. However, from its beginning sound synthesis has focused on flexible and arbitrary sound – which can be limited by the percussive nature of the piano keyboard – leading researchers to separately consider methods of modulatory and continuous control. This is directly related to the issue of how to allow for *expressive* control that helps to convey musical emotion and intent. This term is quite subjective, but key components of expressivity in western musical performance include vibrato, tremolo, legato, glissandi, trills, instrument-specific timbral modulations and timing variations between notes.

1.2.2 Early Electric Sound

Examples of early electronic instruments that addressed timbral variety while giving some element of “added expressive” control include the ondes martenot and the traultonium, which used a wire controller in order to glide between notes or to create vibrato. As early as the late 30’s with the Hammond Novachord synthesizer/organ, designers understood the importance of amplitude envelope in overall timbral output, building in out-of-time controls (knobs) for the attack, decay, sustain and release functions (or, ADSR envelope) that occurred after a key was pressed and released. This can be seen as one of the earlier embeddings of control into a synthesis scheme: by adding the element of note shaping – which would acoustically be a product of breath/finger/hand pressure etc. – as a controllable variable. Additionally, Hammond and other electric organ manufacturers provided vibrato as a controllable (again often knob-based) variable, adding another embedded “performance variable” into the synthesis.

These developments can be seen as precursors to modern sound synthesis, walking the line between electrical instruments and novel sound creation devices. They brought up the issue of *what* is being controlled and *how* this control is happening – what is being

articulated and what is the articulation. While there are many norms in notation-based western musical tradition, these questions arose because in composing the devices, the *nature of the sound itself was being composed*. A very innovative set of instruments that approached the control/sound interplay differently were created by inventor Hugh LeCaine after World War II. His most famous was the Electronic Sackbut, which used finger pressure and position for subtle and nuanced control of the sound [28], keeping this in the human realm.

1.2.3 Modular Synthesis

Design inspirations can also be driven by abstractly imagined sound, or by other technical developments. In the case of Moog's modular synthesizers it was both: attempting to create abstract sounds as described by composer Herbert Deutsch while being inspired by an article on modular voltage control by Harald Bode [25]. This greatly pushed the notion of modular control, with the same functions (sine, saw, square) used for either sound generation or modulation – so that actions associated with liveliness or expressivity such as vibrato or tremolo were once again embedded within the synthesis, this time through voltage-controlled, low-frequency oscillators. While some may argue that this somehow destroyed expressivity, I rather suggest that it shifted this particular element of articulation into the machine, so that human input expressivity could be focused elsewhere – for example through control of the *rate and depth* of amplitude and frequency modulations. This was similarly the case with Don Buchla's modular synthesizers from the 60's, which included a sequencer for the first time as well as sound-sculpting idiosyncracies such as the ability to mix between even and odd harmonics, thereby introducing a timbral control that has persisted in digital music control design. His approach to shaping and control led to separate devices in later years, such as the thunder and lighting controllers.

1.2.4 Classic Vocoder

Perhaps the most complex and articulate control ability lies with the human voice, where there exists a complex interdependence between the excitation (glottis/vocal cords) and modulation (nose/throat filtering). One approach to separating the source

from the modulator was via the vocoder, invented at Bell labs by Homer Dudley in the 30's. This was applied to musical use by Harald Bode, and incorporated by Moog in his synthesizers, adding the articulatory control of the human voice as a new layer of expressive control to shape his modular synthesis. The vocoder has become a classic paradigm for source-filter sound processing, introducing the sort of expressive articulation that was previously reserved in music for singing. Through this device, and later versions such as a linear-prediction based vocoder [29], the notion of control was introduced into the *analysis* stage as well in that the extracted filter parameters acted as control/modulator. The specifics of how control may be introduced is ultimately a product of the source/filter separation method: for example Dudley created a version of his device known as the Voder [30] which articulated through finger control of the bandpass filter gains, foot control of pitch, and discrete selection of voiced/unvoiced input or transient excitation. This clearly must have led to finer articulation (or destruction) of overall speech quality than a more global control of spectral envelope, as evidenced by the 1 year of required training time for articulate speech.

1.2.5 Digital Synthesis

The most influential result to come from Bell labs, of course, was Max Matthews' creation of digital audio in the late 50's. This discovery coupled with his intense musical interest led to the creation of the Music N family of digital sound synthesis languages. A watershed moment came with Music III in 1960, which introduced modularity via the concept of unit generators. Much as with Moog and Buchla, a primary interest was towards composing new types of sound based on imagined abstractions, and with the translation into the digital world this extended into all aspects of composing and performing. As such the system was constructed so that "instruments" could be defined and grouped into "orchestras", played back via a "score". By the time of Music V one had access to several waveform generators, adders and a random number generator which could be used to create an instrument that played discrete events referred to as notes. Thus the paradigm was such that one had to conceive of the types of modulations that would articulate underlying sound generators, and so control was a product of each modular instruments' use. While the language (instrument, orchestra, score, note) and structuring around discrete

note-events surely led to classical usages (e.g. sinusoidal LFO control of frequency), others such as John Chowning extended this control paradigm until it folded back into the creation of new timbres, as with FM synthesis. Therefore modularity – both analog and digital – gave rise to a blurring of the boundaries between articulation/expressivity and fundamental sound alteration. This blurring, I feel, necessitates a reconsideration of the role of control in digital music contexts, which is precisely what I do in chapter 2 for the case of “electroacoustic” music.

1.2.6 Post-Digital-Synthesis

By this heading I don’t mean to give rise to discussion about digital music aesthetics as in [31]. Rather, I now must contextualize my work by briefly explaining the most relevant among the many developments since the early days of digital audio. The Music N paradigm is still quite present as can be seen very directly through the Csound program, as well as being the archetype for Max/MSP and PD - though it could be argued that they have both defined new territory which simply takes off from the Music N platform. Regardless, the research into sound synthesis has become quite sophisticated and blossomed into a widespread global pursuit in the past 30 years with several main parallel developments.

Virtually all of the developments in sound synthesis that have been applied to music can be traced back to telecommunications and speech analysis, synthesis and coding. Two broad categories that exist in terms of focus can be considered as physical models – those that focused on modeling based on the true physics of sound production – and more abstract signal models, which predominantly have focused in the spectral domain in order to access more perceptually-relevant signal information. It is argued in [26] that physical models give rise to more “performable” synthesis algorithms that are more intuitive and expressive, while signal models (dubbed “spectral models” by the author) are more effective for compression, such as that needed for transmission and data reduction. The primary reason for this statement stems from the fact that these models provide control parameters that are directly related to the physical actions that one would measure from a performer (which is again why I noted in the previous section that mapping is trivial in this case). However, it is my conviction – and an idea that persists throughout this dissertation – that when one composes sounds as well as

instruments, the notion of what is articulated and what is articulating is blurred. As such, what may be a parameter for expressive control becomes a product of how a given sound model is parameterized or not, among other things. As one example, this blurring of boundaries is evident in the music and writings of Trevor Wishart, who refers to his personalized sound processing algorithms as “instruments” [32]. This is not to say that physical modeling-based approaches can not be extrapolated and used in sound-focused contexts such as electroacoustic music, but this will not be my primary focus. That said, physical and signal model approaches are not mutually exclusive [33] and should each be discussed briefly.

Physical Models

One of the key developments in this area arose when Kelly and Lochbaum [34] digitized the traveling wave equation for a sequence of acoustic tubes in order to model the human vocal tract, taking into consideration impedance discontinuities that occur between tubes. The two collaborated with Max Matthews in order to create the first digital singing voice synthesis. This laid the foundation for the *waveguide* approach to physical modeling, which has been extended to more articulate vocal synthesis by Perry Cook [35] and other models by Julius Smith [26] and a number of his students. This method provides a very efficient, real-time approach to virtual instrument design. Part of this efficiency comes from the fact that linear, sequential sections can be lumped together, as was discovered experimentally by the now-famous Karplus-Strong plucked string algorithm [36].

When an instrument sound is dominated by its body’s resonator, and when this sound consists of a few long-lasting modes of vibration, one may use *modal synthesis* [37], which represents each vibrating mode as a separate second order system. Strictly speaking one might call this approach physical modeling when the analysis comes from a consideration of the physical geometry of the system in question (rather than strictly perception). However control in this case is less clear and more removed from the physical domain, as one does not typically interact directly with a resonator in an acoustic instrument. In this way, the spectrum between timbral (e.g. roughness) and performer modulations can be explored depending on the modal synthesis implementation. The signal/physical hybridity of the modal synthesis approach is best

exemplified by the state-space formulations of modal synthesis by [38] and [39]. Rather than the specificity of the modal approach, the use of a state-space representation has inspired my own approach to sound modeling in this dissertation precisely because of its ability to represent hybrid signal/physical systems.

A third and canonically different approach arises from representing the vibrating object as an interconnected mesh of fundamental mass-spring-damper systems as in [40] (whose author has spent decades working on this approach) wherein each fundamental interaction is treated as its own vibrating system, with appropriate couplings defined. In this way, there is a continuum between excitation and modulatory control, haptic feedback and sound production. One may model a classic instrumental interaction or create a complex topology of interactions up to the level of composition as Cadoz has done with his piece *Pico...Terra* [41].

While one could also include finite element methods here – in which one discretizes the underlying dynamical system (e.g. the wave equation in a given medium) – I will not discuss this further as it is more of a simulation method for physical systems rather than a sound synthesis approach which makes perceptual and musical assumptions or reductions. That said, I will now discuss signal modeling with a bit more detail as it most directly relates to the work of this dissertation.

Signal Models

As noted there exist purely abstract methods of sound synthesis such as FM synthesis or general modulation/filtering synthesis. However my focus is on methods that seek to transform sound from the real-world in an attempt to perform this sound musically. Within this class of sound processing approaches one can differentiate between *effects* and *transformations*. The former processes input sound without any model or knowledge of the audio structure, while the latter does model the sound based on some analysis method [42]. The most recognizable effects are built up from modulation via LFO of the signal amplitude, a delay line or filter parameters to create effects including tremolo, vibrato, flanging, phasing and wah-wah. Again, as with analog modulation synthesis the control paradigm is comparable to performer modulation of excitation (e.g. string position) or resonator (vocal tract, as in wah-wah) for added expression. As

with FM synthesis, novel timbres are created by extended use of modulation type, rate and depth. Meanwhile, transformations seek to represent either the full richness of input audio or parameterize those elements deemed most relevant, transform this representation and then output the result as a newly modified signal, thus the name analysis/synthesis is used for such techniques. In between effects and transformations are a spectrum of techniques, and where the line should be drawn is not always clear. In this work I will generally use sound processing techniques that are signal-adaptive in some manner, and so I will use the term transformation and sound processing interchangeably, while effect is reserved for the aforementioned types of classic audio effects.

The aforementioned vocoder was a precursor to digital analysis/synthesis systems, wherein the input sound was fed into a bank of bandpass filters, providing the energy for each frequency channel and thus resulting in a representation of the spectral envelope of the input sound. As with the case of the voder, articulations such as that done by the vocal-tract could be achieved through control of the gain on these channels. This is considered a *source-filter model* because the analysis is towards the separation of excitation and resonator, thus control is implicitly separated into excitatory and modulatory actions. In the digital realm (in addition to digital implementations of this same technique [42]), researchers have used parametric techniques such as linear predictive coding (LPC) [43] in order to get more refined estimations of the spectral envelope, very effective in cases when the sound in question is modeled well by an autoregressive process [44]. Often the excitation input itself is discarded, and pulse trains or noise are used for voiced or unvoiced sounds, respectively. These refinements lead to specializations such as better spectral resolution with the tradeoff that the model is more removed from intuitive control parameters (i.e. when filter gains are replaced by filter coefficients). When the model itself does not suggest expressive control, one must examine what is captured in the source or filter, to “find” the signal elements that suggest further expressive deviations. This has been explored extensively out of real time by composers Charles Dodge and Paul Lansky [45]. One inherent quality of LPC-based systems is that they favor smooth temporal motion, leading to a “smearing” of transient audio phenomenon, something that arises in other analysis/synthesis systems as well. Other methods of extracting the spectral envelope

for modification and processing (without a modeling step) have become popular, such as cepstral analysis [42]. The reader is directed to [46] for more on these developments.

Building on the original vocoder, Flanagan introduced the *phase vocoder* implementation [47] which improved on the original in overall sound quality as it could track fine modulations in each frequency channel by keeping track of phase information within channels, for each analysis frame. While the original was also analog (filter-bank implementation), it was later brought into the digital domain [48], with early musical explorations discussed in [49] with later tutorial in [50]. An important result came in linking the phase vocoder with the short-time Fourier transform (STFT) [51], changing the way that it was implemented in practice. This provided explicit parameters for amplitude, frequency and phase, allowing issues to be addressed such as “phasiness” that gives a reverbation-type quality [52] as well as smearing of transients [53]. The phase vocoder itself is simply based on the STFT representation, allowing for perfect reconstruction of audio as long as certain analysis/synthesis criteria are satisfied.

However, when one seeks to build transformation and control on top of this technique, modeling assumptions are made about the slow vs. fast modulation of each frequency channel, the nature of the transients within the signal, and so on. In particular this has led to several high-quality algorithms for separate pitch and time scaling, minimizing artifacts from the signal processing [54][55][56]. In building real-time systems based on the phase vocoder [57], one must construct their own control paradigm and interaction context. One obvious choice is whether to accentuate and elicit some expressive element already in the audio stream, or to shape the audio in a more global way, taking it into a new direction. A control paradigm that can bridge these two approaches is through *scrubbing* – dynamically warping past values of the input sound – while modulating amplitude and pitch. One sound control paradigm that this suggests is to play the “surface” of the sound as one scrubs (I’ll discuss how this surface may be represented in chapter 2). Other ways to articulate continuous modulation can be had by *mutation* between two sounds [58] – more commonly known as *morphing* – wherein amplitude/frequency/phase values are interpolated between two sets, captured from separate sources. Just as analog synthesis brought about the question of composing sounds as well as pieces, techniques such as this bring about the issue of performance dynamics: what is being excited and what is being modulated and nuanced. In

addition to the mapping, this is built into the control structure by decisions such as the precise time-varying nature of the interpolation between STFT parameters.

More control possibilities arise when an additional parametric sound model is built on the underlying STFT/Phase Vocoder implementation. Due to the physical as well as perceptual importance of stable sound partials, additive analysis/synthesis has become a very popular technique. Jean-Claude Risset recognized this fact and constructed an early approach using Music V [59]. However it was McAulay and Quatieri who built the first additive analysis/synthesis scheme based on the phase vocoder [60], in which the “birth” and “death” of partials are tracked over time to define stable trajectories. Many different classes of sound are represented well with this approach, and users can apply time-stretching, pitch shifting and morphing between such sounds in a very convincing way. However there are two main problems: the first is that transients are again not well represented due to an implicit assumption of local stationarity. There have been several attempts to improve transient processing, such as reassigning the energy within each-time frequency window [61] and using the more transient-focused wavelet transform [62]. A second problem arises from the stochastic element of the sound not being well represented [44]. The sound *residual* – the information left if one subtracted the additive synthesis from the original signal – carries important perceptual information such as breathiness, attack characteristic or excitation type.²One popular approach has been to model the residual as white noise passed through a time-varying filter as in Serra’s Spectral Modeling Synthesis (SMS) approach [64]. Other improvements on this have included a separation of transient signals from the stationary sines+noise [65], treating the partials as noise-modulated within a given bandwidth [66] and the concentration of the noise in perceptually relevant critical bands [67]. Just as important as a proper model of the noise, of course, is accuracy and resolution of partial tracking, which can be obscured by limits on time-frequency resolution. Many approaches have been applied to this problem, with two of the more successful approaches being heuristic assumptions about the input sound (e.g. quasi-harmonicity) and the use of Hidden Markov Models for non real-time tracking [68]. For general additive discussion the reader is directed to [69]

²It also assumed increasing importance as a musical element throughout the 20th century, as I discuss in [63]

As with the underlying phase vocoder, one has access to the amplitude, frequency and phase information. The difference here is that – depending on the particular system – one may parameterize attack transients, steady-state partials and steady-state noise separately. This adds quite a bit of flexibility in regards to control considerations, as one can separate excitation elements from continuous partials and noise, where attack articulations or continuous modulations can be specialized for each. For example, the spectral and temporal envelopes for sines and residual can convey different elements of expressive control, and so can be used quite differently to compose what one might call “expressive instrumental dynamics”.

Parameterizing existing sounds is one way to specialize and define control structures at the sound model level, and additionally the interaction between control and fundamental sound qualities can be facilitated through using the sound itself to define control structures that create new musical gestures [70], a method used often by computer music composers and what is sometimes called adaptive effects [71]. Through signal-adaptive behaviors, one can allow for a continuity between sound quality and gestural response. Just as sound analysis and parameterization may give rise to a control type, the choice of “indirect” mapping [72] and the type of features extracted may be designed as a part of the overall articulation and expressive control.

1.3 Originality and Contributions

From a review of pertinent literature, it becomes clear that research into mapping in a digital context is still a relatively new topic. Meanwhile, as I have noted the field of sound synthesis – and in particular signal models using analysis/synthesis – has been developing for some time. Even still, the primary focus has still been placed on analysis model improvements, with less attention paid to control strategies *at the sound level*. Given this, I feel that my dissertation puts forth several significant contributions to the field,³ both in terms of the synthesis of the topics that I address and the novel approaches taken individually in chapters 2-5.

³Note that some of this work has appeared in publication as detailed in the references. In each case, I was first author of the work with co-authors assuming an advisory role. Any and all writing reproduced in this dissertation from these manuscripts is my own.

In terms of the overall synthesis of ideas, this work's originality stems from the fact that it emphasizes a *combined top-down and bottom-up view on mapping and control structuring that incorporates space and time*. That is to say, I rectify the spatial, multi-parametric approach of mapping-as-constraint with a temporal view on control dynamics. This approach is further novel in its *use of electroacoustic music theory as a design principle*, notably through the introduction of a new framework for the concept of *sonic gestures*, which is the focus of chapter 2. Using electroacoustic principles, I present a *sound-first principle for designing control structures*, which is the first (to my knowledge) to apply *Schaefferian theory to feature extraction, with real-time control considerations*. This results in novel signal processing applications, including what I believe to be the first use of the so-called *EMD time-frequency method for analysis of sound texture and modulations*.

In chapters 3 and 4 I will turn my attention to mapping, extending results in this growing body of literature. Chapter 3 contributes a new *mathematical formalization of multi-parametric mapping approaches* focused on a geometric point of view, contributing towards the idea of a *mapping design space*. I present a new *mapping theory* which unifies different concepts related to the subject, including the role of continuous modulation, gestural conditioning and parameter mapping. With these in mind I then apply this to existing work on geometric-focused mappings, extending these by introducing a *mapping toolbox for continuous interpolation and dimensionality reduction*. Building on this, in chapter 4 I will present several *novel control-focused user tests* that are aimed at exploring the perceptual influence of mapping and the *interaction between mapping structure, sonic gesture type and the perceptual control structure* of a given musical performance context. This chapter will then end with some *canonical examples as control archetypes*, followed by two examples of novel *fabric-based instruments* that act as counterpoint to the multi-parametric, solo instrumental paradigm.

While chapters 3 and 4 focus on a top-down, parameteric approach, I do the opposite in chapter 5: focusing on control at the low-level of the sound model. In order to work towards control that is consistent with my electroacoustic-focused interests and musical context, I work on a *novel approach to processing textural and noise-based sounds*. This begins with a use of state-space representation for the model and ultimately adaptive

filtering. While this representation has been used for modal synthesis [38][39] and acoustic parameter analysis [73], to my knowledge this is the first work to use it *towards novel control effects, allowing a mixture of physical/signal approaches*. This leads to one of the first projects which *applies control theory and specifically the extended Kalman filter to create a control-focused sound model*. Therefore, this *applies adaptive control to analysis/synthesis*, allowing for an atomic consideration of the control structure.

The project represents a set of novel contributions as highlighted in the above passages. As a whole, I feel that this work presents a synthesis of ideas that combine multi-parametric mapping, perceptual studies, computational geometry and dimensionality reduction, time-frequency analysis, adaptive filtering and control systems into a unique approach to instrument design and real-time control consideration that puts sound-focused (timbral, textural) control at its forefront rather than as an outlier, using electroacoustic music (and latently, embodied cognition) as its philosophical underpinning.

Chapter 2

From Control to Sound Gestures

In the majority of discussions surrounding the design of digital instruments and real-time performance systems, notions such as control and mapping are seen from a classical systems point of view: the former as a variable from an input device or perhaps some driving signal, while the latter is considered as the liaison between input and output parameters. At the same time there is a large body of research regarding gesture that is concerned with the expressive and communicative nature of human performative action. While these views – of control, mapping and gesture – are certainly central to a conceptual understanding of “instrument”, it can be limiting to mediate ones conception of digital instrument design entirely through them. As an example of an alternate way to view instrumental response, control structuring and mapping design, this chapter will discuss the concept of gesture from the point of view of the perception of human intentionality in sound and how one might consider this in the process of interaction design. I will examine and reflect upon the ways in which gestures – in the sense of dynamics that are either a result of or which suggest human action – are embedded as a trace within a sound signal, and how these might recursively be embedded within a mapping proper. To this end I introduce the notion of “sonic gesture” in order to extend the notion of instrumental mapping design, evoking the latent embodiment in Schaefferian theories of electroacoustic music that will also set the musical context for this dissertation.

Therefore, this chapter is essential to the overall presentation in that it creates a mid-level framework that will serve to both qualify (philosophically, musically) and

quantify (resultant signal processing framework) the output of the top-down mapping approach of chapters 3 and 4 as well as that of the signal-first bottom-up approach to control of chapter 5. It provides a conceptual link between these two and allows for describing not just musical context but *musical control context*. In terms of larger dialogue, an additional goal is to move away from discussions of mapping in isolation – as simply a connective tissue between control and sound parameters – instead viewing it as a process in larger control structuring that is very much an interplay between two complementary aspects of musical performance dynamics, namely human actions and abstracted musical dynamics. In the case of human gestures, the literature is quite vast and often covers communicative visual aspects of performance. In this work, my consideration of actual (as opposed to imagined) human action is similar to the gestural primitive notion of [74] in that the concern is working with the dynamic and organic nature of such control signals in data space. To make this distinction, I will refer to these particular gestural objects as *control gestures*.

At the same time, the majority of this chapter will focus on the gestural nature of musical dynamics and their interplay with mapping and control structuring. In designing an instrument, of course one is considering the sonic material that will be controllable. Beyond this, however, it is the *shaping* of said material as a function of human action that must be designed – and that is so often implicitly embedded within the mapping strategies employed to this end. Many researchers have explored the relationship between the experience of listening to music and creating mental images of human movement [75][76] in the context of classical music rendered from acoustic performances. One hypothesis of this research is that such gestures that evoke a sense of embodiment or “anthropomorphic projection” [77] extend to the experience of listening to electroacoustic music. Without delving too deeply into the discussion on defining electroacoustics, in the context of this discussion let the term come to mean musics primarily focused on formal development through shaping of timbral and textural sound qualities. Just as most instrument design arguably does not focus deeply on sonic sculpting, I will not pretend that this work extends to all instrumental design contexts, but rather specializes to those in which control of timbre and texture is the main concern. What this chapter presents, then, is both an analysis framework and design methodology that are shaped by theories and views normally associated

with the out-of-time electroacoustic tradition. As I wish to avoid delving too deeply into stylistic or ontological discussions about the musical aspect of this established notion of musical gesture, I will instead favor the term *sonic gesture*, which further highlights my focus on signal and sub-to-note-level phenomena rather than larger elements of musical form often described using the former term.

2.1 Control Gesture

My ultimate concern are those gestures that are extracted from performer's actions, as opposed to complementary ancillary and communicative gestures not directly related to sound production. In considering the domain of all performer gesture types, it is worth noting that there are underlying functional gesture types that one may consider as existing both in control as well as human-input gestures, precisely because they are *instrumental* in nature [78] (also know as *effective gestures* in [76]) and so are intentionally acted in such a way as to embed them in the resulting control data stream. In short they are *sound-producing* gestures [79] that may manifest as

Excitation gestures: These can be impulsive actions that result in a discontinuity in the resulting control stream (related to acoustic instrumental actions such as plucking and striking) or continuous actions that result in the continual addition of energy to said resulting control stream (wherein acoustic instrument analogs would include bowing, blowing and scraping).

Modification gestures: Following the typology of [78] , such gestures can be

- *parametric*: characterized by continuous or discrete modulations of one or several features of the sound (with acoustic analogs including shaking, flexing, bending, etc.) or
- *structural*: which are categorical in nature and are related to changes in the physical structure of the instrumental system (e.g. the insertion or removal of a wind instrument segment).

This breakdown is a subset of that given in [78] for the proposed typology of *instrumental gestures*. For the sake of this work it will be safe to consider these as representative of performer gestures, as an analysis of the full spectrum of performer

gestures is beyond the scope of this discussion. Instrumental gestures are the type which directly relate to the current discussion as they, by definition, assume that human contact with an object is taking place, with a skilled manipulation and an “existence of an energy continuum between the gesture and the perceived phenomena” [78]. To further clarify and constrain the situation, I make the a priori assumption that contact with the control device is taking place with a knowledge of the degrees of freedom available due to the underlying sensing technology, so that effective movement degrees of freedom do not have to be learned over time. To put this in the language I’ve just established, the primary interest that this work deals with – in regards to this vast field of gesture research – is to extend the functional aspects of instrumental gestures to the control gestures on which we focus, towards the end of maintaining an instrument-like feel as determined by the chosen mapping structure¹.

Being directly responsible for sound production, control gestures are tied to the world of “perceptually relevant” gestures as discussed in [14], in that they represent actions that (both visibly and audibly) are directly responsible for sound output. Rather than take the approach of the aforementioned authors – extracting and mapping between perceptual gesture and sound spaces respectively – I feel the best approach is to construct the dynamics between action and sound in a feedback loop. This is particularly relevant to an electroacoustic context, wherein “compositional control” (or imagined gestural control) is often geared towards driving elements of sound that are not directly controlled in a parametric sense [80] by sound-producing actions in an acoustic instrumental context. Rather, such sounds arise from the sustain portion of a musical event – not controlled by direct energy transfer to the resonating body, but instead guided in an indirect and nonlinear fashion. For example one may shake an object that includes a spectrally dense resonator in such a way as to influence the roughness of its output. Many other acoustic examples exist (e.g. manipulating bow/string interactions) in which one indirectly controls sustained textural elements that may be clearly perceived as separable in time – acting on a different time scale than timbral elements – and that conjure images of gestural motion.

¹Naturally a complete “instrument-like feel” requires haptic and proprioceptive feedback, but again this is not within the scope of this dissertation.

Following this principle of guiding textural elements leads to actions such as those that [81] have called *sound-tracing* gestures, which can be thought of as outlining general contours, in response to imagined or actual sonic stimuli. I consider such gestures to be central in designing control structures for those timbral and textural sound qualities that are central in electroacoustic music. In this way, *embodied* control gestures for electroacoustic music may find inspiration not only in the direct actions of acoustic instrument performance but also in the indirect human response to musical stimuli in a non-performative context, as in the study of Godoy et. al.

2.2 Sonic Gesture

Through the ubiquitous experience of watching live music, we are able to clearly identify performer gestures upon hearing recordings of instrumental music performance. Going beyond identification, researchers in cognitive psychology have put forth theories of embodied cognition for listening in which it is suggested that perceiving sounds (e.g. vocal utterances [82]) is a process of creating mental images of the generating articulatory gestures. These thoughts have been applied towards a motor-mimetic theory of music perception [83], positing that the perception of musical sounds is inextricably linked to the creation of mental representations of sound-producing gestures.

I assume this embodied cognitive lens in order to examine theories normally associated with “fixed” or non real-time approaches to sonic media. In doing this I take as a starting point work presented in [84] in which Godoy provides a reading of Schaeffer [85] that evokes the gestural and embodied nature of these writings through his proposed notion of “gestural-sonorous objects”. In doing so, the author essentially elucidates an often-overlooked aspect of Schaeffer’s theory of sound objects: that he in fact proposes a link from his typology/morphology of sound objects back to an appropriate sound-producing gesture suggested by a given sound object. This can be clearly seen in Schaeffer’s typological notion of temporal envelope and related morphological notion of dynamical form [86], qualified in terms of human action whose imprint is left on the sound. In the English translation this is presented as “execution” type in concordance with the notion of “*facture gestuelle*” (or, “executive gesture”) as

the sound-producing action. The gestural quality of the sound object is then considered in terms of the overall temporal envelope that Schaeffer breaks down into impulsive, sustained and iterative, which are in turn paralleled with punctual, continuous and iterative gestures by [84]. Considering again the typology of instrumental gestures put forth by [78], we can further draw analogy with these temporal forms and a(n)

- instantaneous excitation gesture
- continuous gesture, either excitation or modification
- periodic gesture, either discrete or continuous, excitation or modification

respectively.

It is quite easy to imagine these various control gestures giving rise to the aforementioned envelope types². However the link with overall temporal envelope, and implicitly with temporal scale, is only a first-order view of the gesture-sound object experience, which we can press further towards the construction of mapping and control strategies. In this context when I refer to “mapping” I don’t mean to suggest that one may map directly between “gestural objects” and note-level sonic objects³, as such an association is determined by high-level functional and perceptual criteria that one can not directly parameterize, but rather that the constant flux between gesture/sound objects must be constrained. Setting the conditions for this interplay between perceived sonic gesture and appropriate control gesture is a global view on mapping – one that I feel paints a more complete picture of the process. It is in contrast to mapping-as-correspondence, which preserves the out-of-time, flowchart interpretation of mapping that neglects the *perception of gestural dynamics*. A more complete view of mapping includes the *perceived result* of a control/sonic gesture causality. I propose that if the state of action and sound dynamics are considered from the point of view of gesture/sonic object qualities, then such a *perceptual, embodied control design* can result through considering this *in a design feedback loop with the*

²Just consider any striking, plucking or pressing (in the case of a keyboard instrument) action of an instrumentalist. Continuous excitations include friction-based gestures such as bowing, rubbing and scraping as well as blowing, while modifications primarily consist of bending, lipping, stretching and various pressure modulations in the acoustic domain. Periodic gestures can arise from any of the above, and are integrally linked to rhythmic patterns and thus are further tied to the emergence of the perception of rhythm on a signal level.

³ Again, this would be more along the lines of the approach suggested in [14].

creation of mapping strategies. Taking such an approach suggests that rather than isolating a perceptual parameter such as roughness or brightness a priori, one engage in a more phenomenological approach, observing the qualitative sonic output that results from a given control structure in order to reveal the perceptually-relevant parameters that are being directly or indirectly driven. The lens through which to view this output is an analysis of sonic gestures. In this way, from a signal point of view “perceptually-relevant” is a product of perceived gestural intent – something that I will discuss a bit further in section 2.2.2.

2.2.1 From Sound Objects to Sonic Gestures

The morphological concepts developed by Schaeffer and associates can suggest new unexplored gesture/sound links that one may use to structure the instrument design process as well as develop sonic gesture extraction techniques. Following this Schaefferian principle, after perceptually “cutting” the continuous sonic stream using the principle of stress-articulation [87], the resultant object’s *morphology* describes both its *form* and *matter*⁴ [86], where the former relates to the global properties of a sound and the latter describes its internal characteristics. Having identified a sonic object one may examine its morphological properties for traces of perceived action. While a thorough analysis of each of these qualities described in the literature by Schaeffer, Chion and others is beyond the scope of this chapter, I will present the essential form/matter properties and how they interrelate, which will help to articulate the boundaries of sonic gesture both in theory and in the particular analysis tools that I have derived.

- **Dynamic Profile**

This quality is directly related to a given sound’s articulation and its energy envelope over time. This relates both to Schaeffer’s first order typology as well as the global form of resultant sounds, and is the quality most obviously related to gestural control. In addition to gesture, it similarly relates to sound identity (timbre) with the most essential information carried in the attack portion of the sound [19][17][18].

⁴This terminology is a combination of that presented in [88], [87] and [89]

- **Mass**

This can be considered as a generalization of pitch, describing the “momentary” spectral profile of a given sound over a perceptible block of time. As Chion⁵ notes, it is the “way of occupying the pitch field”. In other words whether a sound is pitched, noisy, comprised of several distinct pitches or an indiscernible complex of sounds. Therefore we can say that from an analysis point of view, while fundamental frequency is of course an important attribute, this quality also relates to *complexity of the spectrum*, such as pitch strength, the overall span of the spectrum at each sampling instant and the line-likeness vs. flatness of the spectrum.

- **Harmonic Timbre**

This is what Schaeffer refers to as “the additional qualities which seem to be associated with mass and enable it to be described.” [87]. He notes that for sound with a strong sense of pitch this is clearly separate from mass and describes it in terms of “shape”, “color”, etc. while these two qualities become less distinct as the sound becomes more complex, with their separation being dependent on listening context. This ambiguity means that a precise set of qualities may be hard to pin down in general. At the same time, certain properties can be seen to relate to perceived “color” and “shape” of a sound regardless of mass complexity, such as the overall spectral balance, formant structure and envelope type. While studies on timbre [19] have often arrived at three acoustic correlates of timbre perception – one temporal, one spectro-temporal and one spectral – it seems likely that what Schaeffer is describing as Harmonic Timbre can be considered as the spectral axis from this overall timbre space. From this point forward I will refer to this quality as HT.

- **Motion**

This term is translated from the french *allure*, and can be considered as a generalization of tremolo or vibrato [89]. However just as a perceived instrument modulation is a combination of tremolo, vibrato and spectrum envelope modulation [90], motion can be a complex combination of amplitude and frequency modulations. It describes such fluctuations as they characterize a sound

⁵From [87] as translated by John Dack.

during the sustain portion, and so motion as a quality is integrally linked to the notion of modulation gestures acted subsequent to an initial excitation. As Schaeffer notes that for a given sound it describes “what the agent of its energy is, and whether this agent is living or not.” [87]. Therefore motion is closely tied to and suggestive of expressive control, and so becomes a very important feature in this discussion on instrumental control.

- **Grain**

This relates to the micro-level structure of sound objects, including both spectral and temporal properties. It encompasses the notion of *sonic texture*, which describes sound events that are globally stationary with local nonstationary elements [91][92][93][94][95]. Grain is explicitly related to gestural actions by Schaeffer in that it is further broken down into resonance, rubbing and iteration grains. Of course, this seemingly diverges a bit from the Schaefferian principle of reduced listening, yet it is indicative of the gestural imagery of his writings, and that I wish to apply. In a more general and signal-focused view, I introduce the terms *spectral grain* and *transient grain*. The former encompasses the notion of resonance grain and any similar phenomena wherein the causative factor is primarily a spectral feature, such as roughness resulting from dissonant tones. Meanwhile, transient grain refers to grains that are primarily a time-domain phenomenon, resulting from many micro-transients within the signal. It is appropriate to continue using the iteration grain terminology here in that it exists on the boundary with the perception of gestural iteration, as suggested by motion or by an overall iterative dynamic profile. Beginning from an idea suggested in [87][85] and [89], I will further qualify grain phenomena by virtue of quantifying the spectrum, weight and placement of the grain element of a sound.

Matter Profile as Gestural Description

Where these elements describe the overall quality of a given sound object, a consideration of the gestural nature of said object means examining the dynamics of morphological features. To that end, consider a matter profile as consisting of the time-varying aspects of all of the matter-related sound qualities: mass, HT, motion and grain. In his writings Schaeffer referred more specifically to mass and pitch-based

profiles [86][87], but the concept is extended here to all matter criteria, as these contribute to an overall image of a sound gesture and are used (particularly in electroacoustics) to convey a sense of motion. A spectrum exists between phenomena such as mass and HT and between motion and grain, with the focus shifting from one to another throughout the life of a sound. Therefore, examining the co-varying nature of all matter profiles will paint a more complete picture of a given sonic gesture. Further musical implications and justification for this extension of matter profile arise naturally from the dual principles of gesture and texture, as will be discussed in section 2.2.3. From an instrumental point of view, a complete view of the dynamic profile of all form/matter properties can suggest a given control gesture, and a novel way to constrain the mapping design: constructing a system’s “sonic gestural response” in the spirit of analysing an acoustic body’s frequency response.

2.2.2 Perceptual Criteria of Sonic Gesture

Most commonly, in the music technology literature when one speaks of a perceptual feature of sound they refer to either pitch, intensity, or one of a number of timbral attributes including brightness, noisiness, roughness or voiciness. In general, timbral features are quite central to contemporary musical composition practice and in particular electroacoustic music. This attribute has been associated with acoustic correlates through user studies in which the structure of similarity ratings, categorical groupings or comparison to verbal descriptors [96] are matched through methods such as dimensionality reduction techniques to the structure of acoustic similarity [16][17][18][19]. This approach makes several a priori (which is not to say unfair) musical assumptions that equate the concept of timbre to that of instrumental identity and/or ornamentation, or what has been called the “source-cause” aspect of timbre which describes “The natural tendency to relate sounds....because they appear to have shared or associated origins.” [97] That is, the research seeks to find the minimal and orthogonal set of information necessary to differentiate between different sounds and further to group them based on their cause – seeking what has been called the “structural invariants” of timbre [98]. This is quite consonant with a traditional view of

timbre in which this quality is structured and somewhat abstracted by virtue of a musical score: on the one hand speaking to musical schemas that listener's have developed concerning choice of orchestration, and on the other a coloration of note events that are organised relative to this time-tested grid that is the score, giving rise to such concepts as *klangfarbenmelodie* [99]. In either case, the process is one of discretization and grouping of "timbre" and is integral as such to most Western approaches to musical form [98]. Meanwhile, the "concrete" nature of timbre⁶ is recaptured once the process of orchestration extends beyond coloration to composing at the sound level by consideration of instrumental acoustics, as in the case of spectralist music [100].

Timbre similarity studies focus on what Gaver has called *musical listening* [101][102][103], a mode in which one is listening for properties of the sound itself, as opposed to listening oriented towards properties of the source. The latter is an assumption that dominates in the case of ecological perception studies⁷, as is evident from the nature of psychoacoustic experiments and the isolated sound events employed (e.g. [105][106]). Though it may be that the properties of materials are most perceptually salient in a given listening context, in a complex musical situation the interaction between excitation and material may well prove to be more important as it provides information about the "expressive" nature of the performer. Such continuous variations are generally left out of psychoacoustic studies of musical timbre or "non-musical" ecological studies, being attributed to such expressive performer deviations in the former case or simply not entering into discussion in the latter.

In the case of electroacoustic music, continuous variations such as modulations of intensity, frequency or density play an important role in dening musical form⁸. Moreso than in most forms of Western music, this then means that these phenomena – in addition to contributing to performer "expressivity" found in the ornamentation of a gestural-sound object – also act as contributors to form-bearing elements of a musical piece. I maintain that this is highly relevant in the context of a discussion on gesture

⁶To borrow this Schaefferian phrase.

⁷Though one might argue that the work in [104] is an exception that takes more of a "musical listening approach".

⁸From a signal point of view, these qualities which we examine relate as much to the quality of *texture* as to timbre, which can be seen to differ in regards to temporal scales over which changes to spectro-temporal properties occur [107].

and digital instrument models in that if we are to extend these views – in particular the notion of parameter mapping – then we must reconsider our *implicit parameterization of musical as well as perceptual phenomena*. In order to construct a mapping framework that considers gestural intentionality in sound – particularly in the abstract world of electroacoustics – one must consider the continuous interplay between sonic material, perceived human action and perceived musical form. Continuously following (analyzing, tracking) these parallel roles of continuous sound modulation is at the heart of the analysis of sonic gestures, leading to a re-consideration of what a “perceptual parameter” might mean in this context of navigating performer intent at the same time as musical meaning. This, then, takes the place of the standard practice of adopting sound features that arise in a laboratory listening context, having different musical assumptions.

In other words, while known perceptual features related to brightness, roughness, etc. clearly do arise as salient in many listening contexts, they may or may not be the main carriers of information in regards to gestural dynamics. To decide upon such perceptual relevance warrants a *phenomenological* approach in the spirit of Schaefferian tradition wherein it is the *perception of intentionality* (understood in terms of form/matter dynamics) that informs the design of new mapping and control structures. While certainly divergent from most presentations on instrumental design, this “inverse mapping” approach is particularly appropriate for electroacoustics, again because of the interrelation of gesture and form.

2.2.3 Form/Matter and Gesture/Texture: Mapping Considerations and Structure-Bearing Principles

This notion of a global profile for a sonic gesture and “internal” characteristics that generally relate to smaller-scale temporal and spectral changes (i.e. sub-note level) is important for the consideration of sounds that are driven by human action. In the case of acoustic instruments note-level events have an overall shape that is a direct result of performer actions, with textural and timbral characteristics that result indirectly from actions such as slight bow angle/force modications in a cello, or embouchure changes in a reed instrument. Considering the embodied nature of such a global form/internal matter breakdown from a reception/listeners point of view, [108] has written about the

dual concepts of *gesture* and texture in the context of his work on spectromorphology. The author refers to a gesture as “an energy-motion trajectory which excites the sounding body, creating spectromorphological life”, further stating that “when we hear spectromorphologies, we detect the humanity behind them by deducing gestural activity...”. Smalley then goes on to develop the electroacoustic music-theoretic notion of *gestural surrogacy*, differentiated by order ranging from immediate awareness of materials, through standard (acoustic) instrumental sound gestures, to abstracted shapes that cause uncertainty of source/cause through to a so-called “remote surrogacy” wherein the human element is lost as well as perception of source and cause. I maintain (as do many other sources [109][32]) that the interplay between these types of sonic gestures traversing different times scales as well as levels of abstraction from human action are a primary way to create structure in an electroacoustic work. This can be related to the notion of musical gesture in instrumental music that is applied concurrently to temporal scales ranging from note-level events up to larger phrases [75].

In adopting this gesture/texture dichotomy, it is implied that gesture as a form-bearing principle is concerned with propelling (musical) time forward, moving towards (as well as away from) a particular goal. Smalley states that “If gestures are weak, if they become too stretched out in time, or if they become too slowly evolving, we lose the human physicality...(moving towards an) environmental scale.” In the absence of gestural motion we are left with so-called texture, which is the lack of perceived motion but rather focus on inner details of sonic material. While this concerns structural characteristics of electroacoustic music reception, it directly relates to short-term sonic features in that the former regularly emerges from the latter in music of this sort. In this way, the gesture/texture breakdown can be drawn into analogy with the note-level Schaefferian notion of form and matter. Rather than simply equating the two – gesture is form, texture is matter – it is the dynamics of form and matter as well as the interplay between the two that define gesture. These dynamics are the substrate from which imagined (human) gestures arise, and it is in their boundary – what Schaeffer called the “criteria of sustainment” – that Smalley’s texture-as-sound resides. In this instance the signal-model definition of texture [95] – directly related to perception of non-motion – converges to the spectromorphological meaning. It is in the dynamics of grain, and between grain and motion, that electroacoustics offer new boundaries of

imagined human gestures – that themselves are on the boundary between sound-tracing and sound-producing gestures. There is no clear and direct path – in terms of mapping action to sound – that gives rise to such gestural dynamics. Therefore it is worth examining this concept of gesture/texture and form/matter dynamics in the larger discussion of instrument models: to view the design of a mapping/control strategy in terms of the way that it drives these dynamic qualities, realizing that it can have larger implications in terms of composing for the resultant instruments. It suggests an embodied approach to constructing a composed instrument [110], one which considers idiomatic gestural dynamics arising from the often radically different musical production/reception modes of electroacoustic music while maintaining the essential concern with the human element of musical performance practice.

2.2.4 Towards Sonic Gesture Features

In designing a performance system based on these principles – the idea that relevant sound features are a product of perceived gestural intent and that one may “compose” an instrument based on gesture/texture interplay – one must necessarily define the gestural vocabulary a priori, and to some degree begin from observed gestural objects – an instrumental design approach to phenomenology. At this point I will preset one canonical example that gives rise to a particular mapping of acoustic correlates to morphological descriptions, both as a way to further expand upon this notion of sonic gesture as well as to concretely define an analysis system to be used in the perceptual experiments of chapter 4.

To my knowledge, there exists only one project that considers a computational framework for these morphological qualities [88][111]. However, this particular work is concerned with MIR-style sound classification based on a given interpretation of these sound-object criteria with a large importance placed on pitch as a sound feature. Most importantly, there was no attempt to provide perceptual parameters in the sense of linking human perceptual observations (e.g. that of the instrument designer) to these properties. Rather, the classical features of musical and environmental perception are conflated and used as sound object descriptors. For example, these studies equate grain to the quality of roughness, thereby making precisely the sort of a priori musical assumptions that I aim to avoid.

Because they are exemplary in regards to morphological qualities in many ways, and because they relate to my own sonic gestural interest for the instruments I will construct, I will base my analysis on the sound examples that Schaeffer and colleagues at the GRM [86] presented as exemplary (and in fact exaggerated) variants of the qualities of mass, HT, form, motion and grain. In order to discover quantitative measures of each, I explored various signal features in order to find those whose value would correlate directly with a given morphological property based on the “training” sound examples⁹. By definition, the features that may describe Mass and HT vary wildly between different sound sets: a group of pitched instrument sounds have fundamentally different “types of mass” than a group of urban soundscapes, or a group of FM synthesized tones, leading to a different type of HT as well. In fact, from an analysis point of view these qualities by their nature lend themselves to categorial (harmonic vs. inharmonic vs. noise) types of distinction. As this examples focuses on a particular type of sounds (inharmonic, rolling gong), my approach to mass and HT was to find features that were “salient” in the entire set in the sense that they varied with the changing sonic gestures. In a sense these qualities were fixed in the way that one fixes an instrument on which to focus: e.g. a guitar timbre space, or a percussive timbre space. After making this large-scale mass/HT decision, the features of grain and motion became the primary focus of this analysis. Rather than just co-varying in a relevant way, I sought out parameters that were only salient when grain or motion were perceived to be present – and so that correlated directly with the present of these perceptual features. In many ways this is similar to defining a timbre space in the sense that one defines perceptual similarity/difference, and searches for signal correlates, but in this case *the designer is the sole judge of perceptual salience*: defining what grain should sound like, or what type of mass should be present. In this particular case that is not entirely true, however, as I am using sounds that Schaeffer himself defined as exaggerated instances of each feature, so that in sense he becomes a co-designer for the instruments I will build based on the resultant analysis.

The sound examples are based on variations of an archetypal sound([86], CD2, track 89) which is heard as a sustained drum roll (with slow onset) on a large gong or similar cymbal, including the decaying resonances. The tracks that are used provide variations

⁹Specifically these are from CD 2, tracks 90-94 of [86]

on form (shorted and lengthed, both from CD track 90), mass (track 91), grain (92), HT (93) and motion (94). As we will use these sounds throughout the remainder of the chapter, we will from this point forward adopt the notation x_k where k represents the associated track number¹⁰.

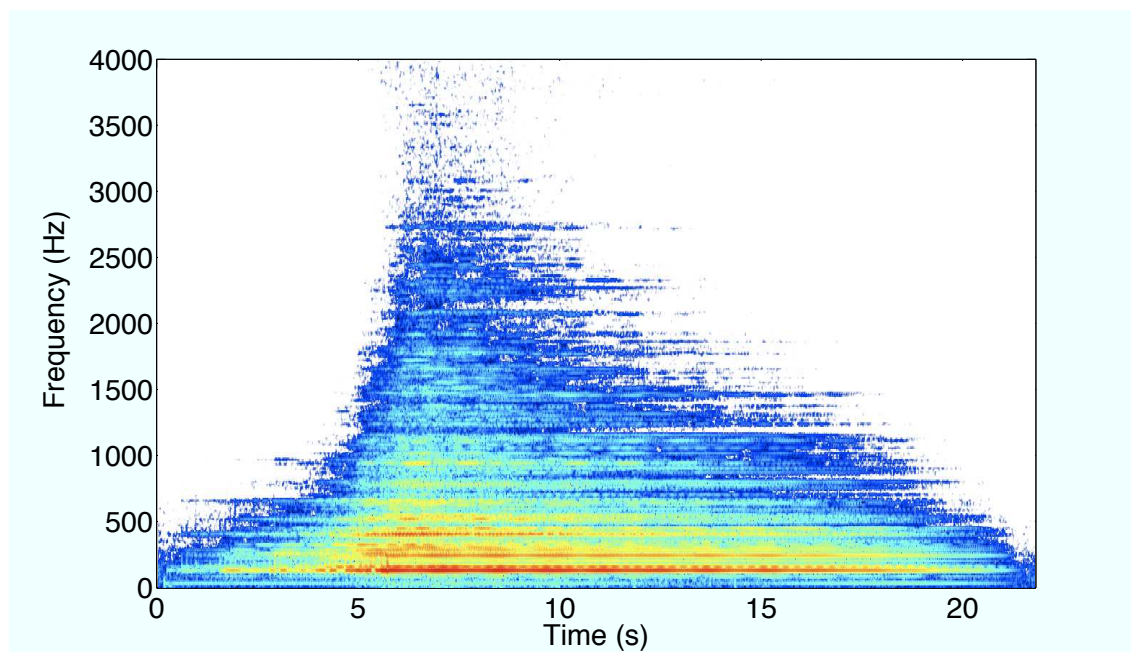


Fig. 2.1 Spectrogram for test sound x_{89}

Having said this, consider the spectrogram¹¹ for the archetypal example x_{89} from this point forward, as presented in figure 2.1. The other sounds are clearly variations on this, and seem to be achieved through filtering, mixing and perhaps effects such as plate reverb. This leads to changes in mass that are not categorical (e.g. pitched to noisy) but rather add some qualitative variations – thus I note these variations and rely on features for this property that are known to be effective in differentiating between mass-based classes such as pitch strength, harmonicity and so on. Meanwhile, I will focus my attention on the other salient properties of HT, grain and motion as well as dynamic and matter profiles.

¹⁰For the two examples of track 90, I differentiate the long and short form variations by x_{90a} and x_{90b} , respectively.

¹¹The relatively low signal-to-noise ratio for this sound set necessitated the use of the given colormap in Audiosculpt to properly show all morphological sound features, as opposed to the Matlab plots that are used elsewhere in this dissertation.

To that end, let us consider the form/matter dichotomy so that we may clearly separate the nature and function of the signal features to the best of our ability. The notion of form directly relates to the temporal energy envelope of a sound object, and it should be easy to convince oneself that this is in turn integrally linked to the “gestural” nature of a sound. However it is worth examining the *dynamics of matter qualities* that are often used to convey musical motion and so become elements of form. I’ll first begin with a discussion of the static qualities of matter, so that we can understand how their dynamic variation plays a role in creating a sonic gesture profile.

Matter Proper

These are the features that relate to a “snapshot” of the spectrum, reflecting its weight, balance and complexity. While something like grain can be considered as “surface”, these characteristics describe the internal makeup of a sound and thus are central to its identity.

In [85] the two qualities of matter – mass and HT – are seen as integrally tied to one another: a sound’s mass defines its central spectral character, and its HT defines the “halo” or “glow” that accentuates this. It is stated [87] that sounds with no mass (defined as a sine wave) or total mass (defined as white noise) have no HT as there is no “room” for this quality. The sound archetypes that I work with for instrument design, and the sound synthesis types used (noise-excited modal filters, granular synthesis) never reach these two limit cases, but they may move within the spectrum defined by these two (particularly granular synthesis). Thus in regards to mass, I would like to be able to continuously track sound output in regards to this dimension, and so I employ signal features that are known to capture this well, namely spectral flatness and spectral spread [112]. While this covers the spectral “expanse” of a sound, it does not describe the relative sense of pitch – which it goes without saying is an important feature. Rather than track fundamental frequency, the interest is in following the continuous variation of pitch-ness, with a focus on temporal domain method that looks for patterns of regularity (thereby complementing the two spectral features). Thus I decided to use a pitch strength measure that is based on a short-term analysis of the autocorrelation function (ACF) [113], and is computed on a blockwise basis with values

ranging from 0-1 over each frame¹². The spectral flatness describes the general peakedness vs. line-likeness of the spectrum, and is a classic feature from speech processing [114]. The flatness F of frame L is computed as follows:

$$F(L) = \frac{e^{G(L)}}{A(L)} \quad (2.1)$$

where

$$G(L) = \frac{1}{N} \sum_{i=0}^{N-1} \log(X_p[i]) \quad A(L) = \frac{1}{N} \sum_{i=0}^{N-1} X_p[i] \quad (2.2)$$

are the geometric¹³ and arithmetic means of the spectrum, respectively, and X_p is the square value of the magnitude spectrum at bin i for frame L . This measure is constructed so that it is close to 0 for tonal sounds and signals with line spectra, and close to 1 for white noise – and so fits our characterization scheme well.

Two measures that together describe the expanse and the balance of the spectrum, respectively, are the aforementioned spectral spread and spectral centroid. I mention the latter here because it is used to define the former. It measures the amplitude-weighted mean-frequency of the spectrum:

$$C(L) = \frac{\sum_{i=0}^{N-1} i(X_p[i])}{\sum_{i=0}^{N-1} (X_p[i])} \quad (2.3)$$

and arises in nearly every timbral perception study as a salient feature related to the subjective notion of “brightness” [18][19]. The spread, then, is a description of the variance of the spectrum, about this mean:

$$S(L) = \sqrt{\frac{\sum_{i=0}^{N-1} X_p[i](i - C[L])^2}{\sum_{i=0}^{N-1} (X_p[i])}} \quad (2.4)$$

Again, the spectral spread describes *how* the spectrum is occupied, and so is related to mass. Therefore my given mass features were chosen apart from the sound examples, to

¹²Since I often focus in this work on non-pitched elements of sound, I only compute pitch strength in certain cases where it is relevant to do so.

¹³Strictly speaking, $e^{G(L)}$ is the geometric mean.

be able to describe continuous motion between pitched and non-pitched elements and spectrally sparse or dense ones. Meanwhile, HT is a very context-dependent idea, and so was something I’ve defined here based on the chosen sound material. It is something that is related by Schaeffer to verbal descriptors such as bright/dull, open/closed and so is similar to words used by listeners to describe instrumental timbres that relate to resonant filtering of an instrumental body (e.g. see [72]), and so HT seemingly relates to the snapshot of the spectral envelope and perhaps further to things such as salience and placement of formants. Given its description, it would seem a priori that “harmonic timbre” describes that element of “musical timbre” which is largely spectral in nature. After analysis of classic timbre-related acoustic features (spectral flux, spectral smoothness, centroid, etc.) across all example sounds, this indeed proved to be the case as the aforementioned spectral centroid – describing the balance of the spectral envelope – correlated very strongly with example x_{93} which presents an “exaggerated harmonic timbre”. This measure was exceptionally higher for x_{93} when compared to the other sound examples. This is illustrated in figure 2.2, where the centroids for the steady-state portion of all sound objects (normalized in time) are plotted together, except for x_{92} which had a very low centroid and so was omitted for clarity of presentation. Clearly the value for x_{93} stands apart from the other measures throughout.

In light of this discussion and the goals for the instruments I will use in the studies of chapter 4, I therefore extract sound features related to pitch strength, spectral flatness and spectral spread to characterize the momentary mass of a sound, while the spectral centroid proved to once again be a perceptual feature in this context, in that it directly correlates with the sound examples identified as having more prominent HT. Certainly other spectral envelope features such as formant frequencies, harmonic spectral centroid, or salience of formant regions may be important for harmonic timbre description; however these chosen features are sufficient for distinguishing different matter profiles while I focus more closely on the true interest of this study: the features of motion and grain.

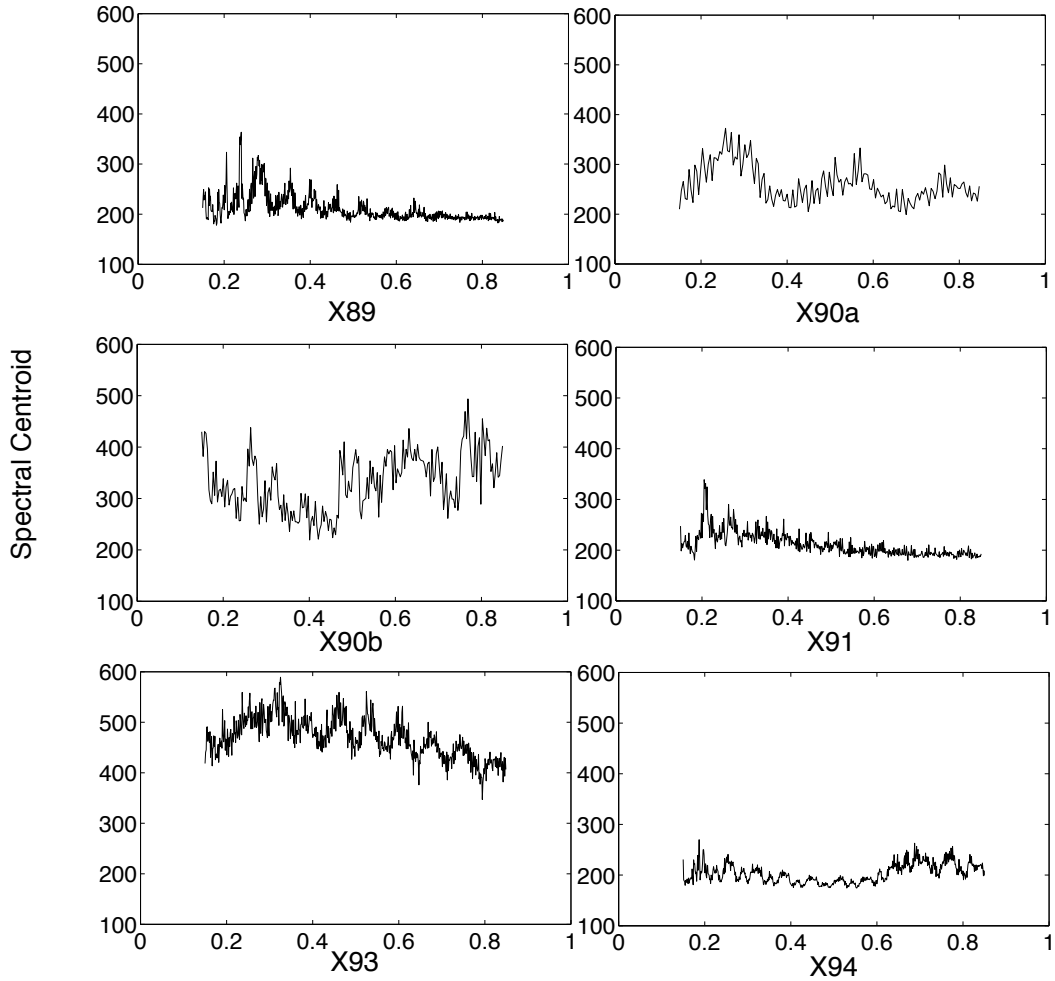


Fig. 2.2 Plot of Spectral Centroid for example sound objects.

Form/Matter Boundary: “Criteria of Sustainment”

The perception of motion and that of grain can be very closely related and in fact can shift from one to the other in the course of a sound’s life. For example a slow modulation of the fundamental frequency of a tone may be perceived as vibrato and/or tremolo , yet at a point after the modulating control signal goes beyond roughly 16 Hz, this movement will fuse with the tone itself and be perceived as a grain quality. This is related to the psychoacoustic phenomenon of roughness, which may result between two tones played at certain intervals, with studies showing a maximal rough quality at particular modulation rates, depths and envelope fluctuations [115][116][117]. Just as

there exists this spectrum of phenomena between the qualities of grain and motion, these two also lie on the boundary of the internal matter of a sound and the dynamic form: grain may manifest as matter properly or an element which ties matter to form, depending on its rate, regularity, spectral placement and amplitude. It can suggest the surface of an object as much as the surface of the sound itself. Similarly, motion may have an intrinsic quality in itself within a sound object, yet it simultaneously suggests the action¹⁴ that caused it. This led Scaeffler to study these phenomena in a different chapter of his morphology, under the heading of a “theory of sustainment”. Indeed these qualities do arise during the sustain portion of a sound event, and in terms of sonic gesture profile are suggestive of continuous excitation and modification gestures. I’ve attempted to look into the deeper structure of a given sound object to extract grain and motion signals in order to characterize them within the context of the given signal. I’ve done this specifically in regards to the dynamics of grain and its relative spectrum, weight and placement [87][89] in order to describe a sonic gesture. The extracted signals related to motion are themselves gestural (thus no need to build a temporal profile), with additional description coming from the depth and rate of a given modulation as it changes over time.

The first measure of grain saliency was created with the idea that grain is intrinsically linked to micro-transients along the “surface” of a sound – that is, not related to the overall trend that is the temporal envelope but rather the quality I have dubbed transient grain. With this in mind, I constructed the *temporal fine structure* measure¹⁵ to describe such micro-variations, while de-trending this measure from the overall envelope and simultaneously removing effects due to overall signal level. This is defined as

$$TFS(L) = \frac{\sum_{i=(L)N+1}^{(L+1)N} (x_p[i] - \frac{(x_p[i+1] + x_p[i] + x_p[i-1]))}{3})^2}{RMS(L)} \quad (2.5)$$

where $RMS(L)$ represents the overall root-mean-square power for block L and $x_p[i]$ is again the power-amplitude of the signal at time i . This describes the average amount of deviation of the signal from its neighbors over a window of N samples, with the

¹⁴These are broken down into mechanical, living and natural in [85]

¹⁵In analogy to the timbral attribute *spectral* fine structure that is presented in [118].

influence of the overall envelope removed. Of course this measure is sensitive to N and can be tuned relative to the temporal scale of variations that one is looking for.

The exaggerated grain example from our chosen sound set consists of many small fast transients infused with the archetypal sound object, which sustain over the life of the sound. This sustainment means that we may take a relatively large window of 185 ms in order to smooth the TFS curve, and may still capture any variation in TFS with a reasonable resolution. If we examine the TFS measure computed across all six example objects, shown in figure 2.3, it is clear that there is a strong correlation with the exaggerated grain example x_{92} , with a TFS that dominates relative to the same measure computed across the other examples (x_{94} had a very small TFS overall and so was not displayed for sake of clarity). Therefore, this suggests that TFS is a strong measure of grain in the sense of a conglomerate of many micro-transients.

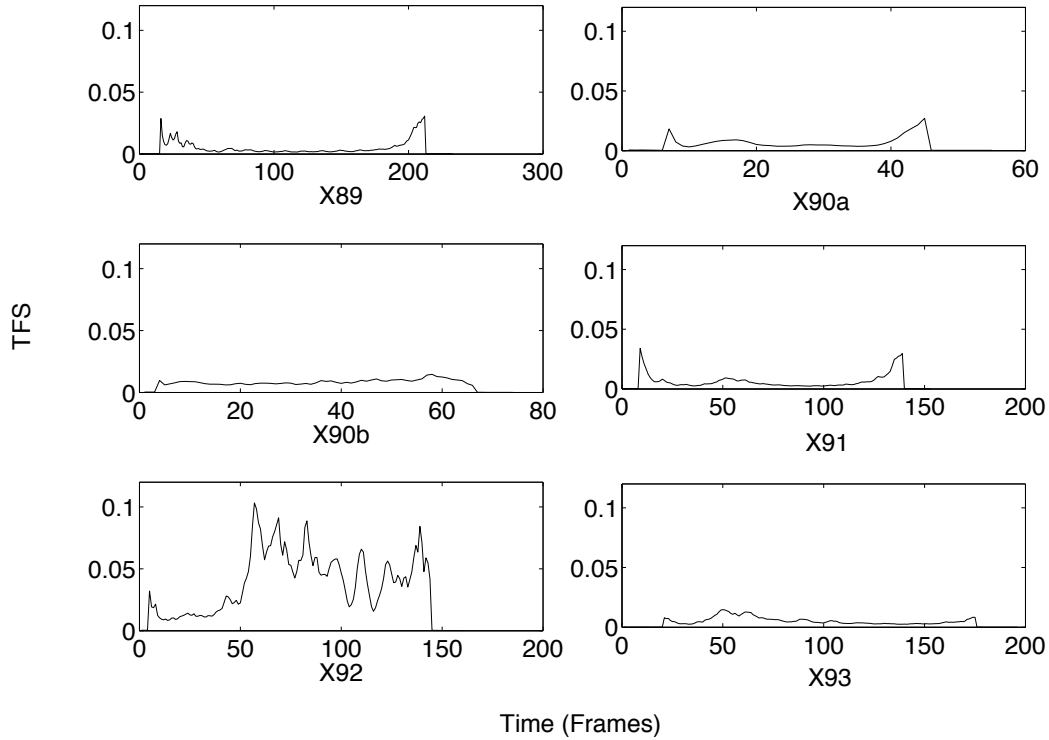


Fig. 2.3 Plot of TFS measure for example sound objects

What this proves is that TFS is a strong signal correlate for the grain quality that

Schaeffer highlights in the *Solfège de l'Objet Sonore*. I would describe his example x_{92} as a sound that is comprised not purely of transients in the sense of immediate discontinuity, but rather an irregular and sharply varying component that has some quality of resonance, with smoother attacks than pure transients. As my goal is to be able to separately describe the three different types of grains, I decided to synthesize an example that would be a much more extreme example in the direction of a transient grain, which would also encompass such sounds as crackling fire or breaking tree branches. The idea of course being that if TFS could describe a grain profile in a complex sound such as x_{92} and maintain a strong (or stronger) value in a synthetic “exaggerated transient grain” example, then it must be a perceptually relevant correlate.

To this end, I constructed the example sound by extracting the amplitude envelope of x_{89} using the RMS of the signal with window size of 256 samples and hop size of 128 ($F_s = 44.1$ kHz). This envelope was time-stretched via interpolation to match the original signal length, and multiplied by a signal consisting of random impulses, as generated by the “dust” noise generator from the Supercollider program¹⁶. This produced a crackling sound purely made of transients, with the same dynamic form as x_{89} . This noise was reduced to a reasonable level of 0.3 times its original, and added to the x_{89} signal. Looking at the result in figure 2.4, we can see that the overall TFS measure is much more than even that of signal x_{92} , and so this genetic experiment further validates TFS as a measure for transient grain.

The fact that TFS correlates with the “dust noise” or the “burning embers” of x_{92} aspect of these examples can be seen further in comparing this to measures from acoustic musical instruments. I have run this measure on a set of examples from the McGill University Master Samples (MUMS) database [119]. Figure 2.5 illustrates the TFS for a disparate set of sounds that represent a selection of the more “grainy” and textural sounds from MUMS. Those depicted are a flute played with flutter-tongue, one with vibrato, a bowed, martelé pizzicato and muted double bass, and finally contrabassoon, contrabass clarinet and bassoon. Note that all of these have a measure much less than the synthetic examples, proving that this models well an “exaggerated TFS”, so to speak. Within the instrument set, TFS reflects degree of graininess in e.g.

¹⁶<http://supercollider.sourceforge.net>

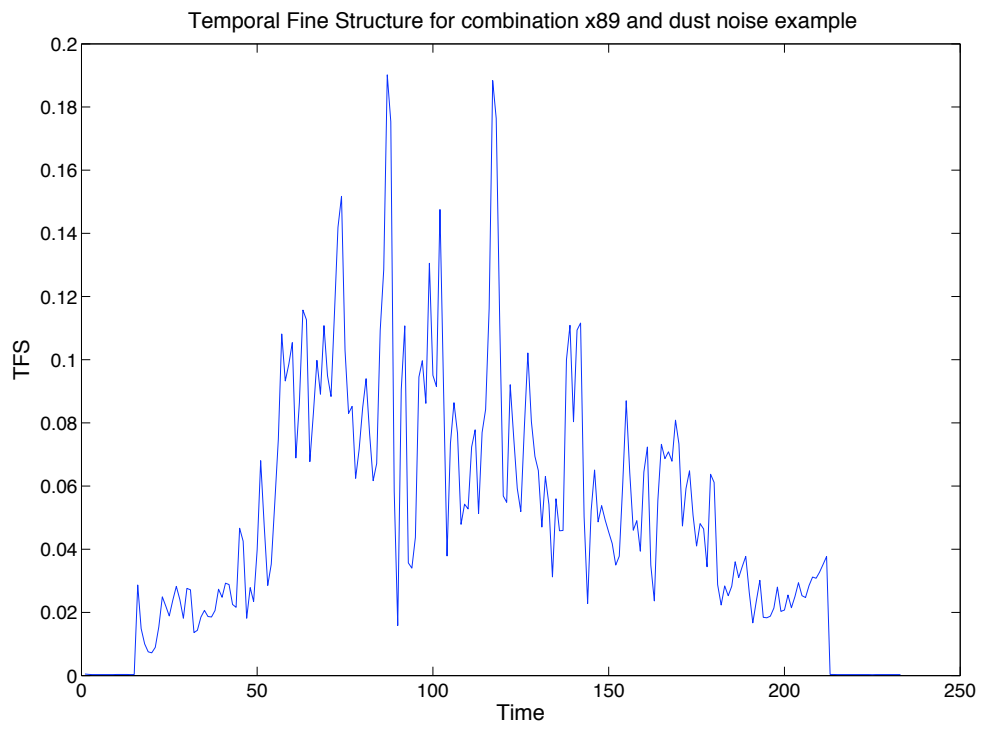


Fig. 2.4 TFS for dust noise example ($N=8192$, $\text{hop}=4096$, $F_s = 44100$)

the flutter-tongue vs. vibrato flute. It is also interesting to note that the measure is a product of the dynamic profile of the sound in some cases, while not in others. While this notion of profile will be explored in a future section, there are a few aspects of this that are worthwhile to note here. First, as noted grain arises during the sustain portion of the sound. The bassoon and contrabassoon have a very short onset and long sustain, while the contrabass clarinet has a slower onset with noticeable timbral shift between onset and sustain. Both of these facts are reflected in the TFS measures, with a very flat curve for the first two while the clarinet TFS follows the amplitude envelope more closely. Further on this point, the fact that this arises during sustain can be seen in the fact that TFS for the pizzicato sound is essentially nonexistent, while the martelé example – very onset-focused but having sustain nonetheless – in fact has the largest TFS measure, with a max around 0.06. Finally, the normal vs. muted double bass example illustrates a clear example of performer influence over grain quality during sustain: each measure follows the dynamic profile, with the latter example similarly muted in regards to TFS. Aside from further validating this measure with these acoustic examples, the reader should keep the influence of the *shape* of the TFS in mind when we arrive at the discussion on dynamic morphology.

Now, our notion of transient grain – in so far as it relates to the idea of a “rubbing grain” – immediately suggests the textural quality of both physical and sound object surface, and the degree of graininess of this particular type of sound is captured well by the TFS measure. However, the notions of iterative grain – such as the drum roll from the sound examples – or that of spectral/resonance grain – such as shimmering and fast modulations that arise from the resonance of dense inharmonic objects – are more interrelated and thus more difficult to decouple. It is spectral grain in particular that most easily blurs the line with the perception of motion-based modulations, differing from grain in regards to rate and depth from a signal point of view. My goal at this point is to decouple these phenomena in order to study them separately and more fully characterize them. The experience of iteration and spectral grain as well as motion becoming perceptually blurred is related to the fact that they all arise – at least partially – as a product of amplitude and/or frequency modulation acting on a given signal. Separating these phenomena becomes quite difficult as these behaviors arise from nonlinear interactions, while transient grains are nonstationary by their nature.

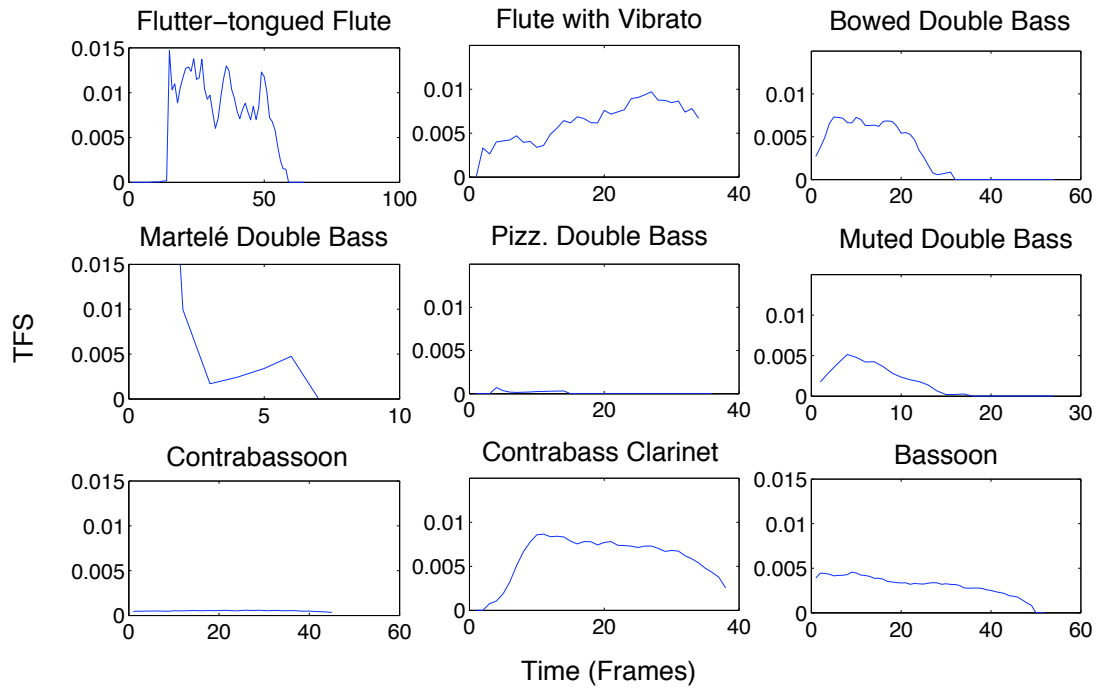


Fig. 2.5 TFS for all MUMS acoustic examples ($N=8192$, $\text{hop}=4096$, $F_s = 44100$)

This makes it challenging to use classic approaches to time/frequency analysis: Fourier methods provide a linear decomposition and don't deal well with transients, while Wavelets are adapted to transient phenomena but still produce a linear signal. As such, I've taken a radically different approach to "grain/motion decomposition", one that forms of the heart of my sonic gesture analysis and that is deserving of its own section for discussion by virtue of its depth and novelty.

- *Grain and its Motion: Extraction through Empirical Mode Decomposition*

To more deeply characterize these criteria of sustainment – and grain in particular – I have drawn upon a very powerful and relatively new time-frequency analysis technique for nonlinear and nonstationary signals first presented by [120] with further development notably by [121]. This technique, known as *empirical mode decomposition* (EMD), breaks a signal down into amplitude and/or frequency modulation (AM-FM) components in the time domain, from the most coarse modulations to the finest high-frequency noise components. This very simply-defined yet complex technique has been applied successfully to highly nonlinear phenomena such as atmospheric wind patterns [122] as well as related work in speech enhancement [123] and visual texture classification [124]. This latter application is particularly interesting to note, as my concept of transient grain is quite similar to and in fact encompasses the common characterization of sonic texture, suggesting that perhaps further analogies between audio/visual texture may be drawn from this work using this one unified approach to signal processing, a direction that very few existing studies have taken¹⁷. Throughout this discussion the reader should keep in mind that I am considering this work in the context of texture analysis; my decision to use the Schaefferian form/matter framework as the context for this discussion arose from the more specialized and in-depth distinction of grain types, and further because it sets a musical context above and beyond an approach to signal analysis.

The EMD method is signal-adaptive, and does not presuppose an orthogonal basis¹⁸. It

¹⁷One example can be found in [125].

¹⁸In this sense it is similar to techniques such as principle component analysis, but is more properly considered as a time-frequency analysis method.

further is defined algorithmically (rather than analytically) by virtue of the following process [121]

1. Given signal $x(t)$ identify all of the extrema.
2. Interpolate across all maxima to produce upper envelope $e_{max}(t)$ and all minima to produce lower envelope $e_{min}(t)$.
3. Determine the mean of the two envelopes $m(t) = \frac{e_{max}(t) + e_{min}(t)}{2}$
4. Subtract this mean from the signal, leaving the local detail $d(t) = x(t) - m(t)$.
 - 4a. Repeat steps 1-4 on the detail signal $d(t)$ until it satisfies two criteria:
 1. The number of extrema and the number of zero-crossings must be equal or must differ by at most one.
 2. The detail signal is considered as zero mean as determined by some relevant stopping criteria [120][126].
5. At this point, the resultant detail signal $d_k(t)$ is subtracted from the input signal and the process begins again on the residual.

At the conclusion of this process (including the variable number of iterations on steps 1-4, known as “sifting”), the signal $x(t)$ will be decomposed into a set of *intrinsic mode functions* (IMFs) $d_k(t)$ and a global signal trend $T_k(t)$ so that

$$x(t) = T_K(t) + \sum_{k=1}^{K-1} d_k(t) \quad (2.6)$$

Again, the IMFs express amplitude or frequency modulation behavior of the signal. A “spectral” interpretation only makes sense in so far as low vs. high modes relate to more vs. less temporal detail and thus, generally speaking, high vs. low frequency content. However there is no direct sub-band filtering, but rather this method is an adaptive time-varying filter. This coarse vs. fine decomposition is what motivated the exploration of EMD for this study, so that signal qualities of grain – related to the lowest (i.e. earliest removed) IMFs – could be decoupled from large scale trends related to dynamic profile, and intermediate modulations related to AM-FM components that comprise motion.

Now, while earlier studies focused on signals that were well-represented by a sinusoidal model, Flandrin et al. [121] described the properties of EMD in the case where the inputs consisted of Gaussian white noise. This prompted a refinement of the EMD algorithm, presented in [127], which presents an ensemble-averaged form of EMD (known as EEMD). This method was introduced to rectify the common problem of *mode mixing* in which oscillations of disparate frequencies jump between IMFs, thus rendering these modes physically meaningless. The authors found that this could be overcome by taking a “noise-assisted” approach in which the IMFs are redefined as the mean of a number of trials that result from “noisy measurements”. In practice, this is achieved by adding a small amount of unique Gaussian white noise to the signal for each trial, and taking the EMD for each. The final EEMD is achieved by averaging across all trials: the effects of the zero mean, uncorrelated random process are cancelled out while the IMFs are smoothed, thereby reducing or completely removing any mode mixing. This approach was applied to the task of extracting long-term musical structures in [128]. Therefore the authors examined the long-term trends of a portion of signal. In contrast, I look to extract these trends for the purpose of isolating the grain portion of the example signals.

In order to achieve this, I constructed an EEMD algorithm based on the freely-available Matlab code for “classic” EMD that accompanies the work presented in [126]. The authors of this work propose some guidelines for the stopping criteria of the sifting process, while at the same time noting that it is possible to over-decompose a given signal by applying too many such iterations. It is suggested that 4 to 10 iterations are often enough for “meaningful” IMFs to emerge from the data [128][127]. I used an excerpt of 18.5 seconds from the canonical example x_{89} in order to compare the case in which sifting iterations were solely governed by a stopping criterion set to measure the “zero mean-ness” of a given IMF (resulting in thousands of iterations) and the case where this was used in conjunction with setting a maximum number of iterations (20). In doing so, there was not a significant change in the structure of the IMFs, yet the computation time was reduced significantly (from 5.5 hours to 5.5 minutes). In contrast, by changing the maximum number of sifting iterations from 20 to 10, a difference could be seen in the lower IMFs: the larger number of iterations produced IMFs that favored sudden emergence of spectral changes, while the lower

number produced smoother IMFs. In order to examine if this was a product of mode mixing or over-decomposition, a number of trials were run for both max iterations of 10 and 20 (call them EEMD1 and EEMD2). After 100 trials and ensemble averaging, this same phenomenon of temporal fluctuations still did exist in certain of the EEMD2 IMFs, most notably in IMF3 from the 17-IMF decomposition as illustrated in figure 2.2.4. This implies that this is not a problem of mode mixing but rather one of over-decomposition. This was further made apparent in listening to IMF3, wherein the example from figure 2.2.4 sounds directly like a part of the input signal, while the random temporal modulations of 2.2.4 sound nothing like the original at the same point in time.

Now, the signal-adaptive nature of EMD means that the process of extracting grain qualities must itself be taken on a case-by-case basis. While a completely adaptive and automatic extraction of grain signals is beyond the scope or immediate interests of this study, I have laid the groundwork for such a system with a series of experiments. These experiments themselves were empirical in nature but proceeded to computational results. As an example, in listening to x_{89} , there is clearly a presence of spectral grain that can be characterized as “shimmering”, arising from the dense modes of the excited gong. There is further an iteration grain that begins as a drum roll and continues as a fast beating quality throughout the life of the sound, obscuring the line between human gesture and a rumbling, iterative decay. Having identified these two unique grains, and following the phenomenological nature of my approach, I then proceeded to isolate and extract these from the overall sound signal.

Of the resultant 17 IMFs from EEMD1, the lowest IMFs occupy the higher part of the spectrum and contain fast enough fluctuations to be listened to directly. In doing so, it is clear that the IMF3 signal very compactly extracts the aforementioned “shimmering” and thus the spectral grain aspect of the input. The full spectrogram for IMF3 is depicted in 2.7. If we look at the overall spectrum of the input signal x_{89} from 1.5 to 20 seconds in figure 2.1, one can see strong energy between 100 and 500 Hz that results from the drum roll excitation and the predominant mode of the gong. Meanwhile, as the strength of the roll and thus that of the signal increases, the overall spectrum envelope “opens up” to include more energy at higher frequencies, correlating with the aforementioned shimmer of the object. Comparing 2.7 to 2.1 one can see that this

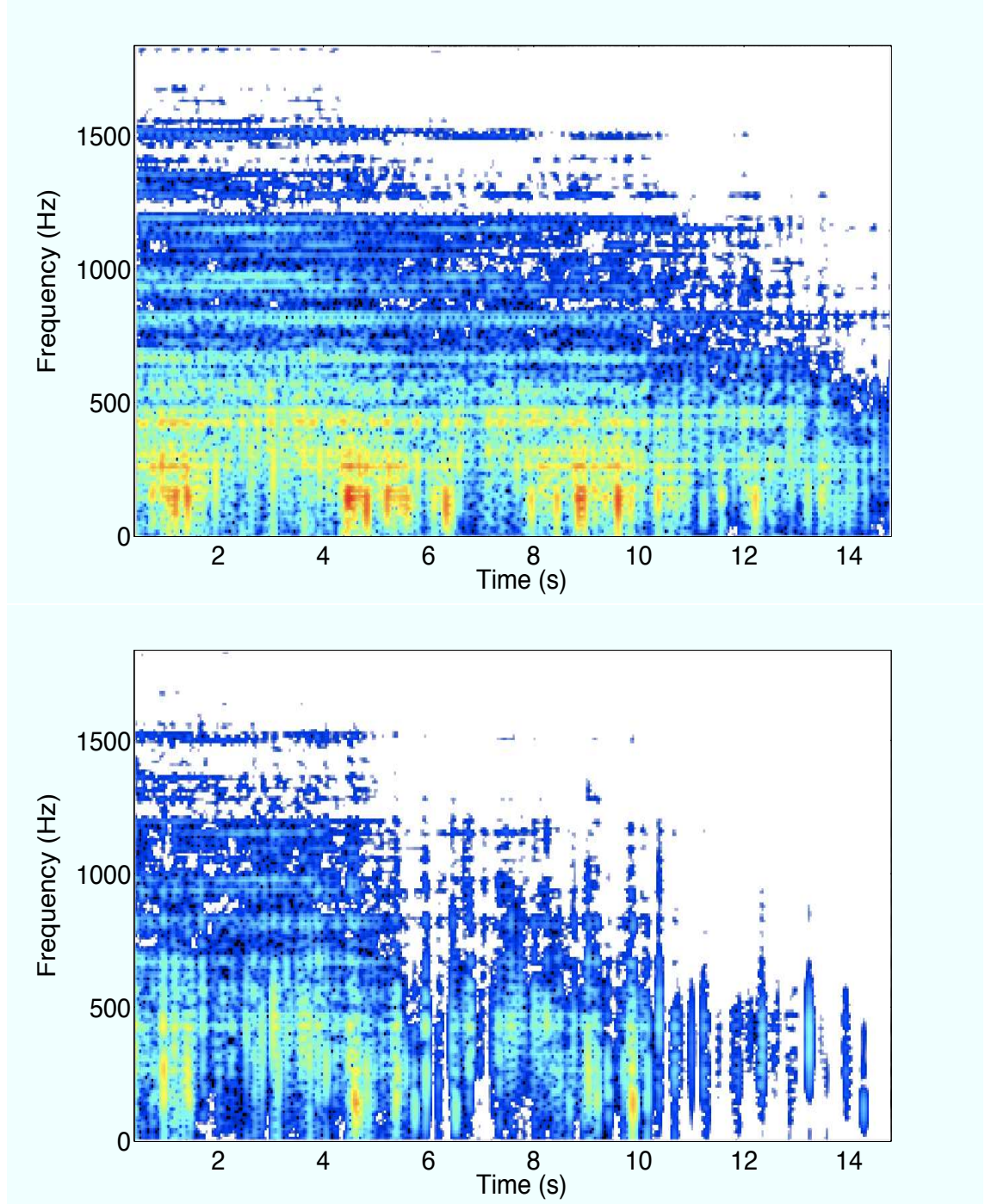


Fig. 2.6 Spectrogram of IMF3 from x_{89} (0-1800 Hz), FFT window size = 4096, 50 trials, with (a) Max iterations = 10 and (b) Max iterations = 20

opening up of the spectral envelope and ensuing shimmer is completely contained within this signal.

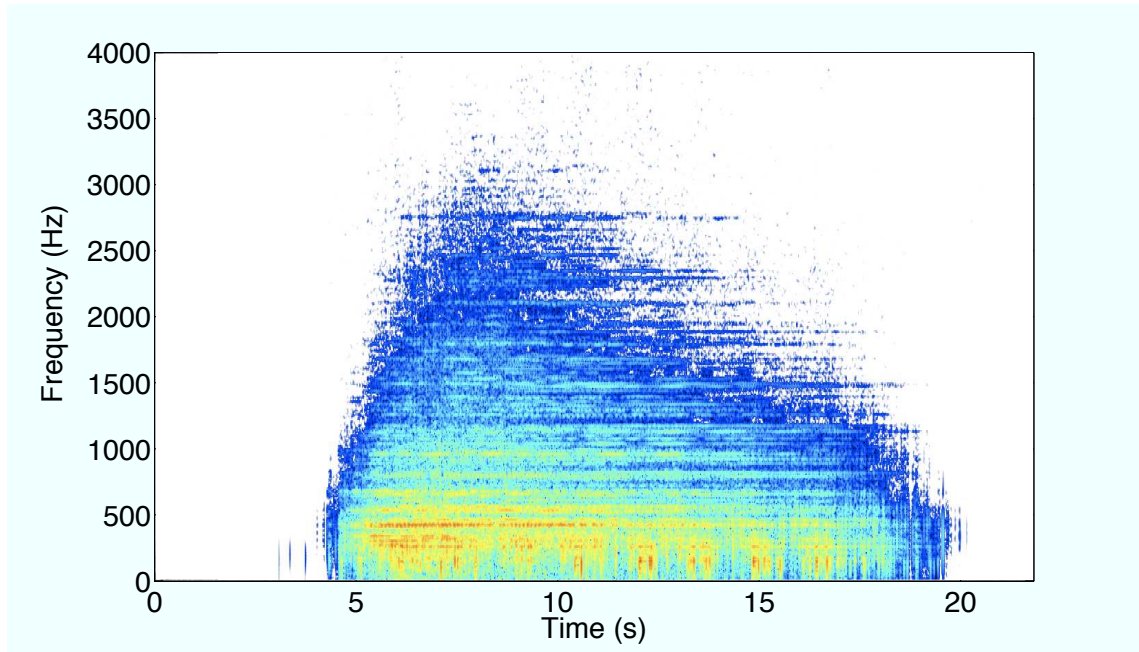


Fig. 2.7 Spectrogram of IMF3 from x_{89} for $f=0-6\text{kHz}$, FFT window size = 4096, 50 trials, Max iterations = 10

At the same time, the low-energy rolls of the first 4.5 seconds and last 2.5 seconds from x_{89} are not present, nor is the high-energy roll/roughness around 150 Hz that can be characterized as an iteration grain. This can be seen more clearly in figure 2.9 where the strong amplitude modulations can also be seen more easily. In this region lies the bulk of the iteration grain that arises from both the excitatory drum roll as well as the ensuing roughness of the gong. This grain aspect is also very compactly removed from the input signal, as is depicted in figure 2.8 which shows the spectrogram for IMF4.

The energy in the 100-500 Hz range has been captured by this mode function, and the iterative nature that characterizes the iteration-grain aspect of this signal can be seen in the spectro-temporal fluctuations in the low, high-energy part of the signal. This fact is further illustrated by looking at figure 2.10, which depicts the time-varying magnitude (non-normalized) located in the frequency bin centered at 161 Hz, having a fluctuation similar to the overall temporal envelope of the signal. The IMFs with the largest amount of energy at this location were far and away IMF3 and IMF4, whose

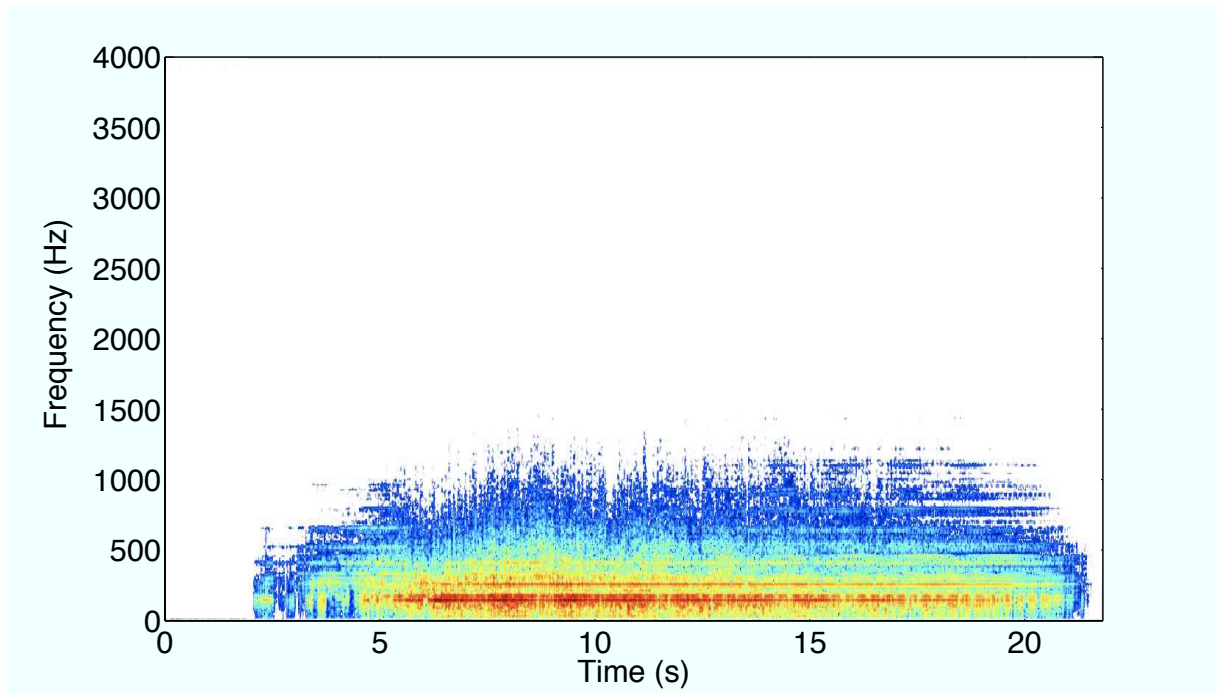


Fig. 2.8 Spectrogram of IMF4 from x_{89} for $f=0-4\text{kHz}$, FFT window size = 4096, 50 trials, Max iterations = 10

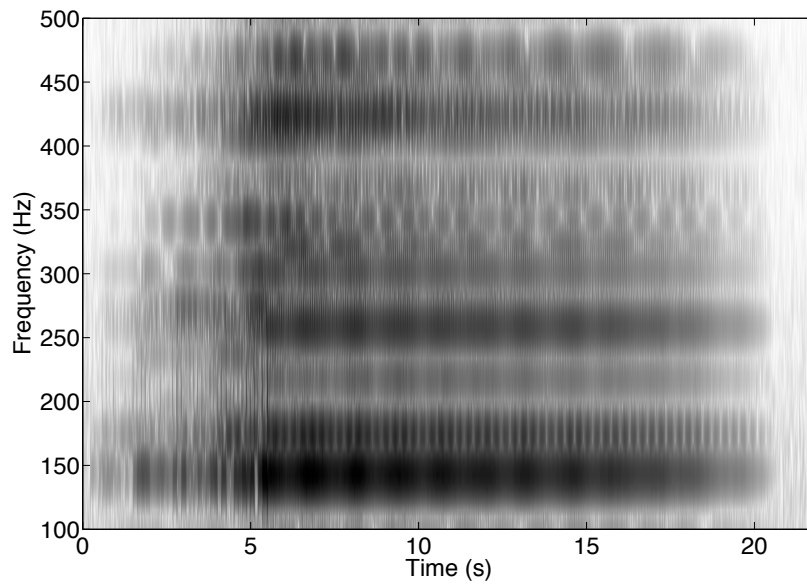


Fig. 2.9 Spectrogram of signal x_{89} from 100 to 500Hz.

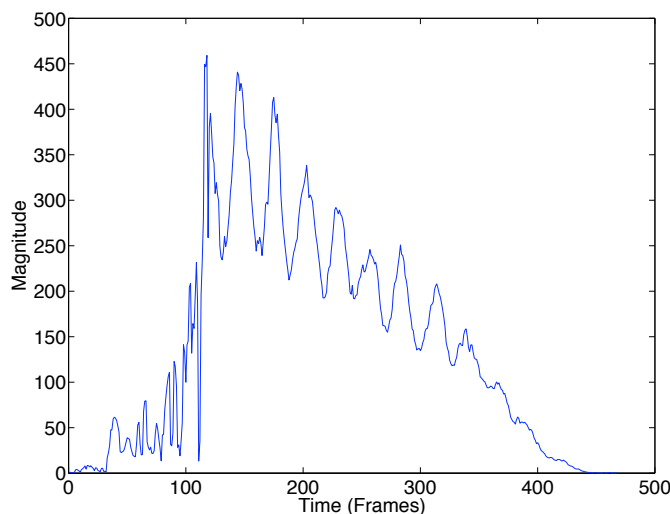


Fig. 2.10 Magnitude of bin located near 150Hz over time as extracted from spectrogram of signal x_{89} . STFT magnitude is non-normalized, with $N=4096$.

magnitudes at this location are depicted in figure 2.11. Note that the energy in IMF4 is not only much stronger than IMF3 – and nearly equal to that of the overall signal – but also the motion follows the form of the overall signal as well. This can be seen in figure 2.12 which shows a concurrent plot of the energy in the given bin for x_{89} as well as its extracted IMF4. This illustrates very clearly that the gestural nature of iteration grain is captured in this IMF, which is further verified by listening to IMF4, wherein one can clearly hear that the drum roll and its concomitant iteration grain quality are contained within this signal.

At this point, the EEMD technique has proven successful – with some human-aided listening and decision making – at extracting the spectral and iteration grain qualities that are embedded within example sound x_{89} in a complex way. The technique has been verified by virtue of informal listener comparisons as well as by visual analysis of the spectrogram. Now, in order to move towards a completely computational system – one that is signal-adaptive at the grain extraction level – I have explored different measures of the IMF signals. In regards to spectral grains, these are defined by dense modes that spectrally interact in a such a way as to cause rapid fluctuations in the temporal envelope. In [87] the shimmer of the cymbal is given as being exemplary of this phenomenon, and in this sense our chosen sound examples are themselves

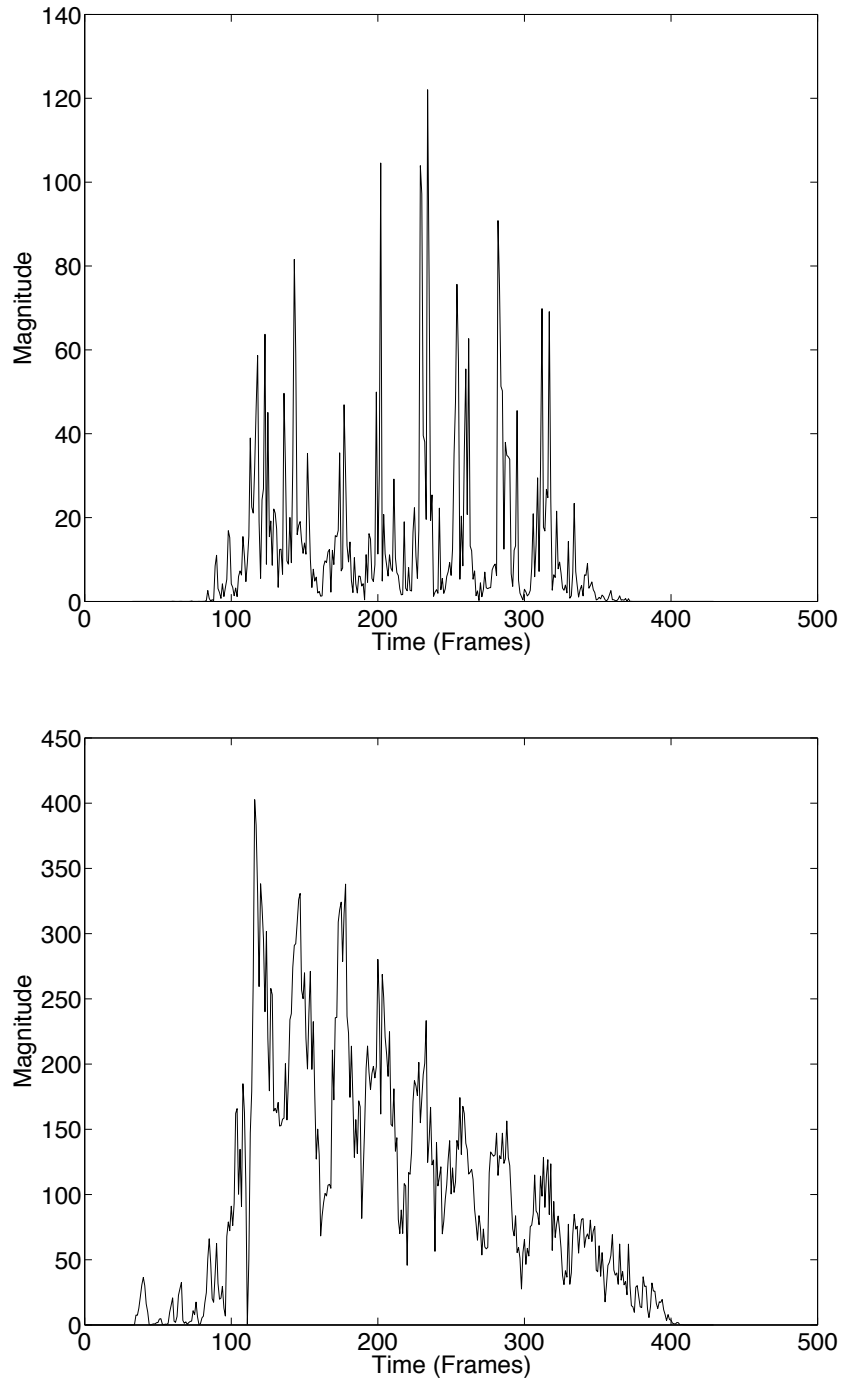


Fig. 2.11 Magnitude of bin located at 161 Hz over time for IMF3 (top) IMF4 (bottom) of signal x_{89} . STFT magnitude is non-normalized, with $N=4096$.

archetypal. The perceptual effect that is closest to this physical phenomenon from the psychoacoustics literature is that of roughness. After analyzing the IMFs from the example signals using the roughness model presented in [129], there were no strong correlations. This is not surprising as the model is based on a parametric sinusoidal approach, and we are working with broadband sounds that may be difficult for such an approach. However, the model of [117] as implemented in [130] was constructed with the consideration of amplitude envelope modulated white noise. An analysis of our extracted IMFs does in fact show a strong correlation between the identified spectral grain IMF3 and the Daniel/Weber model of roughness, as predicted. The roughness measures for the relevant¹⁹IMF functions for signal x_{92} are presented in 2.13. All other examples had an equally salient response for the extracted IMF3²⁰, which would be expected as the shimmering exists in all examples. Therefore, this provides confirmation that the perceived shimmer of the extracted spectral grain is closely tied to the psychoacoustic phenomenon of roughness, as it is determined by the Daniel-Weber model.

At this point I turn the attention towards example sound x_{92} , which again presents the “exaggerated grain” example. This sound is comprised of a drum roll gong, complete with spectral and iterative grain components as in x_{89} , and has a transient grain element added that (generally speaking) sounds like high-pass filtered fire embers²¹ or similar resonant-filtered transients. Looking at the spectrogram for x_{92} in figure 2.14, we can see the “base” sound of excited gong in the band up to 4kHz, with a similar profile and spectral/temporal envelope to x_{89} . The additional transient grain quality can be seen as a cloud, roughly in the range from 4-10 kHz. The EEMD algorithm constructed for x_{89} was run on x_{92} , again trying different stopping criteria to see which would yield the appropriate adaptive basis of IMFs. Once again, a set of 17 IMFs was produced with a maximum of 10 iterations per sifting process.

Now, the interest of course is to see if once again the qualities of spectral grain (aka “shimmer” for this example) and iteration grain (the drum roll/roughness element)

¹⁹Higher IMF modes registered negligible roughness values.

²⁰For all of our EMD measures across all examples, IMF2 was excluded as it produced a very brief sequence of transients that was not noticeable when removed from the signal and thus was considered as an artifact of the sifting procedure.

²¹Quite similar to the same sound from Xenakis’ famous piece *Concret PH* [131].

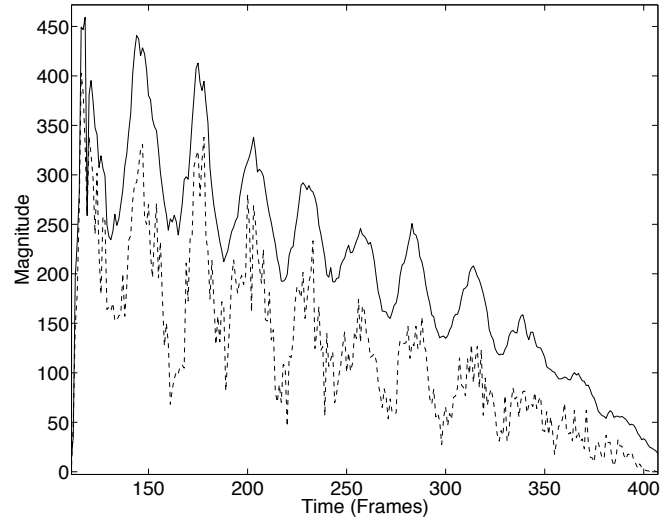


Fig. 2.12 Magnitude of bin located near 150Hz over time for signal x_{89} (solid) and extracted IMF4 (dotted). STFT magnitude is non-normalized, with $N=4096$.

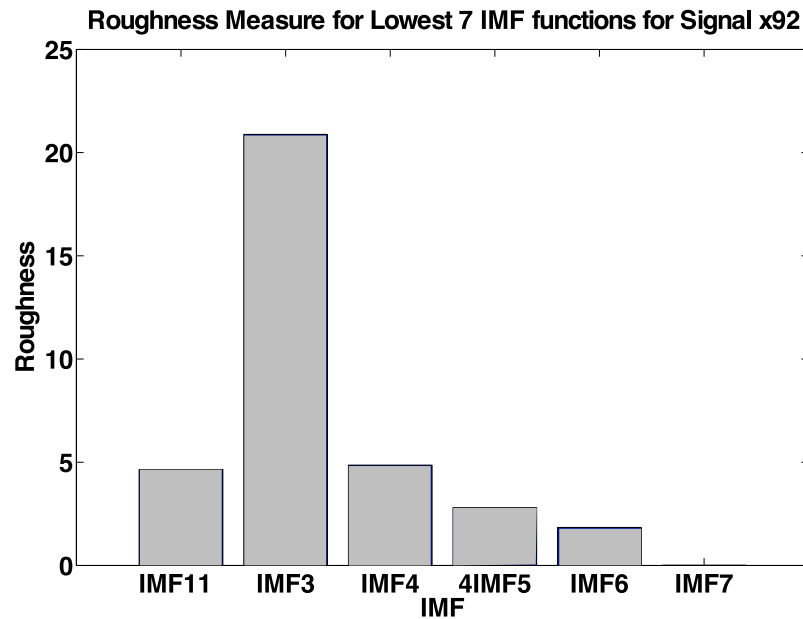


Fig. 2.13 Roughness for seven lowest IMFs for signal x_{92} . Note the saliency of the spectral grain signal IMF3.

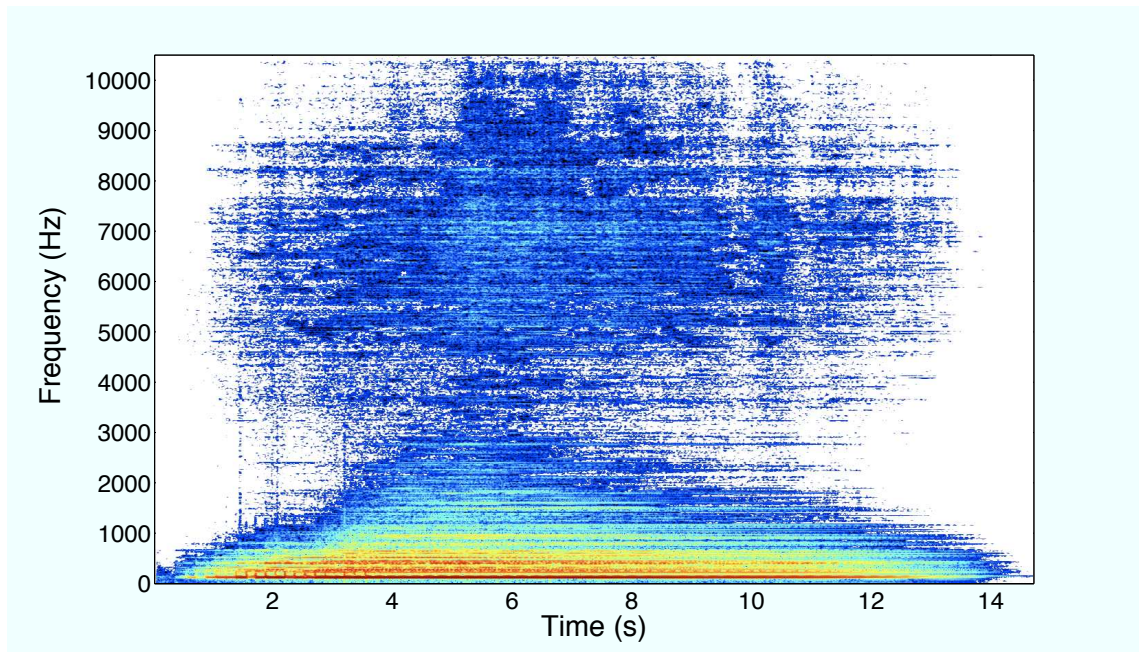


Fig. 2.14 Spectrogram of x_{92}

could be decoupled, with the added interest of extracting the transient grain phenomena. The former two grain types were once again extracted by the IMF3 and IMF4 modes, and their close similarity with those same modes from x_{89} can be seen in figure 2.15, where one can see the same opening-up of spectral envelope and relative spectral energy for IMF3 and stratified shape and concentrated energy for IMF4 that were seen in figures 2.7 and 2.8. In the case of the EEMD analysis for x_{89} , the lowest IMF captured almost entirely the background tape-hiss from the analog recording process – which is interesting in itself as it suggests the use of EEMD for background noise removal, in a similar and perhaps more effective use than that of EMD proper [132]. In the case of signal x_{92} , the added transient grain was quite low and in the high frequency range – thus it was quite “buried” in the noise floor. As a result, the extracted IMF1 for this signal captured both the tape-hiss noise floor and the transient grain. Nonetheless, the transient grain was completely decoupled from the signal, which is a much harder task than that of tape-hiss noise removal. This can be clearly seen in figure 2.16, which depicts the spectrogram for the transient noise. Now, as in the case of the spectral grain, I look for computational saliency to go along with the informal perceptual saliency of listening to the extracted signals. In doing so, the TFS measure

constructed to measure the relative transient grain-iness of a given sound correlates very strongly with the IMF1. This can be seen in figure 2.17 which depicts the TFS for the three perceptually identified grains. The other measured IFMs had a non-negligible TFS measure as compared to the three signals depicted here. This further confirms that the TFS measure was correlating with that part of the sound signal that was perceived as “transient grain”, rather than some other aspect of the examples.

At this point, I have now illustrated a separation of the three types of grain components from our example signals x_{89} and x_{92} . I further have pointed to a signal-adaptive way to characterize this extraction process using the roughness and TFS measures, which can be extremely useful in an on-line, real-time context and is something that I will implement in future work. For this study however, I now focus my attention on characterizing the spectrum of a given grain signal by extracting the mass and harmonic timbre features of spectral spread, flatness and centroid. Having done this, the signal’s spectral mass and its relative placement can be compared by taking the ratio of these measures to the same ones computed on the entire signal. In doing so, I have defined the *relative grain mass* which can be characterized by the two equations given by

$$\frac{F_g(L)}{F(L)} \quad (2.7)$$

and

$$\frac{S_g(L)}{S(L)} \quad (2.8)$$

where F_g and S_g result from applying equations 2.1 and 2.4 respectively to the given grain signal. Similarly, the *relative grain placement* is defined by the same ratio as applied using equation 2.3.

As an example, we consider the relative grain mass and placement from the extracted grains of x_{92} . Looking at figures 2.18 and 2.19 respectively we see the relative spectral spread for all grains and then focusing on spectral and iteration grains. One can clearly see that the spread of the transient grain signal is far greater – which is consistent as it is a much noisier in the broadband sense of the word – and further that the dynamic behavior of these grains differs: the iteration grain is quite consistent throughout, the spectral grain has an arch form that follows the “main” signal, and the transient grain

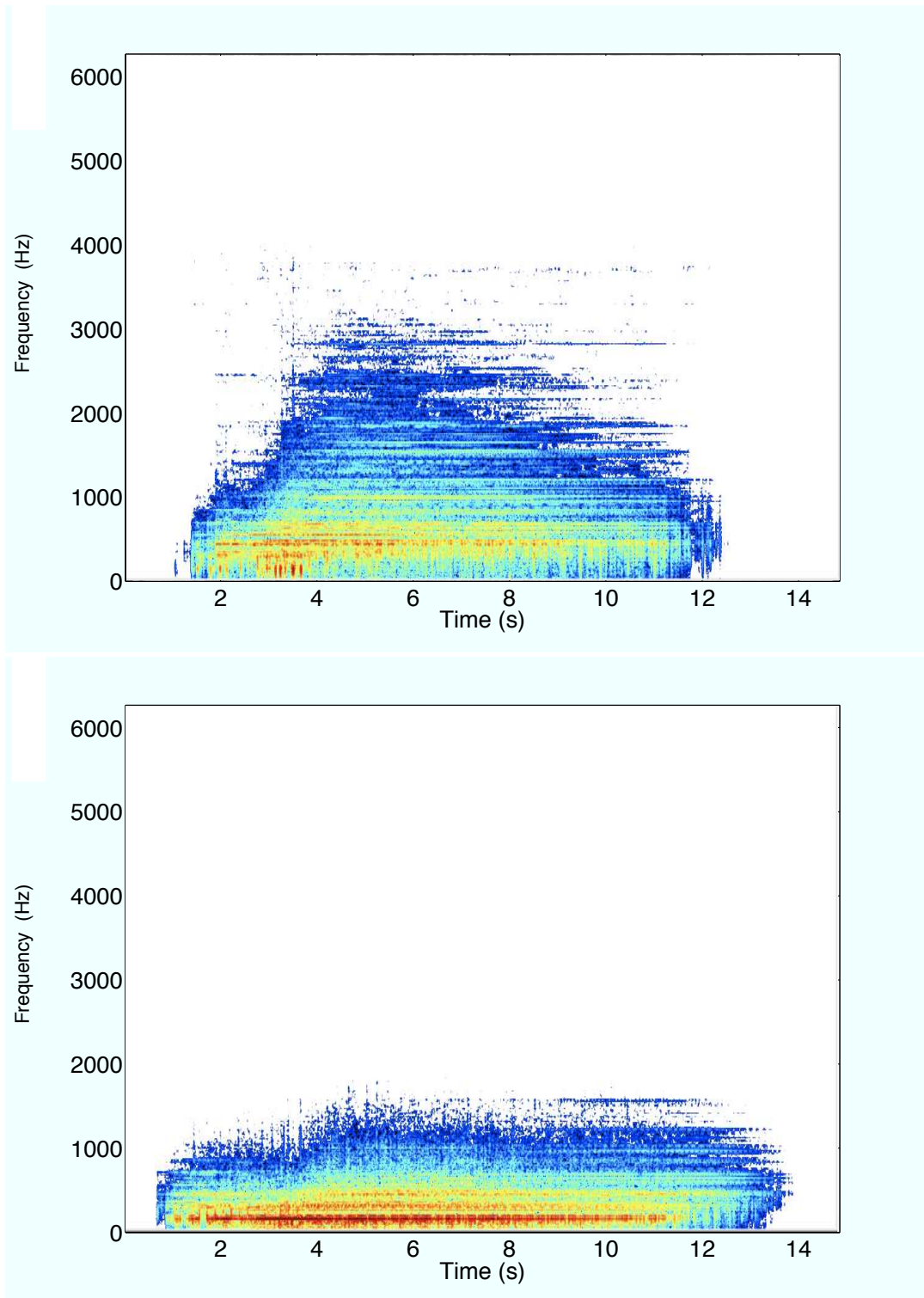


Fig. 2.15 Extracted spectral and iteration grains for signal x_{92} , captured respectively by IMF3 (top) and IMF4 (bottom) of the EEMD analysis with $\text{max sifting} = 10$

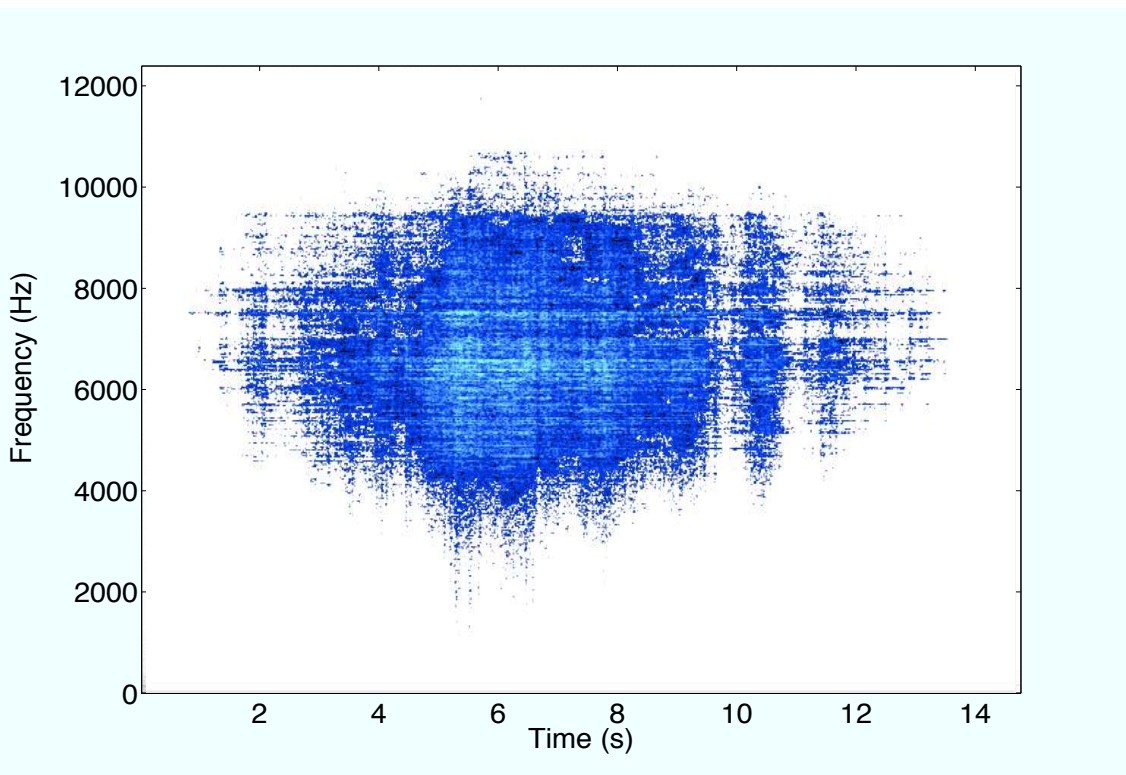


Fig. 2.16 Spectrogram of transient grain for x_{92} ($f = 0$ to 12.5 kHz, $N = 4096$)

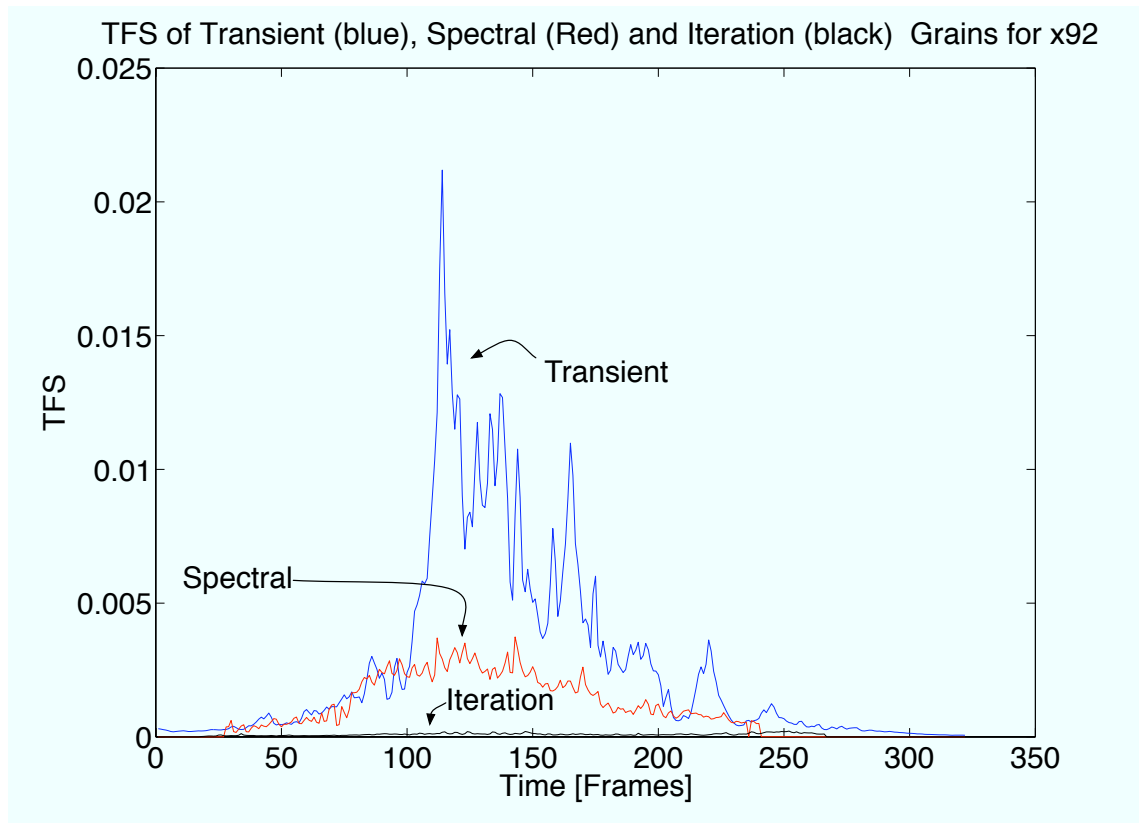


Fig. 2.17 TFS measure for identified transient, spectral and iteration grains for signal x_{92} . The transient grain TFS is much higher, while the iteration grain TFS is nearly zero.

has an opposite “concave up” shape. This serves to illustrate the fact that the “shimmer” is tied to the overall form in this sound, the iteration is a fixed mass and the transient grain is present throughout but accentuates the beginning and end of the sound. Note that this sound set is a limit case in that they share a very similar grain-based gestural contour overall – more distinct sonic gestures produce much more varied relative grain shapes.

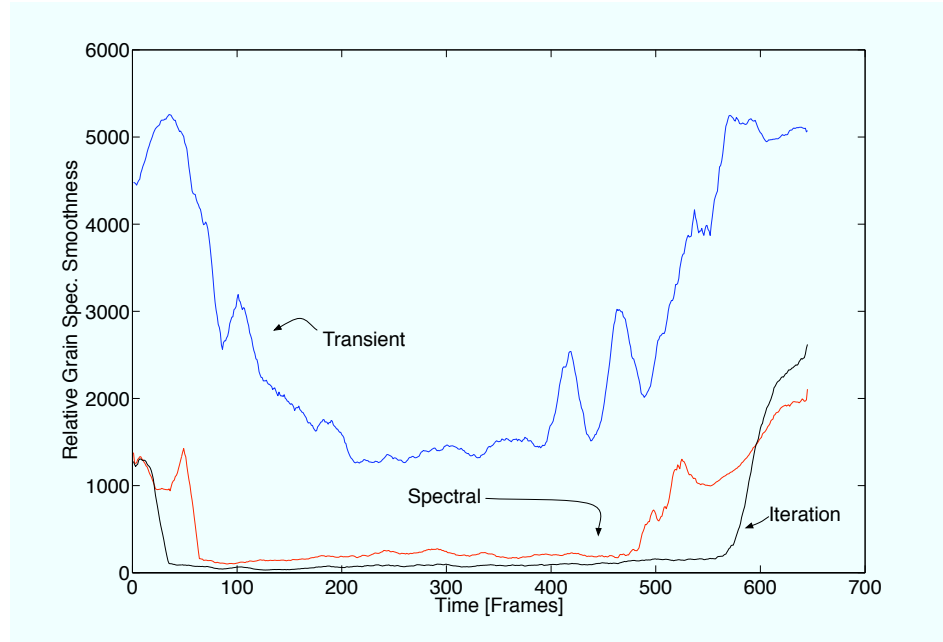


Fig. 2.18 Relative Grain Spectral Spread for all perceptually relevant grain signals.

Similarly, the relative grain placement illustrated in figures 2.20 and 2.21 illustrate the manner in which the transient and iteration grains remain fixed at the upper and lower ends of the spectrum while the transient grain partly defines the form for sound x_{92} .

Now, we can describe the relative “weight” of the grain by taking a ratio of the power of the extracted grain signal to that of the overall signal as a function of time. To do so, we compute the *relative grain weight* (RGW) for the L^{th} frame as

$$RGW(L) = \sqrt{\frac{\sum_{i=(L)N+1}^{(L+1)N} g[i]^2}{\sum_{i=(L)N+1}^{(L+1)N} x[i]^2}} \quad (2.9)$$

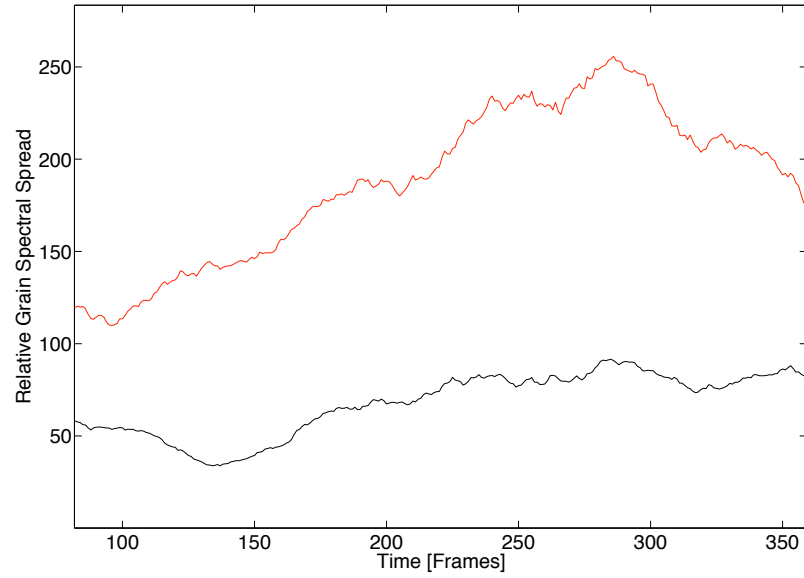


Fig. 2.19 Relative Grain Spectral Spread: Expanded view of spectral (top) and iteration (bottom) grains.

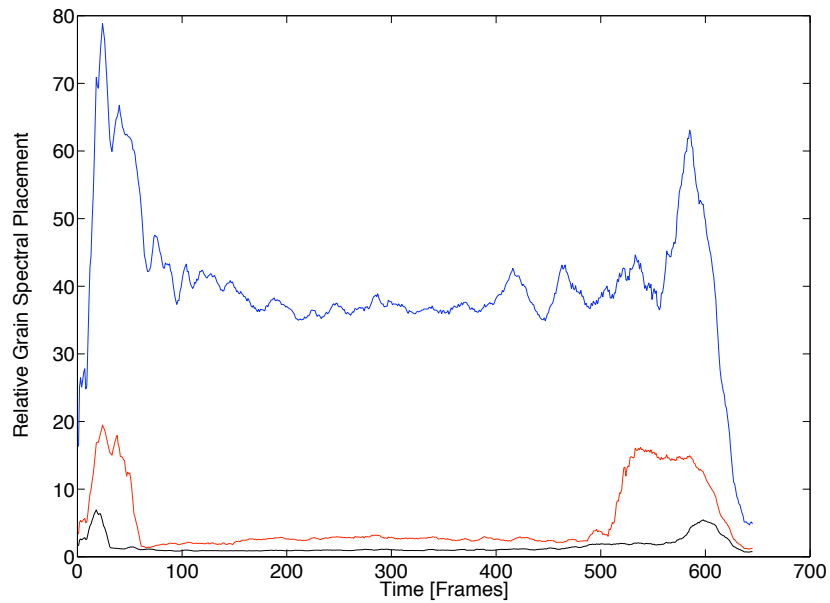


Fig. 2.20 Relative Grain Spectral Placement for transient (top), spectral (middle) and iteration (bottom) grain signals

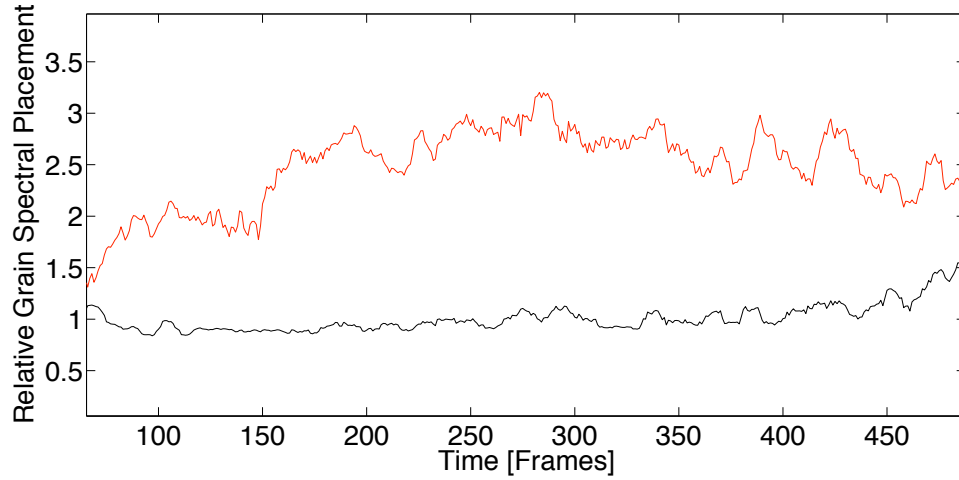


Fig. 2.21 Relative Grain Spectral Placement: Expanded view of spectral (top) and iteration (bottom) grains

where $g[i]$ is the extracted IMF function representing the given (transient, spectral or iteration) grain signal, $x[i]$ the entire signal and N is the window of observation.

The relative grain weight for the extracted transient grain of IMF1, the spectral grain of IMF3 and the iteration grain of IMF4 for the steady-state portion²² of signal x_{92} are plotted in figures 2.22 through 2.24. Notice that the weight of the iteration grain is clearly the most prominent with the spectral grain being about half the power ratio, and the transient grain being by far the weakest. This is certainly perceived when one listens to x_{92} as well. Recall that EMD partitions a signal such that it is a summation of the IMF functions. Thus the power of the grain signals is partitioned accordingly, though it is not done so additively as the RGW measure of equation 2.9 is not distributive. Moreover, the RGW contours from these three figures illustrate the dynamic nature of the relative grain level: the stronger iteration grain has a sharper attack than the spectral grain while both follow a contour that more closely resembles the temporal envelope of the signal; meanwhile the transient grain comes on more slowly and then sustains. Again, this can be heard easily in listening to x_{92} . This *dynamic grain* quality of RGW— and to a lesser extent the dynamics of relative mass and placement — are central to creating gesture/texture motion in the sense of Smalley, and these are the perceptual features that I seek to influence through the control

²²From approximately 1.16 to 12.07 seconds.

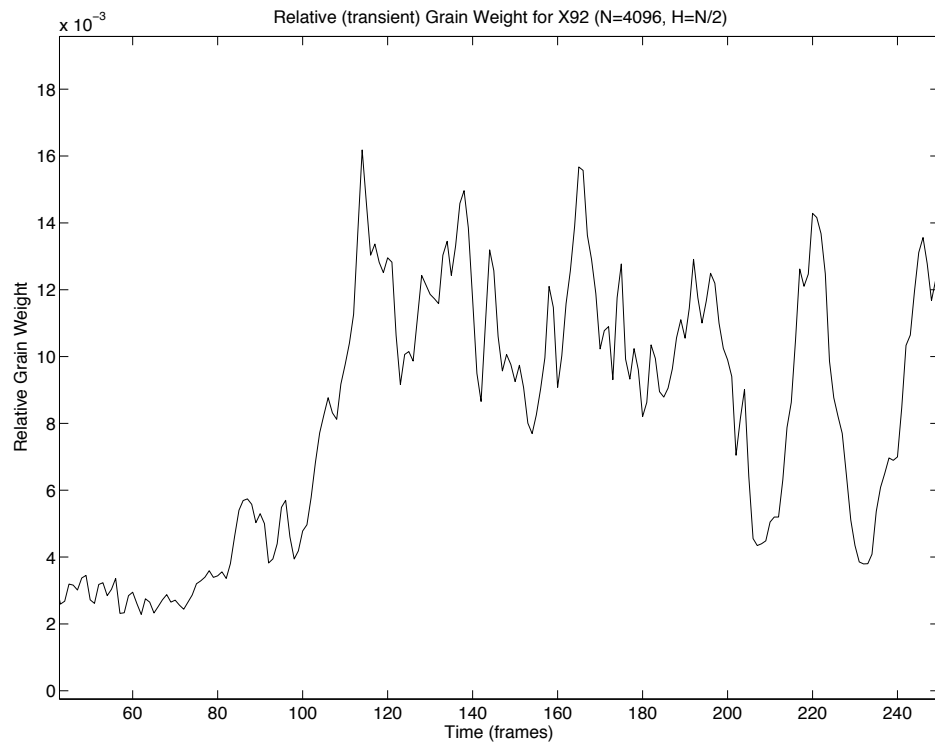


Fig. 2.22 Relative grain weight for transient grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.

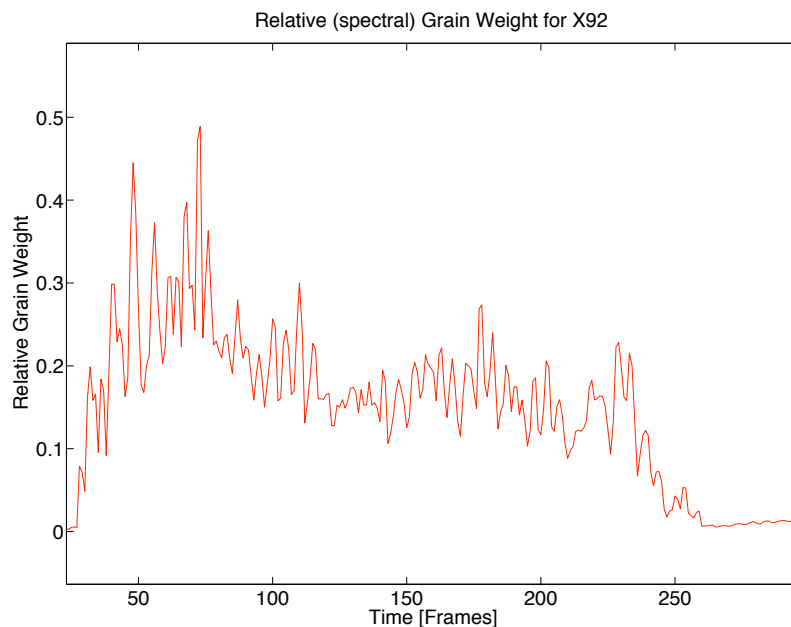


Fig. 2.23 Relative grain weight for spectral grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.

structures presented in chapter 4; it further motivates the sound model defined in chapter 5.

Now, recall that the quality of motion was defined as a generalized notion of vibrato or tremolo. Thus motion arises as one of these two phenomenon, or as some complex spectro-temporal fluctuations that are closely related to grain. Given the definition of IMFs as AM-FM components, one might think them ideal for extracting and representing these modulations. In fact, the more complex modulations that one might call motion – related to spectral envelope modulation [133][90] or cross-channel modulations – tend to be extracted *along with* the given grain that is modulated. For example, there is an element of motion for the spectral and iteration grains of figures 2.7 and 2.8 respectively, and these are both visible in the spectrograms around 400 Hz to 1kHz. Similarly we can see the motion aspect of the x_{92} transient grain in figure 2.16 across the spectrum and throughout the life of the sound. We hear these as *motion of the grain itself* and so it makes sense to preserve them within the relevant grain IMFs, so that an analysis of the dynamic grain will reflect this as well. At the same time, we

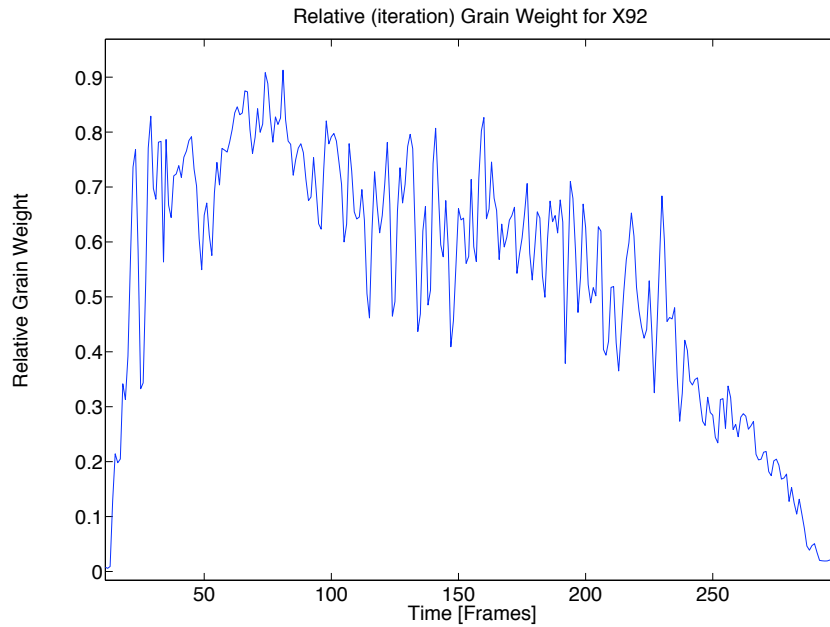


Fig. 2.24 Relative grain weight for iteration grain as determined by EEMD analysis and subsequent power ratio measurement. Window Size = 4096, Hop = 2048.

may indeed extract modulatory behavior utilizing the higher IMF functions, as their definition suggests. However, the signal-adaptive nature of the EMD technique makes this difficult to define an algorithm on the entire signal, or one that extracts precisely the global frequency or amplitude modulation – that we might call motion – within one IMF. For this reason, I compute a parallel EMD on the RMS amplitude envelope of the signal in order to filter out grain-rate frequencies and focus on the large-scale amplitude motion that is directly tied to a sound’s form/dynamic profile as well as directly indicative of modulatory gestures²³. The process is illustrated in the following example: the amplitude envelope of signal x_{89} as depicted in the top figure of 2.25 and based on a window of 186 ms and a hop size of 5.8 ms, is first extracted. The EEMD analysis of this envelope yields a 7-component IMF representation. From this, the sum of IMFs 6 and 7 very clearly represent the overall trend of the envelope as can be seen in the middle figure from 2.25. After removing this trend, we are left with the

²³While we focus on amplitude envelope, this same technique could be applied to e.g. a fundamental frequency curve as well.

modulatory signal as in the bottom figure from 2.25. Using a normalized ACF, and searching for the first non-zero lag maximum, we can clearly find a global amplitude modulation of approximately 0.76 Hz. A visual analysis of the envelope from the top image of 2.25 confirms this rate, and thus we have extracted the modulation rate for the global motion of the signal. The value of this first peak provides us with the “depth” or energy of this modulation, and we may further set a user-defined threshold under which we do not consider the signal to possess a coherent modulation.

This technique is effective for extracting strong tremolos, and can successfully characterize the modulation from signals x_{89} and x_{94} – the two examples with clear amplitude modulations – which can again be verified through a visual analysis of their amplitude envelopes. This simple technique can be applied to the dynamic profiles of spectral features as well, and can be considered as our first global temporal feature. We will now consider several others in order to define a sonic gesture profile and complete this electroacoustic-specific sonic gesture analysis framework.

Form and Dynamic Morphology

The Schaefferian concept of Form relates to the dynamic quality of the intensity over the life of the sound as well as motion, which articulates this. The notion of dynamic form is further articulated through the concept of variation, in regards to melodic or mass-based dynamic profiles. Taking up this concept of sound object dynamics, Trevor Wishart discusses the concept of *dynamic morphology* [134] which he describes as existing in a sound object when “...all, or most, of its properties are in a state of change...”. The author then goes on to consider the quality of motion as a special case of dynamic morphology.

Now, a closer look reveals that the concepts of mass profile and Wishart’s dynamic morphology are quite similar in that they both relate to variation of spectro-temporal properties. The latter seems to focus on any flux or change in said qualities, while the former seeks a sense of profile or form that gives shape and definition to this variation²⁴. This distinction may be a musical one, as much of Wishart’s music (such as [136] or [137]) focus on interpolation and constant movement between recognizable

²⁴We could also relate this to the concept of *spectral glide* from [135], which relates to timbral dynamics.

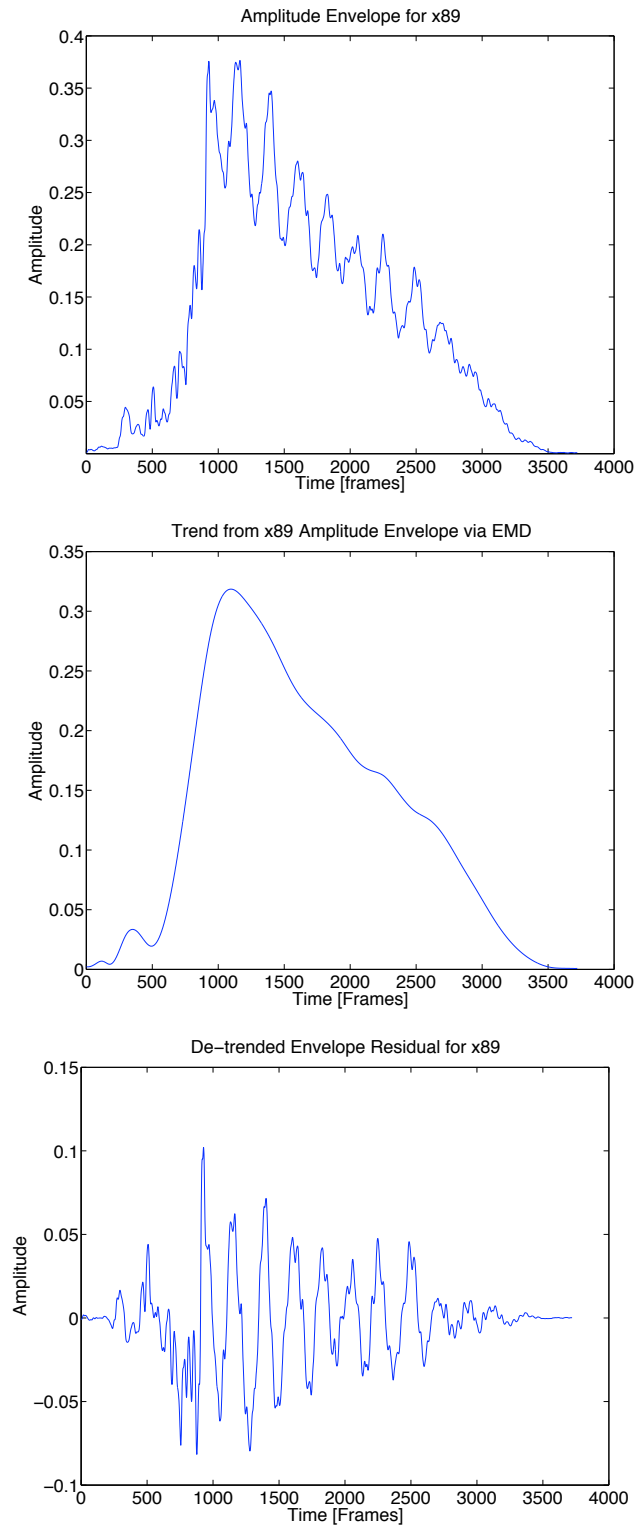


Fig. 2.25 RMS envelope for signal x_{89} (top) the EEMD-extracted trend (middle) and residual (bottom).

sound objects. My intent here is to consider the gestural nature of varying sound properties *within the life of a sound object*. This means a search for form in the sense of dynamic profile and motion (for sustained sounds), but also mass/HT profile as well as the profile/form for grain in all of its definitions as constructed in this chapter. Thus I will adopt the term dynamic morphology with the understanding that I am looking not simply for variation of sound properties (as in the definition of [134]) but rather at the Form of morphological properties in order to characterize sonic *gestures*.

Property	Description	Sub-Type	Features
Matter Proper:			
Mass	Spectral Complexity	Pitch Saliency Line-likeness Expanse	First Peak from ACF Spectral Flatness Measure Spectral Spread
HT	Spectral Balance	—	Spectral Centroid
Matter/Form Boundary (Sustainment):			
Grain	Microstructure	— Transient/Spectral/Iteration Grains Grain Classification RGM RGP RGW	Temporal Fine Structure IMFs from EEMD Roughness, TFS IMF to Signal Mass Ratio IMF to Signal Centroid Ratio IMF to Signal Power Ratio
Motion	Modulation	Rate Depth	ACF of Envelope EMD Peak of Envelope ACF

Table 2.1 Morphological Qualities and their Corresponding Sound Features

Given the complete set of sound features that describe the qualities of mass, HT, grain (in its multiple forms) and motion – summarized in table 2.1 – I will now turn the attention towards the construction of features to describe Form in order to fully characterize sonic gestures.

To this point I have focused on defining sound features that arise from perceived morphological sound qualities using an ideal set of examples, with the intention of using the most salient qualities as measures in defining control structures focused on what we might call “electroacoustic instruments”. The approach to Form-based features must necessarily differ from this, however, as the goal is not to create gestural shapes that are exactly like the given sound examples. Rather, the idea is to examine sonic gestures out of real-time as a way to compare and contrast them. In this way the focus turns more towards classification, and there is large body of work on defining temporal features of note-level sound events for this purpose. In particular I leverage

the work done on the Cuidado project [114]. At the same time my interest in instruments that define gesture/texture through influence of Form/matter properties distinguishes my approach: beyond intensity/amplitude envelopes, I examine the shape of all morphological properties as these co-vary due to the interaction between form and matter. As an example of this, consider figure 2.26 which depicts the RMS-extracted amplitude envelope for all example sounds: the top-most is original sound x_{89} , followed by the two examples of dynamic profile alteration (x_{90a} , x_{90b}), then exaggerated mass (x_{91}), grain (x_{92}), HT (x_{93}) and motion (x_{94}) respectively. Example x_{90b} has an “exaggerated form” in the sense of an extended dynamic profile. Through the transformations that produce this sound (acting on amplitude and spectral envelope, perhaps overlap-adding different tapes such as with the phonogene[138]), other factors are affected as well: specifically, this example has a particularly salient measure of roughness (in regards to the spectral model of [129]), spectral smoothness and spectral flux[118]. These measures correlated only with the HT example, but not with the other matter or sustainment-related features. Therefore, there is an interaction of matter and form that must be considered, and we cannot think of these two properties as truly separable.

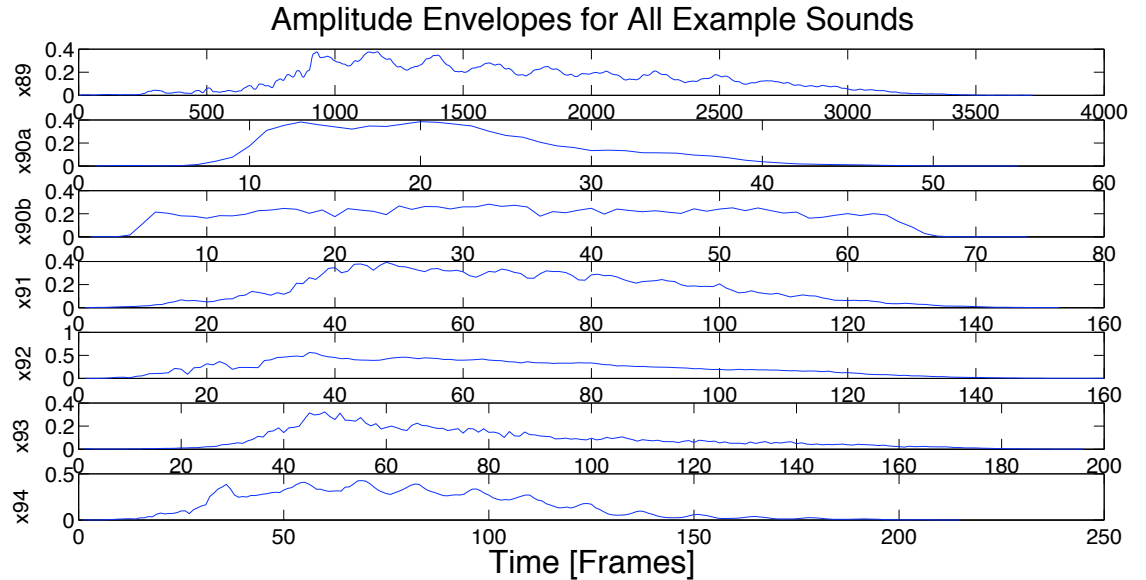


Fig. 2.26 Amplitude Envelope for all example sounds taken from Le solfège de l’objet sonore ($N = 8192$, $\text{Hop} = 4096$, $F_s = 44.1\text{kHz}$)

With this in mind, we consider the dynamics both of the extracted amplitude envelope $A[n]$ for a given sound as well as our matter/sustainment features in order to construct an image of sonic gestures. Looking again at figure 2.26 it is clear that these sounds, irrespective of their lengths, are variations on one general form of temporal envelope. In order to capture this subtle morphological difference as well as more disparate sonic gesture types, I look at a set of global temporal features that anticipate what classes of gesture to expect. The goal of these features is not an exhaustive classification of all gesture-sound shapes, but rather minimal information to allow for a differentiation between fundamentally different sonic gestures, including their internal spectral development. In order to establish this in a way that is considerate of an electroacoustic musical context I return to Denis Smalley who – in the context of his theory of spectromorphology – has differentiated between sounds whose dynamics can be characterized as *attack alone*, *attack-decay* or the so-called *graduated continuant*. The first refers to a sound object-gesture which is purely impulsive while the second is characterized by an attack with sharp decay in the *closed* case and a gradual decay in the *open* case. The third term refers to a focus on the continuant or sustain portion, and is characterized by the presence of this portion as well as onset and termination. Smalley notes that at least three variations exist: (1) a “compressing” of the onset to include more energy near the beginning, (2) a lengthening of the continuant phase, thereby drawing interest away from the onset and finally (3) “increasing the spectral energy towards termination, leading towards, and creating the expectancy of, a new note-gesture.” [108] ²⁵

In order to capture and classify within this general gesture-morphology set, I include certain global temporal features used in [88] to those utilized in [114]. As my intention is not to classify the perceptual nature of the attack per se (though this would be important for a real-time system), but rather to more generally characterize attack-resonance type sounds, I don’t delve deeply into the complex issue of extracting perceptual attack time [139], but rather focus on the time location the maximum amplitude, which pertains to all three of the above temporal shapes. As a first basic feature, I look at the ratio of the max amplitude to total sound length

²⁵We shall refer to these latter three by the short-names GC1, GC2 and GC3 from this point forward.

$$MT(x) = \frac{n_{max}}{T} \quad (2.10)$$

where n_{max} is the location of the max amplitude of the RMS envelope of signal x and T is the total envelope length. If this value is low, then a sound can be considered as either impulsive or a “compressed onset” graduated continuant sound (GC1). In the middle it is a balanced sound with gradual onset and either sustained excitation or long decay. This can be seen quite easily in figure 2.26: x_{90b} is very symmetrical with an MT ratio around 0.5, while x_{92} has a max close to the beginning and the lowest MT ratio, close to 0.2.

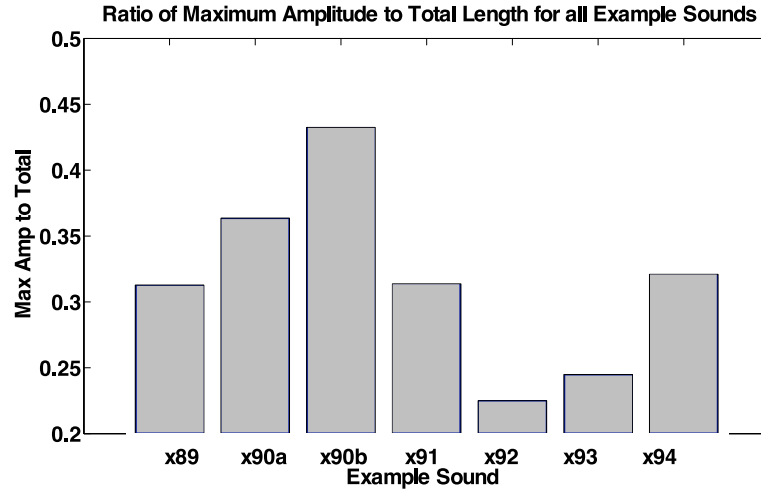


Fig. 2.27 Ratio of Maximum Amplitude to Total Length, Extracted from RMS Envelope

However this measure does not describe anything about the distribution of energy within a sound: the envelopes for signals x_{90a} and x_{94} have maxima more near the middle as compared to the other examples, yet they have a larger proportion of energy before the max as compared to the other sounds. In order to capture this, I utilize the temporal centroid, defined as

$$TC(x) = \frac{\sum_{i=0}^{T-1} n(A(n))}{\sum_{i=0}^{T-1} A(n)} \quad (2.11)$$

Using this measure, I compute the ratio of temporal centroid to total sound length as

$$TCT(x) = \frac{TC(x)}{T} \quad (2.12)$$

The relative difference of envelope balance vs. energy balance can be seen by comparing figures 2.27 and 2.28. While these two measures are enough to differentiate between impulsive or onset-focused graduated sounds, balanced sounds and termination-focused sounds, one can further differentiate between attack, attack-decay and graduated continuant sounds by looking at the temporal flatness, defined as

$$\hat{T}F(x) = \frac{e^{TG(x)}}{TA(x)} \quad (2.13)$$

where

$$TG(x) = \frac{1}{T} \sum_{i=0}^{T-1} \log(A(n)) \quad TA(x) = \frac{1}{T} \sum_{i=0}^{T-1} A(n) \quad (2.14)$$

and so this measure is the temporal analogy for spectral flatness. Figure 2.29 shows the way in which this measure faithfully reflects e.g. example x_{90a} and x_{94} 's brief sustainment compared with the other examples.

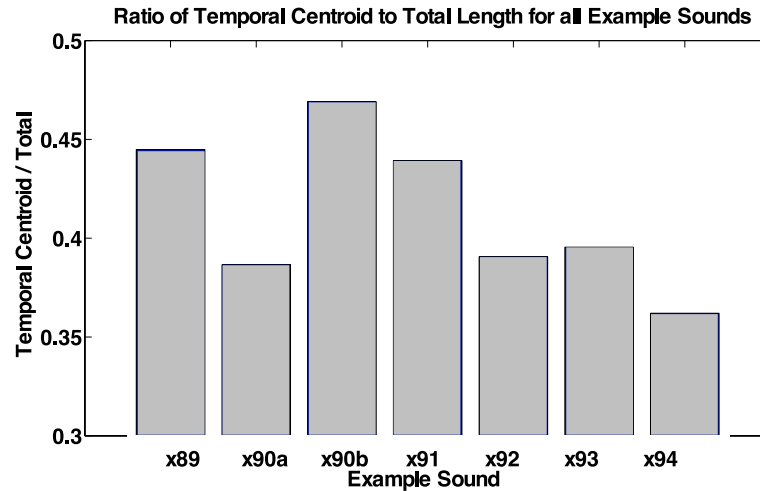


Fig. 2.28 Ratio of Temporal Centroid to Total Length, Extracted from RMS Envelope

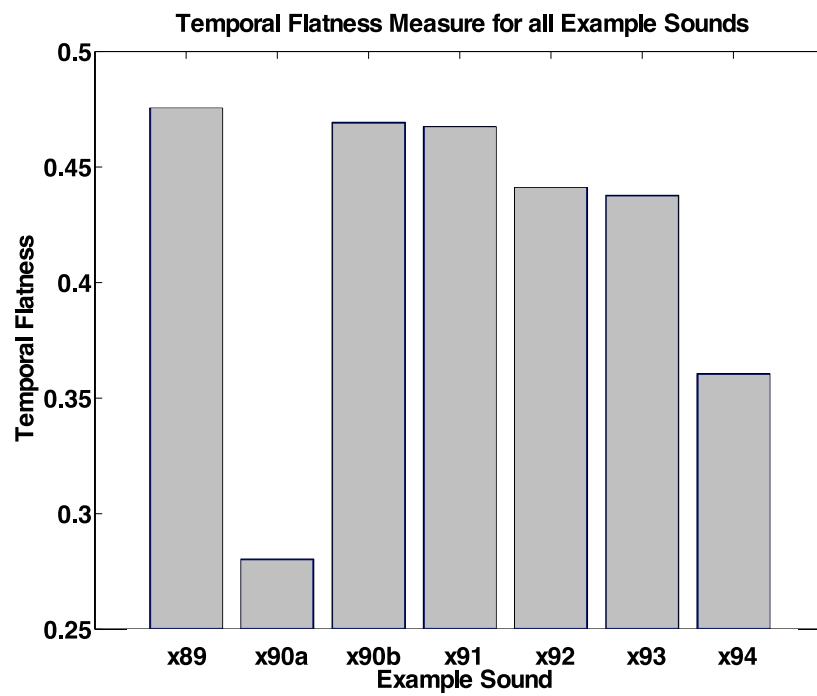


Fig. 2.29 Temporal Flatness for all Sound Examples, Extracted from RMS Envelope

In addition to describing the balance of the amplitude envelope, it is necessary to describe the rate of the onset and decrease portions. To do so, I measure the *temporal increase* and *temporal decrease* of the sound as the amplitude-weighted average of the derivate before and after the maximum is reached, respectively. For example, the temporal increase is calculated as

$$TI(x) = \frac{\sum_{i=0}^{n_{max}-1} A(n)(A(n+1) - A(n))}{\sum_{i=0}^{n_{max}-1} A(n)} \quad (2.15)$$

and the temporal decrease $TD(x)$ is calculated in a similar fashion *after* the maximum amplitude location n_{max} . These measures can greatly help to differentiate between the five types of sound gesture that I have identified. For example, while these example sounds can be considered subtle variations in that they are different types of graduated continuant, we can see from figure 2.2.4 that the drawn out envelope of x_{89} and the shortened one of x_{90a} are reflected in their respective temporal increase/decrease measures.

The dynamics qualities that I have extracted from the amplitude envelope can be used to describe the *spectrotemporal* profile of the sound as well. While the sound examples are, again, quite similar in regards to mass, we can see from figures 2.2 and 2.3 that their spectral qualities do have a different form. Similarly – and of particular interest in this work – we can see from figures 2.18 to 2.24 that the relative grain mass, placement and weight for the respective grains can vary within a given sound as well as across sounds. As another example, figure 2.31 shows the TFS measure of the transient grain portion of a double bass played in several ways (bowed, muted and martelé), as well as the TFS for the next-highest IMFs. This illustrates two important aspects: that the qualitative differences in the gestural nature of the control are articulated in these transient grain measures, and that the relative importance of each IMF extracted from the EEMD changes depending on the player’s gestural actions. Note the swelled onset nature of the martelé (i.e. a GC1-type sonic gesture) causes the grain aspect to be captured in a different mode function – a potentially useful tool for input gesture identification that warrants further investigation in future studies. For this work I maintain my focus on sonic gesture analysis, and in particular spectrotemporal envelopes such as in this example were the focus in designing the mapping/control

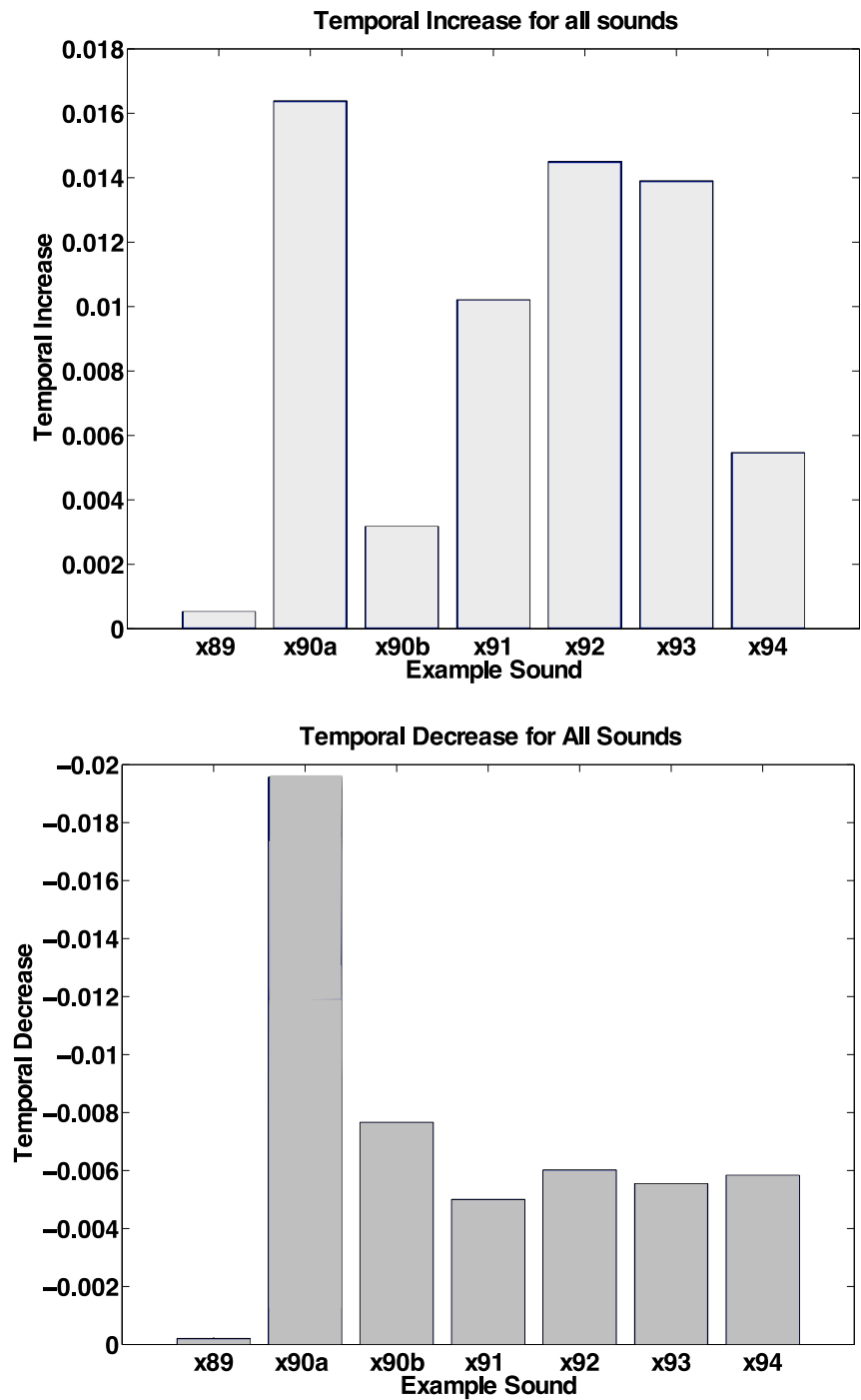


Fig. 2.30 Temporal Increase (a) and Decrease (b) for all example sound objects

structures of chapter 4 – where such grain qualities as well as matter profile and motion helped to characterize and describe a given sonic gestural response, and in so doing defining the musical control context. In particular, movement of grain measures in counterbalance to mass and dynamic profiles are of importance in the context of real-time control of electroacoustic music, as we will see in the upcoming user studies.

Examining sounds (e.g. output of a software instrument) in the context of these five generalized sonic gestural shapes is a means to consider them from a reception point of view as well as understand the gestural nature of their creation. There are many spectro-temporal morphologies possible within each of these classes of sonic gesture, and the process of directing and conditioning this space of possible Form/matter dynamics should be properly considered as part of the mapping process in the context of instrumental design. Certain sonic gestures are indeed suggestive of particular control gestures. Therefore, even while we have thus far re-considered the gestural nature of sound objects from a sonic-phenomenological perspective – not pre-supposing an acoustic instrumental paradigm – it is useful in the context of an instrumental design process to re-consider the physical and perceptual nature of the control of acoustic instruments *from a sonic gesture point of view* rather than a physical gesture point of view.

2.2.5 From Sonic Gesture Back to Control Gesture

The motivation for the theoretical construction of this chapter, in which I've established a more fluid notion of sonic gesture in the context of gesture/texture (from a musico-structural point of view) and form/matter dynamics (from a morphological point of view) is that it simply does not make sense to use sonic objects as some atomic building block from which to create formal elements in electroacoustic music in the way note events are often considered in acoustic music. At the same time, the disconnect between score/note abstraction and subtle timbral manipulation is bridged by accomplished musicians who understand the shaping of sound qualities. As such it is considering my current framework in the context of the many subtle and important nuances that exist in the control/sound relationship of acoustic instruments, particularly the perceptual nature of control as well as sound for a given instrumental gesture.

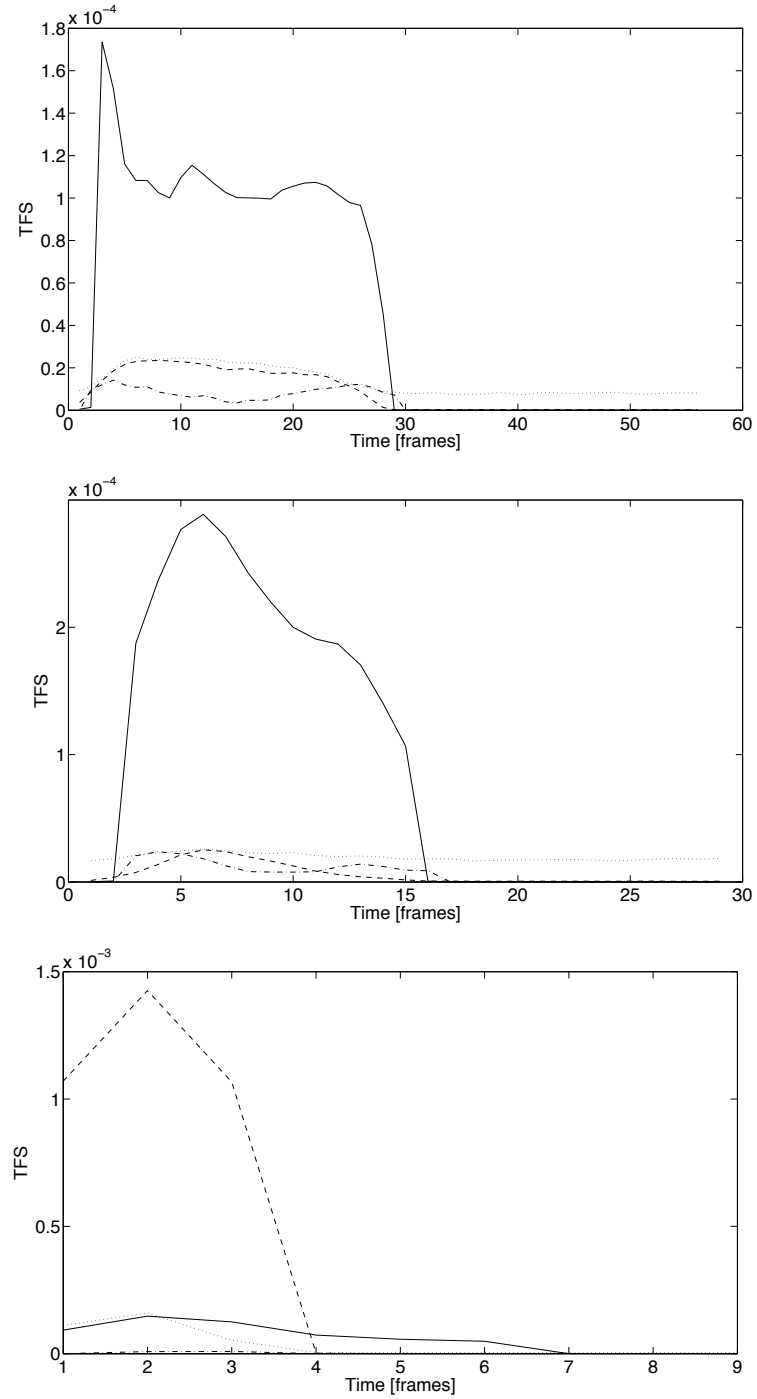


Fig. 2.31 Transient grain and next-largest IMFS for a double bass playing C1 bowed (top) muted (middle) and martelé (bottom). Temporal Fine Structure (TFS) window size was 8192, hop 4096 and f_s of 44.1k.

This is articulated in [140] wherein Levitin et al. discuss the control of a musical note-event from the point of view of the required input gestures to create a given note type and the sonic perceptual parameters that may be affected. They break this performer gesture/perceptual sound feature link down into the beginning, middle and ending of a musical tone - paralleling the classic view of attack, sustain, decay – which I in turn translate into the more abstract and aforementioned notions of onset, continuant and termination. In doing so the authors create a framework that parallels the excitation/modication gesture typology of [78] and expands upon this to include psychoacoustic cues. In the process they introduce the notion that an *explicit* beginning is contrasted with a newly introduced event termed a *state change induction*, described as an *implicit* beginning that arises from some (contextually determined) spectral discontinuity. It is important to note that a key feature of an implicit beginning is that there must first be a continuous excitation from which they can emerge – think of a legato between two notes of a wind or bowed string instrument as a canonical example. Similarly, one may consider the parallel notion of explicit/implicit ending, where the former may result from damping or natural decay of resonances by stopping a gesture, while the latter exists in the same moment as the implicit beginning of another event. In regards to the middle portion of a musical event, the study differentiates between a middle that follows a continuous excitation gesture and one that arises from an impulsive one. In the former case, the user is in contact with the primary resonator (part of instrument to which energy is introduced) and so can modulate pitch, loudness or timbre via primarily continuous modification gestures (e.g. various bow alterations or embouchure changes). In contrast, a middle or sustained portion of a musical event that arises from an impulsive energy source – and so when the performer no longer has access to the energy source – gives rise to potential modifications only if the performer can access the tone generator of the given instrument. The authors suggest that pitch modulations resulting from tension/pressure applied to an instrument’s tone generator are the only relevant control actions in this context, but I argue that the “subtle” timbral effects that can result from e.g. shaking a resonating body are particularly relevant (though not unique) to control of EA-inspired musical instruments where such timbral and textural variations are magnified in importance through amplification. This limit case of acoustic control via sound-producing gestures is an interesting meeting point with sound-tracing type

gestures that might suggest novel control strategies.

In examining the typological thinking of [140] as well as [78] and [84] we see that they consider the universe of possible gestures in a way that is mediated by the vibrating source/resonating body physical aspect of musical instruments. Levitin et al. [140] implicitly suggest a notion of musical gesture through the consideration of separable and perceptual sound parameters that are affected, and what a player does in order to elicit a given response. Translating this into the electroacoustic realm, I propose to consider the given typologies of [78] and [140] regarding the control of the beginning, middle and ending of note events as a starting point to abstract towards control of sonic gestures: what could one have done in order to give rise to a given sonic gesture? This question suggests directions for control structuring. In the course of such an exploration, there is no need to presuppose an attack-resonance instrument model, but rather consider such sounds *in the larger context* of all gestural sound events that one may control.

This more generalized context suggests a modified notion of what constitutes a singular control gesture: in the exposition of [140] there is a supposition of one single gestural action constituting a single gestural unit. However, in the case of a sonic gesture, an iterative sound object may be understood as a unit – one that suggests a continuous or iterative excitation. Among other reasons, causality is extended in electroacoustics through the potential for repetition and automation. Therefore I consider explicit beginnings to include iterative excitations as a separate category from singular impulsive or continuous excitations. Note that this is in contrast to [78], where iterative excitation is considered as a subset of continuous gesture, having discrete excitation. Similarly, the notion of a state change induction that is brought on from a spectral discontinuity is very prominent in electroacoustic music, as sudden changes in matter are often used to signal a new sonic gesture/texture context. Applying the language that we use here, such a spectral discontinuity may exist at the termination of a sonic gesture in terms of the third type of graduated continuant, in which spectral energy is increased just prior to such a state change.

Following the sound-first principle of this article, I have thus paralleled morphological descriptors, sonic gestural shapes and control gestures, as presented in figure 2.32. It describes the onset, sustain and termination phases for a given sonic gesture and what

sort of control is possible at each phase. In the case of attack sonic gestures, an impulsive excitation is the entire control gesture, and the dynamic profile and instantaneous matter features (e.g. pitch) dominate our perception. Attack-decay gestures are also impulsively excited, but there is a chance for control of the ending through damping, and for continuous excitation of the middle through the introduction of the concept of mode changing, which in this context is enacted by an abstracted form of a structural modification gesture. Continuously exciting the middle of an attack decay gesture through mode change amounts to guiding or otherwise affecting the decaying resonances. In this archetype there is added salience from the sonic features related to dynamic matter as well as motion and grain during the decaying sustain. In regards to the sonic gesture archetype of graduated continuant, each of its three variants may result from continuous excitation or from an implicit beginning and by definition extend into a continuously excited middle (CEM). The swelled onset nature of the first type means that it does not result from an iterative excitation – at least not in the sense of repeated application of the same gesture. However this may be the case for the latter two. Meanwhile the first two types may terminate from damping or stopping gestures (i.e. allowing for decay), while the increasing of spectral energy towards the end in the third type leads, perceptually, to an implicit ending. All three of these sonic gesture archetypes possess salient dynamic form/matter qualities including motion and grain, in addition to possibly dynamic grain as well.

There is certainly no need to constrain the universe of sonic gestural shapes by strict instrumental control gestures. However, just as sonic gestures give rise to an imagined embodiment, so too are they mediated by an understanding of musical control. This fact has led me to the current typology as one example of structuring control based on sonic phenomenological principles – I don't claim that it is in any way absolute or complete, but rather one application of the theory presented in this chapter. Remember that these gestural archetypes are very abstract, and do not directly describe instrumental sound or strict control. They relate to my own approach in that my design constraint for instrumental mappings is not instrumental gestures, but rather creating action-sound couplings that maintain a more generalized notion of embodiment while considering the paradigm of acoustic instrumental control. Thus while my instruments tend to have non-physical sounding qualities such as digitally-created transients (as in

the aforementioned Supercollider example), the *shaping* of such sounds is rooted in the physicality of human gesture, resulting from an adherence to immediate response and close coupling of action-to-sound (see section 3.1.1). In regards to this typology, I clearly do diverge from an acoustic instrumental paradigm as can be seen in e.g. the attack/decay gestural archetype: in acoustic instruments one “normally” does not interact with a primary resonator after excitation²⁶, yet I consider this control structure as coherent after a *mode change*. An acoustic instrument parallel would be shaking an instrument or otherwise interacting with a resonating body (i.e. secondary resonator). From a gestural point of view I consider this as switching modes between sound-producing and sound-tracing types of gestural control: the decay portion of the sound is “guided” by the influence of the performer. In order to reduce the cognitive overhead of such an interaction, I will present examples in chapter 4 that encompass a discrete change of control state in order to change the mode of the controller.

Now, at this point I have constructed a sonic gesture analysis framework rooted in Schaefferian theory that combines extraction of local spectro-temporal features and global temporal features. I’ve applied some ideas from the electroacoustic tradition to arrive at five sonic gesture archetypes, examining their control structure as well as the sound features that are most perceptually salient in the context of control. The interest lies in beginning with compositional and sound-theory ideas, and extending these to an actual computation framework to use for instrument design and real-time control. I will put this framework to use in the examination of perceptual control structures, mapping and sonic gesture type in chapter 4.

2.3 Chapter 2 Summary

Bringing these notions together presents a unified framework through which to understand as well as construct an instrument. One may consider an instrument in terms of the ideal sonic gestural shapes that it should produce, or those that it can produce. Desired gestural shapes may constrain control structures as with figure 2.32, which also suggests those elements of sonic gestures that are relevant to examine. Form/matter qualities provide a framework from which to build tools to analyze sonic

²⁶There are exceptions such as stretching of a drum membrane.

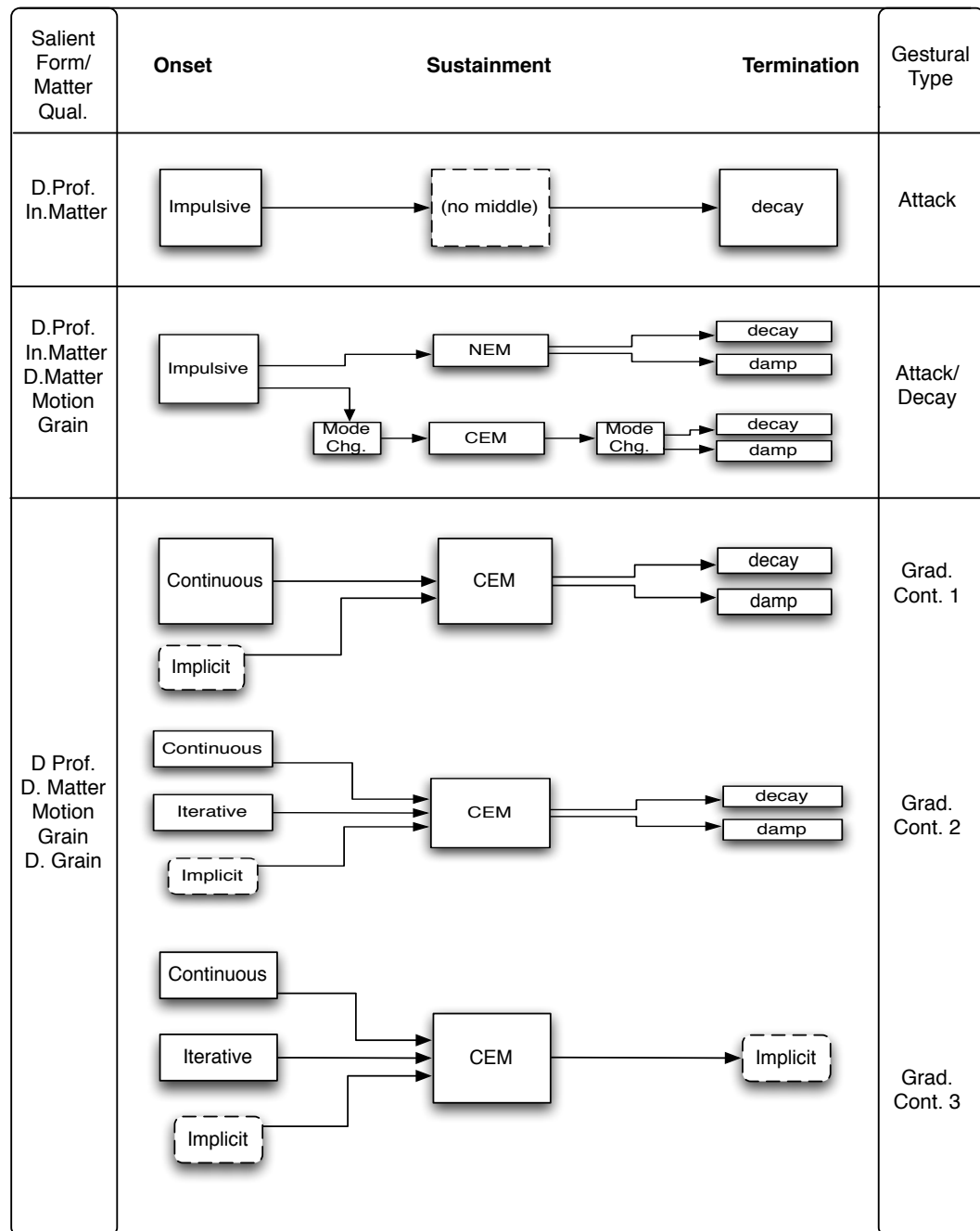


Fig. 2.32 Possible control structures for different sonic gestures types (attack, attack/decay and graduated continuant 1,2,3), including control type for all portions of a sound's life. Also included are the relevant form/matter features that are perceptually relevant and so may be affected (dynamic profile, instantaneous matter, dynamic matter, motion, grain and dynamic grain).

gestural shapes, with the archetype of gesture-shapes suggesting further methods of analysis (e.g. of attack, modulation, etc.). The details of how each dynamic form/matter quality evolves depends on musical context, and the principle of gesture/texture articulates the reception strategies at play in electroacoustic music and why using this meta-framework to design instruments requires choices made by the instrument designer while listening to sonic gestures in context. This is why I stress this phenomenological model of instrument design as a “meta” approach from which personalized and idiomatic design tools may be built.

In addition to this being a methodology for analysis of the gestural nature of electroacoustic music, we may further use this information as feedback during the mapping design process, as we again shall discuss in the upcoming chapter. What is interesting in this approach is that it allows one to directly see the effect that a particular mapping of control-to-sound gesture has on the given sound features, and to derive adaptive mapping strategies that consider an EA musical context. The novelty of this approach – which is at the heart of this exposition – is that it goes beyond “mapping” in the sense of parametric association and influence, towards a consideration of the *form of control and sound gestures* in a given musical context in such a way that properly links these two. The result is to have a more coherent and separable influence on perceived “gestural” effect in the resultant sound output. Note that this approach differs dramatically from existing works that have sought to design mappings based on perceptual criteria such as [70][141][71], as these have focused on higher-level “music” features per se, rather than on phenomenological gesture-based features as I do here, and further seek to drive sound output directly with said features rather than adapt in a feedback control loop. Also note that this work differs from “indirect acquisition” of gestures [72] in that I do not make any a priori assumptions about instrumental acoustics or physical gesture type. In fact it is quite the opposite: constructing the physical reality of instrumental control by “indirect mapping” of performer actions to desired sonic gestures.

Chapter 3

Mapping Theory

The purpose of this chapter is to contribute a conceptual framework for the notion of “mapping”, presenting a formalized and theoretical view of the subject. The results will include a new way that a musical instrument “design space” may be constructed in the context of mapping, as well as an application of the framework to existing works on mapping from the literature. I will end with an implementation (a mapping toolbox) created as a result of this analysis and underlying theory. I will not discuss here the multitude of controllers or transducers that have been used for digital music performance [142] as this is an entirely separate topic unto itself. Rather, the important and often-overlooked question of how control signals from such devices are *structured* and *associated with sound output* (these being two distinct but related sub-topics) will be considered in detail. There are various points of view on this matter, and while I cannot possibly present all viewpoints thoroughly I will discuss fundamentally different ways that control of sound processing may be structured and conceptualized. I will present the software tools whose development were based on some of these ideas, while perceptual studies and more specific applications are reserved for the next chapter.

An important aspect of this theoretical discussion is that it considers mapping both from a user-perceptual point of view as well as a functional context, where the main importance of this lies in the distinction of these two views. The basis for relevant perceptual issues was primarily covered in the previous chapter, while the functional approach will be covered more in-depth here, providing a mathematical point of view. The purpose of such formalization is to move towards a clarity in design concepts and

language, as well as to move towards an instrument design methodology for situations in which one has conceived of their instrument *from a parametric point of view*.

Naturally the particulars of mapping structure for each digital instrument are determined by musical intent and context, and for the purpose of my own work I will relate the relatively abstract realm of mapping design presented in section 3.3 to the physically and perceptually grounded notions of control and sonic gesture that were presented in chapter 2. Mapping can then be more clearly seen as a linkage between physical action and sonic result. In this sense, my goal is to translate the discussion on mapping so that it links an abstract and formalized approach that is intended for representation and conceptualization to a viewpoint that considers mapping in its role as perceived correspondence between physical (i.e. acting on controllers/transducers) and sonic materials, which at its heart is driven by our cognitive and embodied understanding of the acoustic world.

3.1 What is Mapping?

As can be seen in various discussions [7][143], mapping is generally considered as the correspondence one creates between the output control parameters of an input device or control interface [142] and the input parameters of the chosen sound processing technique. Moving beyond the notion of correspondence, the precise meaning of mapping changes depending upon one's intent while its role can similarly change depending on one's relationship to the notion of "gesture" – both human input and resulting sonic gestures – as well as the choice of controller and sound processing algorithms. I use the term "algorithm" purposefully here, as this discussion is restricted to digital-based performance systems in following the discussion of section 1.1.

To my mind, the relevance of this discussion arises from the amorphous nature of computer-based "instruments", where the idea of instrument itself as well as the notions of system, composition or tool are all not well-defined, and where these shifting categories and blurred boundaries give rise quite naturally to questions about the choice of appropriate language, design process and even the reception or understanding of creative works that rely on computer-based performance systems. While I find this a beautiful "problem" to have – and in spite of this variance – there still do exist

commonly-held assumptions and beliefs about musical performance practice and even overlapping aesthetic interests that make a deeper discussion of the conceptual structure and role of “mapping” not only possible but worthwhile. Rather than consider all possible musical avenues, I will maintain a generalized framework for the overall discussion (and tool development) while delving into more specific applications, examples or explorations in the context of my own musical aesthetic. Not wanting to make the discussion strictly mathematical or compositional, rather than strongly privilege one point of view I will strive to balance the discussion by way of mathematical, musical, perceptual and phenomenological perspectives.

3.1.1 Musical Control Context

In the design of interfaces for expressive musical performance and control there are many factors that need to be taken into account. At a fundamental level, one must consider the musical context within which the performance will be presented, and the expressive goals of performer and composer. This will in part determine the *level* of control - whether it be guiding some higher-level musical processes or controlling the moment-by-moment production of sound in response to a performer’s gestural input. The former often necessitates the consideration of completely new metaphors [144] for interaction while the latter can more directly find inspiration in the paradigm of instrumental music performance and/or that of ecologically-based object interaction. These two levels exist at either end of a spectrum with a systems-oriented view of interaction design at one end - wherein the interface as well as the compositional process are together taken as the larger system - and an instrumental viewpoint at the other. In considering the instrumental perspective, the performer has direct control over the smallest details of performance [110], and the system in question is the short-term dynamic human input as well as the energy-dependent sound processing that is activated by this. It shifts the design focus to what we might call the musical performance “transfer function”. Where along this spectrum the musical interaction context lies will further determine, from a designer’s perspective, the role that mapping plays. For example, when one considers an interactive musical system in which gestural control affects a set of probability distributions that in turn affect musical phrases, then “mapping” is further abstracted and relates as much to the composition of a system for

generating new musical works as it does to instrumental design. It is suggested in [145] that the mapping concept becomes limiting in digital “instrument” design in that more novel and contemporary designs tend to have intermediary and indeterministic functions which do not have a clear mapping of human input to sonic output. While this may often be the case, I would disagree slightly in that even highly complex interactive systems are often comprised of global networks of local control-to-sound mappings that behave “nearly” deterministically or causally from a perceptual control point of view. That is, a direct response to physical actions might have a correlated sonic response, which may then evolve independent of the performer’s actions or even of the performer’s intentions. I would further argue that this relative degree of immediacy is closely linked to a system’s ability to produce novel human-like gestures (i.e. organically shaped in the moment and not called from a set list of possible outcomes). In light of this, mapping is a useful concept beyond the design of acoustically-inspired instruments, and its relevance (from a design point of view) is a product of

1. The level of control over the sound production mechanism, from the lowest signal parameters, through symbolic musical parameters up to parameters affecting phrases, movements, sound palette, etc.
2. The temporal window over which control-sound events take place.
3. The perception of causality in action/response and how repeatable this cause-effect relationship is.

The notion of mapping in relation to temporal scale will be discussed in more depth in upcoming sections, as it closely relates both to the *quality* of a given sound signal one intends to control – be it timbral, textural, pitch-based or dynamic form – and also to the notion of *hysteresis* in a sound synthesis method or in the mapping itself. The temporal scale and musical “level” over which one notices sound qualities or control response is where the focus of this discussion lies and delimits the theory presented going forward. This is important to be aware of, so that it is clear I am not extending to those discussions in the computer music literature that use the language of mapping in the context of compositional practice. One clear example of this can be found in [146], wherein the author regards mapping as a creation of relationships between musical gestures and larger compositional forms. This highly abstracted view does not

directly come to bear in this discussion of real-time control, centered around the creation of a link between performer and sonic gestures. While compositional and music theoretic ideas did arise in the last chapter in the context of the gesture/texture discussion, “mapping as composition” does not relate in that it exists out-of-time in the sense of acting in the space of musical ideas, and in actualization it is well outside the physical realm of human action/response – beyond the bounds of *perceptual temporal support* for immediately-perceived human action.

I want to stress this at the outset of the chapter, as the more formalized discussion to follow (particularly section 3.3) at times moves away from explicit discussion of time or musical context. I don’t want the less mathematically-oriented reader to become bogged down in details of the formalism, so I will ask such a reader to keep in mind that I will revisit these notions of time and musical context more extensively in the next chapter. For now, suffice to say that where the discussion is leading is towards a musical performance context that can be considered as *instrument-like in feel, but novel in control and sonic response*. That is, my design goal is to retain some of the immediacy of acoustic performance discussed above with a particular focus on so-called novel controllers [2], with the intention of controlling inherently digital aspects of sound control: timbre and texture rather than classical musical parameters such as pitch, note-level harmonization, rhythmic timing and so on.

3.1.2 Perspectives on Mapping

In the context of instrument design there are several different ways that we might conceptualize the mapping component, which in turn will affect the strategies we may use for associating control and sound processing parameters. For example, in [145] Chadabe differentiates a *hierarchical* point of view – in which a few high-level parameters unidirectionally control many low-level parameters – from a *network* point of view in which all parameters can equally send and receive control information in a non-deterministic manner. This viewpoint assumes the perspective that mapping is a series of correspondences, or the out-of-time snapshot of input/output control potential. I refer to this as a *systems* view on mapping, and would relate this to two other distinct views as follows:

- **Systems** point of view: the liason or correspondence between control and synthesis parameters. This view is represented by the classical “flowchart” paradigm that is ubiquitous in engineering.
- **Functional** point of view: defined by the operations that associate variables from a domain set with some desired target or range set of variables, possibly with intermediate steps. These sets can be endowed with properties themselves, which may be inherited, constrained or altered by the operations that map between them.
- **Perceptual** point of view: the sensation of human intention leading to some representative sonic result. Therefore this is integrally tied to the notion of gesture, including ancillary gestures. Unlike a perceptual parameter (e.g. pitch) a perceptual view on mapping can only be identified *after an action has taken place*, and so there is no quantifiable description a priori.

Quite naturally all of these things interrelate, and this division itself can be seen as hierarchical with the latter term being more “high level” than the first two. The perceptual viewpoint is purely phenomenological in the sense that it is determined first by what one *experiences* and then extends to the underlying mapping that is imagined. In contrast, the first two points of view are the means by which a particular mapping is explicitly constructed, and so one looks first to the operations and actions that give rise to a desired response or experience. It is quite possible that two mappings may be the same from a perceptual point of view but differ greatly from a systems or functional point of view. For example, one may map from the vertical velocity of a spatially-manipulated controller such as the Polhemus or Radio Baton through an integrator and finally into the playback position of a sample buffer, yet mapping from acceleration to this same sound feature can have the same perceived mapping if the integrator of the first functional mapping is defined properly.

It is the interplay between the perceptual and the other two more formalized views on mapping that I am interested in here. The systems and functional views act as duals of one another, and shifting between perspectives helps one to have a complete picture of a performance system. In some instances, the views are vastly different and give unique

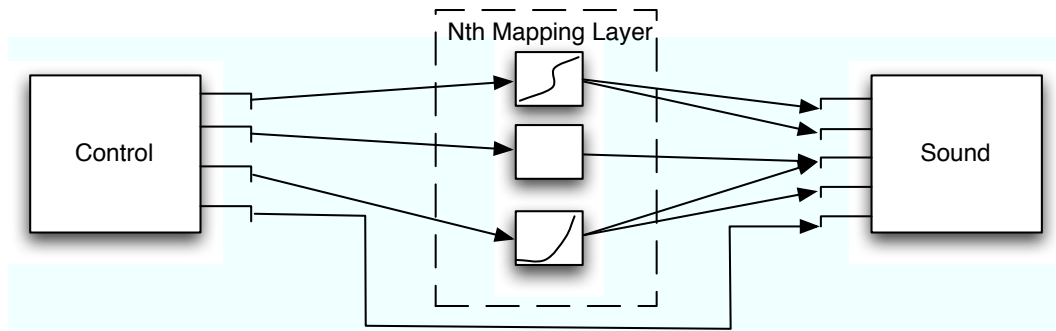


Fig. 3.1 A generalized systems-oriented view of mapping from control to sound parameters.

insights or design possibilities. To explain this further, consider the diagrams of figures 3.1 and 3.2. The first gives a generalized representation of a systems view on mapping, and the latter a functional view. While they both describe a control-to-sound flow of control, the first figure considers the number of *degrees of freedom*¹ afforded by a physical controller or synthesizer, and how each of these degrees or parameters can be associated. Also essential is the issue of *inter-connectivity* in terms of the degree of complexity of the parameter association: often discussed in the literature in terms of one-to-one, one-to-many or many-to-one mappings [7]. Further, each association may be modified, conditioned or otherwise warped in order to tune the response of a given parameter, as is illustrated by the transfer function-like representation of the “Nth mapping layer”. This fact leads to an important point regarding this view of mapping: that in addition to parameter association complexity, many layers of mapping may be necessitated in order to extract relevant gestural or perceptual phenomena, as discussed in [3], [13] and [14].

Meanwhile, a functional view is concerned with the *dimensionality* and *structural properties* of the set of input/output parameters that a mapping acts on, as well as the properties of the mapping that are inherited by or endowed upon said sets. This is illustrated in figure 3.2, which shows a *composition* of mapping functions f and g , mapping between control set X and sound synthesis set Y . Rather than individual parameter conditioning or inter-connectedness as in a systems view, the focus here is on the properties of X, Y, f, g , etc. and how they are defined within some underlying

¹In the engineering/physics sense rather than the statistics sense.

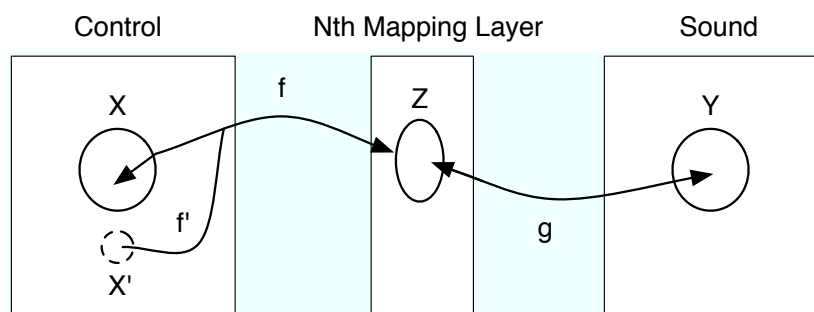


Fig. 3.2 A generalized functional view of mapping from control to sound parameters.

set that encompasses them all. One may consider similarity between sets or mappings relative to a given metric, as is illustrated by set X' and mapping f' . Further, a mapping may act on a parameter set's *algebraic structure* if it is concerned with transformations of the symbolic inter-relations within a given parameter set, such as a permutation. Such transformations are common in the case of mappings acting on musical structure, an act taken to its extreme by serialist composers. Classic examples of this can be found in [147], wherein the author relates such compositional mapping ideas as permutation and inversion in a 12-tone context to similar 12-step rhythmic transformations using the author's notion of "time point" sets of event onset times. In this discussion therefore mappings exist between measure of pitch material, and from pitch structures into rhythmic structure. One can imagine an instrumental mapping equivalent wherein controls are mapped into different configurations of *potential* pitch structures (rather than compositionally realized). In fact, this is the approach taken by the authors of [148], who present mappings that allow for dynamic tuning of a given "isomorphic controller" by altering the pitch mappings of a specialized keyboard on the fly.

In contrast a mapping and its underlying sets may be interpreted as *spaces*, giving rise to certain spatial concepts: the distance between parameters, or the idea of a "neighborhood" of parameters come into play. In this case one is concerned with a mapping's *geometric and topological structure*. Rather than focusing on symbol-based mappings that may or may not directly affect signal-level parameters, the focus becomes the definition of mappings that act *dynamically on parameter values themselves*, resulting in a signal-focused mapping approach. In this case the geometric

and topological properties of the mappings can exhibit a strong influence on the perceived response – including the excitation of a sound process, continuous modulation of parameters, morphing within a “timbre space” [18] or switching between sound states. I will focus from this point forward on this particular type of functional mapping, delving into more mathematical details in section 3.3 and later examining the influence of geometric structure on the perception of control in sections 4.1 and 4.2.3. While the functional nature of mappings will be the immediate subject of these discussions, it will become clear that a systems viewpoint must be adopted at times in order to fully explain an instrument design, while a perceptual point of view arises whenever sonic or musical intent is described or reflected upon.

How vs. What

Yet another way to distinguish the meaning of “mapping” arises in the course of designing an instrument when planning both the influence and the response of a given control variable or action. Looking again to figure 3.1, the direction arrows signify the given parameter association or “what to map where” - that is, the association of control and sound synthesis variables. Meanwhile the transfer/warping functions of the middle layer describe the (potentially) continuous response extended to any given gestural input. This distinguishes the “what” aspect of mapping in the former case from the “how” aspect in the latter.

From a functional viewpoint, a musical instrument can be seen as a collection of discrete and continuous control variables and their sonic effect. While it is certainly true that gestures acting on discrete controls can be quite a complex system in and of themselves, the underlying *control gesture* itself is often towards the end of selecting among a set of options (e.g. buttons, keys), states, or exciting a dynamic sound model which then contains most of the complexity of the system. Meanwhile, continuous control tends to produce loudness or timbral variations of the system once it is excited, moving more of the complexity to the control variables themselves and the initial control-side mapping. Therefore, to focus on the most complex cases of control structure in general, I will assume an underlying set of continuous control parameters. In this case, a set of N continuous control variables can be seen² as a subspace of the

²This implicitly assumes that all degrees of freedom of the given controller can be moved separately.

N-dimensional Euclidean space R^N , with each possible combination of the N variables as a point in this space. Acting on the controller can then be seen as moving through this given parameter space. Adopting this view, the aforementioned “what” aspect of mapping becomes the point-wise association between an N-dimensional controller space and an M-dimensional space of sound synthesis parameters. At the same time, the “how” aspect can be seen as the rules governing the association of control/sound points that are not explicitly mapped in a pointwise fashion. Instead, this is concerned with the association of entire subregions of the respective parameter spaces. The “what” aspect dominates discussion on mapping and clearly is of great importance; it also encompasses the important issue of perceived control complexity [4]. However the “how” element also influences perceived complexity particularly in the case of signal-based mappings concerned with modulation-type gestures [149]. I will explore this fact in the context of continuous control of textural sound features in chapter 4. However, it is first important to be aware that not every continuous transformation of control or sound data is a mapping proper. Rather, there is a fluid boundary between the conditioning of data, the design of temporal mapping dynamics and the extraction of static gestural control features.

3.2 Boundary of Signal Conditioning, Mapping and Control Gesture Design

To be more precise regarding this “how” aspect of a mapping strategy, we must remember that mapping per se is ultimately tied to perception and more directly to intentionality. As such the continual linear/nonlinear transformations of input parameters to intermediate or output parameters may be an element of the mapping if they have a direct affect on the perception of performer intention or establish a link between input and sonic gestures. If this is not the case, however, such transformations can be considered as relating to *signal conditioning* – that is, the processing of a signal so that it meets input compatibility requirements of a system, most commonly in regards to linearity, boundaries or bandwidth. In the case of designing musical performance systems, this manifests either as the conditioning of input control signals or the creation of transformations that ultimately affect the sound parameters.

Thus while the boundary between operations on signals that can be considered mapping or ones that we consider as signal conditioning is blurred, one separating factor is whether such operations act on data that is considered part of the control data or part of the sound parameter data. Given the unidirectional control-to-sound nature of most music performance systems, it is most common that “control-side” signal operations manifest as signal conditioning, while similar operations on the “sound-side” are more properly a mapping. A similar distinction is made in [71] between control and effect in the specific context of digital audio effect control, but “signal conditioning” is used throughout to mean both actions that simply condition data as well as those that can be considered as mappings themselves. Naturally these things are a product of design context, and so I simply raise this ontological question of signal conditioning vs. mapping in general, to be considered on a case-by-case basis.

Assuming the situation in which a mapping is defined on a continuous and multi-parametric space of sound parameters, we can consider the “what” and (more centrally to this discussion) the “how” aspect in terms of a Euclidean geometric framework so as to characterize a given mapping strategy. To this end, I’ll first cover a few conceptual and formal aspects of mapping from a functional viewpoint, defining a set of relevant geometric and topological properties.

3.3 Functional View on Mapping

We now return to the discussion on the functional point of view of mapping, begun in section 3.1.2. Both this view and the systems perspective maintain a similar basic assumption in that for both we consider a digital musical instrument as a multi-input multi-output system in which a collection of control parameters, through a series of complex and potentially hidden relationships, affect a set of sound processing parameters that ultimately produce a one-dimensional output sound signal. One primary difference, however, is that the interest in assuming a functional view is to consider the *properties* of the mapping-as-function, including the way it relates to the underlying set of input/output parameters. Such properties are of interest here insofar as they provide an understanding of the usefulness of a mapping function for a given musical context. A functional analysis of mapping further addresses – in a unified

fashion – the issues of dimensionality reduction, parameter interpolation and signal/gesture conditioning.

Now, the purpose of this section is not to propose *the* taxonomy of musical mappings –as this is very much fluid and based on musical intention– but rather to develop a language and conceptual apparatus by which one may more easily create a personal instrument design space. The usefulness stems from the fact that, as we will see in subsequent sections, there is a mutual influence between functional properties of a mapping, gestural response of the instrumental system and sound dynamics of the output. Thus understanding the tradeoffs of a given mapping structure can lead to a better overall design.

Formalization

Following the assumption of an instrument as a general multi-parametric control system, we may begin by expressing a given mapping as a function f that associates every element within an input set (or domain) X to another element within the set $f(X)$ known as the output set (or range). We then write $f : X \rightarrow Y$. The output is generally considered as a subset of some larger set Y which may endow $f(X)$ with certain properties, the most musically relevant of which we will discuss in a moment.

In order for the mapping to be *well-defined*, for every $x \in X$, there must exist a *unique* element $f(x) \in Y$. That is, there cannot exist multiple elements $\{b_1, \dots, b_n\} \in Y$ such that $f(x) = b_1 = \dots = b_n$. In some cases one control variable x_1 may be mapped to several output variables $\{y_1, \dots, y_n\}$, which may seem at first contrary to the well-definedness of the mapping. However, the input and output sets can be (and most likely are) multi-dimensional with $\dim(X) = n \neq m = \dim(Y)$, so that in this instance the value $f(x_1) = (y_1, \dots, y_n)$ constitutes a single element mapped across dimensions of the range set $f(X)$, and so does not violate the well-definedness of f . Rather, it constitutes either a “one-to-many” or “many-to-many” mapping [7] if we consider this from a systems perspective. This notion of inter-parameter complexity can be expressed in terms of spatial attributes of a mapping function. Therefore, in order to draw these parallels we must introduce the concept of space into the discussion.

3.3.1 Geometric and Topological Framework

For the sake of conceptual clarity and discussion, consider all sets of control or sound parameters as subsets of a vector space [150]. In doing this, each possible *state* of a controller or synthesizer – that is, any given list of parameters that describes the output, orientation or physical properties at a moment in time – becomes a vector in this space, and the concept of distance arises naturally in our musical domain of concern. This is due to the assumption that all parameter sets³ are a subset of N-dimensional Euclidean space \mathbf{R}^n . In other words if we have n different parameters, whose set of all possible values are represented by $\{\mathbf{P}_1, \dots, \mathbf{P}_n\}$, then the collection of all possible combined values becomes

$$X = \{(x_1, \dots, x_n) | x_i \in \mathbf{P}_i \subset \mathbf{R}, i = 1, \dots, n\}$$

For the purpose of the ensuing discussion, assume from this point forward that $X \subset \mathbf{R}^n$ is our control parameter space and $Y \subset \mathbf{R}^m$ is our space of sound synthesis parameters. There may also exist a mapping into some intermediate space(s) $Z \subset \mathbf{R}^d$ in order to further condition an instrument's behavior, such as by extracting higher-level or perceptually-relevant parameters. We consider this another *layer* to the mapping as in figure 3.2, and we express the entire mapping h formally as a *composition* of mappings by writing

$$h = g \circ f$$

where

$$f : X \rightarrow Z, f(X) \subset Z$$

and

$$g : f(X) \rightarrow Y.$$

³Real numbers are most likely to be encountered in computer music applications of continuous control. The assumption of an underlying Euclidean space parallels the classical conception of one's local, physical space. Regardless of the physical truth of this representation, it is nonetheless very useful and important for constructing a mental representation of the design process. Even in the case of a standard dynamical systems expression of a given instrument, the parameter space defined by a system's physical states can be seen to lie on a *manifold*, a space that is Euclidean in regards to its local structure [151].

The degree to which a given mapping layer defines control or sound structure is similar to the issue of signal conditioning vs. mapping in that it is contextual, as we will discuss in 3.3.2.

Now, expressing the state of all collective control/sound parameters at a given time as a vector within \mathbf{R}^n gives rise naturally to the notion of a *metric* or distance between states. How one measures distance may be relative to perceptual criteria, or more standard measures such as Euclidean distance depending on ones definition of similarity. In my work presented in this chapter and the next, the concept of distance is implicit and perceptual in that similarity is defined experientially by the user during the mapping design process. Regardless, once a notion of distance is in place, it gives rise naturally to the concept a given set's *topology*, or the nature of its interconnectivity. This defines a set's "neighborhood" structure in the sense of local relationship between elements or points. It is relevant to this work in that we often want to design mappings which construct and further preserve the topological nature of a given parameter set in the mapping between variables. We would like for topologies to be preserved in this particular case of *user-defined perceptual spaces* such as those that are the focus of this dissertation. In order to explain this further, we need to distinguish between the dual concepts of a mapping as geometric structure and as a transformation of such structures.

Dimensionality

In section 3.1.2 I discussed the degrees of freedom of a given controller or sound synthesis algorithm in the context of a systems view of mapping . In a functional mapping context, this same idea translates into the dimensionality of the underlying control or sound synthesis space. In many cases, the number of sound synthesis parameters is far greater than those of a given controller or input device, and so *dimensionality reduction* of some sort must take place. However, while we map between parameter spaces of different dimensions, note that the degree of freedom of motion within any given parameter space is *bounded by the lowest dimensional space through which we map*. For example, the control space of a laptop trackpad is two dimensional

if we consider only positional input data⁴. If we map this continuous control surface in a continuous fashion to a higher-dimensional set of sound parameters – be it additive synthesis, granular synthesis, etc. – the dimension of the *effective* parameter subspace itself will exist in two dimensions *within* this larger space. The manner in which this control surface “occupies” the space will be a product of how its control data is *embedded*.

Mapping as Embedding

Strictly speaking, a mapping $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is an embedding⁵ if it is a *homeomorphism* from \mathbf{R}^n onto $f(\mathbf{R}^n)$. Generally speaking, this means that the essential shape of the domain set is preserved under the mapping. More formally it means that

- **f is continuous:** Essentially, a mapping is continuous if (possibly infinitesimally) small changes in position of input space correspond to similarly small changes in output space. That is, there are no jumps in output space when none are present in movement through input space.
- **f^{-1} is continuous:** The *inverse* f^{-1} of mapping f , assuming it exists, is precisely that for which $f^{-1}(f(x)) = x$ for every element x of a given set X . In this instance we further require this mapping to be continuous.
- **f is bijective:** This means that for every y in $f(\mathbf{R}^n)$, there is precisely one x such that $f(x) = y$.

That being said, I use the term embedding a bit more loosely here to mean any geometric surface (parameter space) continuously mapped into another space of the same or larger dimension. The mappings that I will present for musical use in section 3.4 will all be continuous and bijective⁶. Whether they have a continuous inverse will depend on the *pointwise* mapping.⁷ This particular aspect of topological preservation is

⁴While the parameter space would be higher if e.g. velocity or acceleration data were extracted as parameters.

⁵From a *differential* topology viewpoint, there would be further requirements of smoothness and compactness to fulfill. Here we mean the term in the fundamental context of *point-set* topology.

⁶With respect to their target space that is. They won’t map to the entirety of \mathbf{R}^m .

⁷In agreement with the language from [1].

not as essential in the classical case of control-to-sound flow of control. An invertible mapping would be essential in control systems based on inverse modeling wherein some ideal output becomes the domain for an inverse mapping to the given control input [152]. There is similarity between such work and my approach as they are both inherently phenomenological in nature; however the implicit “inverse mapping” here is done via human perception, as part of the design process.

As noted earlier, the idea of pointwise mapping can be seen as the functional mapping equivalent to the aforementioned “what” element of mapping, in that it arises from a direct correspondence between input and output points. For example, in the case of a direct controller-to-sound synthesis mapping, it would be the association of one discrete state of all control variables to one discrete state of sound synthesis variables. It differs, however, in that a pointwise mapping in this spatially-oriented context suggests a holistic or integral manner of conceptualizing an instrument [4][153], while the previous discussion was concerned with a more separable cognitive mode in that it focused on the correspondence of individual *parameters* rather than overall *states*. Taking this analogy further, the “how” aspect of mapping from a functional viewpoint corresponds to the geometry of a given mapping – that is, how it fills the parameter space. While a systems view would be concerned with conditioning of single-parameter dynamics, this functional interpretation examines the influence on the inter-dynamics of all parameters. In this way, the geometric structure of a mapping can influence instrumental dynamics and sonic gestural response immensely, and it is this influence that I will focus on for the remainder of this chapter.

From a practical standpoint, the mapping geometry is determined by the manner in which a given technique performs interpolation, extrapolation or regression on the known control or sound states. The relative high or low level of control depends on the nature of control/sound parameters: for example, parameters directly related to action or sound perception would be considered more high level than raw sensor values or synthesizer parameters. Similarly, the *nature of the embedding* for a given mapping influences the relative closeness to “perceptually relevant” control. In particular I distinguish between a *direct* embedding and a *mixture* embedding. In the former case, the usable input space – with boundary and interior defined by a given mapping technique – is “placed” directly within a target space (of the same or higher dimension)

through a simple linear mapping that aligns the axes of the underlying input/output spaces (e.g. \mathbf{R}^n). This can be of interest if one would like to “hear the geometry” of a shape that is embedded in sound parameter space, such as with the sonification of data [154]. As an example, consider a two dimensional control device such as a trackpad or other 2-D position sensing device. I have mapped this continuous control surface into a 3-D sound synthesis space as follows: take 49 equidistant points, and map the x-values pointwise into the bandwidth of a second order IIR filter in a similarly equidistant manner. Similarly, the y-position values are mapped into the fundamental frequency of a bank of harmonic oscillators, used as input to said filter. Now, I embed this surface into \mathbf{R}^3 by assigning each point to a value that corresponds to the center frequency of the filter. This is the aforementioned “simple linear mapping”, which in this case dilates (i.e. stretches) and translates each point from \mathbf{R}^2 into \mathbf{R}^3 . Finally, in the sound synthesis space I interpolate the values using a bilinear mapping function (to be discussed in detail in the following section) that defines a 2-D surface “directly embedded” in three dimensional space, as depicted in figure 3.3.1. In this case, what we hear in regards to the geometry is simply the cross coupling due to this embedding – depicted in figure 3.3.1 – and the nature of the interpolated geometric surface. The filter parameters then implicitly map into the lowest-level filter coefficient parameters by virtue of the sound processing algorithm. Note that this mapping is axially aligned between spaces by construction, but needn’t be and in fact could be rotated into higher dimensions via matrix operations, such as those presented in [12].

Now, while this example is an embedding in the strict sense⁸ of the word, it is less musically interesting in that that it is an object which retains its shape at the expense of being *insensitive to perceptual metrics or musical interrelations within control/sound space*. Therefore this type of embedding may be more useful for sonification, for very high-level control spaces, or as visual representation/feedback of parameter navigation. In general, therefore, I find a mixture embedding to be more relevant for designing useful control structures, particularly for timbral control.

The difference between a mixture and a direct embedding lies in the pointwise mapping: rather than preserving the geometry of an input space, the structure is created by user-defined associations between input and output states. In other words,

⁸That is, the mapping is a homeomorphism.

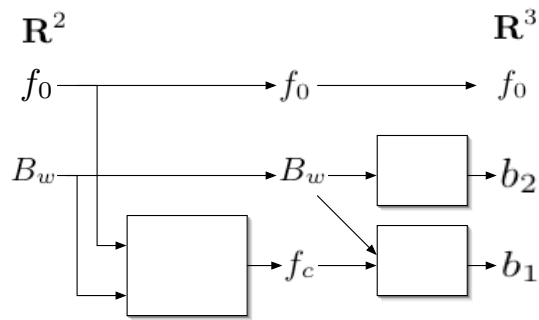
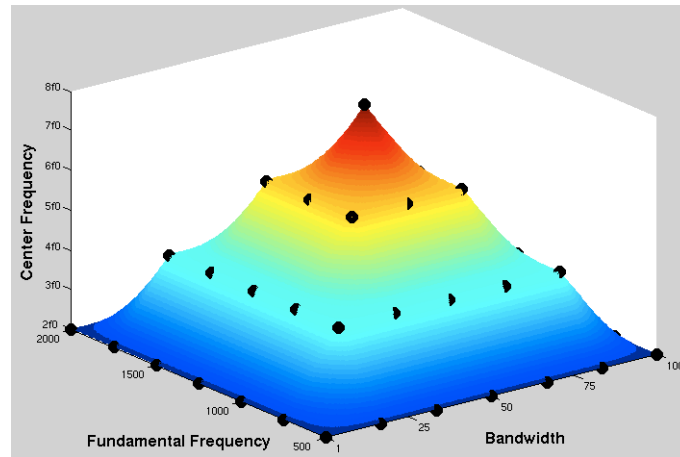


Fig. 3.3 (a) Direct embedding of 2D control surface into 3D sound synthesis space, including pointwise mapping. (b) Interrelation of parameters due to direct embedding.

in the case of a single control-to-sound mapping a system designer would decide that at one state of a given controller it should sound a certain way, with this sound being defined by the given parameters of the sound synthesis algorithm. In doing this, there is an *implicit perceptual distance* imposed on the two spaces: something that is considered by the designer as “close” and grouped accordingly may be quite far in the Euclidean space of sound parameters. The result is that movement of a controller may induce a trajectory in sound space that covers a large distance and potentially a great deal of intermediate timbres. Another consequence of this is that this “stretched” mapping, by virtue of affecting the speed of parameters, affects greatly the *dyamics of the sound parameters and therefore the resultant sonic gesture*. From a functional point of view, we also cannot guarantee that an input control space is homeomorphic to its sound output counterpart, as this new space may “collapse” in on itself or otherwise self-intersect. In other words, two different inputs to a controller could potentially result in the same sound. For example, take the two-dimensional input space from the above example: if we map the input points to drastically different sound states, with the same mapping we may arrive at a drastically different output topology as I have done through the example in figure 3.4. I don’t consider this as undesirable behavior in general, and in fact it could be very much wanted to have the same sound occur in several different control states. By contrast, this would be a problem in fields such as the aforementioned inverse systems modeling, wherein a mapping must be invertible in order to find the input needed for some ideal output. In such cases, an algorithm such as a *self-organizing map* [155] would be needed in order to guarantee the preservation of the input/output topology so that the two are homeomorphic.

Relevant Properties

While we cannot guarantee the nature of the embedding, other more musically relevant qualities can be preserved if we know something about the underlying properties of the entire mapping structure – including the extension from pointwise association to the entire input/output spaces.

In knowing the functional expression, regardless of the global shape of the output space one can better understand the local geometric and topological structure that influences the overall control structure. I’ve already discussed the property of *continuity* in the

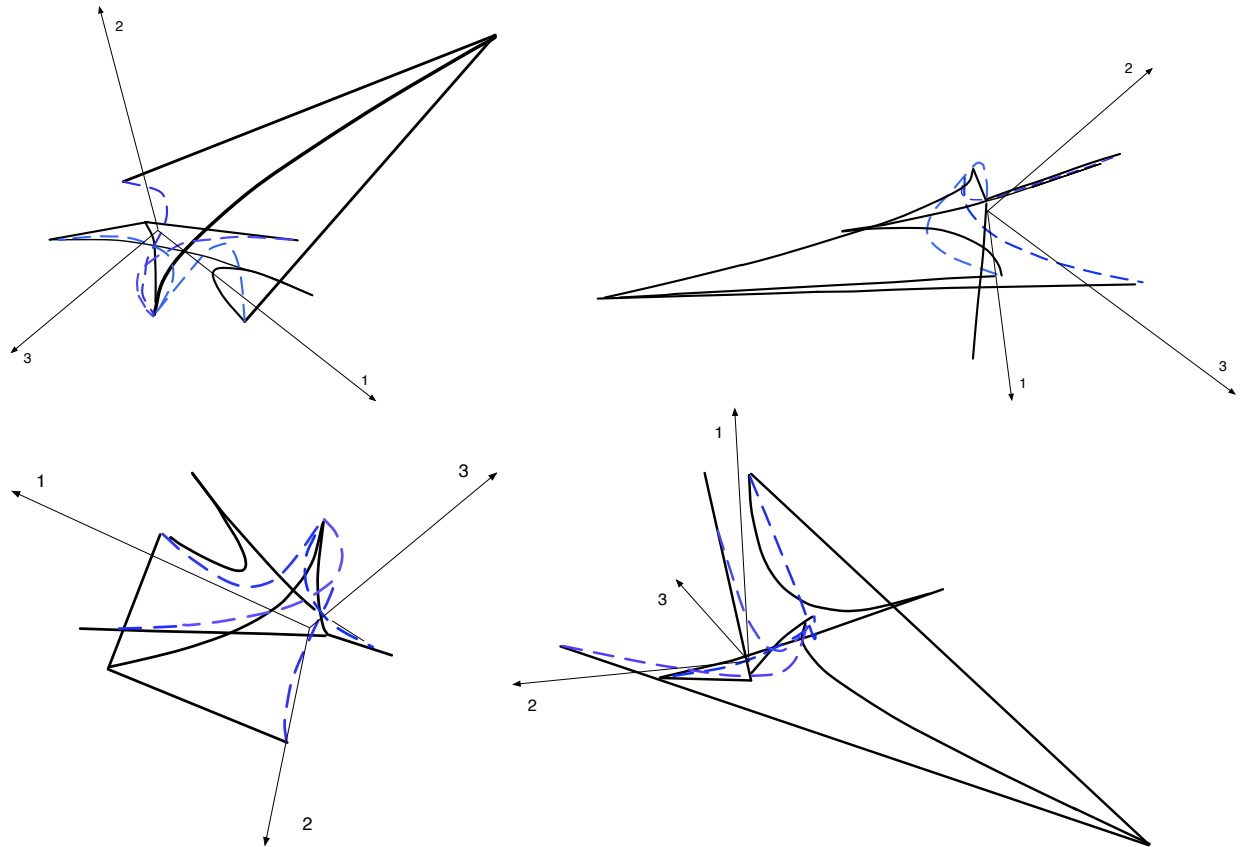


Fig. 3.4 Various views of a mixture embedding between two dimensional control space to three dimensional sound output space. The continuous input trajectory, sampled every 20 ms, travels around the borders of the rectangular input space (solid, black) and then across both diagonals (dashed, blue) to provide the given output trajectory. The output is continuous, but the space is split and passes through itself, thus the mapping is not a homeomorphism. Dimensions are indexed along the axes as a point of reference.

previous discussion, which is quite important from a musical point of view for control contexts such as morphing between sound models (at a high level of control) or timbral modulations (at potentially a low level). Continuity is ontological in the sense that a mapping either will or will not have it, by virtue of its existence. In the same way, a mapping will or will not have these additional properties:

- **Differentiability** A mapping is differentiable if one can describe its rate of change for all values of the domain. If this can be described iteratively on the resulting expression of change, the mapping is considered to be of higher order differentiability. Not being differentiable can lead to a perceptual discontinuity in continuous control situations, depending upon the nature of the embedding which may stress this feature.
- **Smoothness** At what point a mapping is considered “smooth enough” is contextual, but this property is directly related to the order of differentiability of the mapping, and whether this is continuous or not. If a mapping has a continuous n th derivative, we say that it is C^n . This property may be directly linked to expected or perceived dynamic qualities of a mapping and again can be accentuated under a mixture embedding.
- **Linearity** Following standard mathematics, a mapping is linear if it obeys the principles of superposition and homogeneity. In the context of geometric mappings, this influences the complexity of cross-coupling between parameters.
- **Explicitness vs. Implicitness** Extending the definition given in [7], a mapping is explicit if it can be represented analytically as a function of its input space, so that an output can be found for any element of a given input set. An implicit mapping lacks this property, and requires additional knowledge of a dependence between input and output elements (with possible restrictions on input and output sets) in order to find the corresponding range element of a given domain element. For example, the use of neural networks as mapping [8][9] are implicit as they require training to take place.
- **Exactness** A mapping is exact if the entire mapping input/output space agrees with the underlying pointwise mapping. This is important if, for example, a certain sonic response is required at a particular state of the control input.

- **Global vs. Local Definition** A mapping is globally defined if all states from the underlying pointwise map exert influence on the entire geometric structure of the mapping. If the influence at a given point is isolated to certain *neighbors* that are close – relative to some distance metric – then the mapping is locally defined. This is the topological property of mapping that we are most interested in preserving here.
- **Parametric** A mapping is parametric if it depends on variables that can be influenced in a time-varying way. As I will show in the examples that end the next chapter, this can be important for adapting a mapping to achieve a certain *gestural response*. This is particularly coherent when the parametric control influences the above-named properties.

With this set of properties in mind, I have created a modular library of mappings that have complementary qualities and that can be combined in different musical contexts, as I will present in section 3.4. These tools are then applied in chapter 4. The modular combination of different mapping structures is key to this approach, which implicitly means the use of multiple layers of control. As such it is worth considering, in this given functional context, what it means conceptually and practically to utilize multiple layers of mapping strategies.

3.3.2 Multi-Layered Approach

In [13] the concept of a “composed instrument” is introduced in order to describe a system in which different controllers and sound synthesis methods may be combined modularly. The key idea that is presented is that of having two mappings: one for transformation of control data into so-called “abstract parameters” and then one for mapping from this parameter space into an appropriate set of sound synthesis parameters. A similar concept is implicitly discussed in [3] by presenting the concept of navigation through an “expressive timbral subspace” which then determines a morphing between additive sound models. In this case there are two mappings as well: from control data input to this timbral space, and an implicit mapping determined by the model definition as well as the interpolation rule. In both of these examples there exist two *layers* of mapping in the resultant system. Further, in each system –

particularly in the first example – the idea is to abstract and parameterize the essential characteristics of controller and sound synthesis separately, in such a way that they “meet in the middle” with an equal number of control and sound abstract parameters. Building on this idea, in [14] the authors introduce the notion of *gesture perceptual* and *sound perceptual* spaces, wherein a mapping is defined between these two high-level and abstracted spaces. Other mapping layers exist on control and sound side in order to extract perceptual information from the low-level parameter spaces. The heart of this idea is that – while a mapping between “raw” data and perceptual parameters may be complex – a mapping between gesture/sound perceptual parameters is direct, of the same dimension and perhaps linear. This is therefore a three layer model which focuses on two sets of extracted perceptual – rather than “abstract” – parameters.

These three examples represent a line of thought that is very much top-down in its approach to mapping design. Such an approach seeks to build up from many lesser parameters to a few relevant and salient parameters, directly linking the two as in our abstracted figure 3.1. However, as I briefly discussed in section 3.2 and will illustrate in section 4.5, there may exist many layers that deal to varying degrees with conditioning the raw data, associating parameters, cross-coupling parameters, musically-relevant warping functions or with defining a time-varying gestural response. The number of layers proper may be arbitrary and a decision of the interface designer in some cases, but it may also be determined by the desired system response as in my upcoming fabric examples of section 4.5. If the desired perceptual mapping consists of holistic control over some abstracted parameters having d (where $d \neq n, m$) degrees of freedom, then the multi-layered mapping must address the *effective topology* of the entire control structure. In the embedding examples, there was a topology determined by a single control-to-sound pointwise association and a single functional mapping structure. However, the dimension of “topology space” need not be the same as the underlying control or sound spaces. For example, consider again a two-dimensional controller, and suppose that we wish to drive output sound parameters in such a way that the two degrees of freedom influence differently one localized area of sound space. In this case we may map entire regions of input space onto a 1-dimensional topology of points in an intermediate space by continuously *projecting* onto it, which in turn map into sound synthesis space, as illustrated in figure 3.5.

The important point that I wish to underscore through this discussion and the brief example⁹ is that just as available degrees of freedom determine underlying dimensionality of parameter spaces, the *number of intermediate mapping layers and their topology* will influence the perception of a system’s immediacy of control: whether this is over one aspect of the sound or a more expansive set of sound features. Therefore in building “downward” from a desired set of control degrees an understanding of the effective topology space is important; at the same time, this spatially-oriented view doesn’t paint the entire picture, as concepts such as “response”, “feel” and overall perception of influence ultimately relate to the temporal quality of mapping as well.

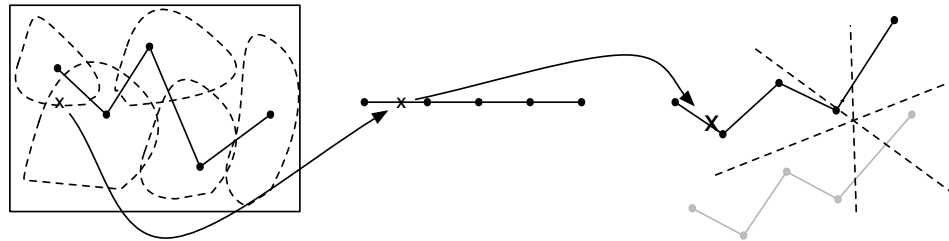


Fig. 3.5 A projection from regions of two dimensional control space onto a one dimensional intermediate space, which determines the ultimate topology when mapped into a three dimensional sound synthesis parameter space (with grey “shadow” on X-Y plane).

3.3.3 Spatial vs. Temporal View of Control

Throughout this chapter the focus has been on control and mapping as a spatial phenomenon. However it is important not to lose sight of the fact that perception and control of sound and music is inherently temporal, as was emphasized through the development of sonic gesture analysis in chapter 2. Therefore, this spatial *representation* of mapping must be reconciled or otherwise considered such that time is not lost in the process. The “hidden temporality” of a spatial view on mapping becomes evident when observing that, as previously mentioned, the geometric structure of a continuous mapping has a strong influence on the dynamics of sonic gestures that result from a given input gesture. Similarly, in constructing a pointwise mapping one is

⁹The overall control structure for which will be presented in section 4.3.3.

not just defining a steady-state response of the instrument (i.e. this one action makes this one sound) but the topology that one constructs will define a certain gestural response for the instrument. In [156] the author’s focus is spatial in that he constructs pointwise mappings so as to give maximal timbral variety within the sonic space. However, this may constrain the gestural response, and so adjusting the input/output topology after the fact may be necessary in order to tune the overall gestural and temporal quality. These adjustments are acted on the mapping topology precisely because such information is not directly represented in a functional mapping expression defined from a multi-parametric (and thus spatial) point of view.¹⁰

At the same time, the mapping itself can be parametric – a key element of my own mapping design work – and so the dynamics may also be influenced by the mapping as it varies over time. Having a time-varying mapping – as for example in [157] – can strongly influence the perceived “feel” or control. It is not clear a priori what the influence of parametrically controlling a given mapping will be, and there are no hard-coded “rules” for dynamically altering a mapping’s geometry in order to achieve a certain gestural response. The influence is very complex and indirect, depending on the mapping geometry and the nature of control and sound synthesis. In light of my goal to create a modular set of mapping structures that may be combined and layered, the input controls to a given mapping are simply treated as free parameters to be explored in a given musical context. Generally speaking, from my experience it has been the case that in order to define a system that feels instrument-like and playable, I tend to modulate such parameters in order to construct mappings that are themselves dynamic, but which provide a repeatable gestural output and so are static from perceptual point of view. This has been my solution to the difficult and intrinsic problem of space vs. time in defining mapping structures; another approach would be to *begin* by parametrically describing the actual dynamics of control or of sound as in [151] and [158], and to work *towards* a spatial structure in order to understand “near” vs. “far” control states. Regardless of the “direction”, the essential point that I want to convey here is that there exist these two dual representations of a control structure, and that one must explore external relationships such as human perception in order to

¹⁰To further this point, note that slight changes in pointwise mapping may vastly change the distance between two states considered “neighbors”. This then affects the speed with which a gesture travels through the underlying space, thereby greatly affecting the output sonic gesture.

understand a coherent link between the two.

3.4 Towards a Toolbox of Mapping Functions

To this point, the chapter has focused on building a theoretical basis for mapping in instrument design, covering different definitions and perspectives while privileging a functional point of view and developing this in the process. Some key ideas that can be taken from this discussion include

- Multiple layers of mapping may be required to condition signals, to arrive at the desired degrees of freedom and effective topology and to condition a control or sonic gestural response.
- Changing a controller or sound synthesis method further necessitates a potentially different coupling of multiple mapping strategies.
- Functional properties of an explicit mapping can suggest its suitability for different musical control contexts.
- With the top-down parametric approach to mapping taken in this chapter, the geometry of a given mapping strongly influences the sound dynamics and gestural response.
- A temporal view of mapping is also important, and one solution from a top-down perspective is to use parametric mapping structures.

I regard this theoretical development as the first contribution of the chapter; the next set of ideas that I will now present apply this theoretical framework to existing mappings from the computer music literature, placing them in a unified framework and leading to a sort of “mapping design space”. From this analysis, I have developed a modular set of mapping tools for designing control structures that address the issues from the above points.

3.4.1 Parallel Developments

MnM: Multiple Linear Regression

In [12] the authors consider mapping as an operator that expresses each point in sound parameter space as a linear combination of the N parameter values of a given control input. This has the geometric interpretation of a hyperplane which represents control space mapped into an M -dimensional sound space. In the language of this chapter, it is a linear and direct embedding. The exactness of such a mapping is bound to the number of stored presets: if this number is less than or equal to N the plane will pass through the preset sound values exactly, but if it is greater than N the mapping becomes a multiple linear regression model and the plane passes somewhere between the preset sound values. The constraint of a linear approach to mapping in this case is traded off with the ability to draw on many results from linear algebra regarding matrix operations in general, and from an implementation point of view the FTM library [159] for matrix data processing.

Metasurface: Natural Neighbors

In [24] the author presents a technique for mapping from two dimensional control space (on-screen mouse position) into M parameters of sound processing. It is implemented within the Audiomulch sound processing software [160], and so was created with this particular tool in mind. The approach is based on a pointwise mapping in a mixture embedding fashion, and utilizes natural neighbor (NN) interpolation [161]. This technique computes a so-called Voronoi tessellation of the known input states across the entire control space¹¹, which creates “patches” that define a neighborhood as all space that is closer to a given state than any other. When an input is given to the control space, a new tessellation is computed based on the previous states and this new input value, with the overlapping areas between old and new patches determining the influence of each neighbor state, thus interpolating an intermediate state. This technique is continuous everywhere, and is C^1 at all points but at the known states themselves. Therefore there is a discontinuity in smoothness at these locations. Also,

¹¹More precisely, within the *convex hull* of the known states.

the mapping implementation is constrained to two input space dimensions, with extensions to higher dimensions less straightforward.

Colorblobs: Gaussian Kernels

A technique for geometric control is presented in [23] in which a mixture of Gaussian kernels map a two dimensional control space into an M-dimensional sound space. Users may interact with an on-screen display and determine the spatial layout and position for steady-state values of sound by mapping from screen position into a vector of sound parameters. The mixture embedding is determined by a continuous mapping that is a linear combination of L normalized Gaussian kernels such that:

$$z = \frac{W(x, y)}{S(x, y)}$$

where

$$W(x, y) = (e^{-\frac{x^1-x}{\sigma_{x1}}-\frac{y^1-y}{\sigma_{y1}}})z^1 + \dots + (e^{-\frac{x^L-x}{\sigma_{xL}}-\frac{y^L-y}{\sigma_{yL}}})z^L$$

and

$$S(x, y) = (e^{-\frac{x^1-x}{\sigma_{x1}}-\frac{y^1-y}{\sigma_{y1}}}) + \dots + (e^{-\frac{x^L-x}{\sigma_{xL}}-\frac{y^L-y}{\sigma_{yL}}})$$

z is the final output in sound parameter space, (x, y) is the input to control space and $\{(x^1, y^1), \dots, (x^L, y^L)\}$ and $\{z^1, \dots, z^L\}$ are the respective input and output states that determine the pointwise mapping. Note that all output vectors are of dimension M.

This normalized form guarantees that the sum of all contributions from the preset sound states will sum to one. However, this method is not exact: the Gaussian kernels are defined everywhere, and so the influence affects values even at the precise location of preset states in the input control space. Rather, the mapping follows what I refer to as a *gravitational system model* in that each “blob” exerts its mass over objects that hover around it. To receive a weight of 1 from a single blob, it must spatially be located between the input and all other blobs. In addition to moving/placing blobs (i.e. changing the pointwise mapping), this approach can be tuned by changing the $(\sigma_{xL}, \sigma_{yL})$ values to exert influence more or less in a given x or y direction. This important parametric quality is articulated by the authors in [23] as well as in the available software [162].

I present these three approaches in a separate section for the dual reasons that they were developed in parallel with my own mapping library but also to explain why I have not considered including them into my own work. The first reason simply is that in the case of MnM and Colorblobs (CB), efficient implementations already exist, and are even in the same software environment as my own work so that they can be used together modularly in the spirit of this work. Beyond this, I avoid inclusion of the regression approach of MnM and the NN technique for deeper structural reasons. The former creates a single linear surface in control space – making it cumbersome and difficult both for modular mapping design and tuning of gestural response – while the latter is not easily extensible to higher control dimensions and its properties can be encapsulated by other techniques, as I will discuss.

3.4.2 Previous Work and My Extensions

Grid-Based Simplicial Interpolation

In [20] a mapping $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is defined which interpolates points spaced in a grid. The known states in control space need to conform to a shape which is topologically equivalent to an n-dimensional lattice. These presets are associated with states in synthesis space Y . This gives us a pointwise mapping between control and sound space whose exactness is preserved by an embedding that is defined as below. To interpolate to other points within the grid, a scheme is developed which first partitions each hypercell into simplices.¹² When a point p is input into the lattice, the algorithm determines which subdivision p lies in. If it does not lie at a vertex of the lattice (representing a control-to-sound pointwise map), the vertices $\{v_0, \dots, v_n\}$ of the simplex in control space which contains p determine the point as follows:

$$p = v_0 + \sum_{i=1}^n \alpha_i v_i$$

where

$$\sum_{i=1}^n \alpha_i = 1, \alpha_i \in \mathbf{R} \forall i$$

¹²Plural for simplex, the generalization of a triangle to n dimensions.

In other words the point p is expressed in terms of barycentric coordinates with respect to its containing simplex. This scheme is piecewise linear, and it requires that control points are spaced relative to some grid. Therefore, the technique has a storage requirement determined by the 2^n vertices of a hypercell multiplied by the number of such cells needed to cover a control space at a resolution set by the user. From a computation point of view, the number of necessary operations¹³ is $O(mn)$. In terms of the geometric structure, this method is non-differentiable at the edges between subdivisions.¹⁴ That is, there are sharp creases at the join between simplices.

Scattered Simplicial Interpolation

Another mapping technique that utilizes simplex-based interpolation is presented in [1]. This method, referred to as *simplicial interpolation*¹⁵ extends the results from [20] in several ways. As with the grid-based technique, it is based on a pointwise mapping between $X \subset \mathbf{R}^n$ and $Y \subset \mathbf{R}^m$. The difference between the two is that this approach allows the set of preset states within control space to be scattered. The author achieves this by creating a triangulation of the points in X rather than fitting them to a grid. In particular, the method used to do this is the *Delaunay triangulation*, which has been widely used for spatial interpolation.

This triangulation in \mathbf{R}^n induces a similar one in \mathbf{R}^m , giving us a mesh embedded in the higher-dimensional space in analogy to the embedded lattice in [20]. As in the previous method, the point's barycentric coordinates are used, which determines a piecewise linear surface defined by the triangular mesh in X . This likewise interpolates the surface induced in Y , as the pointwise mapping gives rise to a mapping of the simplices as well. Therefore, as with the above an algorithm exists to locate the surrounding simplex for a given input value, and the mapping then determines the associated point in sound space based on this simplicial index. As discussed in the

¹³It is claimed in [1] that the actual time to compute is $O(n^3 + mn)$, the same as for the scattered Simplicial Interpolation method.

¹⁴It is suggested in [21] that this be dealt with by introducing B-spline blending functions. See [163] for more on this technique.

¹⁵While I apply different forms of this name to both this and the previous technique due to their similarity, the term was coined in [1]. I will often refer to this technique by the abbreviation SI from this point forward.

section on embedding, this technique – as a mixture embedding – causes the piecewise *simplicial complex* to retain its continuity, linearity and local topology, but depending on the user-defined pointwise map this complex may be stretched and “crinkled” in sound space.

This technique allows for scattered data points, can be edited locally (i.e. inserting new states only changes the neighboring pointwise mapping), and the number of points is not constrained by the structure of a grid, making this technique more flexible than [20] in general. Otherwise there are many similarities between [20] and [1]. Both are continuous, piecewise linear and are non-differentiable at the edges between simplices. Thus, the former technique may be appropriate for data points which already lie in a grid due to some system constraint. However, the Delaunay triangulation has several properties that make it a desirable choice, including regularity of angles (less singularly sharp points) and the ability to give a best approximation of smoother functions (see [164] for more on this). Thus, for sufficiently large number of points this approach may give an acceptable approximation of a smooth surface while keeping computational load relatively light. However, if one is creating a pointwise mapping by hand, it is unlikely that there would exist a large number of data points. Therefore this mapping may need to be used in tandem with other techniques having complementary properties.

Bilinear Interpolation

In [3] and [13] a bilinear mapping is used to interpolate between additive models aligned in a two-dimensional grid structure. In the first work, two axes represent pitch and timbral dynamics with a user-defined ordering while the stored states of the intermediate space represented different models of a clarinet derived from additive analysis. In the latter work, it is suggested that this approach can be augmented by a higher-dimensional space of abstract parameters (density, inharmonicity) that describe the additive models. The bilinear mapping scheme makes sense in the context of these works because it is defined relative to a lattice structure, is locally defined (by the four corners of given area of the space) and is continuous as well as differentiable – important for moving between additive models. However it is not smooth in that it is not continuously differentiable at the boundary between cells, which could be a limiting factor for something as perceptually salient as additive data that may require smoother

interpolation [60]. Further, the geometric structure of this technique is actually hyperbolic, and so the mapping surface would be comprised of curved patches that vary little when far from known states and a lot when moving closer to these states. I will describe an extension of this technique to a mixture embedding in N dimensions in the implementation section to follow.

RST

In considering the above techniques and their musical use, it became clear that they overlap in terms of their attributes, in particular all are locally defined and are relatively rigid at the intersection between local neighborhoods. As such, I looked into work that examines mapping from a fairly different perspective: geographic information systems (GIS) and related computational approaches to cartography. In doing so, it became clear that spline-based techniques were ubiquitous.

It is well known that spline functions are quite useful for smooth interpolation, but are not necessarily exact and may cause overshoots depending on the regularity of the data points [163]. An interesting class of spline interpolators may be defined if we assume the associated function f should be smooth, and define a *smoothness seminorm* (SS) S for f for the purpose of minimization. In N-dimensional space, this problem has been formulated with an SS that expresses smoothness based on integration of higher-order partial derivatives, and which gives a unique solution [165] that we may express as

$$z(x) = T(x) + \sum_{j=1}^L \lambda_j R(x, x^j) \quad (3.1)$$

where x is an arbitrary point within the boundaries of the defined space, $T(x)$ is a "trend" function, the $\{\lambda_j\}$ are real-valued weights and R is a radial basis function which is determined by the requirements of the given application. Once the form of R is determined, the weighting functions are found by a system of linear equations that express the requirement of minimum bending energy (i.e. maximum smoothness). For two-dimensional surfaces, the so-called thin plate spline (TPS) was introduced [166] in order to approximate a surface that acted as a thin metal plate by expressing the bending energy of such an object, forced to pass through the data points. However, the

stiffness of the surface may cause many overshoots if there is a large change in gradient between points, and so the regularized TPS with tension was introduced in [167] to be able to parametrically alter the stiffness of the resulting plate. In later developments [168][169] the same authors extended this technique by considering derivatives of all orders in the SS, leading to the completely regularized spline with tension (RST). A generalized expression was found, and an explicit form for two and three dimensions was derived so that the trend function $T(x)$ simply equals a constant a_0 , while the radial basis function takes the form

$$R(r_j) = -\{ \ln[(\frac{\phi r_j}{2})^2] + E_1[(\frac{\phi r_j}{2})^2] + C_E \}, d = 2 \quad (3.2)$$

or

$$R(r_j) = \frac{1}{\phi r_j} \operatorname{erf}(\frac{\phi r_j}{2}) - \frac{1}{\sqrt{\pi}}, d = 3 \quad (3.3)$$

where r_j is the Euclidean distance between input value x and data value x_j , $\operatorname{erf}()$ is the integral error function, E_1 is the exponential integral function and C_E is the Euler constant [168]. I first presented the RST method as a musically inspired mapping technique in [143]. One of the key points was that it complements and improves upon (in the proper musical context) the above methods in that it has smoothness and exactness that can be parametrically tuned (in a tradeoff), with smoothness up to C^∞ , and is globally defined. The tension parameter ϕ controls the exactness of the surface, moving between a thin plate and a flexible membrane, while a smoothing parameter is used in the system of linear equations to determine the $\{\lambda_j\}$ and thus the global smoothness of the surface. These two parameters interact and must be tuned manually. Some guidelines are given in [168].

3.4.3 Mapping Design Space

From this exposition we can see that the functional properties of a mapping structure can vary quite a bit between techniques, as exemplified by these seven examples from the computer music literature. Following the conviction that the geometry of a given mapping is a determinant of the gestural response of an instrument, I presented these examples in order to highlight the tradeoffs that exist when choosing a mapping strategy for a given musical context. Having done this, the idea now is to move towards

a *design space* of multi-parametric, spatial mapping strategies – not for creating an objective taxonomy, but to inform designers of different techniques available to them. To that end, consider the comparison of the seven aforementioned mapping techniques in table 3.1. Here we see that all mapping strategies are continuous, and all but the simplicial techniques are differentiable (they are not at the boundary between simplices). Further, most are not smooth: the MnM is in the sense that it is a simple hyperplane, but this in and of itself is limiting. The Colorblobs (CB) and RST techniques are highly smooth. Most can be defined by scattered control/sound states, except for the multilinear and grid SI techniques, however the scattered SI method is the only one which can be edited locally – that is, one can define new pointwise mappings without changing the global structure of the mapping. Only the RST technique allows for variable control of smoothness and degree of exactness, however it will never be purely exact per se, and so it is considered as an approximation technique as are the Colorblobs (see gravitational model discussion) and MnM (assuming number of states L is greater than control dimension N). The geometric nature of each mapping varies from Hyperbolic (HB) to Piecewise Linear (PL), a general Nonlinear Curve (NC), a Gaussian Mixture (GM) and a simple Linear (L) surface. Finally, the computational complexity is rated, with the run-time computation requirement of the MnM matrix mapping being fastest (speed rank=1) and the radial basis computation of the RST method being the most costly (rank=7).¹⁶ All techniques have been implemented efficiently, and I have coded RST as a Max external such that the tension and smoothness can be adjusted in real-time.

In light of these qualities, it became clear that I could span the mapping design space quite well if I chose to focus on the multilinear scheme, the scattered SI scheme and RST. Given these three techniques, one may define curves with rapid variation (Multilinear), with parametric smoothing to a high degree (RST) and work with scattered data that can be edited locally and on the fly (scattered SI, which I will refer to as simply the SI technique from this point forward). Further, efficient implementations of CB and MnM exist already in the Max/MSP environment, and may be used in conjunction with these mappings. The NN technique, while smoother than SI, is not straightforward in terms of complexity for extending to higher control

¹⁶This is based on an assumption of many data points. In fact, the computational complexity of the RST technique is bound to the number of points and not the control or sound space dimension.

dimensions, while its other traits are covered well by the other three mappings. Having decided on the three aforementioned mapping strategies as a basis, I have implemented (multilinear) or ported with many architectural modifications (SI and RST) these strategies in the Max and Jitter software environments.

Property	Multilin.	Grid SI	Scatt. SI	RST	CB	MnM	NN
continuous	Y	Y	Y	Y	Y	Y	Y
local vs. global def.	L	L	L	G	G	G	G
differentiable	Y	N	N	Y	Y	Y	Y
C^k	$k = 0$	$k=0$	$k=0$	$k = \infty$	$k=\infty$	$k=1$	$k=0$
grid vs. scattered	G	G	S	S	S	S	S
locally editable	N	N	Y	N	N	N	N
parametric smoothing	N	N	N	Y	N	N	N
exact vs. approx.	E	E	E	A	A	A	E
non/linearity	HB	PL	PL	NC	GM	L	NC
Bdd. Control Dim.	N	N	N	Y	Y	N	Y
complexity	5	3	4	7	2	1	6
citations	[3], [143]	[20]	[1]	[143]	[23], [162]	[12]	[24]

Table 3.1 Mapping strategies and certain relevant properties

3.5 LoM Toolbox

This is the third, application-focused result of this chapter, and provides the tools that I use throughout chapter 4. As such, I will describe these implementations in order to give the reader a sense of their functionality. These externals, which I collectively refer to as the Library of Maps¹⁷ or LoM Toolbox, were written in C/C++ for the Max/MSP environment. The way that I have tended to utilize them is explained by way of the musical examples of chapter 4; here I will focus on the core externals that were first presented in [170].

lom.si

The SI technique was available previously as standalone C++ code, which I have since modified and ported to Max and Jitter. The computational geometric algorithms were likewise ported by the original author, leveraging work done at AT+T labs [171]. These

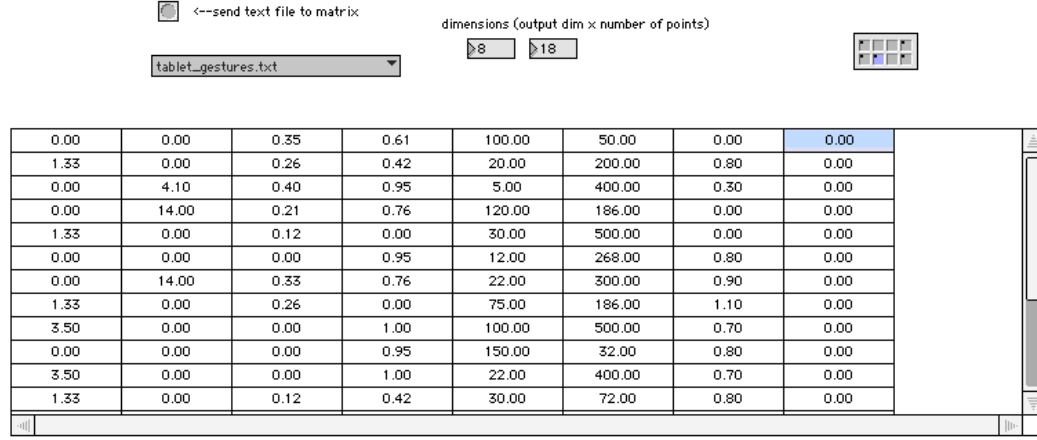
¹⁷This title is inspired by the series of texts of the same name from Moira Roth.

algorithms for the Delaunay triangulation were kept intact, as well the code which performs a search to determine which simplex within control space an input point lines within. These have been integrated into a max external by way of a Max wrapper that includes the combined C and C++ projects. The way that data is handled has been modified for Max. Also, when an object is instantiated, the control and sound dimensions are defined by arguments to the external. A third argument is a flag for stating whether one wants to utilize a self-organizing map to determine the pointwise map or to do it by hand. Assuming this is done by hand (as I normally do), then in order to define the pointwise map, text files are loaded into a matrix table either automatically or are edited by hand. The input and output tables are labeled accordingly, and the external receives the commands “inpoints” and “outpoints” in order to differentiate the data. The table editor for output points is illustrated in figure 3.6. Once the pointwise map is defined, the triangulation is automatically computed, and the object is ready to use. There are certain constraints on what constitutes a “legal” triangulation of the pointwise data, and the warnings from the Goudesene/Hull code for this have been kept intact. In this object, they are posted to the Max window as error messages. Assuming the data is valid, the object is ready to use and input lists are interpreted as control space values; upon receiving a list, the proper enclosing simplex is discovered and the input is mapped to a new output in sound space.

A previous version of this object was created for Jitter in which triangles – in the case of a two dimensional control space – were computed in OpenGL and displayed in the `jit.gl.window`. The processing and computation was heavy in this implementation, and so later version have included an abstraction *lom.jit.si* that manages all the handles to draw OpenGL triangles before passing this information off to the native Jitter objects who receive all drawing commands. An illustration of this is depicted in figure 3.7.

lom.multi

I have extended the aforementioned bilinear interpolation into a *multilinear* mapping for higher-dimensional use. This is a mapping $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$ which, for a given input state $x = (x_1, \dots, x_n)$, may be expressed as



0.00	0.00	0.35	0.61	100.00	50.00	0.00	0.00
1.33	0.00	0.26	0.42	20.00	200.00	0.80	0.00
0.00	4.10	0.40	0.95	5.00	400.00	0.30	0.00
0.00	14.00	0.21	0.76	120.00	186.00	0.00	0.00
1.33	0.00	0.12	0.00	30.00	500.00	0.00	0.00
0.00	0.00	0.00	0.95	12.00	268.00	0.80	0.00
0.00	14.00	0.33	0.76	22.00	300.00	0.90	0.00
1.33	0.00	0.26	0.00	75.00	186.00	1.10	0.00
3.50	0.00	0.00	1.00	100.00	500.00	0.70	0.00
0.00	0.00	0.00	0.95	150.00	32.00	0.80	0.00
3.50	0.00	0.00	1.00	22.00	400.00	0.70	0.00
1.33	0.00	0.12	0.42	30.00	72.00	0.80	0.00

Fig. 3.6 Table for editing text values or inputting new output sound states to lom.si object.

$$f(x) = \sum_{i \in H(x)} \omega_i \prod_{j=1}^n (1 - |x_j - x_j^i|) \quad (3.4)$$

where $H(x)$ is the set of grid points of the hypercube H that contains x , $x^i = (x_1^i, \dots, x_n^i)$ is the set of known control states, and the $\{\omega_i\}$ are the weights for each of these respective states. The above expression is used for a direct embedding into a higher dimensional sound space, or for controlling a single scalar value such as loudness within a high-level space, as with the bilinear map into a space of sound models in [3].

The lom.multi external is instantiated by providing a list that corresponds to the resolution of the control grid for each dimension. For example, if (3, 3, 2) is the given list of arguments, the object will create a 3 x 3 x 2 grid in three dimensions, and so there will be 18 preset states. Another list of arguments can be optionally given, which define the resolution for each control dimension. If none are given, the control grid is assumed to be a normalized hypercube. Optional messages include a list preceded by the message “weights” in order to set a weight value for each state, where this list can be dynamically altered in real time. If no list is given, all weights are assumed to be 1.

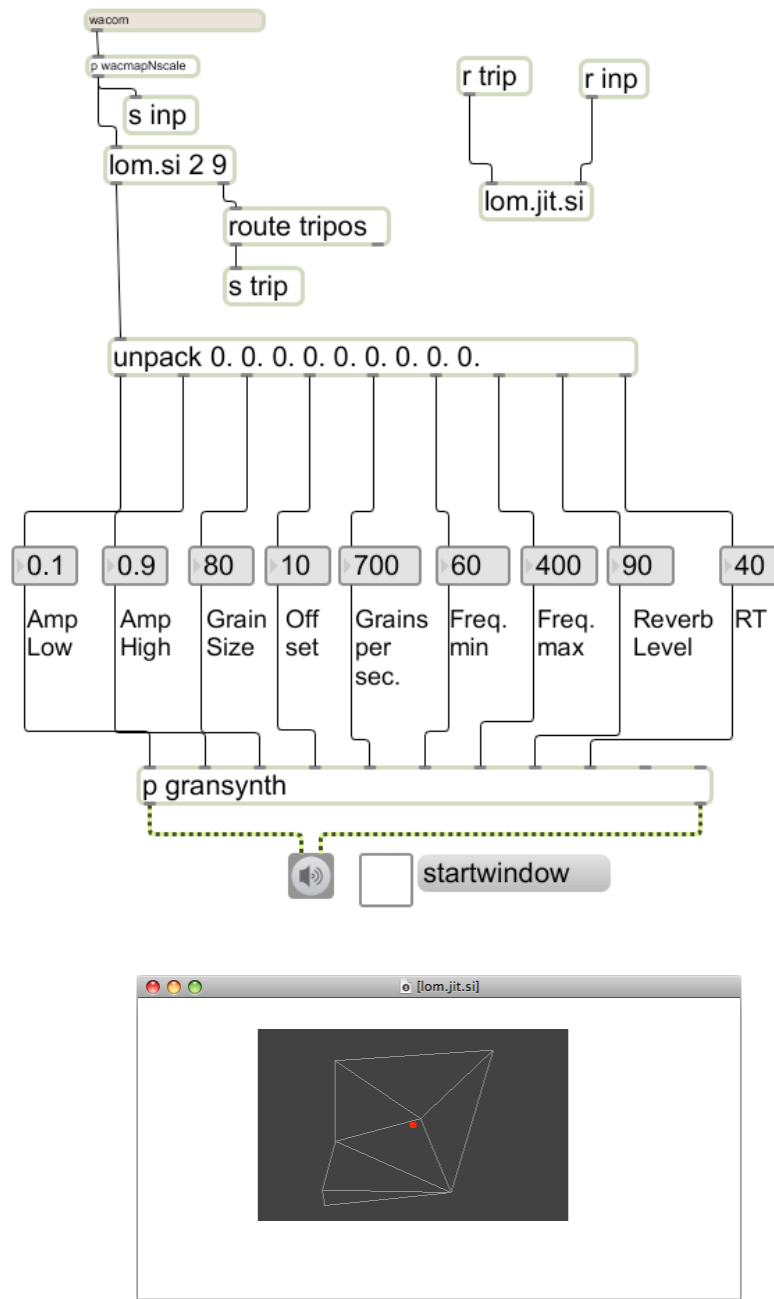


Fig. 3.7 In this example a two dimensional control space (Wacom position) is mapped to a nine dimensional space of granular synthesis parameters. The control data is sent to `lom.si`, which interpolates this input vector and outputs a list of sound parameters. A second output from `lom.si` sends triangle information to the `lom.jit.si` abstraction which then draws the parameter space using the native OpenGL rendering tools (bottom window). As with the implementation from [1], nodes of the triangular complex refer to stored parameter sets, and the input point lets the user know where they are in parameter space. This visual feedback is most useful for initial learning of an instrument.

lom.multi2

I've built an extension to the lom.multi object in order to define mixture embeddings. The expression of equation 3.4 is modified to become

$$f(x) = \sum_{i \in H(x)} \omega_i y^i \prod_{j=1}^n (1 - |x_j - x_j^i|) \quad (3.5)$$

where the set of $\{y^i\}$ are sound states that are defined through a pointwise map with the respective $\{x^i\}$. This mapping retains its hyperbolic nature, despite being warped by virtue of its mixture embedding, as can be seen in the previous example of figure 3.4. The Max external accepts the same control messages and is instantiated the same way as with lom.mutli, except that an extra argument is given for the sound space dimension. Further, this object accepts a list of output values (i.e. sound states) from a textfile – as with lom.si – that define the pointwise map between these and the control space grid structure. As an alternative to outputting sound state values, an object by the name *lom.mutiw* was created in order to provide the weights that correspond to the given sound states, which are output as a list and can be used for another purpose by the user.

lom.rst

I have implemented this mapping in Max/MSP based on code from the GRASS system for GIS modeling. The object accepts an argument for the number of points as well as initial values for smoothing and tension. An optional fourth argument determines the control dimensionality – either 2 (default) or 3. In order for a meaningful surface to be defined, a list must be given with coordinates for each point as well as a weight value for that point. This can be entered from a text file or manually using an interface similar to figure 3.6. As with lom.multi, this object may be used for direct embeddings, as it outputs the interpolated weight of a given point in control space. Further, the smoothness and tension parameters can be dynamically tuned in real time, in order to adapt the surface.

lom.rst2

The RST technique can also be extended to a mixture embedding as well. However, we cannot guarantee that the associated scalar weight values are positive and sum to one as with the SI techniques barycentric coordinates. They are further sensitive to scale as well smoothing and tension, with the physical relevance varying as a function of these. Through empirical means, I have found that when the $\{\lambda_j\}$ weight parameters from equation 3.1 become irrelevant, the output of the radial basis function R often becomes more well conditioned. For this reason, I wrote another object, lom.rst2, in order to output a list of weights as well as the direct output from the basis function. Therefore, due to the sensitivity of all associated scalar values, the user can condition the weights before applying them to desired sound states outside of the object, and thus creating a mixture embedding that retains the properties of the RST technique.

3.6 Chapter 3 Summary

This chapter was focused on developing a theoretical basis for mapping in digital instrument design. It began with a discussion of what precisely constitutes “mapping” per se, challenging the notion that this strictly describes parameter association. Rather, the relevancy of mapping was seen to be a function of musical control context – ranging from interactive systems that control high-level musical events to the immediacy of digital instruments with a close coupling of action to sound. Assuming the latter context, different perspectives (systems, functional, perceptual) on mapping were identified, its boundaries (with signal and gestural conditioning) discussed and sub-parts of the mapping process outlined (how vs. what). Taking off from this, I delved more deeply into a functional approach of mapping, formalizing this with a geometric interpretation and suggesting a framework for analysis. I then applied this theory and analytical framework to existing mappings from the literature and constructed a design space through which to understand the deeper structures of past approaches with the goal of combining and extending these. The chapter then ended with such a set of extensions in the form of the LoM toolbox. Now that this groundwork has been laid and the tools explained, the following chapter will discuss a set of studies and examples aimed at exploring these tools, examining their influence on

perception and discovering interesting as well as unique musical control structures.

Chapter 4

Mapping: Perception, Practice and Gestural Dynamics

In this chapter, I will examine the ways in which mapping is indeed contextual and how the choice of gestural and musical “vocabulary” determines the role and influence that certain mapping strategies may have on musical experience. I will illustrate this by virtue of perceptual experiments in which users were asked to subjectively rate their experience of various “instruments” defined by different mappings, as well as to engage with musical control “tasks” that allowed me to quantitatively compare and contrast their performance. While these mappings were designed to preserve the complexity and multidimensional nature of a musical performance experience, for the purpose of experimental design and control they were restricted to one mapping technique at a time, mapped in a fixed way to a given sound synthesis method. In order to move towards completely unrestricted and musically flexible performances systems, I will then extend these same mappings using modular design techniques and multiple layers of mapping that build upon the basic strategies presented in the previous chapter, resulting in new musical control structures that are more than the sum of these modular pieces. These examples were constructed with musical applications in mind, which again grounds them with musical intention.

While the user studies examine the influence of mapping structure on perceived musical control structure somewhat in isolation, the examples that follow this illustrate (by

example) the fact that mapping design is an interplay between musical intention/aesthetic, desired control/sonic gestural response, functional qualities of the mapping strategy and the choice of parameter association. Further, this interplay must be considered in a design feedback loop, an overarching theme to this dissertation. Finally, as a sort of counter-example I will present some examples in which the issue of mapping was approached from a complementary point of view wherein the qualities of signal and control gesture conditioning were of primary importance. This brings the issue of *dynamics within mapping vs. the dynamics of mapping* to the forefront, and leads naturally into the following chapter that will deal with control dynamics more explicitly.

There are thus three main results of this chapter:

- To the best of my knowledge, the first attempted studies that examine the influence of mapping strategies on the perception of control, sound quality and effectiveness in a digital instrument.
- The creation of several canonical musical control structures in which multi-layered mapping strategies can be used in an adaptive and user-controllable manner.
- A presentation of mapping design in the unique context of fabric-based controllers, including a discussion of how this design context differs from the more classic paradigm of multivariable instruments intended for solo performance.

4.1 Influence of Mapping Trajectory on Perception of Control

The authors of [4] articulated the multiparametric and immersive nature of musical performance. They refer to this situation as a *holistic* mode, in contrast to an *analytic* mode that arises from breaking down tasks into subsets and dealing with each as separate objects or entities. This latter mode of thought is aligned with the paradigm of the WIMP interface and most human-computer interaction (HCI) contexts. In a similar vein, the authors of [153] stress the importance of matching a controller to the perceptual nature of a task in a standard HCI context. They define *integral* and *separable* as fundamental descriptors for the perceptual structure of tasks, where these terms can be seen as parallel to holistic and analytic cognitive modes, respectively. The

authors of this work demonstrate the importance of matching controller and task through a series of user studies informed by the HCI literature, in the spirit of similar work such as [172] and [173]. Meanwhile, an objective comparison similar to that being done in classic HCI design contexts has been applied to the choice of transducers for musical control in theory [174] and in practice [175].

The use of comparison standards from HCI and the fundamentally different nature of musical interaction (when compared to the WIMP paradigm) raises the question of what defines an appropriate musical task. This issue is addressed in [176], in which the authors put forth suggestions geared at adapting HCI tasks to a musical context. In section 4.2.3 I will further address this for the context of laptop-based performance of electroacoustic music. The approach taken by the authors of [4] was to change the level of complexity in terms of the correspondence between controller and synthesis parameters. While the results were informative and the chosen task appropriate given the holistic nature of musical interaction, it is important to note that the controllers themselves were changed as well. Thus, one must consider that the perceptual nature of these controllers contributed to the perceived difference between the interfaces. This fact was acknowledged in [5] and a new study was briefly mentioned in which the controller was fixed (MIDI fader box) while the association of the position and/or velocity values was mapped to sound parameters in various ways. In the context of the recent mapping discussion, this approach examines the effect of *what* is associated. However I maintain that it is equally important to examine the effect of *how* this association takes place in the context of conditioning gestural response as well as continuous control type. In particular, taking a multi-parametric and spatial approach to mapping and timbral navigation, a holistic mode of control and musical interaction is that much more dominant. To examine more precisely the interplay between the geometric structure of mapping and musical control, I conducted a preliminary perceptual study (first presented in [149]) that examined the role of mapping structure both in terms of subjective response as well as quantitatively in regards to visual feedback, in order to examine the usefulness of a geometric strategy doubling as a visual mapping representation.

4.1.1 Preliminaries

In order to isolate the effect that different mapping geometries can have on an instrument, a user study was constructed consisting of a simple interface comprised of a Wacom tablet and FM synthesis. The (x, y) position of the tablet was mapped to the intermediate FM parameters of carrier frequency f_c , harmonicity H and modulation index M . Thus we can think of a two dimensional control surface directly embedded in a three dimensional intermediate space, with another mapping directly to the lower-level synthesis parameters for modulation frequency f_m , modulation depth m and again carrier frequency f_c . In this example the synthesis parameter space is of the same dimension as the intermediate space, and the second mapping layer is automatic in that it is defined by the use of FM synthesis. In this way, the mapping defines the same situation as that which was illustrated in figure 3.3.1 during the discussion on direct embedding. However even with a single geometry defined as such, note that the way in which we embed the control surface and in what parameter space defines the complexity of the mapping. The resultant FM interface is such that moving in one direction along the tablet affects change in all three sound parameters, and thus even this simple example is not simply a one-to-one mapping as is illustrated in the systems view flowchart of figure 4.1.

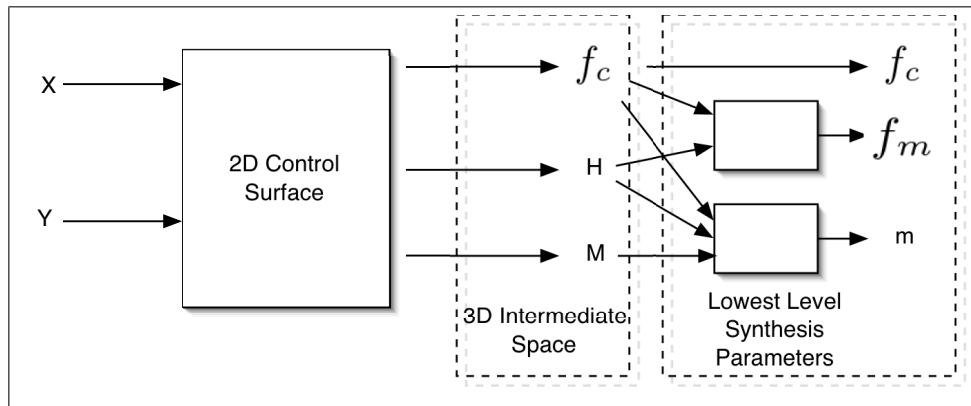


Fig. 4.1 Tablet controller as surface embedded in a three dimensional FM synthesis space.

Using the same parameter mapping between tablet and FM synthesis (e.g. X position controls H , M), I constructed two different mapping surfaces – one based on a

multilinear map and another based on the regularized spline with tension (RST) technique. Both of these are continuous and differentiable, but only the latter has higher-order smoothness while the former is comprised of hyperbolic “patches” that produce more variation between states [143].

4.1.2 Qualitative Assessment: Expressivity and Ease-of-Use

Eight subjects were presented with different incarnations of the interface as defined by the two mappings. Every combination of parameter mapping between controller parameters (x, y) and synthesis parameters (f_c, H, H) was presented in random order. That is, the two dimensional surface defined by the tablet controller was directly embedded in a three dimensional space in a fixed manner, and the dimensions of this latter space (adjusted for proper scale/resolution) were assigned randomly to the three synthesis parameters. For each set, both the multilinear and RST mappings were used to create the surface passing through (or very near) stored preset points.¹ These mappings were also randomized in terms of initial order of presentation, and once presented were fixed for a given parameter set while being differentiated by the labels “number 1” and “number 2”. Thus subjects were presented with a given parameter set defined by a given parameter mapping and a different mapping function within each set (e.g. interface A-1/A-2, B-1/B-2, etc.). The subjects were told in advance that there were potential differences within and between each set, and were given time to explore each interface. After familiarization they were asked to move the stylus along particular constrained paths on the tablet surface (again, this was done for each possible combination of parameter/spatial mapping). The subjects could move back and forth through this trajectory indefinitely and in any fashion, and could switch between the two choices within a given parameter set at any time using a key on the computer keyboard. Each subject was asked to give subjective reactions to the two interfaces, citing any differences in general and in particular were asked to comment on the expressivity and ease-of-use. These two terms were further defined to mean “musically interesting” or “instrument-like” in the former case, and in the latter case was related to the idea of repeatability and being able to “find” a point in sound space

¹While the RST technique is approximate, the surface was tuned so that the difference in steady state preset sounds between mappings were imperceptible in informal listening tests.

or create a desired musical gesture.

The reactions of the participants were quite consistent. While it was true that each mapping was constructed in such a way that varying one parameter affected several synthesis parameters, the perception of this was not equal in all cases. Regardless of the order of presentation or the orientation of control surface in sound synthesis space, people consistently found the multilinear mapping scheme to be more interesting from a musical standpoint. In fact, the majority of those interviewed stated that this mapping “added another dimension” to the interface (though this was not actually the case). Other comments ranged from “it is more non-linear” to “this one sounds more gestural.” Thus, in this musical context where absolute position (e.g. specific pitches) was not important, the relative motion of trajectories through sound space characterized by patches of smooth and sharp transitions between points was favored over the highly smooth surface. In other words I found that *the dynamic quality associated with the transition between sound synthesis parameters – as determined by the mapping – strongly influenced the perceived expressiveness of the interface*. Further, the how mapping aspect was the dominant influence regardless of the what (e.g. parametric) mapping configuration.

There was a different reaction when subjects were asked about the ease-of-use, defined in terms of one’s ability to “find” a sound and repeat this. The majority of users found the RST mapping better in this regard, saying that it was “more direct” and was, for example, “easier to find a specific pitch”. Thus there existed a conflict in terms of the mapping preference: the multilinear technique was deemed more expressive and musically interesting while the RST mapping was deemed easier to navigate the sound space with. This shows that in choosing such a mapping strategy *a tradeoff exists between the expressive potential and its ease-of-use and repeatability*.

4.1.3 Quantitative Analysis of the Effect of Visual Feedback

When used with a two or three dimensional controller, both the multilinear and RST mapping strategies have an inherent visual representation by their nature. Thus in the case of navigating through an abstract sound space, visualizing the mapping itself might conceivably help one to know precisely where they are in some

appropriately-defined intermediate space. On the other hand, the perceptual nature of control might be different than the perceptual nature of the sound space, in which case perhaps the visualized surface should reflect this control structure while a second mapping translates this representation into sound synthesis space. This latter approach was taken in [21], in which the mapping from control to sound synthesis space was achieved by a piecewise-linear interpolation over triangularized regions, but the visualized mapping surface was based on a spline interpolation. In this case movement of a cursor over a smooth rubber sheet-like surface resulted in a potentially jagged movement across piecewise simplices in sound synthesis space. In screen-based interactions such as this, the perceptual structure of control space and sound space should be explored separately as well as in tandem.

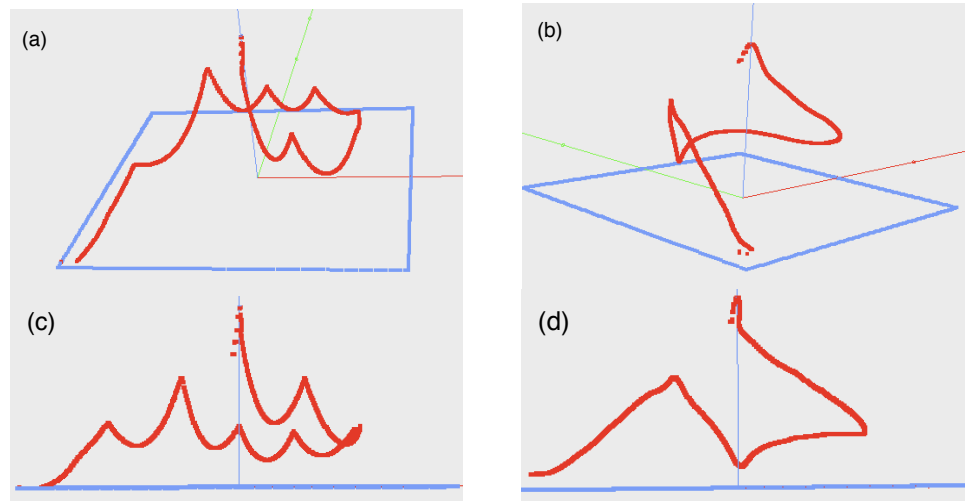


Fig. 4.2 Visualization of trajectory across a mapping surface (a) Multilinear 3D trajectory and (c) 2D Projection. (b) RST 3D trajectory and (d) 2D Projection.

In order to explore such visualization of control space, a quantitative “target acquisition” task [172] was constructed as a way to examine the effect of using a single mapping surface for both parameter mapping and visual feedback. The same Wacom tablet/FM synthesis interface from the previous test was employed. For this experiment, a box was presented on screen containing an ‘x’ placed at one of two locations. The presentation order of location 1 vs. 2 as well as the given mapping were randomized. When the stylus was moved across the screen, a “trace” was left of the

trajectory as in figure 4.1.3. For the test, the subject's view of the controller was obstructed, so that the only visual feedback they were relying on was screen-based (as well as inherent proprioceptive and sonic feedback). The participants were instructed to "acquire" the target x by using the visual feedback in the form of the trajectory, and were informed that this would be timed. The timer began when the subject pressed a button on the stylus and stopped as soon as the target was reached – at which point an on-screen button flashed and a distinct bell sound could be heard indicating the successful acquisition. For each test the stylus began in the lower left corner of the tablet controller. The time to target acquisition was recorded for each subject across all tests, with the results shown in figure 4.3.

The graph of figure 4.3 displays the mean and standard deviation for acquisition time in ms for the multilinear and RST-based mappings at both locations of the target point (note that the range is different for the two graphs). In both instances, the RST mapping took less time to acquire the target. This difference was exceptionally large for location 1, and was considerable for location 2 as well. I attribute added difficulty of acquiring location 1 with the multilinear based mapping to the fact that its position was at a global maxima on the mapping surface, whereas location 2 was situated at a local maxima. This further seemed to affect the variance, as this was quite large for multilinear location 1. Overall, the variance was considerably lower for the RST based mapping surface. Thus, from this test we see that the RST mapping was more intuitive and consistent as visual feedback in comparison with the multilinear scheme.

4.1.4 Preliminary Study Conclusions

The role that mapping plays in determining the "feel" of the instrument was explored in these two initial studies. In particular my interest was in the *how* component of mapping as embodied by its geometric structure, and the manner in which this determines parameter trajectories and thus influences the musical gestures that are possible with an instrument. I isolated the effect of this aspect of mapping by varying interpolation strategies while keeping controller, synthesis algorithm and pointwise mapping fixed. The results indicate that the dynamic quality of the mapping surface does alter perceived expressiveness. However, a tradeoff was found between the increased musicality afforded by a mapping and the ease with which one can explore

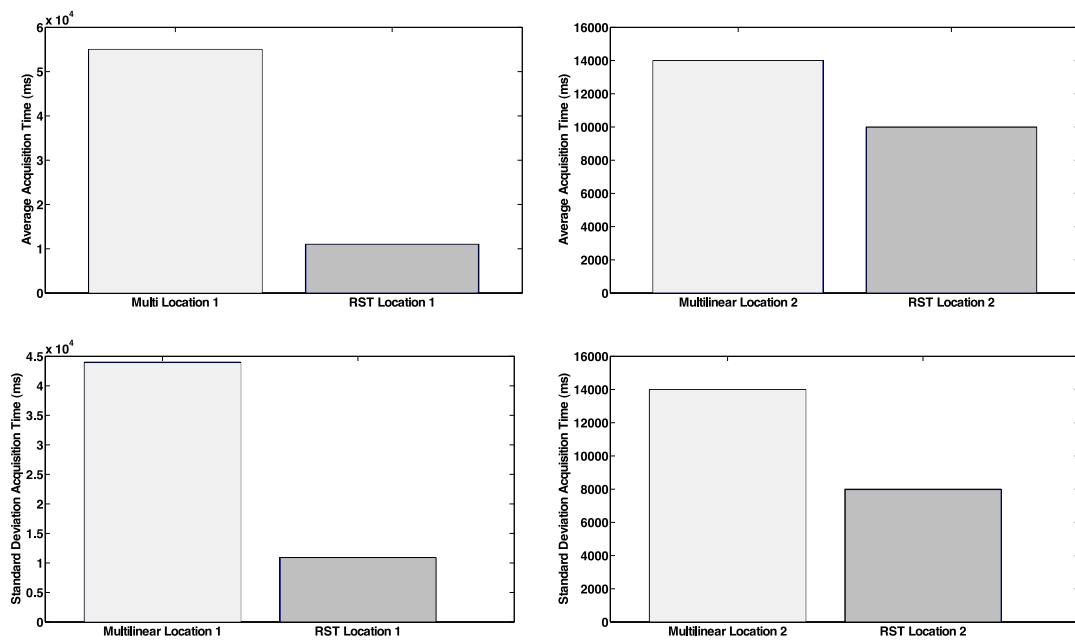


Fig. 4.3 Mean (top) and standard deviation (bottom) for acquisition times from test 2. Graphs depict mean of location 1 (top left) mean of location 2 (top right) standard deviation of location 1 (bottom left) and standard deviation of location 2 (bottom right).

the sonic terrain of the instrument. This tradeoff extended to the use of the mapping for visual feedback as well.

4.2 Analysis of the Interaction Between Mapping Structure, Sonic Gesture Type and Perceptual Control Structure

In order to explore the influence of mapping structure on the perception of control more deeply, a second more formalized and in-depth study was conducted that introduced more holistic musical performance “tasks” for expanded quantitative analysis. The musical context was further considered and defined, in order to situate the results more meaningfully. This second perceptual study sought to extend the preliminary results in several ways: it was more formalized in its methodology and execution, it quantified and expanded the subjective ratings of subjects – learning from the verbal responses of the previous study, and it introduced a test that would measure the ability of subjects to create sonic gestures with a given mapping in a holistic and musical environment.

In extending and formalizing the exploration from the initial test, I wanted to further examine a few key points that arose: the manner in which the perceived complexity of an instrument is affected by a mapping as well as how mapping structure influences the perceived aesthetic and musical quality of an instrument.²One thing that was clear – in terms of subjective response – in the preliminary study was that gesturing along a constrained path seemed too separable and unmusical. Therefore, I focused on designing a more holistic and overall more musical task. In the process, I extended the experiment to include an examination of the role of sonic context.

As a result of these updates and modifications, in designing this second set of experiments some key changes were made. In regards to examining subjective perception of overall musical feel and aesthetic quality, I chose to omit the term “expressiveness”, as I felt this was too strongly tied to a classical notion of emotion expressed by way of embellishments such as vibrato or tremolo. Returning to the

²Rather than extend the quantitative visual target acquisition test, it was simply re-run during this more formalized testing. Similar results were found that strengthen the above conclusion regarding the tradeoff between musical preference and ease of visual navigation. Having said this I will focus on the updated, sound-focused experiments for the rest of this section.

discussion of chapter 2, I feel that modulations such as these may become more intertwined with musical structure in electroacoustic music, and the qualities that contribute to expressiveness are less straightforward. As again, “electroacoustic instruments” are my primary interest throughout this work, and so I focused on terms that translated into this realm in a more neutral fashion. As I will discuss in the test description, I focused on the overall subjective sound aesthetic preference, the feeling of control one had with the instrument and finally the perceived complexity one felt the system possessed.

4.2.1 Experimental Design and Preliminaries

Space and Presentation

All phases of the experiment were carried out in the perceptual testing laboratory of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT). Subjects were seated in a standard ergonomic office chair, between two Dynaudio BM15a active monitor speakers, with a flat-panel screen, keyboard, mouse and Wacom tablet presented in front of them. Software for the test was written in Max/MSP and Jitter, and run on an Intel-based Macintosh laptop. The sound volume from the laptop and into the speakers was set to a normalized level of playback.

Subject Backgrounds

Nineteen subjects³ were recruited for the experiment, and compensated five Canadian dollars for each fifteen minutes of their time. The breakdown included four female and fifteen male subjects. Each was presented a questionnaire prior to the experiment (listed in appendix A), asking several questions to determine skill level or background that would provide any relevant advantage or training. Subjects were asked whether they had any musical training (defined as lessons in music theory based on playing or listening) or if they actively played an instrument – and for how long. Of course this was to separate the groups of musicians from non-musicians, and to help in identifying particularly trained subjects. While this is quite common in music perception experiments, I have yet to encounter another such study that has inquired into subject experience in electroacoustic music. As this study and larger work is targeted at this

³For test 1, data from three subjects was corrupted and could not be used.

sub-discipline, I asked subjects if they created EA music or listened to it on a regular basis (nearly every day). EA music was defined for the purposes of this experiment as “music that is primarily focused on shaping sound (timbre, texture) rather than pitch or rhythm structures (these being secondary), and that requires the use of electronic technology (recording, processing or synthesis) for its creation.” Finally, a third set of questions sought to detect any advantages based on the tablet controller or the nature of the control action, and so subjects were asked whether they were familiar with the Wacom tablet or if they were active as drawing artists.

Sound Synthesis

The two methods of sound synthesis employed were granular and a modal source-filter approach. The first one was defined by an abstraction that I created in Max/MSP, while the latter utilized the resonators and res-transform objects [177] that allow one to provide a bank of two-pole IIR filters with a list of frequency, amplitude and decay values – one for each mode – and to transform these with higher-level parameters. In particular my implementation controlled eight such parameters: two parameters for a global scaling of the decay rate and amplitude of all modes, two others – spectral corner and spectral slope – for a simple spectral envelope control, and finally four parameters to create “clusters” of resonances around a given mode. This includes control of which mode to center the cluster around, and the “spread” or spacing around the given mode in terms of frequency, decay rate and amplitude. Moving these clusters has the effect of indirectly controlling the roughness of the resultant sound. I thus used this approach in order to influence the overall roughness and timbre via the control of spectral envelope and relative contribution and decay of modes. The bank of modal filters are driven continuously with white noise, further centering the focus on the resonance qualities and control of these.

The granular synthesis method is actually a mixture of two grain streams, defined by different implementations: one in which grains are synced by a signal-rate (SR) sawtooth wave and one in which control-rate (CR) events determine the individual grain playback. The two streams share all relevant synthesis parameters, which allows for a more diverse granular sound that nonetheless sounds like one source; the implementation is described further in [178]. The individual amplitude levels for each stream define two parameters for the synthesis, with the others being grain duration

(5-100 ms), input buffer position randomness (0-500ms) for both streams, gain for a per-grain feedback delay, grain density for the control-rate stream and finally grain overlap for the signal-rate grain stream. The input waveform is a vowel-like fragment taken from a vocal performance.

These two synthesis models define two different sonic contexts. The modal approach allows for control of high-level spectral and spectro-temporal properties that contribute to timbre and roughness in the sense of spectral grain (as defined in chapter 2).

Meanwhile, the granular approach controls relatively lower-level signal parameters that affect the transient grain, the overall temporal smoothness, and spectro-temporal dynamics by controlling grain size and the variations on the input waveform. This is encapsulated in table 4.1, which describes the primary sonic gestural quality that is being affected for each sound parameter.

Using these two disparate approaches allowed me to construct two difference sonic worlds that were distinct enough (while remaining musically interesting) to allow for the influence of sonic context on mapping and control perception to be examined in a separable manner. This desire to control different grain and roughness qualities, in two different sonic contexts, necessitated the move from FM synthesis (as in the preliminary experiment) to granular and modal synthesis. The 3-parameter FM instruments, in my opinion, did not produce the same level of musical richness, and did not allow for the same degree of timbral and textural manipulation as these two. However in the modal case, there is more of a clear separation of timbral and roughness control, while the lower-level granular parameters have several interdependencies. For example, in the granular instrument the mixture of what is more noticeably controlled: short-term timbral attributes or separable textural ones depends on the relative balance of the amplitudes for the two grain streams. These differences made it clear that the two sets of instruments were rich enough to present sufficiently different sonic contexts to be suitable for a perceptual study.

Mapping Structure

While the sound synthesis algorithms alone defined the *sonic context*, in terms of the perceived action/sound behavior the addition of a mapping structure defined the overall *sonic control context*. That is, the perception of the sound result and overall control dynamics were not simply a product of the synthesis used, but were also

Synthesis Type	Synthesis Parameter	Primary Sonic Feature
Modal	Global Decay Rate (all modes)	Matter Profile
	Global Amplitude Rate (all modes)	Spectral Envelope, HT
	Spectral Corner Spectral Slope	
	Cluster Center Frequency Cluster Frequency Spread Cluster Rate Spread Cluster Amplitude Spread	Roughness, Spectral Grain
	Output Amplitude	
Granular	SR Stream Level CR Stream Level	Balance Timbre/Texture Control
	Buffer Position Randomness Feedback Delay Gain	Timbre: Mass and HT
	Grain Duration Grain Density for CR Stream Grain Overlap for SR Stream	Motion, Transient/Spectral Grain
	Output Amplitude	
		Dynamic Profile

Table 4.1 Sound synthesis type, sound parameters controlled and sonic gestural qualities affected.

determined by endowing these with a mapping structure.

As with the previous experiment, a Wacom graphics tablet was used for control. The resulting interface utilized five absolute degrees of freedom: two dimensions of position, two for stylus tilt and one that sensed pressure on the tablet. These five parameters were mapped to the two sound synthesis spaces using two distinct mappings based on simplicial (SI) and multilinear (MI) interpolation strategies. Recall that the former can be defined from scattered pointwise mappings, while the latter cannot. For the sake of a controlled experiment, the preset instrument states were aligned in a grid in control space. The geometric mapping structure was applied to the four position/tilt dimensions as follows: a pointwise mapping was defined between nine control states at the extreme of the control space and nine respective sound states of the given synthesis model. As both of these mapping strategies are exact, it was straightforward to ensure that the sound states were precisely the same between mappings. Therefore the steady-state response for a given sound set was exactly the same at preset states regardless of the mapping, and so it was only the *dynamics that resulted from the mapping geometry that varied between the two*.

In the case of the granular set, the sound states were a particular combination of the

eight aforementioned synthesis parameters. Thus the boundaries of instrumental space were different snapshots of a given timbral quality, determined by a particular combination of the low-level granular parameters. The modal set differed in that the states arose from transformations of a single model, defined by modal analysis of a bass string. Therefore the sonic control context, in the case of the modal set, is best described as smooth timbral interpolations between instrument models with a sound similar to a continuously-driven bass string (e.g. as with an *ebow*). The granular set provided a sonic control context that is best described as navigation between smooth and grainy patches of a vocal timbre space. Finally, I mapped the pressure dimension directly to global amplitude level for both sets, aligning with a commonly-held principle that a mapping should require continuous energy to produce sound if it is to be musically satisfying [4][179]. A nonlinear transfer function was applied to the pressure scaling in order to condition the pressure-to-loudness response, and this same function was used for all pressure mappings.

Given this mapping of all five input parameters, four different instruments are thus defined: two mappings (SI and MI) each for the granular set, and two for the modal. Each of these provide a control context in which continuous pressure shapes the overall dynamic envelope, and the other four degrees of freedom transform various timbral and textural qualities. In the language of chapter 2 this structure is well-suited to creating graduated continuant types of sonic gestures.

As I wanted to explore the perceptual nature of the attack-decay gestural type with continuous modulation, I created two more instrumental archetypes⁴ by slightly modifying the overall modal mapping: rather than mapping position to global *output* level, I mapped this value to the level of the input noise source. Therefore, the energy transfer in this instance is an excitation rather than continuous input, which partially determines the timbral nature of the attack. This attack-resonance control context – a natural fit for the source-filter nature of the modal set – therefore defines a fifth and sixth instrumental archetype for use in the studies. While this change in parameter association was slight from a systems point of view, the perceived effect on control was much greater, as can be seen in the first of the user tests. The six different instrument

⁴The former set will be referred to as the GC instruments and this latter as AD instruments from this point forward.

archetypes then, are four GC instruments – defined by the possible combinations of modal and granular synthesis with SI and MI mapping structures – and two AD instruments defined by the aforementioned change in parameter mapping to the two modal GC instruments.

4.2.2 Subjective Rating of Instrumental Qualities

The entire experiment consisted of two sub-tests ran in succession. The first of these was intended to examine the subjective response to certain defined qualities of each instrument. However, in order to encourage holistic “free play” – and thus a more musical interaction – subjects were not given the qualitative labels until after they had experienced each of the six different instruments. One issue that arises in taking this approach is the effect of memory: subjects need a minimum amount of time to explore the instrumental space adequately, but of course the more time added the more difficult it will be for subjects to recall the first instrument. Based on informal pilot experiments in which subjects played with the instruments for varied lengths and were asked to recall those earliest presented, I chose one minute for each instrument. Even with this short duration, I did not want the experiment to be a test of memory, and so subjects were allowed to briefly write notes to themselves between tests in order to facilitate recollection.

Similarly, I did not want the subjects to spend time in the free play period learning the controller. Therefore, the functioning and degrees of freedom of the graphics tablet were explained to subjects a priori, with physical and gestural examples presented by the test administrator in order to demonstrate the limits of the controller. While the language “instrument” was used for each set, the fact that the study explored the influence of mapping on perception of control was made clear at the outset, in the preamble that each subject read and signed (included in appendix A). As another neutralizing test condition, the six instruments were randomly presented to each subject.

The process for test 1 was as follows: a key was pressed on the computer keyboard to begin the timer and randomly load an instrument, while an on-screen display alerted subject as well as administrator of the start and end of the timer. The subjects were

asked to explore the instrument freely during this time – taking notes only between instrument sets – and were informed that they would be comparing and contrasting the instruments after experiencing them all. After all six had been presented, a window was displayed showing a bank of on-screen sliders as in figure 4.4, allowing for a relative rating along a continuous scale. The first quality that was rated was “sound preference”, which was defined as one’s subjective preference for the sound output in general of a given instrument. After this, subjects were asked to rate “controllability”, defined as the strength of a feeling of coherence or correlation between physical action and resulting sound output. The third quality was “transparency”, which was defined as the strength of a feeling that a given movement degree of freedom affected one element of sound in a separable way (as opposed to many elements, in a complex way). This could also have been defined as “one-to-oneness”, but this language was avoided for the sake of those not familiar with this terminology. I collected the responses of all participants and examined their general trend as well as specificities in order to better understand the perceptual control structure that arose from a given mapping structure.

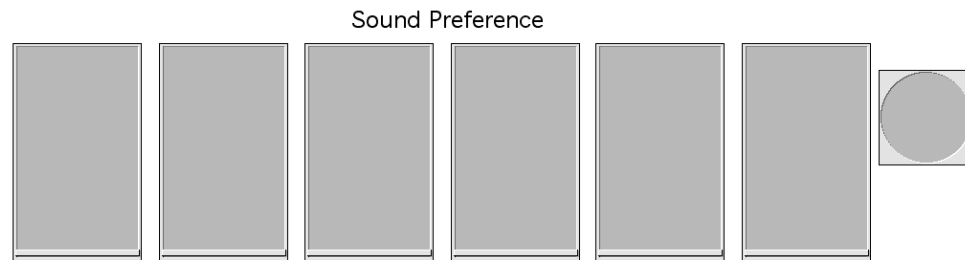


Fig. 4.4 Side-by-side ratings for sound quality in test 1.

Global Response Results

Sound Ratings

In terms of the sound preference ratings, one might expect that the different sound sets – modal and granular – would cause people to consistently prefer the sound world of a given synthesis technique. That is, one might expect that the sound of the modal-AD instruments would receive ratings similar to the very closely-defined modal-GC

instruments. This was not the case, as subjects varied their ratings such that the four modal synthesis-based instruments did not co-vary when compared to the granular instruments. However, the majority of subjects preferred the sound output of an instrument defined by the *same mapping* for both modal-GC and for the modal-AD instruments. In other words, if a subject preferred the sound of the MI-based modal-GC instrument, they also preferred the MI-based modal-AD instrument. In other words, for the overwhelming majority of people (15 of 16), if they preferred the sound arising from a given mapping for the modal set, they stuck with this across the differing control structures (GC vs. AD).⁵ Therefore, the preference for overall sound output quality *followed more strongly along mapping lines than across the changing control structures* – in this case the changing control structures meant moving from continuous pressure-to-amplitude into an attack-resonance type of control.

While the subjective preference of sound was not strongly tied to a given sound synthesis technique in nearly every case, there was a much stronger correlation with mapping type: 12 of 16 subjects consistently preferred the sound of an instrument that was defined by one particular mapping, across all six. The particular preference of mapping type varied between subjects: a few consistently preferred instruments defined by MI (5 subjects) while others preferred SI-based instruments (7 subjects).

Given this majority, it is not surprising that this trend of preferring the sound consistently with mapping type (across sound synthesis types) also was true in the mean of all subjects, as can be seen in figure 4.5. On average, the SI-defined instruments were preferred in terms of their sound output. Taken as a whole these results are highly interesting and, to my mind, counterintuitive: *while the sound synthesis type did not have an isolated and direct impact on preferred sound type, in the majority of cases – and on the average – the mapping structure did.*

Controllability Ratings

In terms of the controllability ratings, there was again a coherence with mapping structure – however this was only one of two coherent groups for rating type. Of the 16 subjects analyzed, 14 of them belonged to one of two coherent groups – split evenly with 7 members apiece. The first group preferred the controllability of a given instrument strictly along mapping lines, as with the sound preference ratings. That is, for a given

⁵The sound and controllability rating data can be viewed in appendix B.

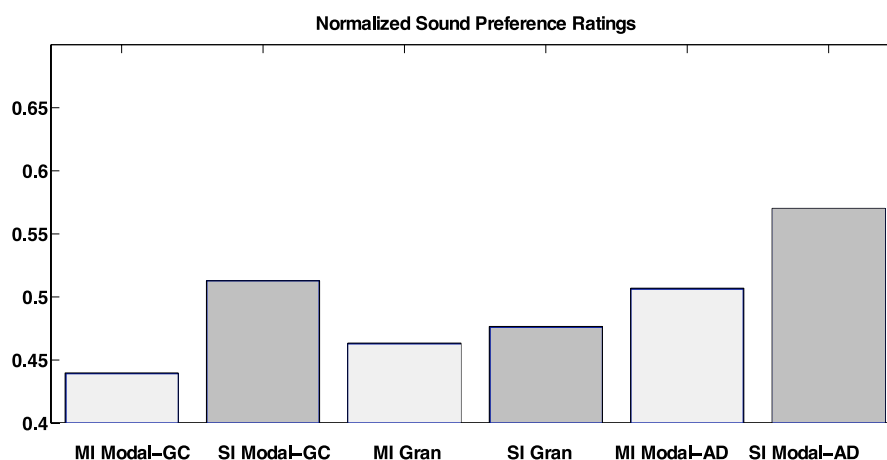


Fig. 4.5 Mean Normalized Ratings for Sound Preference.

sound synthesis type, the one with a given mapping structure was always preferred. Meanwhile, the other group preferred controllability of a mapping based on sound synthesis type: one mapping was preferred for the four modal instruments, and another for the two granular instruments. Of the 7 that rated in this way, 5 of them preferred the SI technique with modal synthesis and MI for the granular synthesis. I will discuss more on the background of these two groups in a moment.

While it would seem from this grouping trend as though there was less consensus in this rating set, the important thing to note is that in both cases the perception of controllability varied with the mapping type. In one case this was outright while in the other this changed with sound synthesis type – in other words as a function of sonic context. On the average, the group that rated controllability of mappings differently along sound synthesis lines rated this difference much more strongly. So much so that this defined the global average, as is clear when examining the average controllability ratings from figure 4.6, where the overall preference for mapping make a clear change between sound synthesis types. Therefore while there is less consensus than with the sound preference ratings, on the average we can say that the group preferred a given mapping structure’s controllability based on sound synthesis type, and that this preference was clear: SI felt more controllable for the modal instruments while MI was more so for the granular instruments. This presents the interesting and (to my mind) counterintuitive result that suggests *the mapping structure directly influenced perception of sound quality*,

while the mapping influenced the feeling of being in control in a way that changed with sonic context. My musical hypothesis regarding this global phenomenon is that the former is due to the mapping geometry largely defining the sonic gestural response while the latter may be due to the relative high vs. low level of control between the modal and granular synthesis methods used. The changing controllability ratings with sonic context was mostly influenced by the strong rating pattern of an EA listener-heavy group, so that EA listening experience may very well be a form of training for interaction with these instrument types. This idea will be explored in a moment.

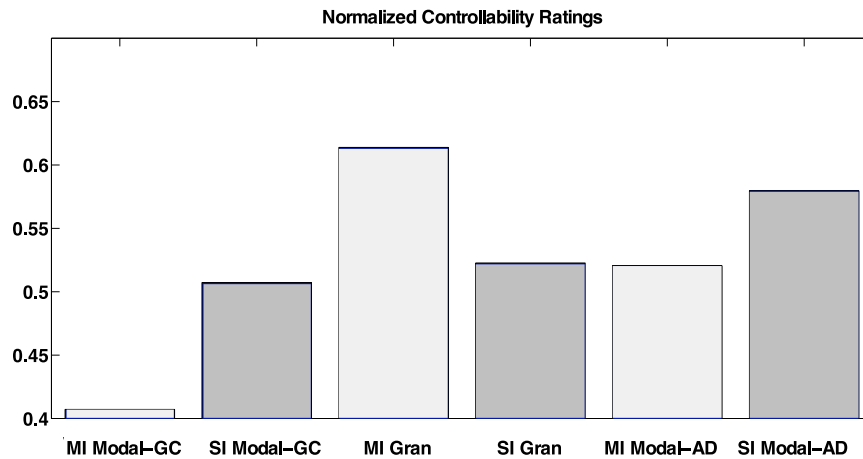


Fig. 4.6 Mean Normalized Ratings for Controllability.

Transparency Ratings

In terms of the global rating average, we can see once again in figure 4.7 that *the group in general felt the SI technique was more transparent with modal synthesis and that the MI technique was for granular synthesis*. What is most interesting about this is that it occurred with a lack of consistency between subjects. Specifically, only 3 subjects consistently rated a given mapping as more transparent while only 4 did this in regards to sound synthesis type. This is perhaps due to the fact that the perceived complexity/simplicity of a mapping can not be reduced to its geometric structure in parameter space or to the sound synthesis type in nearly as “clean” of a way as it can in regards to sound preference or controllability. The fact is, by nature of this approach the mappings are all many-to-many and complex. An exaggerated test would be to vary a simple parametric *what* mapping and see how perceived complexity is affected, though

this may create a less holistic and musical control situation – something that I aimed to avoid altogether with this second set of tests. Perhaps what should be taken from these ratings is that neither mapping structure particularly dominated or failed in terms of mapping transparency – so that perhaps the multi-parametric nature of the test design already defined mappings that were not very transparent. Something else to note from this rating set is that people did not correlate their controllability and transparency ratings, suggesting that they were able to separate these concepts in their mind. Noting these phenomena as interesting for future study, I will primarily focus on the sound quality and controllability ratings for the remainder of the discussion.

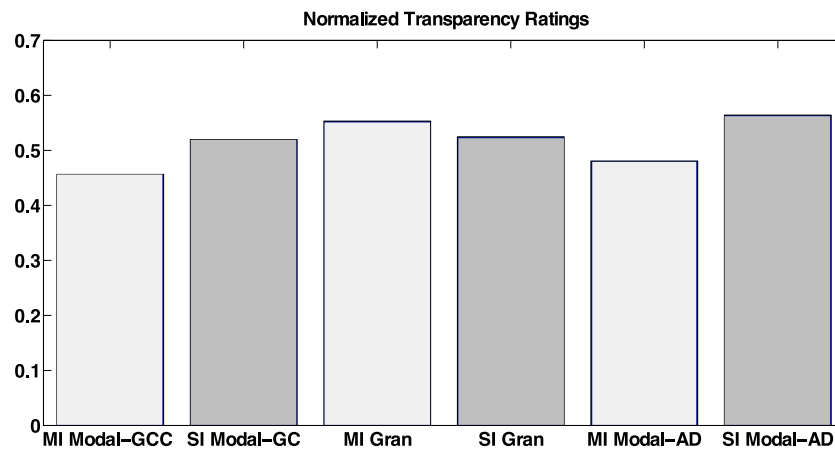


Fig. 4.7 Mean Normalized Ratings for Transparency.

Test 1 Global Response Summary

For both transparency and controllability ratings, the average subjects rated more highly SI for modal instruments and MI for granular instruments. Thus the subjects generally rated these qualities such that the control/transparency rating varied with mapping type, with particular choice being a product of sound synthesis type. At the same time – and I find this one of the most interesting result of the study – the sound preference ratings did not change consistently with sound synthesis type. Rather, the group rated their preference strictly along mapping lines, with a preference for SI-based instruments regardless of sound synthesis employed. This suggests in general that *the overall perception of sonic quality was integrally tied to the sound gestural dynamics as defined by the mapping structure.*

It is also interesting to note that for both sound preference and controllability, the preference for mapping did not change across the GC vs. AD control structures for the modal instruments. 15 out of 16 people rated more highly one mapping type (specific preference of SI vs. MI varied) regardless of the control context, while 14 out of 16 did so for the controllability ratings. Therefore, this illustrates that *subjects were able to translate their experience of the mapping/sound structure beyond the immediacy of continuous dynamics vs. attack-resonance*. At the same time, these control structures were a strong determinant of rating when it came to transparency where many people changed their mapping preference along these lines. This could be more due to the difficulty of that task – as well as a confusion between *how* and *what* mappings – rather than any specificity of the ratings, as discussed.

Test 1 Group Responses: Questionnaire Based

Another interesting set of phenomena arises if we look at the average response of different subgroups. From the questionnaires, we can define four groups: Those with formal musical training (8 subjects), those who play an instrument regardless of formal training (11 subjects), those that regard themselves as electroacoustic (EA) composers (4 subjects) and those that listen regularly to EA music (10 subjects). If we look at the means of these groups (depicted in figures 4.8 and 4.9), there exists different sub-tendencies: the musical training group, instrumentalists group and EA listeners all prefer the MI mapping across the board in terms of sound preference. Therefore, the overall sound ratings are largely biased by the EA composers, who themselves did not rate coherently along mapping lines. Further, 3/4 of the EA composers categorically did *not* rate sound preference strictly along mapping lines, representing 3 of the 4 who did not rate in this way. One hypothesis is that these subjects were more used to separating the sound qualities of an “instrument” – which could be a studio-based real-time sound processing algorithm – from the dynamics of the mapping structure.

There is further indication that “EA experience” defined a certain perceptual rating strategy in the two groups of controllability ratings: all 7 of those that rated the controllability of a mapping as a function of sonic context had identified as someone that listens to EA music on a regular basis, while only 2 members of the other group did. This might further suggest that some sort of “training” in regards to being an avid listener of EA music came into play, causing subjects to associate controllability of a mapping within a certain sonic context, while the other group focused on the controllability of a

mapping abstracted from this. Taking this further, if we look at the averages based on background, the musical training and instrumentalist groups rated controllability strictly along mapping lines. Essentially the musical training group was a subset of the instrumentalists group, and so combining these groups we can see that *those who were skilled in instrumental control rated controllability along mapping lines, rather than as a function of sonic context as in the global average*. One question that arises from this is whether experience in playing music changed the focus from sound algorithm to mapping structure, or whether experience with EA music changed the focus towards control/sound qualities of a mapping in given sonic context. To examine further, I will examine grouping tendencies from a different point of view.

Test 1 Group Responses: Cluster Analysis Based

These group averages arose from backgrounds that were self-defined with various degrees of overlap. Therefore, in order to examine grouping tendencies based on the rating data itself, I conducted a hierarchical cluster analysis, using a single-pass nearest neighbor approach and Euclidean distance metric. Clusters were examined in Matlab (statistics toolbox) both with measures of tree inconsistency and per-rating distance for various threshold levels in order to “prune” the tree and create consistent groupings. In regards to sound preference ratings, two distinct groups formed that were robust to these variations in clustering. These included all but two subjects, who formed their own groups and were discarded as outliers. Regarding these two coherent groups: the first (6 subjects) had only 2 members with musical training and 1 EA composer, while only half listened to EA music or played an instrument. Meanwhile, the other group (8 subjects) had 3 of the 4 potential EA composers (1 being discarded), with 7 of 8 members playing an instrument as well as listening to EA music regularly. Therefore, this latter group had a great deal more experience in musical training and performance in general, as well as specifically in regards to EA music.

Interestingly, while 10 out of 14 subjects from these two groups individually preferred the sound of an instrument strictly along mapping lines, these two distinct groups both change ratings with mapping as a function of sound context (see figure 4.10). The group with less general musical and EA listening experience preferred MI for modal instruments, SI for granular, and for the more musically and EA experienced group, it was the opposite. In both cases, however, the groups were consistent in preferring the sound of instruments across control structures (GC vs. AD) for the modal instruments.

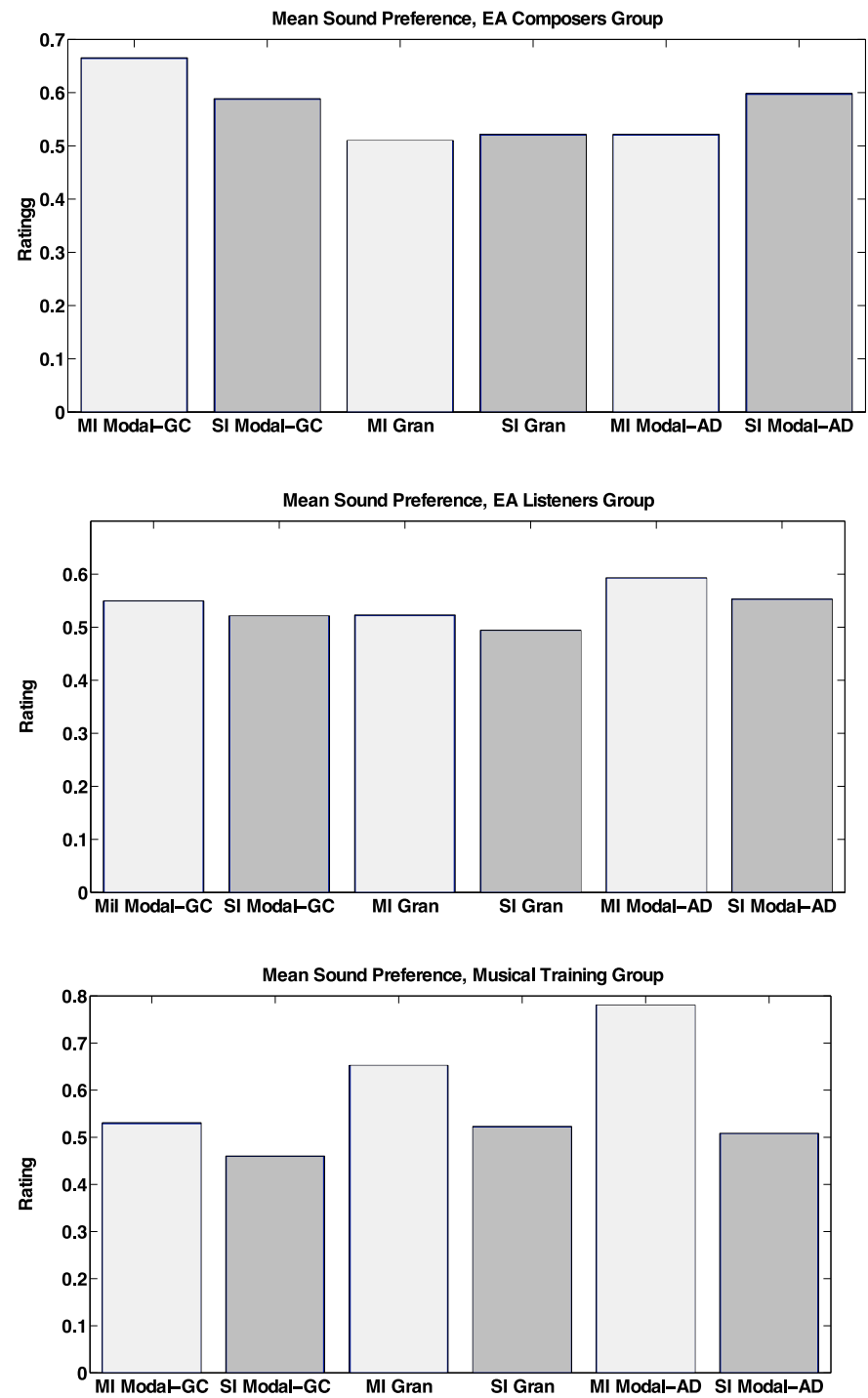


Fig. 4.8 Sound preference rating mean for Composers of EA Music (top) EA Music Listeners (middle) and Subjects with Musical Training (bottom).

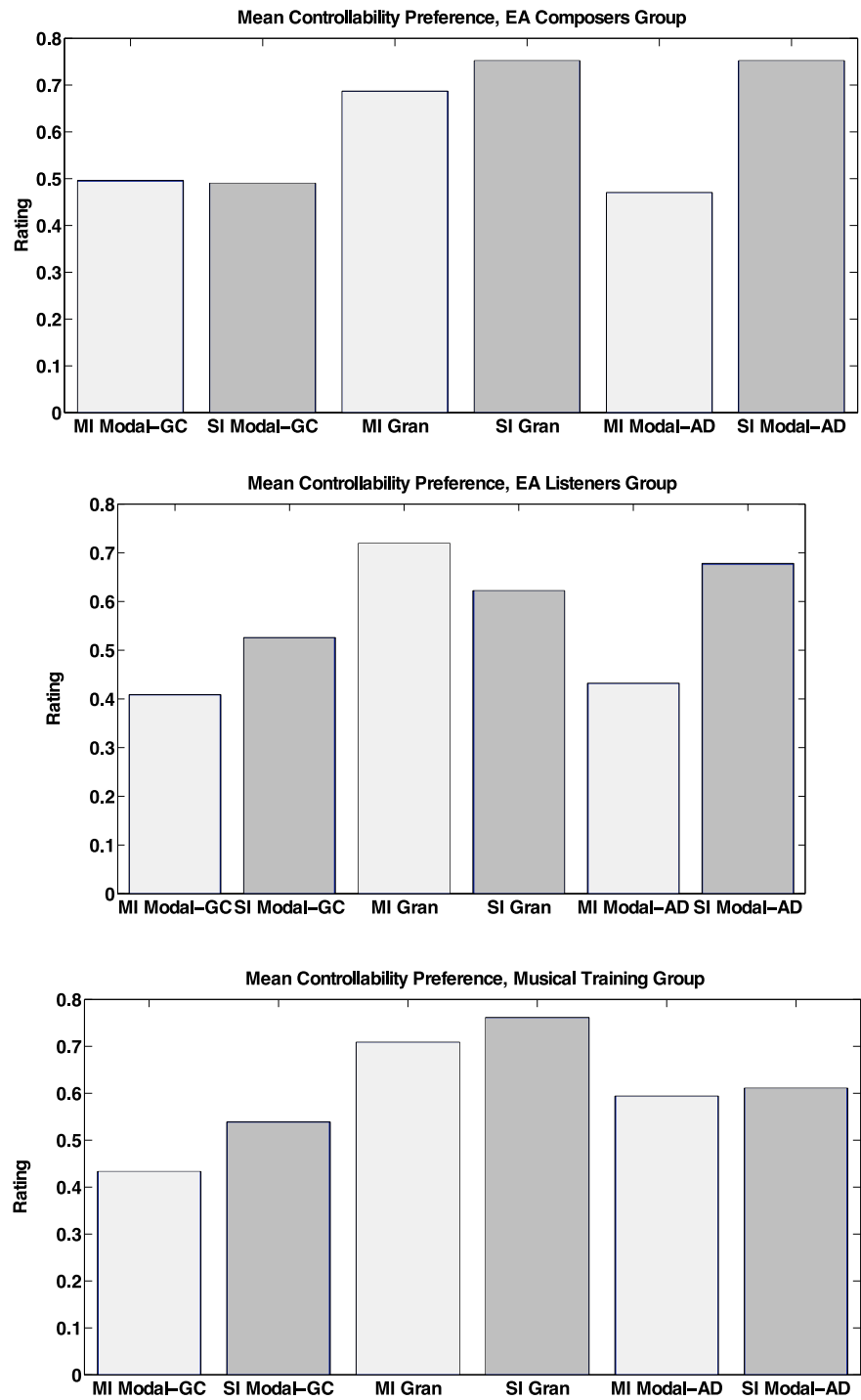


Fig. 4.9 Controllability rating mean for Composers of EA Music (top) EA Music Listeners (middle) and Subjects with Musical Training (bottom).

Therefore, this reinforces the previous conclusion that *users can separate sound quality from control structure type, while mapping influences perception of sound quality*. Indeed, these ratings showed that mapping preference was a product of sonic context for both groups, with a difference of opinion between groups as to whether MI or SI were better in different context. This is likely attributable to a group aesthetic decision, as the two differed greatly in their opinion of the SI-gran instrument's sound quality.

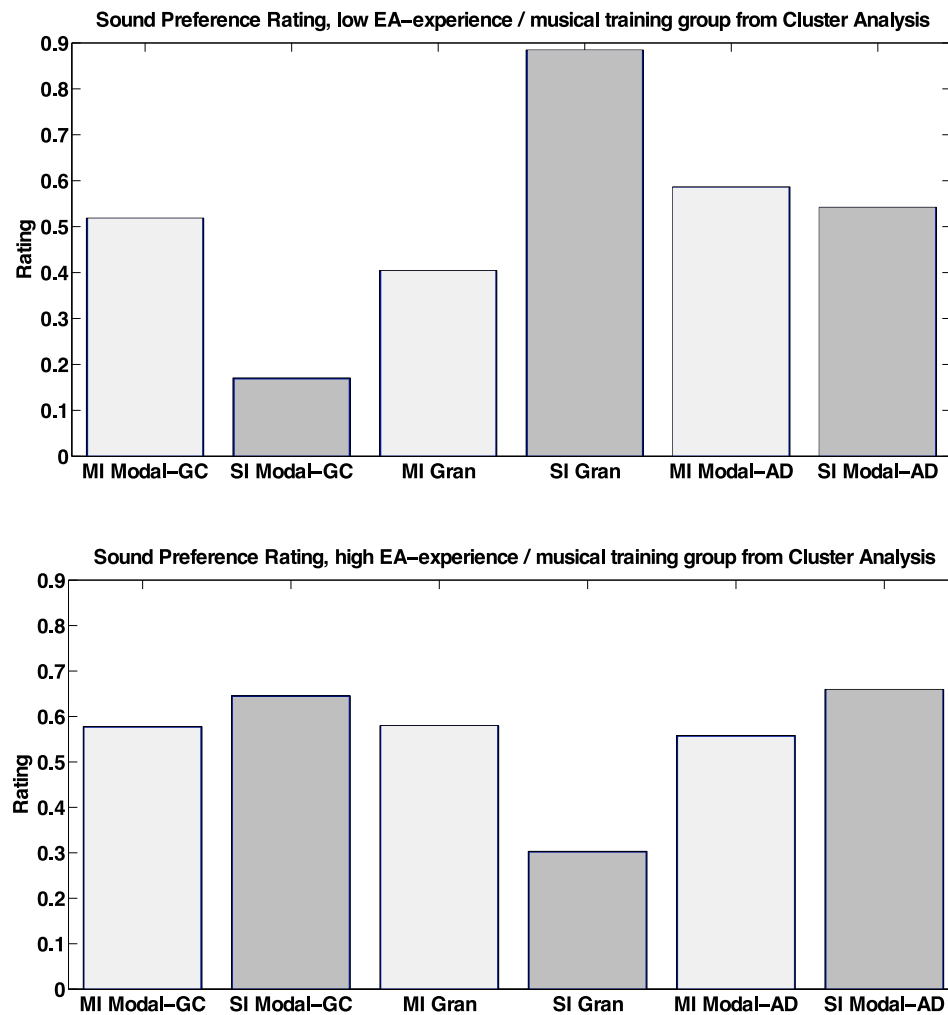


Fig. 4.10 Sound preference rating mean for the low EA-experience / musical training group (top) and the high EA-experience / musical training group (bottom).

Applying this same cluster analysis to the controllability ratings, we once again see two

groups forming (three outliers were discarded). The first group consisted of 4 subjects, all of which listen regularly to EA music with 3 of 4 having no musical training. The other group consisted of 9 subjects, 7 of which had musical training and 4 which listened to EA music. Thus we have one group with lots of EA listening experience and little musical training, and the second group which has strong musical training and less experience with EA listening. The group with high degree of EA experience and low degree of musical training followed the same strategy of preferring the controllability of instruments of a certain mapping structure, changing specific preference along sound synthesis lines. This group preferred the control of a given mapping in a given sonic context. Surprisingly, the low EA experience/high musical training group was not consistent, with little variance in ratings between mapping types. Further, the high EA/low training group had less variance between subjects with their ratings, while the low EA/high training group showed more variance between subjects for 5 out of 6 instruments. The mean and variance, respectively, for these two groups are displayed in figures 4.11 and 4.12.

Test 1 Conclusion and Significance

The first thing to take from this study – and which gives weight to the claim of my hypothesis – is that the *perception of the sound quality as well as overall control feel could not be separated from the geometric mapping structure*. This is evident in the way that ratings varied along mapping lines, whether in absolute or as a function of sonic context (i.e. sound synthesis type). Looking at the global averages, it seemed that sound quality preference was directly a product of mapping outside of sonic context – a very counterintuitive result, illustrating that the *continuous interaction with a given mapping structure influenced the perceived sound quality of the instrument*. This extended to the overall perception of controllability and transparency as these were also a product of mapping, changing with sonic context. That said, it is important to note that this first study did not possess sufficient power in order to achieve statistical significance. This can be seen by analyzing the observed power from the study, which gave values of 0.267 and 0.07 for sound and controllability ratings respectively in regards to the mapping/synthesis interaction effect. In other words, the chances of finding a p-value with this data set were too low in order to make a decision on significance. All the same, there is an important debate in the psychology literature as to the true importance of dichotomous significance testing. It is argued in [180] that continuous effect size measures paint a more accurate picture of the statistical relevance of findings, relative to the sample size in question.

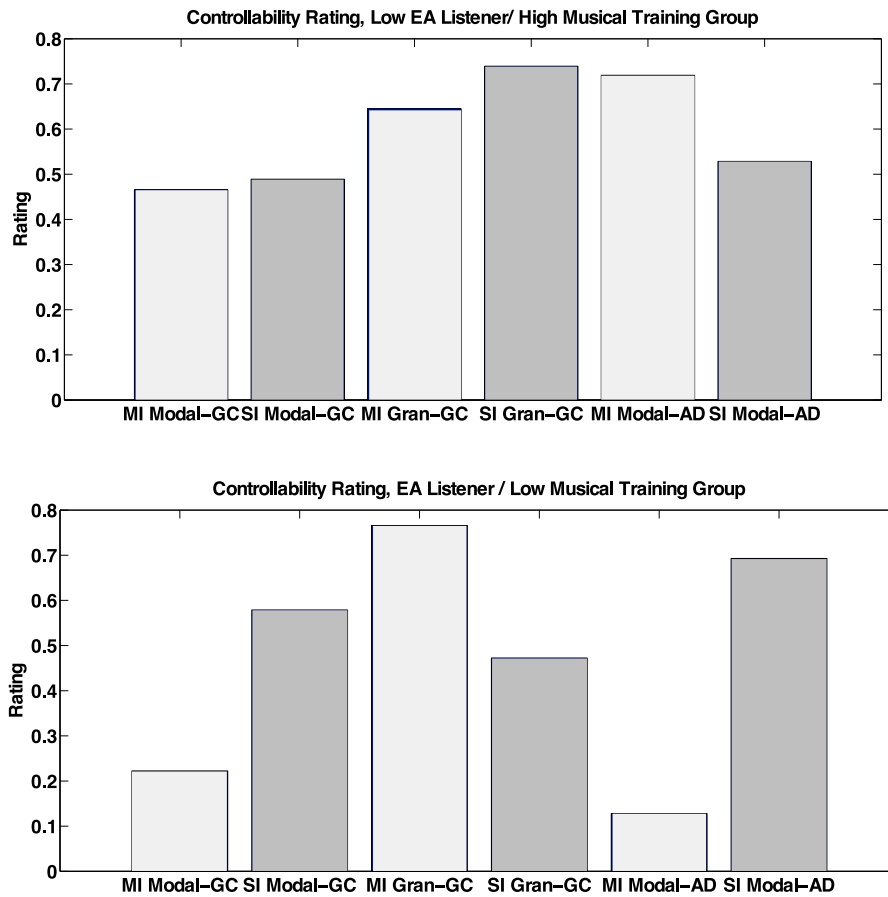


Fig. 4.11 Controllability rating mean for the low EA-experience / high musical training group (top) and the high EA-experience / low musical training group (bottom).

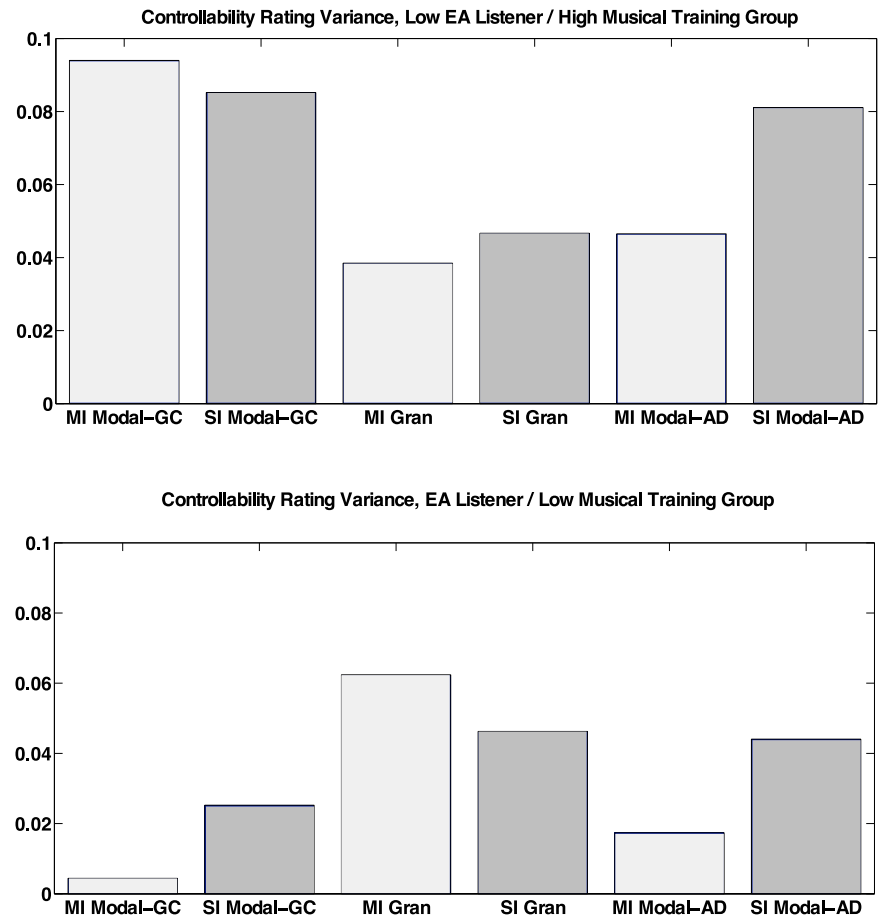


Fig. 4.12 Controllability rating variance for the low EA-experience / high musical training group (top) and the high EA-experience / low musical training group (bottom).

Following this argument, I have calculated partial eta squared values for each rating set. The values for sound ratings ($\eta_p^2 = .025$) and controllability ($\eta_p^2 = .054$) suggest a small-to-moderate sized effect [181] in support of my claims. Transparency does not follow this trend, which again might be due to the ambiguity of this term. Overall, I feel that this effect size-based measure is a strong enough statistical indication that this initial study does indeed possess validity and warrants further investigation, so that these experiments should be followed up with a larger subject group.

Looking more closely at subject backgrounds, it became clear that this had some influence on rating strategies and perhaps aesthetic biases in certain cases. Transparency ratings didn't show any group tendencies, which I attribute to the low level of mapping transparency of all instruments employed. In regards to sound quality rating, one clear division was that while 12 of 16 subjects preferred the sound output of instruments from one mapping type, 4 others – 3 of which being EA composers – preferred the sound of an instrument for a given mapping that changed with sonic context. Similarly, in the controllability ratings there were two coherent groups: those that rated highly the controllability of instruments with a given mapping (7 subjects), and those that again changed this with sonic context (7 subjects). The former group had few members (2) who listened to EA music, while the latter was fully comprised of EA listeners. Looking at these two results together suggests that EA music experience informs ones perception of sound as well as control, causing ratings to be made in a given sonic context, rather than absolutely.

As these were self-identified groups with overlap, I decided to examine how the responses themselves were organized through cluster analysis. Doing this revealed a more complex-yet-coherent structure in regards to background: one group with a lot of musical training and EA music experience, and one with little musical training and EA experience. Interestingly, both of these groups preferred the sound of instruments from one mapping type, changing with sonic context. The former group liked the SI mapping for modal instruments and MI for granular, while the other group was the opposite. This is interesting because it showed that taking musical training and EA music experience together, there were similar strategies employed, only changing in regards to specific preference. However as noted above if one looks only at EA music experience (in this case, composers) then this changed the rating strategy from rating along mapping lines to rating as a product of sonic context, which suggests that *EA music experience was a*

stronger determinant than musical training of the perceived sound-mapping interplay. This was further reinforced in the cluster analysis of controllability ratings, wherein two groups were formed that decoupled these two backgrounds: one with high EA listening experience and low musical training, and one with low EA listening experience and high musical training. As above, the EA-experienced group rated coherently the controllability of instruments along mapping lines, changing preference with sonic context. Interestingly, the musically trained group produced very flat ratings overall, with less consistency between subjects. Again, EA music experience – whether composing or listening – was a stronger factor in determining consistent rating strategies.

Most importantly, we can see that the mapping structure consistently was the salient feature in perceived sound as well as control quality. This underscores the importance of this aspect of mapping – as I had espoused in the previous chapter. It was further interesting that sound synthesis type outright did not influence subjects’ sound/control perception, yet it was the case that experience in EA music influenced ones perception of the nuanced changes in “mapping feel” as a function of sonic context. That is, mapping type was the strong determinant of sound/control rating in any case, and this changed with sonic context as EA experience was added. This suggests that the example instruments – focused on timbral and textural sound manipulation – were successful as “electroacoustic instruments” in that it elicited coherent responses from those with familiarity with the field. This is not to say that the set of responses from the EA-experienced were “better” – rather I hypothesize this set of instruments allowed for the nuanced changes in mapping feel across sonic context to be perceived, as this was done more consistently by this EA group. The fact that EA listeners – not only composers – succeeded in this way also gives some weight to the notion that such listening is an inherently embodied phenomena, as this experience seemingly carried through to these action-sound tasks.

By no means am I suggesting that I have proven the embodied nature of EA music listening, or that I have shown any mapping structure to be better outright (though there was clear preference for different mappings in different contexts, depending on the sub-group). In fact one could rightly say that this holistic approach to perceptual testing – in such a nascent field – perhaps raises more questions than it answers. However, I do feel that these results are quite promising in showing the influence of EA music experience on perceptual structure of sound manipulation – something not well

understood and which should be studied in more detail. Further, this study has illustrated the powerful influence of mapping structure/geometry – not simply parameter mapping – in perceived sonic/control response, as well as the importance of defining mapping structure *for a given sonic context*. Thus while there are many questions to be answered, these three primary areas have been shown to be of relevance and influence, providing a basis for more refined perceptual studies in the future as this area advances. For now, I would like to continue the perceptual testing whilst shifting the focus a bit. While these results examine one’s perception of the *entire instrumental potential*, I now want to more closely examine the interplay between mapping structure and sonic gesture response in order to see how this influences performance *in an actual musical control context*. To this end I designed a “sonic acquisition task” to explore subjects’ ability to trace sonic gestural contours.

4.2.3 Navigation and Sonic Gesture Target Acquisition

The second test in the experiment applies the analysis framework of chapter 2, focusing on sonic gestural dynamics and the way that mapping structure influences control of this. This test was designed in order to quantitatively examine subjects’ ability to create sonic gestures that matched some target gesture in regards to overall dynamic, mass, HT and grain profiles – in short the sound’s overall dynamic morphology. While some might say that “target acquisition tasks” [172] are not a truly musical context, I would argue that this test design does promote holistic, musical thinking by virtue of the fact that subjects can freely play before deciding a posteriori what a “good” gesture was. Though the test does ask subjects to interact with the computer keyboard as a secondary discrete controller – playing back recorded loops – this is in fact a core musical performance feature of many performative styles of electroacoustic music, and in particular in the case of “laptop music” [31].

Test Design and Procedure

For this experiment, subjects were asked to mimic a given target sound. They were informed that the sound was created with the instrument in front of them, and that while it would therefore be possible to recreate the exact sound, this was unlikely. Subjects were further instructed that – rather than focusing on matching pitch, volume or any single sound feature – they should try and match the contour of all changing sound

parameters to the best of their ability. They were presented with sixteen target examples to try and mimic: four different types of control gestures input to instruments defined by combining MI/SI mappings and modal/granular synthesis. Subjects were informed that they would have up to four minutes for each attempt.

The user determined the pacing of each attempt, using the computer keyboard. When the “a” button was pressed, a new set would be loaded and a timer would begin. At this point, subjects could press the space bar in order to hear the target sound. They could further press the “2” key in order to stop this playback. The target sound could be reviewed as many times as one wished – though after each stoppage the sample would play from the beginning. This was important in order to maintain focus on the overall gestural contour of the sound, and not some momentary details. After hearing the target, rather than immediately producing an attempt subjects were allowed to navigate through the instrument space in order to get a feel for the possible sound output – thereby creating a more holistic performance experience.

As an a posteriori parsing of the collective audio attempts would be a very difficult problem, the subjects were asked to do their own phrasing in regards to recording their target-matching attempts. In order to keep the selection overhead low, I chose to use the button on the tablet stylus for this. Specifically, subjects needed to press this button to begin the recording, and let go in order to stop. An on-screen flash informed the subjects separately of the start and stop recording actions, so that they could be aware of the mechanism at work and could practice this if they wished. In a uniquely electroacoustic fashion, the true matching here was ultimately between *sound files*, and so subjects were allowed to review their most recent attempt. For this purpose, the “1” key began the previous attempt playback, while the “3” key stopped this. As with the recording process, an on-screen flash informed subjects’ of the beginning/ending of this file as well as the target sound file. In this way, subjects were truly comparing the ground-truth sound files, and were not misled by the playback system or things such as excessive silence at the beginning or end of the file.

Subjects were allowed as many attempts as possible within the four minute period. An on-screen display informed them when there was thirty seconds remaining, so that they could ensure that they registered at least one good attempt. When subjects were satisfied with their attempt, they rated their confidence along a continuous scale, thereby stopping the timer. Pressing “a” would then advance to the next target set. The four instrument

sets were randomized, and within this each of the four target sounds were randomized as well. Subjects were informed that the instrument set would change every four attempts, so that they could anticipate this change and focus on the task. A screen display allowed them to keep track of the attempt number. Subjects were informed that they could take breaks in between each attempt set (i.e. after rating and before advancing to the next set via pressing the “a” key), and that this would not affect the test timing.

Therefore, the interaction was primarily a single-handed continuous navigation with the tablet-based instrument, with secondary “cue and review” functions with the other hand, using the computer keyboard. The first is a holistic musical control context focused on timbral/textural control, while working with keyboard to control sampled material is commonplace in laptop music performance. In regards to the overall musicality, the “task” was similar to the act of mimicking a musical phrase (in this case sonic gesture), which of course is quite common in styles of music performance such as jazz improvisation.

Target Gestures

In order to isolate the effect of the mapping on the sonic gesture, I generated a given control gesture, and drove each of the four instruments with this same data. In this way, one control gesture would create four sonic gesture targets after being input to the MI modal, SI modal, MI granular and SI granular instruments. I wanted to create target gestures that spanned the categories presented in figure 2.32, and so I created control gestures with one of the gestural archetypes in mind. While this was straightforward in the case of overall dynamic profile (given the mapping of pressure-to-amplitude), several attempts were needed in order to create sonic gestures that fit into the given categories. In other words this was done in a design feedback loop. For example, graduated continuant gestures were used for type 1,2 and 3 (GC1,2,3). In order to achieve e.g. the proper swelling of spectral energy towards the end (GC3) of a gesture (see figure 4.13), several trajectories through control space were explored. The primary criterion was that these needed to sound “right” relative to my conception of the given sonic gesture, but further I applied the sonic gesture analysis framework to see that this also followed suit. After arriving at adequate attempts for GC1,2,3, this process defined twelve of the sixteen sonic gestures: three gestures applied across all four mapping/synthesis instrumental combination. The final four came from variations on the input control gestures. I wanted to examine the attack-decay gesture type, and so the MI Modal-AD and SI Modal-AD instruments were again employed with an appropriate excitation control gesture. As this

sonic gesture type is not possible with the granular set, I decided to balance the gestural set between modal and granular targets by including a second GC2 target whose control gesture had position modulation throughout the continuously excited middle. Therefore, the target gestures for the Modal-GC instrument were defined by two instances (SI and MI) of GC1,2,3, with two Modal-AD gestures. The Granular-GC instrument produced sonic gestures in the archetype of GC1,GC2,GC3, and also included a GC2 gesture that possessed continuous modulation.

Analysis Framework

The interest is to examine subjects' ability to trace contours in regards to texture and timbre. In the language of chapter 2, this means continuously modulating the sound in regards to dynamics, matter/HT and grain. In other words the question is whether one can trace the gestural morphologies such as that in figure 4.13 utilizing a given instrument, and to what extent mapping structure is an influence in this performative action. Continuous pitch salience variation was only strong for specific regions of one particular instrumental set (Gran-GC), and AM/FM modulations existed as a product of the synthesis methods themselves (rather than solely as a function of control input). Therefore neither pitch contour nor motion (modulations in intensity or frequency) variations were examined in the analysis stage.

It is important to note that the interest here is not to test one's timing in the sense of being able to recreate the exact duration of a gesture. Rather, it is the temporal morphology that is under analysis, and so only duration-independent features were used, drawing from those discussed in section 2.2.4 of chapter 2. The "temporal morphology" vector, then, was comprised of the following nine features: ratio of maximum to total length, ratio of temporal centroid to total length, temporal flatness, temporal increase (TI), temporal decrease (TD), non-weighted TI, non-weighted TD, maximum derivative during onset and finally minimum derivative during decay. For more on these features, the reader is directed back to the definition and larger discussion from the end of section 2.2.4. The essential point here is that this vector is duration independent: if two gestures had the same temporal shape but different durations, they would produce the same temporal morphology vector.

In order to examine the gestural contour of several different attributes, I extracted this nine-dimensional feature vector from the RMS amplitude envelope (dynamic profile), the spectral centroid (HT profile), spectral smoothness and flux (mass profile). These define a

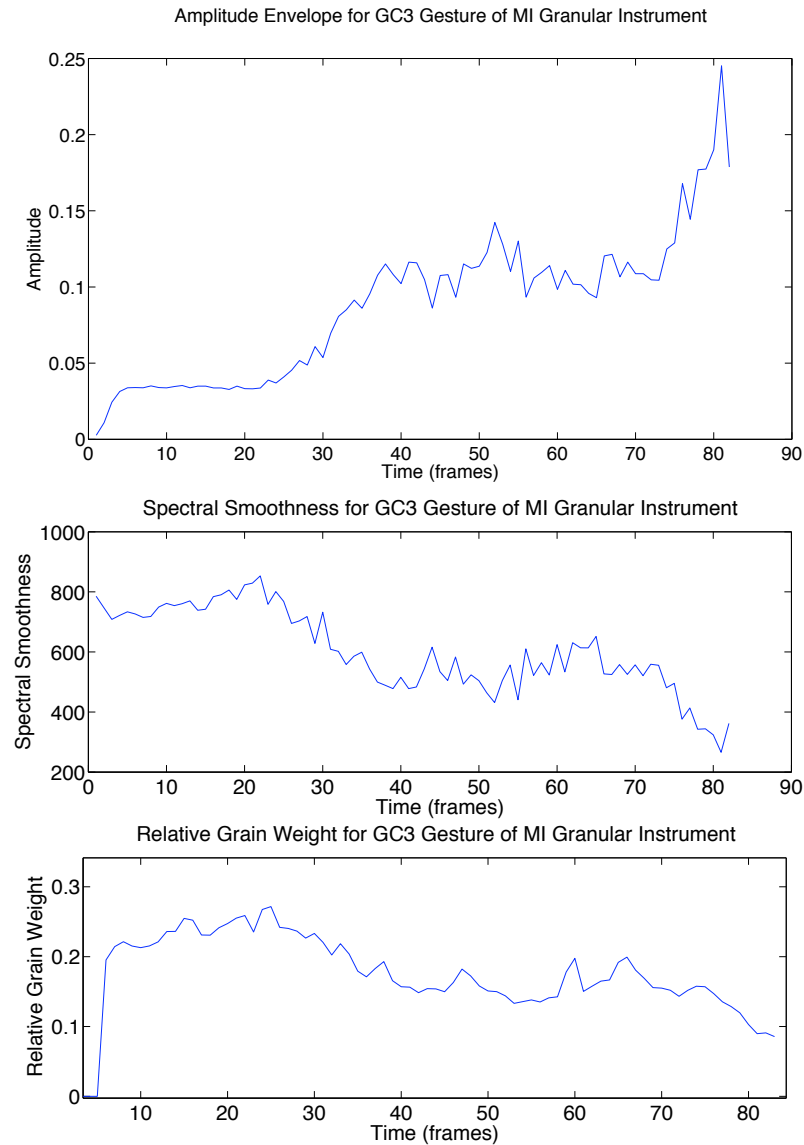


Fig. 4.13 Spectrotemporal Profile of GC3 Gesture for MI Granular Instrument. The overall envelope (top) increases in amplitude while the Spectral Smoothness (middle) decreases as the sound's spectrum becomes continually more jagged. Meanwhile the Relative Grain Weight (bottom) decreases at a similar rate as more widespread spectral energy overrides the fine-grain fluctuations.

global form and matter profile for the sound, and their relevance to timbre perception and control is well-documented ([19][182][183]). Taking this further I wanted to examine the limits of perception in regards to finer matter qualities, and so the dynamics of several grain features were examined as well. Within this local/fine category of grain profile I extracted three fundamental features: relative grain weight (RGW), overall graininess by way of temporal fine structure (TFS) and finally the relative grain spectrum (both mass and placement) through extracting relative grain spectral smoothness (RGss), flux (RGsf) and centroid (RGsc).⁶

Therefore, this analysis framework provides nine profiles that cover dynamic form as well as timbral (mass/HT profiles) and textural (grain profiles) dynamics with a nine-dimensional vector for each. As a whole, this set of 81 features provides an out-of-time and scale-independent description of the temporal morphology of amplitude and spectral dynamics. However, in order to use these in any sort of metric for perceptual comparison, one has to rectify the fact that each feature has a dramatically different scale. As the ultimate goal is to compare the relative distance between user attempts and target sounds, all features were normalized to 0-1. This was done by extracting all of the features for every attempt, across all users. Extreme values (several standard deviations above the mean) were filtered out, and the maximum of the remaining values $p_{k,max}$ was used for normalization of the given temporal feature k . Therefore all values achieved a maximum of 1, and the relative distance to a given value could be compared between users. After normalization, I examined separately the distance between all 16 targets and the 16 attempts made by 16 subjects⁷, for each of the 9 form/matter feature profiles. The distances were computed using a standard Euclidean metric, so that for a given form/matter quality the distance between target profile $P = \{p_1, \dots, p_9\}$ and subject profile $S = \{s_1, \dots, s_9\}$ is defined as

⁶Recall that the relative grain parameters are based upon the EMD analysis method, which requires some decision in regards to grouping of intrinsic mode functions (IMFs). An analysis of a variety of sounds output from the modal and granular synthesis methods revealed that the former produced high values of spectral grain (as determined by roughness-based measure) while the latter produced transient grains (from the TFS-based measure). Further, the spectral grain was consistently confined to IMFs 2-4 of the analyzed modal output while the transient grains were located in the first two IMFs of the granular output. Therefore, in the feature extraction across all users the IMFs were fixed for the two granular sets, and so comparison of relative grain phenomena implicitly means spectral grain for the modal set and transient grain for the granular set.

⁷Data from 3 subjects was discarded from test 2 as well.

$$D = \sqrt{\left(\frac{p_1 - s_1}{p_{1,max}}\right)^2 + \dots + \left(\frac{p_9 - s_9}{p_{9,max}}\right)^2} \quad (4.1)$$

so that the final description of a subjects' performance is thus a scalar value for each of the nine form/matter profiles. After this entire feature extraction process, I examined the individual and collective performance of all subjects in order to discover the influence of mapping on performance in regards to global trends or grouping tendencies, as with test 1. This study produced enough data to fill another chapter of this dissertation. Due to space and time constraints – and because a full treatment of all variables is beyond the scope of this work – I will present the most interesting results regarding the interaction between mapping structure and sonic context for different sonic gestural features.

Results

Overall Trend: Tracing Sonic Gestural Contours

The reason that the test included different target gesture types was to span a variety of different sonic control contexts, as well as provide for the possibility of analyzing subject's performance as a function of these gestures. However, examining all 16 target gestures would require many more than 19 subjects, and so this is left for future experiments having more participants. For this experiment, the four attempts that corresponded to the four gesture types within each instrument (defined by mapping/synthesis pair) were averaged for each subject. In doing this I am examining the average performance across these disparate gestures, which provides a more generalized view of the mapping/synthesis interaction and influence on performance. At the same time, recall that the AD gestures define a different control structure type. I therefore also examine the means after removing the AD gesture from the modal targets and the modulated GC2 gesture from the granular targets, leaving only GC1,2 and 3 gestures for the two granular and two modal instruments. In a sense, this is the subtractive equivalent to analysis of the influence of control structure on subjective ratings from test 1. After averaging the distances across target gestures, each subjects' performance was represented by four values (corresponding to the SI Gran, MI Gran, SI Modal, MI Modal instruments) for each of the 9 form/matter features. As with test 1, in order to examine the overall performance trend I took the average across all subjects and compared them with respect

to these four instruments and 9 sonic gesture features.

Going into this analysis, I originally hypothesized that overall performance would be tied to the sound synthesis outright – that one mapping would work better across the board for a given sound synthesis type, with the other more suited to the second synthesis type. In other words that mapping efficacy would be tied to sonic context. A deeper influence was illustrated, however, in that the performance with a given mapping changed as a function of subjects' ability to trace a particular sonic gestural feature. Overall, subjects fared much better at tracing certain sonic contours with one mapping than another.

Global Profile Tracing

Recall that the sonic gesture profile features were related, in chapter 2, to the qualities of dynamic profile (RMS envelope), mass/HT (Spectral spread, flatness and centroid) and grain (TFS, relative grain features). These features were chosen through a reading of Schaefferian theory, from psychoacoustics literature and from an analysis of a given set of sound materials. It is not a given, however, that these features would group together in terms of performance relative to the different instrument types, as this is an experiment into the perceptual structure of sonic control context – and not just a passive timbre perception task or static sound feature analysis. However, it turned out that all of the global temporal and spectro-temporal features exhibited the same trend: for amplitude envelope (AE), spectral centroid (SC), spectral spread (SS) and spectral flatness (SF) the *performance in tracing these features was consistently better with a given mapping for a given sound synthesis type*. In particular, the performance in tracing all of these features was better across the board using the *SI mapping with granular synthesis and the MI mapping with modal synthesis*. These mean values are depicted in figures 4.14 and 4.15, where we can see this trend in all graphs. In regards to AE and SS, all distance values were fairly close, and so are graphed together. However, the granular instrument distances for SC and SF were considerably higher, and so I've split the ranges of the graphs between modal and granular sets for 4.15. The important thing to note is the relative change: that in each case there is a considerable difference in distance between MI and SI for each set. The range in values between granular and modal showed that some specificity of the sound synthesis type was a factor as well (e.g. transients in granular case leading to high SF), as the granular instruments were more conducive to tracing AE and SS profiles while modal instruments were more so for SC and SF.

Grain Feature Tracing

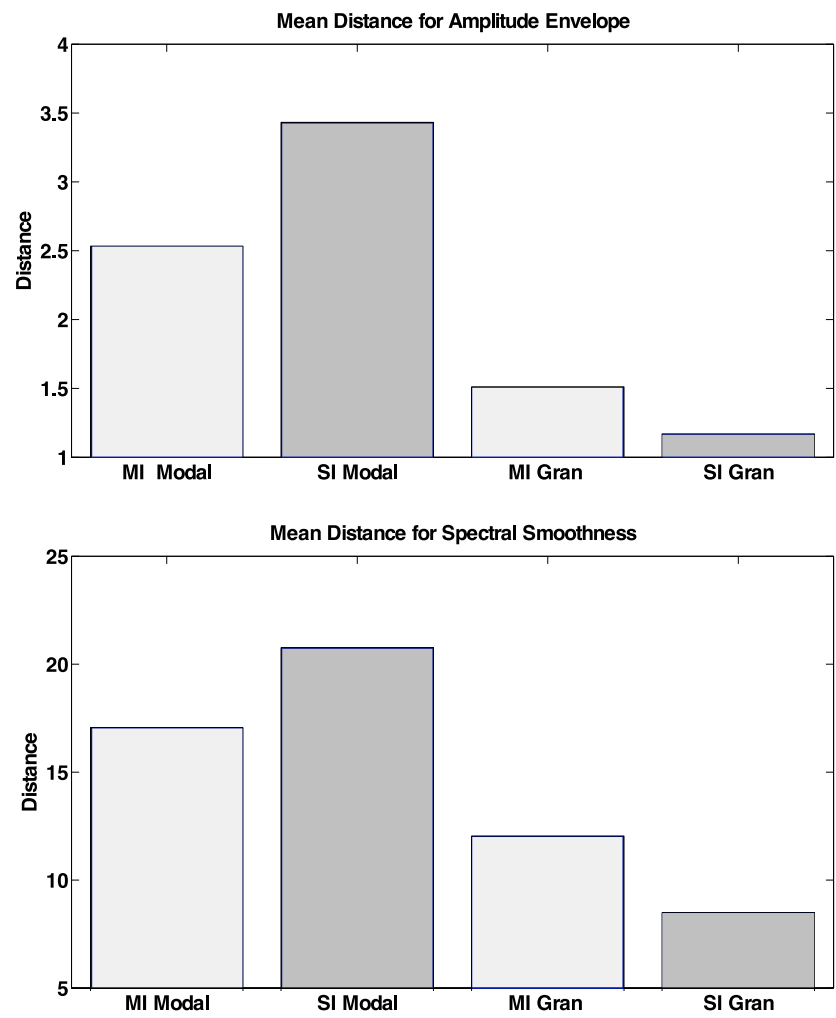


Fig. 4.14 Average distance to target for all subjects in regards to Amplitude Envelope (top) and Spectral Smoothness (bottom).

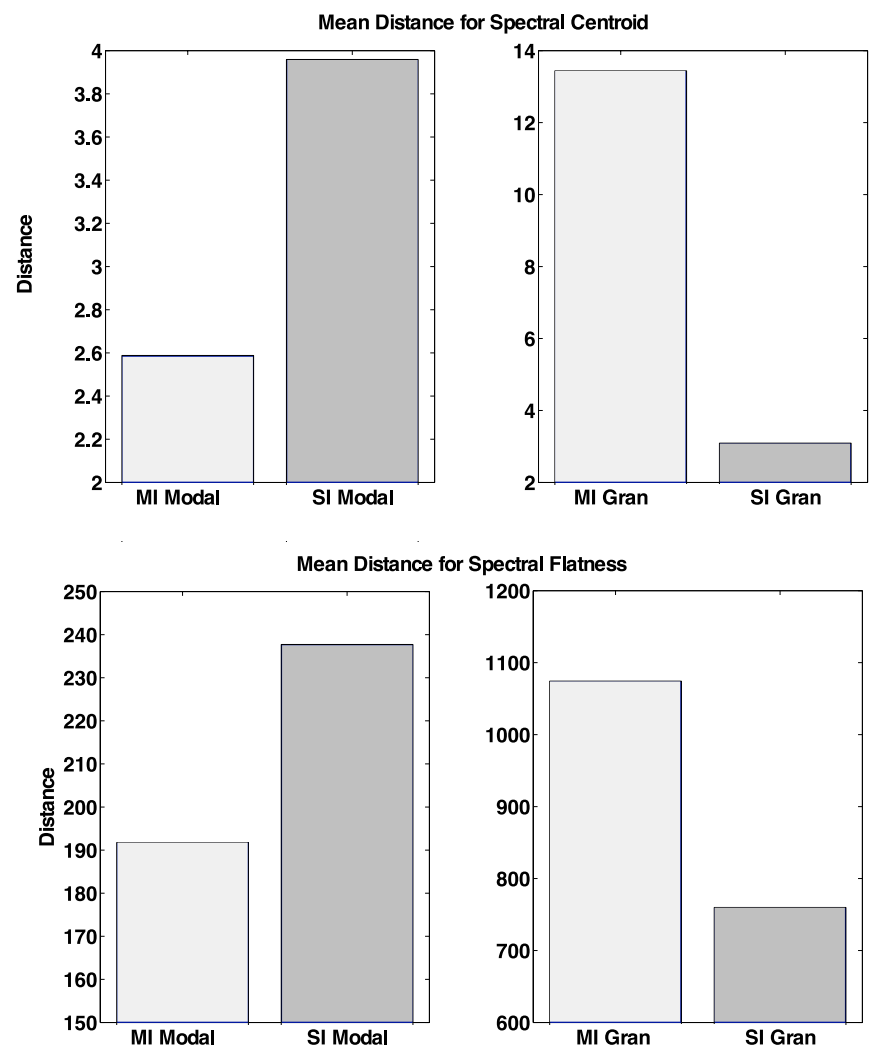


Fig. 4.15 Average distance to target for all subjects in regards to (a) Spectral Centroid and (b) Spectral Flatness.

Looking at the more detail-oriented tracing of grain-type feature profiles, there was a different influence of mapping type. In three of the four feature profiles, *one mapping produced consistently better distance measures*. Specifically, subjects performed better across the board with the MI mapping in matching RGsf and RGss profiles (figure 4.16), and better with the SI mapping in regards to TFS and RGW (figure 4.17). From my a priori analysis, I found that the grain signals extracted via EMD were of the transient grain type (high TFS) for the granular instruments and of the spectral grain type (high roughness) for the modal instruments. Thus it is interesting that one mapping type (SI) was better at tracing the RGW profile regardless of sound synthesis type. Another interesting aspect of this mapping-consistent performance was that overall, subjects *performed better at tracing spectrally-defined grain features with the MI mapping, and temporally-defined features with the SI mapping*. One interesting exception to this was the RGsc feature (figure 4.18), in which subjects performed better with the SI mapping in the gran instrument set and MI in the modal set. This followed the trend with global SC and suggests that *RGsc and SC were less separable as control features compared with the other global/relative feature pairs*. This phenomenon bears further exploration, as it may provide insights into choice of perceptual features for controlling dynamic/mass/grain profiles separately. As a first step in this direction, this study establishes that the choice of mapping can have influence on controlling different aspects of sonic gestural features, and that this influence happens differently between global profile parameters (dynamic/mass/HT) and properly-defined fine-scale parameters related to grain.

Control Structure Influence

Recall that each subject's attempt was the average performance in regards to four target gesture types. This included an AD attempt for the modal instrument set, which had an overall different control structure and which required a different amplitude control (attack/excitation). Therefore it is possible that this gesture biased the overall results with the modal set. In order to examine this, I removed the AD target gestures for all subjects, as well as the modulated GC2 gestures from the granular instrument targets, before once again finding the mean. Interestingly, *all of the global means remained the same*. Therefore – as with the control/sound ratings of test 1 – the *control structure influence was minimal, while the mapping structure imposed a strong influence on performance*.

At this point one could question what the influence of each target gesture type was –

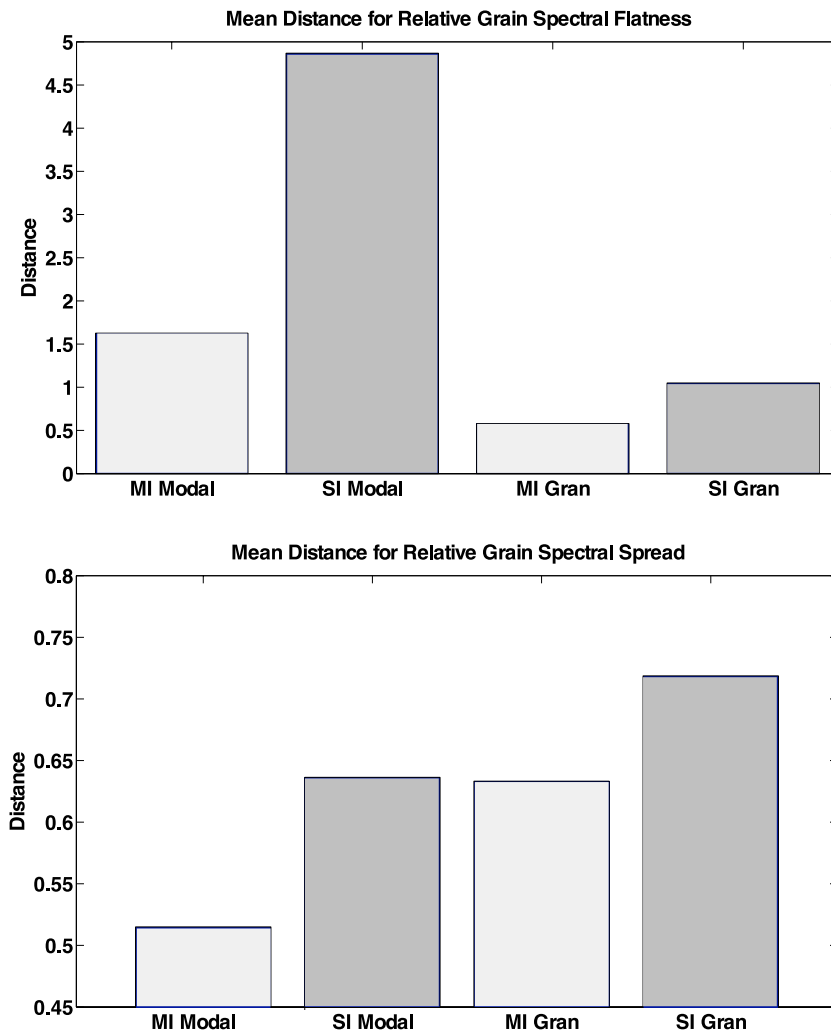


Fig. 4.16 Average distance to target for all subjects in regards to Relative Grain Spectral Flatness (top) and Relative Grain Spectral Smoothness (bottom). MI mapping structure leads to better performance for both instrument sets.

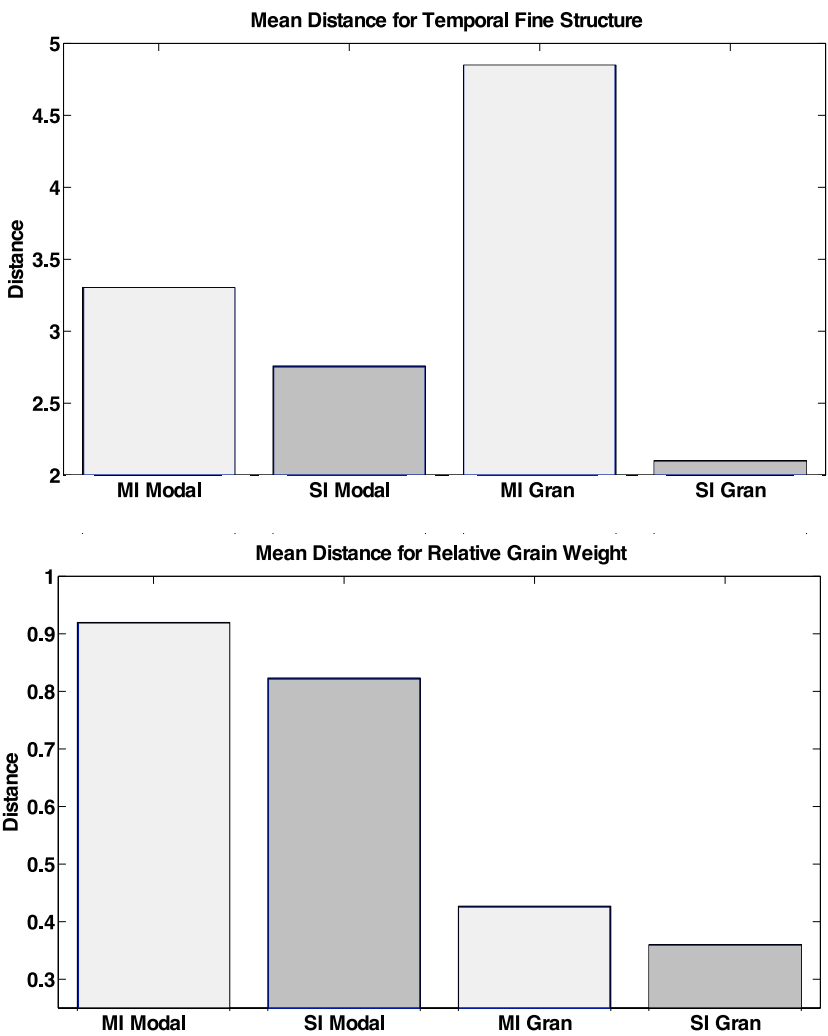


Fig. 4.17 Average distance to target for all subjects in regards to TFS (top) and RGW (bottom). SI mapping structure leads to better performance for both instrument sets.

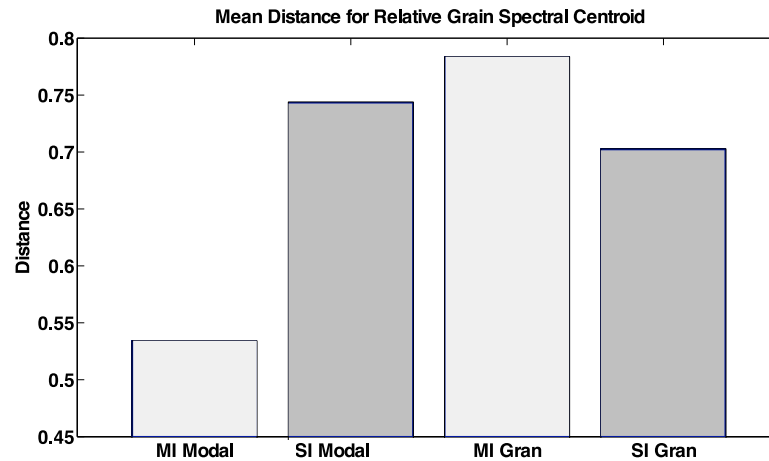


Fig. 4.18 Average distance to target for all subjects in regards to Relative Grain Spectral Centroid.

something that presents a lot of degrees of freedom to analyze and which would require more subjects to extract meaningful trends. However, in examining the data one can see that strong variations occur more noticeably between subjects than within subjects, and so a worthwhile analysis that does fall within the scope of this study is to examine any grouping trends that occurred in regards to mean performance values of each subject.

Background, Grouping and Performance

As with test 1, I once again conducted a hierarchical cluster-based analysis on the average distance measures for each subject using the same criteria, with respect to each of the nine sonic gestural features. The overarching trend that emerged was that one large group (8-11 subjects) formed in regards to the global dynamic/mass/HT features as well as with the grain features RGW and TFS. The other 5-8 subjects forming their own group (i.e. were outliers) or in some cases one other group of size 2. Meanwhile in regards to the relative grain spectral features, two distinct smaller group formed in each case (with a few outliers). In all of the global non-grain features, there was a subset of 7 subjects that belonged to the larger coherent group, while 6 of the 7 belonged to the main RGW group and additionally 4 of them did for TFS. Therefore this subgroup performed in a surprisingly consistent fashion across these different features – though it remains to be established whether this was in regards to overall performance or in terms of the structure of their relative mapping/synthesis performance. In a similar fashion to test 1,

this group was heavy in both EA experience and instrument playing (6/7 subjects in each case). Interestingly only 3/7 possessed musical training, meaning the group consists primarily of musicians without formal training.

As with the clustering of test 1, I compared the means from this group to the overall trends beginning with the global profiles of figure 4.19. Two things are immediately apparent: the first is that performance for each spectrally-defined (SC,SS,SF) sonic gestural profile changes with mapping type consistently, rather than as a function of sonic context. The second is that the performance in regards to these spectral profiles is considerably better than the overall mean. Meanwhile in regards to AE, RGW and TFS we can see from figure 4.20 that this group performed with a similar structure to the overall mean and with similar performance. In light of test 1, a question that this raises is whether EA experience or musical training alone (or together) led to better performance. In order to examine this, I looked at the average performance of three groups: those that were musically trained without much EA listening experience (MT/Non-EA), those musically trained with EA experience (MT-with-EA) and those with EA experience and no musical training (EA/Non-MT). Looking at the mean performance across all groups⁸, *the coherent cluster-based subgroup performed considerably better than any of the individual background-based groups in regards to all spectrally-defined features*. Meanwhile, all four groups performed with precisely the same structure and the same performance as compared the overall mean for both AE and RGW. The TFS distances followed roughly the same structure in each group, with no clear improvement in performance across these. The relative grain spectral features showed little consistency between groups in terms of structure of performance or overall performance itself.

Test 2 Conclusion and Significance

The general trend shows us that mapping did in fact influence performance in all cases, but that the specific manner varied with sonic gesture profile. The AE, SC, SF and SS profiles all showed an influence of mapping, where this changed in each case with sonic context. The significance of these findings was backed by a three-way repeated measures ANOVA, which demonstrated a significant interaction ($p = 0.007$) across terms. At the same time the grain profiles for RGW, TFS, RGsf and RGss showed an influence of mapping that was consistent across sonic context, with SI proving better for the temporal grain features (RGW and TFS) while MI gave better performance for the spectral ones.

⁸Due to the large number of figures, these are displayed in appendix B.

A three-way repeated measure ANOVA and subsequent cross-term analysis revealed this mapping/gestural feature interaction to be statistically significant as well ($p = 0.001$). This result quite interestingly illustrates the fact that *geometric mapping structure coherently influenced performance* and that this happened such that *this particular performance changed with sonic gestural profiles: being consistent across dynamic/mass/HT profiles on one hand and across grain profiles on the other*. Comparison across the different groups (one cluster-based, three background-based) suggests that the combined experience of regular listening to EA music and playing an instrument among these subjects led to better overall performance in regards to tracing the global spectral contours (SC,SS,SF) than either EA listening alone or musical training alone. While the mapping type still influenced performance in these groups, it seems that the better-performing EA/instrumental group was able to “overcome” influence of sonic context, to the point where only the mapping influenced performance. At the same time, the AE and RGW were highly consistent across all groups in terms of cross-instrument performance while TFS was to a lesser degree. Therefore *these temporally-defined features – both dynamic and grain profile based – were easier to control in general with a given mapping type across the board, with SI performing better overall*. While there was no sub-group trends in regards to spectral grain features, the overall trend suggested that *MI produced better performance with spectral-based grain profiles*.

Test 1 and 2 Summary

Both of these tests have illustrated that the “simple” change in *how* mapping that arises from changing the geometric structure of a mapping can coherently influence the perceived sound quality and controllability of an “electroacoustic instrument” as well as the performance with this instrument. This was my primary hypothesis at the outset of the experiments. This influence included the counterintuitive result that in general perceived sound quality changed only as a function of mapping structure, not with sonic context. Conversely, controllability and transparency changed along mapping lines, but the mapping influence changed with sonic context. Meanwhile, there was an interesting general trend along these lines in test 2: performance in tracing dynamic/mass/HT features all varied with mapping structure in different sonic contexts *in the same way* – MI being better for modal instruments and SI for granular ones. This preference *mirrors the overall perception of controllability* from test 1 in terms of general pattern as well as specific mapping/synthesis pairing. The final general trend showed a mapping-dominated

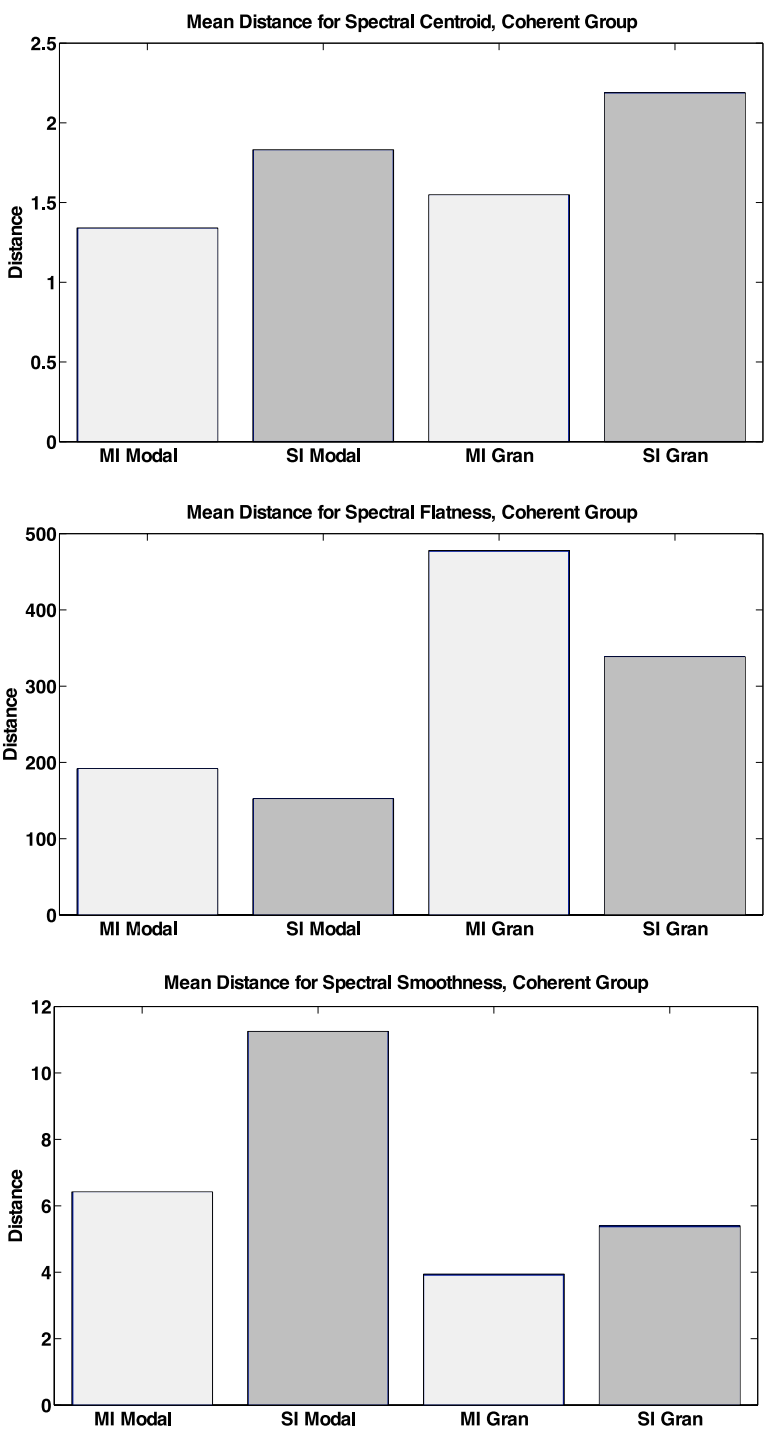


Fig. 4.19 Average distance to target for Spectral Centroid (top) Spectral Flatness (middle) and Spectral Smoothness (bottom) for the coherent subgroup formed across AE, SC, SS, SF and RGW contour tracing.

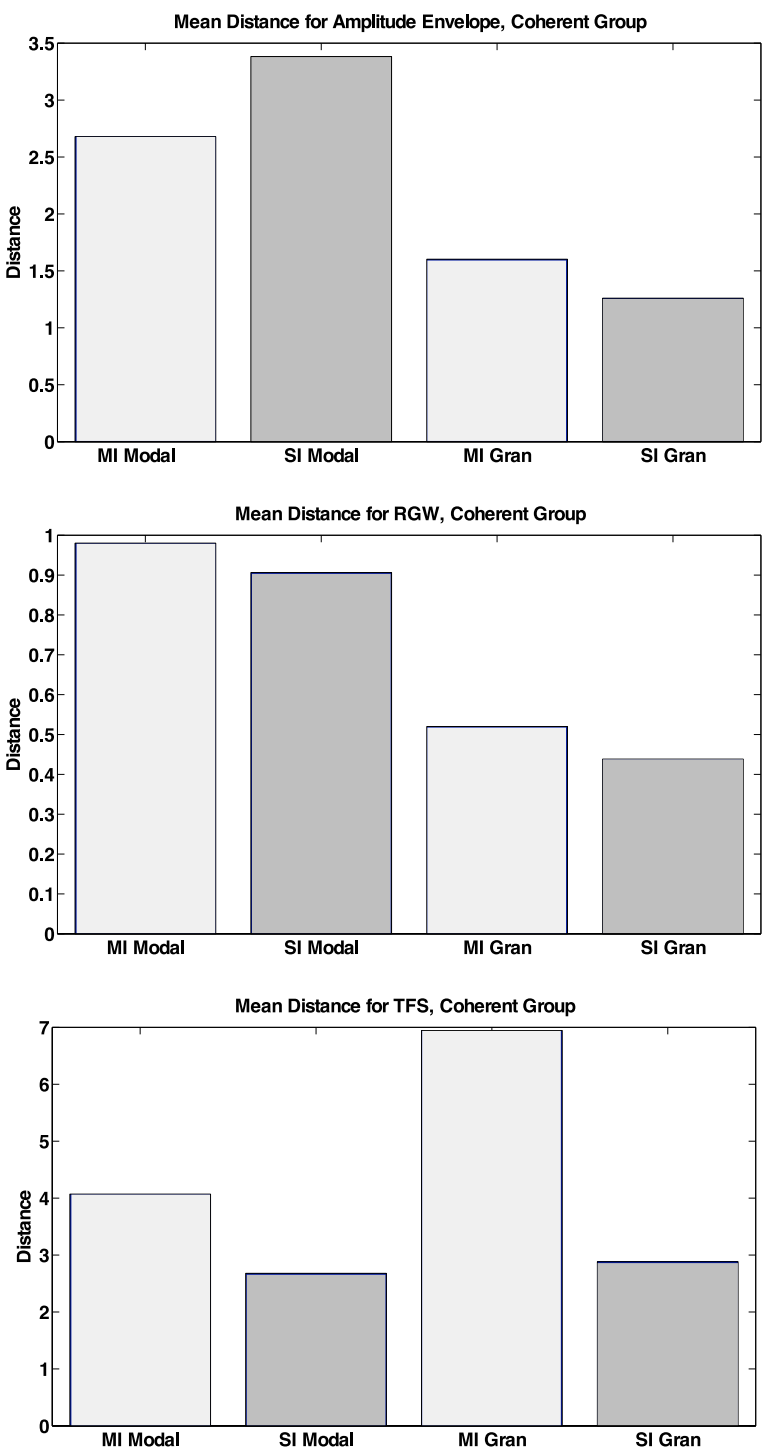


Fig. 4.20 Average distance to target for Amplitude Envelope (top) RGW (middle) and TFS (bottom) for the coherent subgroup formed across AE, SC, SS, SF and RGW contour tracing.

performance in regards to tracing grain profiles, where MI led to better performance outright with spectrally-defined grain parameters and SI doing so for temporally-defined ones.

Recall that in the preliminary experiment I discovered a tradeoff between mappings in regards to expressivity and ease-of-use. In this second set of experiments, the overall trends point to a much more complex-yet-coherent set of tradeoffs. The first depends on the sonic context: the smooth, spectral-grain heavy modal instruments were perceived to be more controllable with MI and the transient-grain and sudden granular instruments with SI. This itself is counterintuitive in that MI by its nature is more sharply varying and less smooth when compared to SI. The second tradeoff depends on what sonic gestures one feels are important in producing an instrument: if one wishes to trace matter-based features related to mass and HT and the overall profile, these results again suggest that MI works better with modal instruments and SI with granular. However in tracing the finer textural details of grain parameters, the tradeoff switches to whether spectrally (MI) vs. temporally (SI) defined features are more important. Naturally an instrumentalist wants subtle control over both of these aspects of sound, and from these results it seems clear that *multiple layers of adaptive mapping* are needed to change control modes. The next section will thus present comprehensive musical mapping/control structures that address the ability to balance these two aspects of performance.

One final thing to take from these studies is that background seemed to have a coherent influence on perception and performance. Clearly more subjects must be run to find more conclusive evidence of background influence, but from these 16-19 subjects it seems first that EA music experience was the most coherent influence in rating for test 1, and combined EA experience and instrument playing generally led to improved performance in test 2. An interesting find was that the more EA experienced subjects were the only ones to rate sound preference as a function of sonic context, while subjects with combined EA experience and instrument playing were the only ones to perform better at tracing global matter strictly as a function of mapping type. This suggests another mapping tradeoff in performance vs. perception. While we can't know the exact mechanisms at play, I will offer one final hypothesis: that EA experience on the one hand allowed for perceiving differences in sound quality that changed with mapping-in-sonic-context, while combined EA/instrument experience allowed subjects to perform more effectively such that they were able to overcome the influence of sonic context.

Regardless of the particular reasons for the differences in performance, to my mind this background grouping trend suggests that researchers should at least consider that there might be a subset of musical tasks (such as playing the EA instruments of these tests) in which EA music listening as well as instrument playing (even self-taught) may be more appropriate knowledge than formal musical training. As I've noted, this experiment is presented as a new direction in examining perceptual control structure, and thus raises many new questions for future analysis, including this issue of background as well as obvious interactions with pitch, rhythm (classic timbre perception questions) and so on. From this work, suffice to say that mapping geometric structure can have a noticeable influence on perception and performance that needs to be considered in regards to both sonic context and the sonic gestural response of the given design. At this point I will present several designs that consider precisely these two things.

4.3 Modular and Parametric Control Structures

The instruments used in the previous section are comprised of mapping structures that move continuously and with holistic control through a parameter space. This structure was isolated for the purpose of the experiment, in order to focus on the influence of the mapping itself. However in an uninhibited musical context, an instrument will most likely have either discrete controls and discontinuous jumps in/between parameter spaces, or the nature of the parameter space will change over time. Navigation between or alteration of such spaces itself can be a controllable variable, or can be determined by the requirements of the sound output as with the excitation-based approach of [13] where the system in question switches between steady-state models upon attack, in order to simulate the attack/sustain of an acoustic instrument.

Beyond switching between parameter spaces, using modular combinations of mapping structures can be used to determine gestural dynamics. There are tradeoffs and differing qualities between mappings as I have presented both mathematically and from a perceptual point of view, and these can be combined in order to span disparate control/sound spaces that have unique requirements in terms of dynamic behaviors. These dynamics are not simply achieved by modular mapping combinations, but also potentially through parametric control of the mapping layers themselves in order to adapt and condition the gestural nature of the input or output. In my own performance systems

I have used these approaches in a few canonically different ways as warranted by the specifics of the control context in question. Rather than show every implementation I have done, I will present the underlying structures that make them unique – and quite naturally there are many variations on these examples, as there should be given our still-experimental era of digital music.

Generally speaking the musical control context for the presented examples is similar in that for each the focus is on a “laptop music” performance paradigm, which – as I noted in previous chapters – does not have the communicative aspects of human gestures as its primary concern [31]. However real-time control and organization of sound materials *is* of paramount importance – yet the ability to organize complex sound materials such that one may continuously move amongst these in order to repeatably and reliably evoke sonic gestures is difficult and few performance systems achieve this. This is a particularly challenging problem given that often the sound processing utilized will have many controllable parameters that are not immediately suggestive of a particular physical gesture, as with the granular instruments of the previous section. A third defining factor in these mapping designs is that they make the aesthetic choice of controlling sounds that have rich textural properties. Following the chapter 2 analogy to acoustic instruments, the control of such micro-variations in these examples is generally not direct, but rather is a product of simultaneously affecting many other parameters. All of the examples use a standard Wacom tablet – a good choice due to its ubiquity, accuracy [184], cost-effectiveness and suitability both for laptop-centered performance as well as other, more overtly gestural styles of computer-based music.

4.3.1 Transformation of Resonant Models

Before getting into more complex structures, I will elaborate a bit on the modal synthesis model from the previous user study – in particular two variants that arise from the basic instrument design.

Organization of Sonic Space

As this modal parameter space is relatively high-level, it makes sense to associate points in space with steady-state sounds, and to morph between them. This is comparable to the classic conception of a timbre space, wherein one wishes to discover sounds

“in-between” known sound models. To this end, I designed a control structure that was locally defined and locally editable, in order to place particular sounds that were generated from preset models in close proximity on the tablet surface. Further, new models may be defined at a given location, thereby changing the tablet response only in neighboring regions. In this way, the sonic character of these regions are defined by the manner in which the stored models are scattered across the plane of the tablet, as well as how said regions are connected. In this particular example, the preset points were triangulated on the tablet surface using the SI mapping from the LoM toolbox.

While I use the analogy of timbre space for this example, it is not truly a perceptual space that is under control in that the degrees of freedom do not correspond to perceived sound qualities per se, and linear changes in sound parameters do not result in perceived linearity of timbral sound transformations. That said, this high-level nature of the sound parameters – coupled with the proper placement of model states – does lead to a situation in which one can easily find steady-state sounds while being able to make repeatable sonic gestures. The distinction I am making is that rather than have some high-level axis of transformation, these repeatable sonic gestures come from a combination of control parameters that “cut across” control space and do not always arise from the same control gesture. The primary reason for this is that while the parameter mapping itself is fixed, the *perceived* mapping varied due to the hysteresis present in the modal synthesis model (itself due to the inherent memory of the resonant filters employed). Because of this subtle temporal element, the process of constructing a particular sound space and a predictable coupling of physical/sonic gesture means exploring the speed and ordering of different pen/tablet trajectories, leading to different sonic gestures at the same tablet location, which in turn prompts the movement or insertion of different sound models in a design feedback loop. Thus already from this simple SI/Modal structure one can see that *the mapping choice as well as the mapping design process itself is determined by the control context (sound model interpolation), the relative high vs. low level complexity of the sound parameters and the (potentially subtle) time-based behavior of the chosen model.*

Tuning of Sonic Space

In a different approach to the same material, I laid out the model states in a grid around the tablet boundary, and utilized the non-linear MI mapping to generate the sound space.

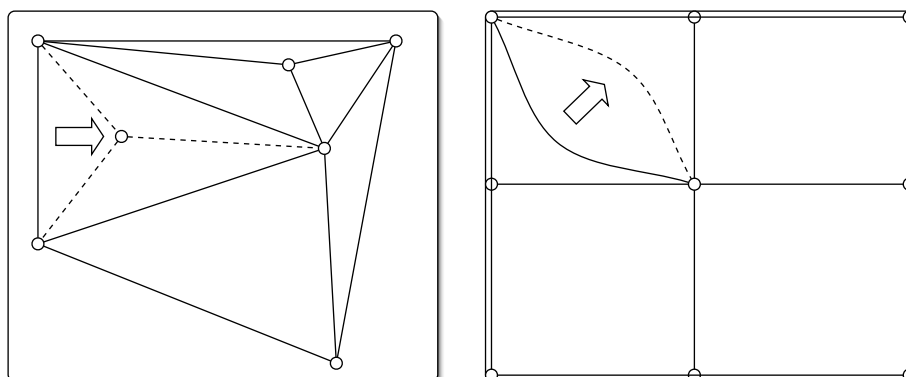


Fig. 4.21 Different spatial layout of sound model presets on tablet surface, and their respective approaches to model interpolation for example 4.3.1: movement or insertion of new sound models, creating new localized regions (left) vs. varying weightings for each fixed sound model, creating a different geometry for the model interpolation (right). The former allows for more defined sonic gestures while the latter makes it easier to define a global sound quality within a given region.

In this case, rather than move preset points around in control space the modification comes from tuning the weighting of each model. Given the nature of this mapping, this amounts to warping the geometric “shape” of the sound space (see fig. 4.21). While it becomes more difficult in this case to define a sound that will occur at a precise location, it proves to be much easier to define regions having a given sonic character. That is, due to the mapping structure it is easier to construct regions of the tablet having a global feel, but more difficult to construct repeatable sonic gestures. The functional properties of the given mapping contribute to this in that the ability to “tune” this mapping technique compensates for its inability to define model presets at arbitrary locations in control space. Further, the globally smooth nature of the mapping makes it easier to create long, smooth sonic gestures that work well within a slow-moving and dense sonic space such as that defined by the modal synthesis technique. This tradeoff must be weighed in regards to the performance context: whether the music calls for many different action/sound gestures that return often, or whether a general sonic context (i.e. slowly varying, dense modes) is more important.

While the input device, sound synthesis method and the underlying sound models were the same in this example, the mapping/control structure was different. The creation of this structure was a product of - and suggested the use of - a certain approach to

mapping. The design process focused on the tuning of global sound qualities for a given region of the tablet. In this way the focus was on the “spatial response” (i.e. physical layout of sonic materials) and the perceived smoothness of this across the tablet, in contrast to the temporal dynamics that were of primary importance in the previous example. Furthering this point on the “design feedback loop” that arises in mapping creation, note that this variation in approach arose from a difference in musical control context, which both determined and was informed by the design process as well as the choice of mapping strategy.

4.3.2 Multi-Layered Approach

The qualities of the modal synthesis technique and its high-level nature suggested a certain approach: spatially laying out the known sound models and interpolating between them – a sort of user-defined timbre space. Other instruments that I’ve created start from a low-level set of parameters and try to build up towards the ability to shape sonic gestures. One pair of examples begin with the SI granular instruments from the user studies (though with sine waves rather than voice as source material), and build up from this by combining different mapping layers in order to define a more dynamic temporal response.

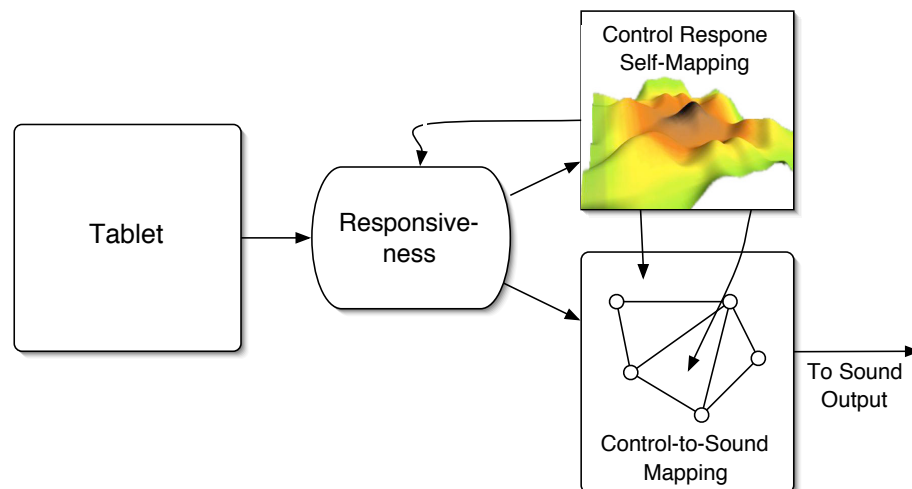


Fig. 4.22 Control structure for example 4.3.2: two concurrent mappings control the possible sound combinations (control-to-sound) as well as the control/sound temporal response (control response self-map). Combined they determine the resultant sonic gestures.

Concurrent, Parallel Mappings

Granular synthesis is capable of generating vastly different sounds [185], making it full of possibility yet very difficult to create coherent sonic gestures in real-time. One approach would be to find particular and interesting trajectories in sound parameter space, and to constrain a mapping to only produce these sounds. However this would seemingly not make for an interaction design with a very high ceiling on virtuosic use [144]. My approach here is to use concurrent mappings which “occupy” the same controller domain (i.e. they are both defined on the two-dimensional tablet surface), wherein one acts to map from control to sound parameters while the other acts on the behavior of the control data itself in a self-controlling fashion (see figure 4.22). For this I use the SI-based method to control the mixture of sound parameters (as in the user study instruments), and overlay another mapping – the RST technique with its tunable tension and smoothness – which controls the responsiveness of the tablet pen position as determined by low-pass filters. The filtering of control data serves two purposes: to adjust the “speed” of sonic gestures through the sound parameter space, and to smooth the transition between triangulated regions of control space, which may be necessary as this mapping scheme is not globally smooth.

As with the previous set, this design results from thinking spatially by laying out preset sound points across the tablet. Another dimension of spatial thinking is added in this case, as the second mapping is concerned with adjusting the responsiveness of the control at given points along the tablet. However, this approach adds a temporal element in that I consider the musical gesture-dynamics that result from the physical input motion. After some tuning adjustments to both mappings, interesting dynamics do emerge from the instrument. At the same time, in playing with this instrument another level of mode change or similar modification seems necessary, as the control-side mapping layer plays the role of a simple adaptive filter of control data which has little variance in its response. This realization in turn led to the following mapping/control structure.

Meta-Control of Mapping Layer

Towards the end of creating an interface with more hidden surprises and control possibilities, I added a layer of complexity to the above mapping scheme. Rather than define the controller responsiveness at given points in space, I defined a mapping to affect

this parameter, taking as input the tilt values of the tablet pen. From a geometric standpoint, the space of control parameters in this example is four-dimensional (two separable two-dimensional planes of control) rather than a single two dimensional surface as in the above example. With this approach to control structuring, one may choose to lay out a mapping “surface” over the tablet as in the above example (i.e. to think spatially about the control of temporal response), or define the control based on particular physical gestures. Taking this approach further one might examine ancillary gestures [78] and design a meta-layer that best adapts to a given performer’s unique motions. For my part, I composed control dynamics that felt most natural yet not overly sensitive – tuning the resonance of control response much like the resonance of a filter. Not surprisingly this control structure was more difficult than the previous in regards to repeating or maintaining a given sound, but it afforded the most diversity of response to a given set of input gestures.

While the first multi-layered example was a bit too constrained (in spite of perceptible changes in control response), the second added a potential cognitive overhead that was not trivial to manage in regards to changing the control response in real-time. Ancillary gesture analysis is one possible approach, but without a repertoire of common gestures this might not be the most effective strategy. One thing that *is* certain is that these two examples underscore *the complex role that mapping plays in the structuring of subtle and articulatory control, including the added potential of time-variant mappings through meta-control and/or feedback control.*

4.3.3 Granular Scrubbing

The previous examples presented four distinct augmentations of the basic instruments from the user studies. They illustrated different aspects of spatial and temporal mapping design that I will build upon in the next three examples, resulting in more complex mapping/control structures. This next set in particular focuses on the gestural act of “scrubbing” through waveforms, which is quite popular in computer-based music re-synthesis, normally achieved through the use of additive synthesis, the phase vocoder or granular synthesis. I have used this technique with granular synthesis for several years now, in the context of a larger performance system that I use regularly in performance [178]. The nature of granular synthesis (e.g. artifacts from windowing and overlap) often

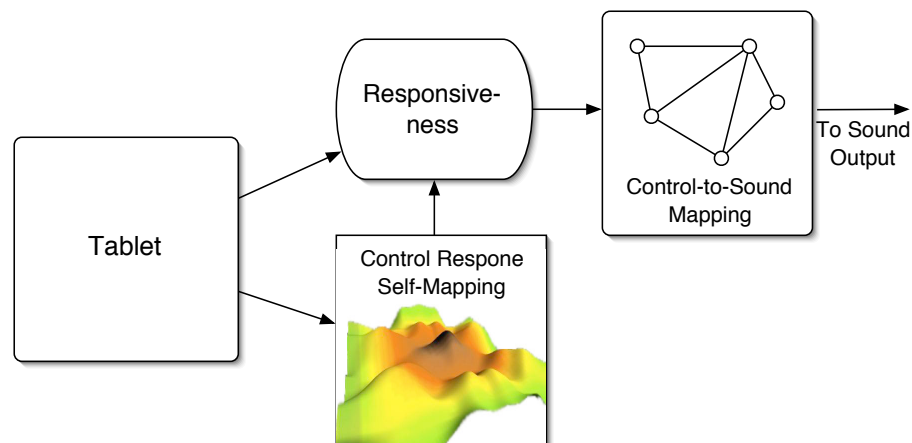


Fig. 4.23 Control structure for example 4.3.2: the control response self-mapping acts as meta-control, affecting the responsiveness in parallel with the mapping from control to high-dimensional sound parameter space.

gives rise to textural effects when used this way, which is further consonant from a control point of view as these sound qualities suggest the surface of an object [86] – in particular the perception of raking or scratching across its surface. In this interaction, there are in fact three gestures that must be considered in the design process: the control gesture, the sonic gesture output and the inherent gesture of the material to be scrubbed. Therefore, to have repeatable gestural response an effective strategy is to construct multi-layers of mappings to be used in a controllable and adaptive way, in order to change these gestures with context as I will discuss via three canonical examples.

Designing Topology Space

In scrubbing through sound samples, an initial consideration is how one wants to access this material: laid out along one dimension for continuous control, as a function of input control time, or controlling position in the sample in a more indirect way. The first is quite common and is one that I use most often. Taking this approach and mapping the input sample to one of the positional dimensions of a Wacom tablet provides for immediate access to any point in the sample's playback, so that the material can be played at variable speeds and starting points: a more complex and flexible version of turntable scratching.

This type of control becomes most interesting, however, when other sound parameters are modulated continuously as the material is scrubbed. To do this in a holistic way is

difficult as the mapping from linear position to sample time dominates the control possibilities. However, musical source signals (such as vocal samples) generally have different parameter combinations that are appropriate at different points in the file. Also the transition between relevant states of sound material (e.g. between vowels and consonants of speech) varies across the life of a sound, and this can be built into the mapping design as well. One approach is to utilize a continuous global scheme such as the aforementioned Gaussian mixture method. While this is a global approach, recall that it behaves in a “gravitational-system” fashion in that insertion of states with suitable mass can block the influence of more distant states. In this way, the two-dimensional control surface of a tablet can be projected onto the one-dimensional topology defined by the left-to-right states of a given sample. This projection was depicted back in figure 3.5. Note that the final mapping into three dimensions shown in this figure is for demonstration only, as the actual dimension of sound space is much higher.

In order to produce such a continuous one-dimensional topology in sound space, I have used a design in which parallel control spaces are mapped (one-to-one) into contiguous and user-defined frames of the given input sample. In doing so each space naturally provides duplicate information for the states that exist at the boundary between frames. In order to merge this information and maintain continuous control, I created an overlap-and-min approach as depicted in figure 4.24. For this design, the position along the tablet dimension that is mapped to sample time is segmented, with each resulting 2D “slice” being mapped into a different two-dimensional control space. States in this space are created such that overlapping areas between segments on the tablet share the same states – thus the mapping is duplicated and in the same spatial location in each control space (see the middle layer of figure 4.24). These duplicate states then map into the same high-dimensional vector of granular synthesis parameters. Using this structure, one can define different elements of the sample – such as vowels, consonants, transition areas of speech – to have particular sound processing characteristics based on the layout of each control space for a given frame. In regards to global scrubbing across frames: taking the minimum between the two duplicate versions of a given state (from each of the contiguous control spaces) results in the correct number of states for the resultant control space, and ensures that the final output is continuous. Further, the variance of each Gaussian kernel can be changed individually, and so the transition leading *into* a given state can be different when approached from the left side of the tablet or from the right. This structure

thus allows one to define the modulation potential based on particular regions of an audio sample, with continuous control of the sound arising as one scrubs across the sound. For example, if scrubbing is mapped to X-position, then every left-to-right trajectory with the same X-speed will give rise to the same scrubbed source material – however a unique overall sound will be produced by virtue of the particular path taken across the tablet surface, which in turn creates a unique path through sound space. Thus one can design an intuitive modulatory control that is based on the inherent sonic gestures of the source material, defined by the transitions between appropriately-selected sonic states.

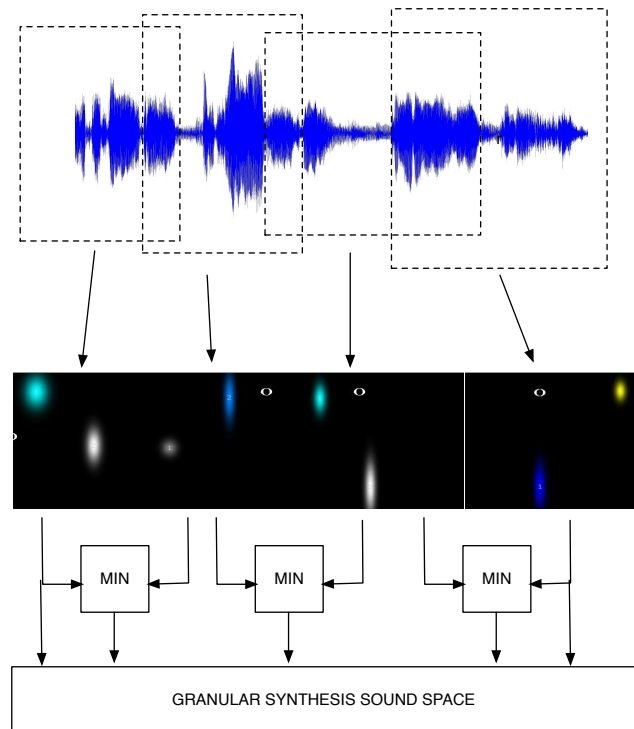


Fig. 4.24 Overlap-and-min of parallel control spaces, in order to construct one-dimensional topology in sound space.

Controlling Topology Space

For a variety of reasons, one may want to vary from the one-dimensional topology designed above. For example, if the input samples are changing dynamically, then a system must adapt to the changing gestural nature of this material. Allowing for influence from non-adjacent states can make the control gesture adapt to the changing

material by dynamically mapping to different trajectories in sound space and allowing for “alternate paths” between sound states. Further, the speed of scrubbing influences the resultant sonic gesture in several ways depending on the nature of the input sample, particularly in regards to the rate of change for parameters such as grain size that may cause pitched artifacts if varied too quickly. Therefore, I have designed a version of the Gaussian overlap-min mapping structure that moves continuously between a one and two-dimensional topology, while additionally adapting the dynamics of a given control gesture. The result is that the path in sound space is parametrically varied as well as the its speed. In order to account for the contextual effects of scrubbing speed, I map from the “phase space” given by velocity in both X and Y directions in order to modify the Gaussian kernels. This is achieved by influencing the size of the windows as in figure 4.24 in order to indirectly control the topology of the space: larger windows result in more influence from non-adjacent states. Also in this mapping layer, the X and Y variance of the Gaussian states are driven by X/Y velocity, which both influences the 1D/2D topology as well as the between-state transitional nature of the resulting control dynamics. The mapping structure used for this layer is the SI strategy as it provides continuous control while allowing for constant mapping in localized sections of the velocity space. As a further alteration of the dynamics, an exponential-windowed moving average (MA) filter (or leaky integrator) may be applied to this mapping⁹. The addition of this particular dynamics conditioning is interesting if it is tied to physical input gestures: by gating the velocity conditioning so that this mapping layer is active only when the stylus is in contact with the tablet. Therefore looped control gestures only map through the Gaussian mixture space. In adopting this structure, when a sudden scrubbing motion is looped the initial processing will occupy a localized two-dimensional region of sound space, but as the leaky-MA filter dies down the granular processing will fade back to the one-dimensional frame-based path that may tailored for the given sample. Therefore the initial excitation energy is mapped to a unique continuous transformation, while the machine-like repetition traverses a part of the space that is designed to work well for the given sound material. This control structure is depicted in figure 4.25.

⁹As I will discuss in section 4.4, note that a similar filter is used on the velocity data, but that this functions as signal conditioning and is not used towards designing a control gesture.

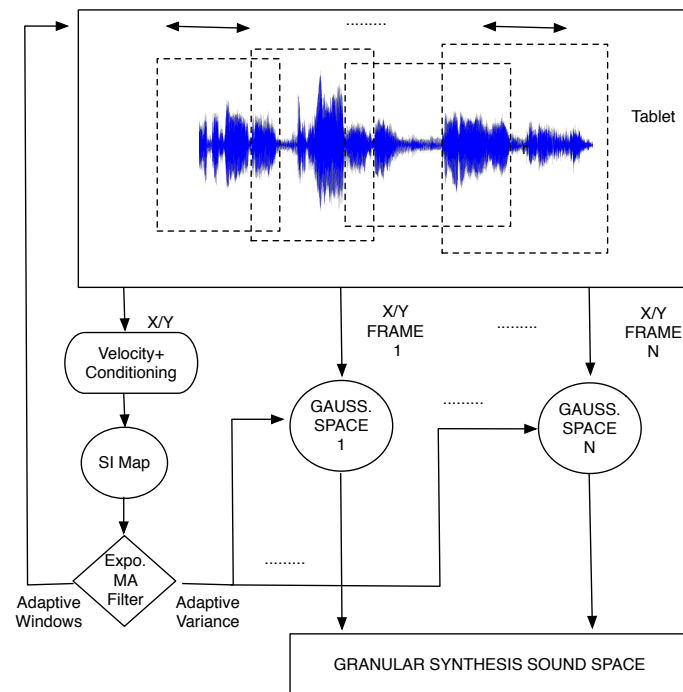


Fig. 4.25 Control structure is dynamically altered as input velocity influences the intermediate topology space as well as shape of mapping-to-sound space.

Control Gesture Design Through Control of Mapping Layer

Rather than scrubbing based on mapping tablet position to sample position, in some contexts it is more desirable to scrub continuously through the input audio from beginning to end, with possible looping. For example, when improvising between acoustic and electronic/tablet musicians, the latter may want to modulate the sound as it happens, with scrubbing rate being controlled indirectly as a product of the input gestural energy (rather than a direct left-to-right sweep). In these instances, the one-dimensional topology space and continuous control of this does not make sense. Instead, it is more musically relevant to use input gesture time as the temporal position variable itself, and utilize speed of input gestures to define the resultant control gestures. In other words in this structure I map the time – beginning from the point of contact with the tablet – into the scrubbing playback time of the input sample (or intermediate buffer, if audio is streamed). In this way, the player can move in any direction and this will scrub through the material in the same way, allowing for direction independence on the tablet. A physically intuitive mapping layer results if we map the speed of X/Y movement (taking the max of these two) into the speed of scrubbing. This allows the entire area of the tablet to be mapped into sound parameters for the underlying granular synthesis. At this point, one could simply map the tablet surface area using a mixture embedding into sound parameters. However, just as the left-to-right scrubbing of sampled material suggested the definition of frames in the audio, so too does the variable-speed nature of this control gesture suggest that the type of granular processing be tied to the control dynamics. With granular scrubbing, certain parameter values are more coherent and produce less artifacts at particular speeds (e.g. larger windows and less grain density for very slow scrubbing); therefore, I once again map from the X/Y velocity space, in this case directly into granular parameters. I utilize the SI mapping for this, as I require the mapping to be exact, local and scattered. Further, a piecewise linear response between speed and parameter change is coherent with the linear scrubbing through playback time. Through initial explorations with this instrument, it became clear that some latency and feeling of mass were needed between fast gestures and the ultimate granular modulation. As such, an exponential MA filter was used to modify the mapping from X/Y velocity into the SI granular mapping. This greatly improved the feel of the overall scrubbing. However, the feel from this structure is somewhat static in that certain speeds are

perceived to be more stable with more smoothly-decaying modulations, owing to the perceptual distance between parameter sets in granular parameter space. This in turn is a product of the particular mixture embedding. In order to compensate for this and to accentuate the resonance of certain scrubbing speeds, I have added an RST-based mapping from the X/Y velocity to the responsiveness of the MA filter (defined by window size and exponential damping coefficient). In adding this conditioning/response layer, a given scrubbing speed will result in a trajectory through sound space that varies in a nonlinear way. Therefore certain speeds can be tuned to move more slowly or quickly as needed. In the language of chapter 3, this response mapping is therefore a two-to-one direct embedding from velocity space into the responsiveness of the control filter. Now, this velocity-to-response mapping creates a complex control gesture that varies based on speed and overall gesture time. The order of speed variations greatly affects the overall control dynamics, and is something that can not be easily planned out by a human performer. Therefore, in order to make the response itself more controllable, the velocity-to-response mapping can be varied by *changing the smoothness and tension of the RST mapping*, which as we recall from chapter 3 are two control parameters for this technique. For this structure, the control of smoothness and tension is altered either by X/Y position or tilt, so that the time-varying response can be mapped either to different areas of the tablet or to different orientations. In doing so, it is not actual sound material that is being modified, but the action-to-control gesture that changes, which indirectly influences the overall sonic gestural output. This resultant control structure thus illustrates how a highly dynamic mapping can give a coherent and repeatable gestural response through a modular construction that acts primarily on the control-side parameters. The overall mapping structure is depicted in figure 4.26, in which the dotted-line arrows illustrate the gestural control of the RST mapping layer. The velocity-to-response mapping is illustrated in figure 4.27, wherein the actual mapping curve is determined through control of smoothness and tension, causing an interpolation between the possible response curves.

Sonic Gesture Design Through Feedback Control of Mapping Layer

The parametric control of the mapping layer in the last example was inspired by the need for a control gesture that varied with the temporal behavior of the scrubbing. This

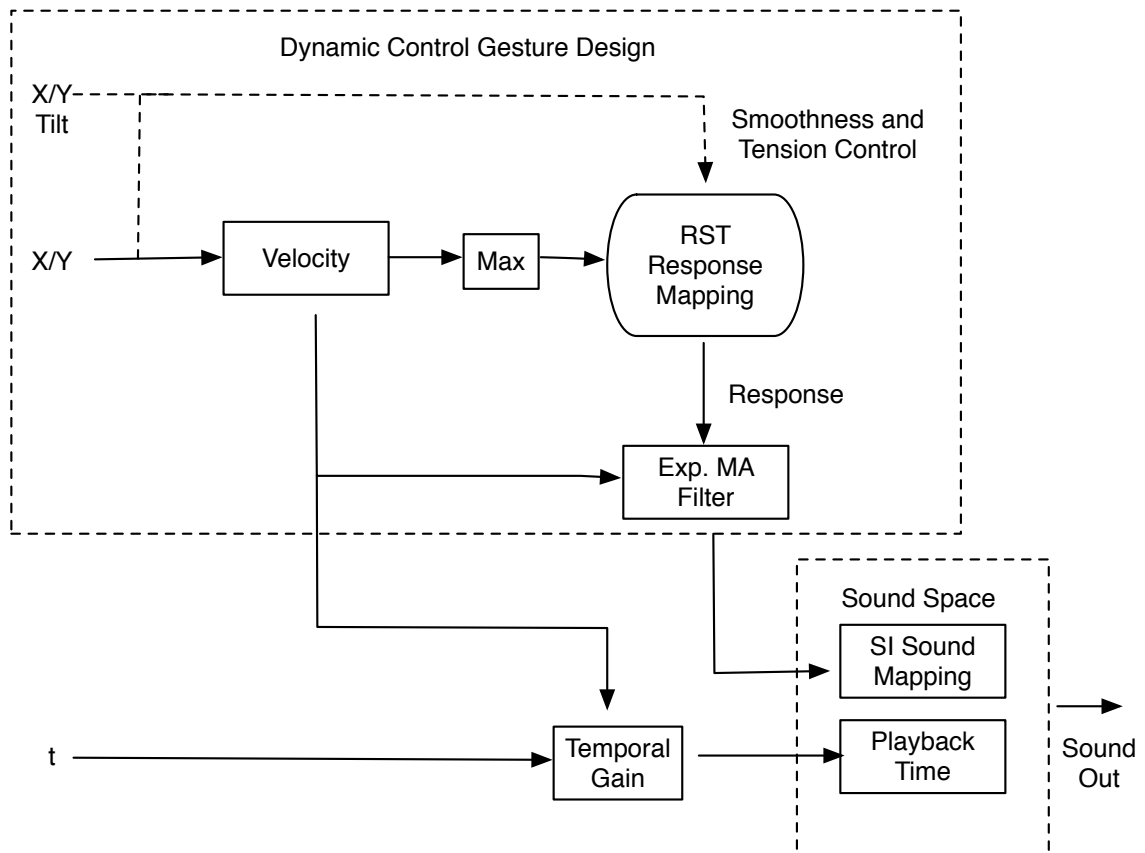


Fig. 4.26 Example mapping structure in which scrubbing is based on temporality of gesture. Speed of input action determines granular processing through SI mapping, and response and dynamics of control gesture are conditioned by another filtering layer that is itself gesturally-controlled by position or orientation.

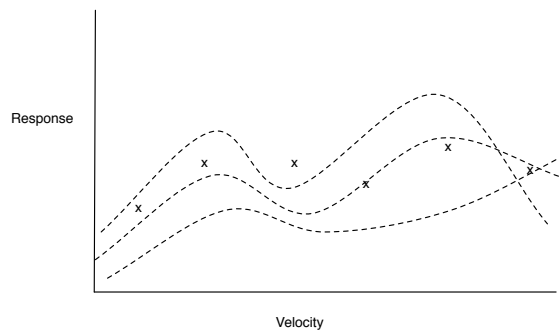


Fig. 4.27 Velocity-to-response mapping. The actual path in this space is determined by control of RST mapping's smoothness and tension parameters.

behavior and the need for adaptation arose from the desire to link control and sonic gestures more clearly and in a contextual way (changing with sonic source material). By identifying certain sonic gestures, this process can be extended by designing a control structure in which sound features drive this adaptation in a feedback control loop. In particular one effective control structure has resulted from my desire for a mapping that would focus more on the transient nature of slow scrubbing, eliciting different control feels depending on the spectral profile of the source material. In order to create such a design I have again drawn upon the morphological features from chapter 2. While a sound's entire morphology is not known in real-time processing situations (e.g. ratio of maximum time to total length), it has proven effective to extract temporal envelope and derivative information of matter features to drive the mapping response. In particular, for this structure I extract TFS as a transient grain measure and use this to adapt the scrubbing speed. Further, the RST-based control-response mapping (which was driven by gestural input in the last example) is adapted by first extracting a short-time window of amplitude and matter profile features¹⁰ and mapping this back into RST smoothness and tension. Following such a design, a mapping/control structure can be designed that takes into account the sonic gestural response of the instrument (as discussed in chapter 2) in order to link control and sound gestures in a more perceptually meaningful way. I've chosen to end the section with this particular structure, depicted in figure 4.28, because it arises not only from a consideration of the gestural aspect of control design, but also takes a synthetic view of mapping from the functional (considering parameter space structure), systems (considering parameter complexity/association) and perceptual (considering ultimate action/sound response) points of view.

4.4 Signal Conditional, Gesture Conditioning

An interesting lens through which to view this boundary of signal conditioning/mapping – first brought up in section 3.2 of the last chapter – is by assuming a perceptual view of mapping that considers the sensing of intentionality, shifting the concern to a proper match of control gesture and sonic gesture. In order to define a proper “perceptual gestural response” one needs to design and condition control gestures, something that was

¹⁰The temporal variation and flatness for amplitude, spectral centroid, deviation and spread are generally most effective.

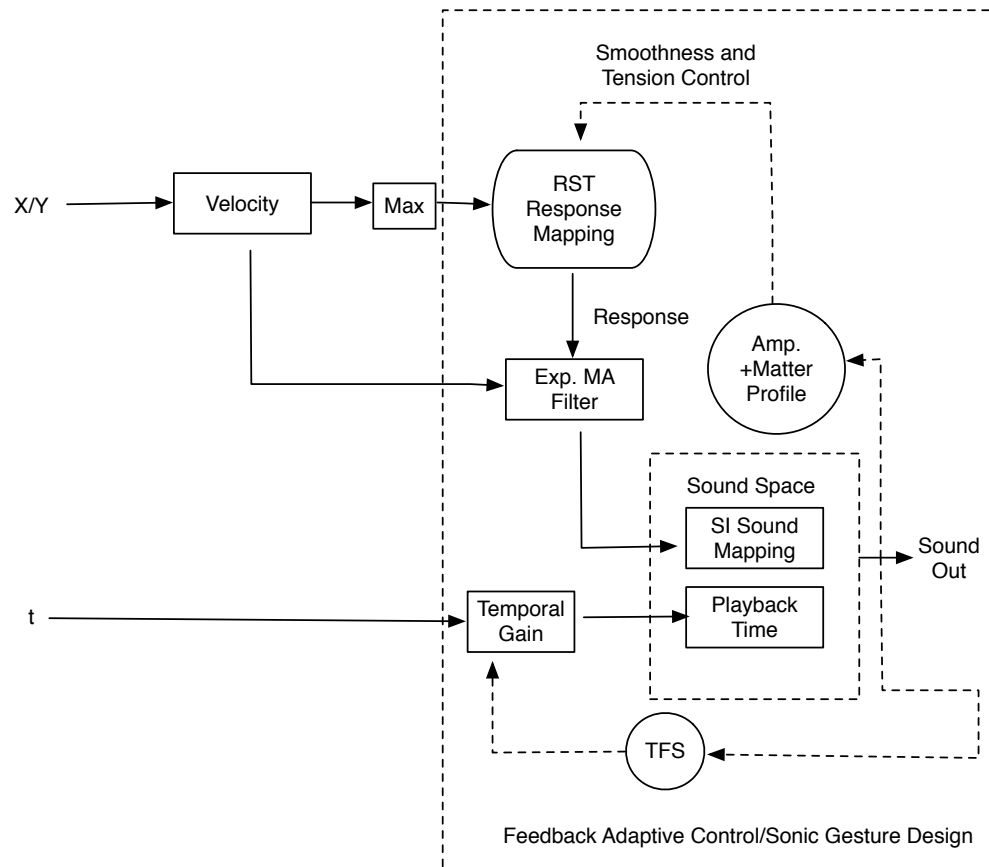


Fig. 4.28 Feedback adaptive mapping structure which changes mapping layer based on sonic gestural output.

implicit in the latter three control structure examples. By considering the perceptual aspect of mapping, the issue raised previously then extends to the boundary between signal conditioning, mapping and gesture conditioning.

Clearly, perceived temporality is an issue that must be addressed in designing an instrumental system. The user studies illustrated that mapping structure can define and alter perceived sonic gestural response, while the example control structures suggest ways that gestural and adaptive control of the mapping layer itself can give rise to coherent control and sound gestures. However, there is more to gestural dynamics than just multi-parameteric control through time: the nature of controller and interaction help to determine *how and where* perceived gestural dynamics are defined in the overall system design. The Wacom tablet as a controller generally does not move, and has orthogonal degrees of freedom with high precision and accuracy. Therefore gestural conditioning in the example control structures focused on inserting a mapping layer on the control side and dynamically controlling this by gesture or sound features. However in other control contexts there is not such a clear multiparametric structure, and one must be more careful to conceptually separate those functions and operations whose role it is to manage data from those that create dynamic behaviors that coherently respond to a given physical action. As the final discussion of this chapter, I will illustrate the challenge that this poses by way of two examples that present a considerably different control context than the tablet-based instruments which have been my primary focus.

4.5 Case Studies: Fabric-based Interaction Design

These instrument designs were created in the context of the *WYSIWYG* project¹¹ whose aim was to create novel fabric-based instruments intended for improvised play by the general public. In the case of many fabric instruments, a priori cognitive models of interaction are often physically introduced by an instrument's form factor, or the use of sensors and controllers (buttons, sliders, etc.) that are themselves separable or constrained to certain specialized actions and degrees of freedom [186][187]. In contrast, this project's design objective was to augment improvised play through fabric-based interfaces that do not rely on knowledge of "instrumental gestures", use any segmentation

¹¹A collaboration between the *Input Devices and Music Interaction Lab* of McGill University and the *Topological Media Lab* of Concordia University.

or recognize motion based on underlying musical structures. Rather, the goal was to promote the salient features of the textile (flexibility, stretchiness, texture, etc.) as determinant of the possible modes of interaction without directly referencing everyday cloth or fabric-based artifacts, or other human-computer interfaces [188]. For my part, I created two very distinct types of sound-based interactions, beginning from two different fabric controllers. In each case I'll begin by discussing the controller design, as by its nature it is a very strong determinant not only of the control degrees of freedom, but also of the signal and gesture conditioning required for a given interaction.

4.5.1 The Tapestry

The controller for this instrument – known simply as “The Tapestry” – is a 20' x 4' (6 x 1.2 meter) fabric that is designed for collaborative, public interaction. It was woven on a digital jacquard loom in the XS lab [189] by Marguerite Bromley. The weaving was such that conductive thread was used for a set of bird patterns scattered across its surface, with one example depicted in figure 4.29. There are twenty bird patterns which serve as electrodes for an integrated capacitive circuit¹² that can sense body/hand proximity in a range of approximately two feet (depending on bird size). The circuit outputs a digital PWM signal that is then converted by an Arduino sensor interface [190] into a usable signal before being sent into Max/MSP for all subsequent analysis and mapping.

Now, from purely a systems point of view the controller provides twenty separate channels of continuous proximity control that are similar in principle to twenty separate theremin antennas. However, given the flexible nature of the control object – which users can bend, stretch, flap, etc. – there are many more *implicit degrees of freedom* and ways in which users may affect the sensing. Further, the conductive thread provides a very volatile and varied response when ranging from proximity motion to slightly touching the birds, through fully touching the birds or finally to bunching them. The mapping from these intuitive physical gestures into usable control dynamics was a key aspect of the design, and brought the issue of gestural conditioning to the forefront.

¹²Designed by Elliot Sinyor and David Gauthier.



Fig. 4.29 Sensing bird from the Tapestry, woven with conductive thread.



Fig. 4.30 Users interacting with the tapestry instrument.

Interaction Context

The Tapestry was suspended lengthwise with the upper edge approximately 5.5 feet (1.7 meters) from the floor, as can be seen in figure 4.30. My design intention was to extend the presence of the object away from the surface and out into space, while inviting users to approach and interact further. This became the initial design constraint, and the sound and interaction were developed from this concept. I'll first briefly describe the overall interaction possibilities, before describing the conditioning and mapping used to achieve this in more detail.

When a viewer approaches the Tapestry and comes within a certain range, a low level of ambient sound fades in. The specific type of sound depends on which bird a user is in proximity to. As one moves closer, there is more detail uncovered in the overall sound, and it seems a bit louder still. As a bird is touched a sound begins to play and fades if touched momentarily. However, if the bird is touched continuously, the sound becomes louder and more varying, and takes longer to fade away. The “release” aspect of one's gesture also has a different quality depending on this. This behavior – temporal response as well as the manner of the sound variation – changes depending on whether a bird is touched, held, stroked or bunched. If more people interact with birds, this behavior is different still. Finally, the overall state of activity for all people in general proximity determines the quality of sound that is produced as well as the global response characteristics of the hand/fabric interaction.¹³ Here I focus on the design of the immediate interaction between users acting intentionally on the Tapestry, but this activity level information does affect immediate response in a type of indirect mapping, as I will mention.

Designing Control Gestural Response

The conductive sensing thread in the Tapestry is highly nonlinear, and the manner in which the electronics queries each bird results in data rates between 7-20 Hz. As a result of these two constraints, initial signal conditioning played an important role in the creation of usable control gestures. In regards to the nonlinearities, different ranges of

¹³At an ever larger scale, the gallery space and adjoining hallways were tracked via cameras and the software Jitter. This data (including amount of energy in the space and ratio of energy inside to outside) was broadcast as “state” information that could be used by the individual installations.

proximity exhibit more or less exponential increase, with different output ranges. My solution to this was to utilize overlapping versions of the same signal, with different log functions and different normalizations – a fairly straightforward case of signal conditioning. This parallel set of ranges also allows for defining thresholds: when users are 2 feet away, a low-pass filtered sound fades in, and once a second threshold is crossed, proximity controls the cutoff frequency of this filtering.

Compensations made for the low sampling rate of the control signal had a strong influence on the perceived mapping and temporal response, with this conditioning in turn influenced by higher-level control data. In particular, I produced continuous data response by the use of a bank of leaky integrators for each bird, combined with a single-pole lowpass filter. The result is an output value that increases while receiving the continuous (positive-valued) input, but dies down when this input ceases. Not only was this used to smoothly produce continuous output, but it reflected a desire on my part to produce a sense of “mass” when moving towards a given bird, rather than simply treating the Tapestry like a bank of cloth buttons/triggers.

This sense of pushing into the fabric was achieved by triggering the integrator once contact with the birds was made. Due to the inherent nonlinearity of the fabric (e.g. dependent on the degree of contact between threads), moving ones hand causes a more rapid swelling of the integrator, and bunching increases this effect much more so. Thus, in this instance idiosyncrasies of the controller were used in order to differentiate types of gestural input. These integrated control values were mapped to the amplitude of the sound file associated with a given bird. Thus, if one touches a bird then a sound is triggered, and if contact is kept the sound increases with accelerating volume. The perceived effect (and so mapping) is of a location that gets too “hot” if one’s hand is left too long – much like a stovetop burner. At the same time, a lowpass filtered version of this integrator output controls the perceived *release characteristic* of a given touching gesture as follows.

When contact is made with a given bird, the triggered sound file plays and is simultaneously sent through a real-time granulator¹⁴ that processes synchronously with the sound file, creating a texturizing effect. The accumulated value from touching a bird, after lowpass filtering for smoothness, controls the amount of gain on a scrubbing [144] action that allows the user’s hand position – after releasing – to affect the past values of

¹⁴Whose implementation is described in [178].

the granulated buffer. As the sound volume falls off more or less rapidly after release, the perception is of a decay envelope instead of a new mode of control. Given these operations, the simple perceived mapping of moving towards an object (the bird), touching/rubbing/bunching (with its varied response) and the sound of release (perceptibly different than the “attack” of first contact) is quite complex from a systems view of the mapping. The overall control structure centered around this gesture includes actions that share responsibility between conditioning and affecting the perceived mapping.

Interplay Between Control and Sonic Gestural Response

Given this physical-to-control gesture mapping and conditioning, the sonic gestural response becomes a product of the resultant control gesture, the hysteresis of the sound processing, the state of the Tapestry and the larger room-state information. The hysteresis from the granulation arises both from time-stretching and from a feedback delay implemented on each of eight grain streams. This allows for a variety of textural effects, depending on the collective delay and gain parameters for each stream, as well as window type and size, grain rate and frequency transposition.

As a result even basic, static input files have a noticeable time evolution that is driven by the conditioned control gesture – the speed of approach, length/type of contact and speed/position of release influence these memory-based granulation parameters through the mapping to scrubbing position as well as global volume and filtering. A single control gesture does not directly control the granular parameters however. In order to design a multi-user interaction, the global state of the tapestry was used to control this sound processing. In particular, the birds were divided into regions, and the normalized sum of the integrators from each region was used to interpolate between two states of the granulator: low energy was mapped to a synchronized granulation with no feedback and imperceptible processing, while very high energy moves towards a lag in granulation, feedback delay and window sizes that results in a texturizing effect.¹⁵ As a result, the pushing and building of energy on the scale of one bird was mapped out to regions of the

¹⁵In its initial showing, this modular concept was extended to a meta-control on the scale of the entire room in that the amount of energy controlled a point-based movement in a “state space” defined by a triangle. Each node represented a set of granular parameters and different source material, while an SI mapping determined the global sound quality.

tapestry, thereby extending this interaction metaphor and encouraging improvised collaboration. The result is a perceived mapping on a larger time scale: the sense of building up a “composition” from individual bird “instruments”. There further is a sense of “sections” in that a different granular stream exists for each region (normally divided into three).

4.5.2 The Blanket

The Blanket was constructed from a 3 x 3 meter piece of highly stretchable Lycra fabric, on which was sewn a 5 x 5 grid of light-dependent resistors (LDRs) that span its surface. It is hung from the ceiling (or surrounding walls) through its corner rivets and positioned horizontally 6' (1.8 meters) above the ground, as in fig 4.5.2. The sensing surface of the Blanket faces the ceiling and an array of lights are projected in parallel so that the intensity increases from faint at the Blanket surface to intense several feet above this. Players are underneath or to the side, and interact with the instrument by using their hands or upper body to push upwards or by shaking the instrument from the boundary. This changes the shape of the surface and consequently the output of the light-dependent resistors, whose output is transmitted to an Arduino [190] sensor interface for sampling and conversion. While the response of the LDRs was not linear, a reasonable approximation was obtained by utilizing a vertical array of stage lights whose intensity increased from bottom to top. As with the Tapestry, the electronics for this controller were designed and built by David Gauthier and Elliot Sinyor.

The focus for this instrument is the *result* of participants' individual and collective input gestures rather than the ancillary or expressive gestures of individual users. Thus gesture here is considered not from an instrumental or communicative perspective, but from a third point of view: those gestures that are embedded in the fabric itself. The concern then is how to capture these “Blanket gestures” and to represent them in the resultant sound. The design challenge is to map the sum total of all individual and collective gestures as they emerge from the fabric, which requires appropriate gestural extraction approaches.

Interaction and Intentionality

The extraction of gesture was driven by the search for *intentionality* [191], which for the Blanket I assumed would manifest as repeated or concurrent motions. Therefore, rather than focusing on raw positional sensor data, I developed a correlation-based analysis. This decision was made so that more meaningful gestures could be found than via simple position, yet this information is “low-level” enough that novel behavior may emerge from the time-varying and continuous sensor topology in a way that may be overlooked by methods that overspecialize to a given gesture, such as pattern recognition. As the Blanket LDRs are distributed uniformly across its surface there is no directional bias, and one is free to select any subset of sensors from the grid for analysis.¹⁶The analysis is based on extracting various spatial and temporal correlation sequences from a subset of the sensor grid.

The correlation features were designed to respond to certain Blanket gestures that arise from modes of interaction that I considered to be indicators of intentional action, including periodic motion and general wave-like or repeated movements. The three main techniques used were a multi-dimensional cross-correlation between entire sensor regions (as in figure 4.32), an instantaneous correlation between a set of different sensor locations and autocorrelation extracted from each sensor location. The first are spatial correlations that provide information about the contour and direction of blanket gestures: concurrent motion and the phase relationship between different areas of the Blanket. The third feature set consist of temporal correlations, and give information about regularity and smoothness of motion. A more complete discussion of this analysis can be found in appendix C.

This extraction of salient features from temporal wave patterns using the temporal autocorrelation function (ACF) is analogous to a sound/music analysis context, where this has been used to measure the ratio of odd/even harmonics or of voiced/unvoiced parts of speech, among other features [42]. The Blanket gesture analysis extends this approach in that I look for idiosyncrasies unique to the two-dimensional fabric control surface using the spatial correlation analyses as well.

¹⁶A separate WYSIWYG Blanket project by David Birnbuam and Freida Abtra experimented with directly mapping from sensor subsets into sound [192].



Fig. 4.31 The Blanket instrument (a) sans human interaction (b) collective play along the interior. The colored lights projected from the top are for theatrical effect.

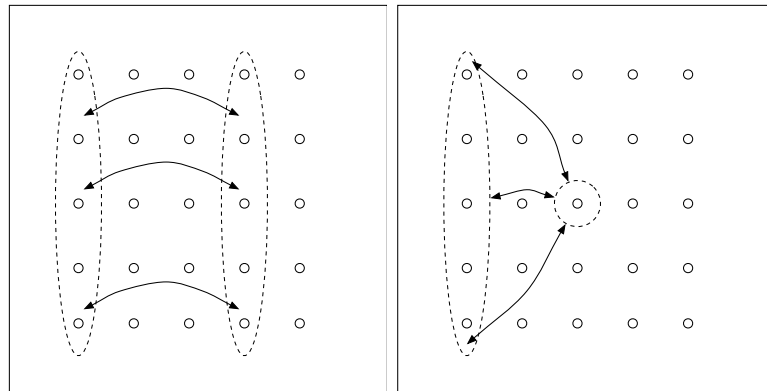


Fig. 4.32 Different Approaches to the Blanket topology include (a) the rectangular interaction between columns and (b) the circular interaction between boundary and center.

Gesture Features and Mapping to Sound

There were several higher-level features extracted from the correlation sequences, while continuous cross-correlation values and filtered positional values from different sub-regions were used as lower-level control values. Using an instrument analogy, these two sets of low vs. high level continuous features were partitioned (by virtue of the mapping) much like many bi-manual instrument interactions – such as playing a bowed string instrument – where a dominant hand controls excitation and finer timbral details and the other provides modification.

The most effective higher-level features were ones that clearly responded not only to regular motion, but the *process of moving towards regularity*. In addition to fundamental frequency, this included the degree of *periodicity*, taken from the (non-zero lag) peak value of the ACF and the degree of *harmonicity* extracted from the combined values of all peaks that exceed a user-defined threshold.¹⁷ Further, taking the harmonicity normalized by the total power of the signal, gives a measure of the *harmonic-to-noise ratio*, which is a useful measure of how many people (in a multi-user context) are engaged in producing collective gestures.

These measures were augmented with a relatively lower-level value of *roughness*, taken as the vector-valued RMS difference between consecutive windows of a given autocorrelation function. To my knowledge, this is a novel use of the *continuously varying autocorrelation function itself* as data to drive a mapping. In order to have more immediate continuous feedback to users, the positional sensor data from the Blanket was grouped in regions, and these were conditioned by an FIR filter and leaky integrator in series. This was used to condition the temporal response and to parallel the physical motion of the Blanket itself as it rose and fell – the first example of an action on the boundary of signal and gestural conditioning.

For sound output, I utilized the GMU granular synthesis implementation [193] which allows for control of many grain streams with variable control of the grain source for each. Those features that take longer to develop on the Blanket are mapped to more “global” and perceptually salient features of sound: the Blanket was divided into N regions, and the periodicity of the N th region was mapped to the mean frequency of the N th grain stream. Meanwhile, the harmonicity was mapped to the variance of these same streams.

¹⁷Where this threshold is a free parameter that can tune the system response and is highly composable.

The harmonic-to-noise (HNR) ratio and the roughness were mapped in a convergent fashion to the ratio of noisy-to-sinusoidal buffers that were used as source grains. In particular, a higher HNR populated the pool with more “pure” grains, and roughness continuously controlled the probability of noisy vs. sinusoidal being selected. This part of the mapping is in analogy to a slower modification gestures, which modulates the average frequency and noisiness. At a lower-level, the smoothed/integrated positional sensor data was mapped into the amplitude for each stream, so that pushing up on the fabric would build up the sound in amplitude like an excitation gesture that moves on the same time scale as the Blanket surface. Finally, the temporal response of the integrators as well as select ACF windows were influenced by the multidimensional cross-correlation values. The effect of this is that more in-or-opposite phase motions (as opposed to noisy motion) encourage a tighter temporal response of the action-to-sound coupling. This was a clear case of the interplay between conditioning and mapping, where the parameter mapping was fed back into the mapping layer in order to condition the Blanket gestural response.

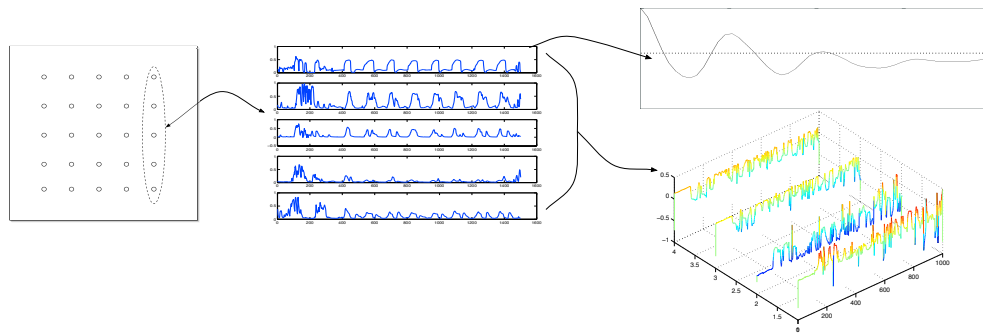


Fig. 4.33 Sensor data from Blanket (left) is “tapped” as time series (center) and autocorrelation (right top, function of lag time) and cross correlation (right bottom, function of spatial lag and time) series are extracted in order to provide information about regularity and phase relationships, respectively. Above data results from flapping the Blanket edge closest to the chosen column.

Interplay of Blanket Signal/Gesture Conditioning

A key element of this instrument is that any gestural input will result in a continuous response. This allows a participant to explore the system’s response, finding the “resonances” that arise from my loose modeling assumptions about intentionality as well

as other meaningful responses that have not been considered, thereby adhering to a continuous human-in-the-loop system design. This instrument illustrates a strong interplay between conditioning signals and gestures, as “gestural analysis” per se is a conditioning of *control gesture response* rather than a symbolic-level gestural feature extraction. In this case the fabric itself – its position and orientation in space – can be considered a mapping of sorts in the sense of an embedding in a control parameter space. The gesture of the blanket then becomes both a conditioned version of this movement through parameter space as well as movement through an intermediate “gesture space” defined by the correlation-based measures.

In addition to the feedback mapping from correlation into responsiveness, there is also an interplay between signal/gesture conditioning in the tuning of the correlation parameters, where there is a dependency in regards to how the sensor streams are windowed across time and space. For example, the window size for each incoming stream in the case of temporal autocorrelation has a strong effect on the detection of periodicity, on the temporal response of this measure and also on the time scale over which information propagates across the surface. There is a tradeoff between bias and variance in the data streams depending on the estimate used (see equations C.2 vs C.3 in appendix C), as well as depending on the implicit assumptions of what occurs outside the given window of observation [44].

These factors of windowing vs. instantaneous estimate, window type and length are treated as free parameters that are used to adjust the system response (for example, in variable lighting situations). Thus on one level these parameters are used for conditioning of the signal integrity, but in the process this also conditions the response of the system and moves the “resonances” towards different types of input gestures. Given the continuous sonic feedback of the instrument, the control gesture design was strongly informed in a feedback loop with the mapping to sound, which paid particular attention to sonic response over multiple time scales: slow modification vs. fast excitation Blanket gestures.

4.5.3 Tapestry and Blanket Projects: Reflection

This Tapestry illustrated a control context where in order to achieve a given perceived mapping, I had to determine the time-varying nature of the input data, the separate and

coupled roles of control and sonic gesture, the influence of mode or state changes, and the role of feedback in design. Operations such as filtering and integration here play the parallel roles of conditioning a control curve, of regulating state changes through feedback control and of mapping the input to a sonic gesture. This system illustrates how different levels of perceived mapping may arise from an underlying system of gestural conditioning. In this case the relative importance of control vs. sonic gesture design was determined primarily due to the presense (or lack) of memory in the system. Temporal response was carefully considered based on the controller and the constraints of the interaction type, so that the focus was on conditioning a control gesture that was tightly coupled with the performer's input gesture. The dynamics of this conditioned gesture then fed through into the sound processing, with the scale of this temporal response controlled by the global activity of all bird sensors.

In contrast, the Blanket was not about conditioning one gesture and feeding this dynamically upward through the system to create a perceived mapping to sonic gestures. Rather, the focus was on the different temporal scales and dynamics motions of the object itself. Rather than conditioning a volatile sensor array to react to certain gestures, the design focused on observing the common gesture types, and reinforcing these through the correlation-based measures that were mapping in combination with filtered, low-level position information. Conditioning the signals through tuning window size and responsiveness coefficients was towards finding a good coupling of Blanket-to-control gesture, while the mapping to sound arose from an observation of two distinct time scales (modification/excitation) based on immediate pushing/flapping vs. building up of coherency (e.g. periodic motion). In the first case the interest was on feeding the pushing gestures into a larger sound process, while in the latter is concerned with sonifying the gestural nature of the moving fabric. In both cases, however, there was no simple system mapping from controller-to-sound, but instead a series of conditioning steps that served to regulate the signal, define a dynamic behavior, or both. Further, in each case a mapping from control output back into these conditioning parameters was needed to defined the overall perceived control-to-sound gesture mapping.

4.6 Chapter 4 Summary

This chapter applied the theory and addressed the subsequent questions that arose in chapter 3. The interplay between mapping structure and sonic gestural dynamics was explored through the set of user studies, which suggest that this structure alone can indeed influence perceived sound quality and control feel. It was also the case that mapping structure influenced performance in the musical acquisition tasks in a way that deeply related to sonic control context: certain mappings were more conducive to performance in tracing certain sonic gestural contours.

While the instruments from these studies were complex in regards to the multidimensional, cross-parameter nature of the mapping, they were still relatively limited for the sake of the perceptual study. In order to define more musically relevant instruments, I therefore augmented them by adding modular mapping layers, aimed at creating a more natural feel and subtle articulation possibilities. This included adaptive control of mapping parameters themselves, altering the geometric structure of the mapping in order to change the control/sound coupling dynamically. This illustrated the way that the desire to create certain control or sonic gestures is built into the control structure itself, subject to the musical control context.

These control structures began from a multi-parametric system in which mapping was viewed first and foremost from a spatial point of view. Gestural dynamics were built in to the system through dynamic control of the mapping layer, either from control-side or in feedback from the sound output. This approach is top-down, as the mapping structures are imposed on a space and the behaviors of control/dynamics are governed by this structure.

Moving away from a purely spatial, top-down approach to control structuring I ended the chapter with two examples in which gestural dynamics needed to be conditioned at a lower level, by virtue of the controller type and the interaction context. Parameter mappings were constructed in order to defined a certain perceived temporal response, which brought the issue of *dynamics within mapping vs. the dynamics of mapping* to the forefront.

In all of the systems of this chapter, sound synthesis is considered through the lens of how it related to the overall control structure. While this is very relevant to many instrument designs for the reasons I have outlined, one may want to design more intimate details of

signal behavior or low-level control dynamics. For this reason, I will now invert the discussion in order to focus more directly on time rather than space, and to consider mapping from the point of view of the signal-level control dynamics – a bottom-up perspective.

Chapter 5

Sound Modeling and Control

5.1 Defining Control/Sound Gesture Dynamics, Considering Sonic Context

The examples of the previous chapter describe methods to dynamically change a given control structure in order to achieve an overall gestural response. The resultant sound behavior was tuned in general, high-level ways: making the sound become more or less grainy for a given gesture, or more or less bright. One thing that is hard to do using this approach, however, is to link the specifics of certain control gestures to certain sonic gestures. One way to achieve this is to express the temporal dynamics of such gestures, and how control affects the sound over time. This requires one to consider control in a lower-level part of the instrumental design hierarchy, and to *embed this in the description of the underlying sound process*. A novel way to achieve this, as this chapter is about to explain in depth, is to design a state-space model of the control/sound interaction.

The dominant paradigm for understanding a linear system (audio filters, the Fourier transform, etc.) is the transfer function representation, which describes the input/output relationship that the system will produce. Any such system can also be represented as a recursive difference equation that expresses not just the out-of-time effect of the system, but the dynamics of this system over time [194]. Therefore, one can express and modify the dynamics of a sound transformation system, building a control structure around this representation. As we will see, methods for statistical estimation can be applied to this framework, and can be extended to modeling in non-stationary and nonlinear

environments. A further advantage is that a physically-inspired model of a control/sound system can be used to constrain the interaction, so that physical and more abstract signal models can be used in a hybrid fashion. This potential for parameter estimation, accounting for nonlinearity and model hybridity are the primary features that I exploit in creating control/sound structures using this representation, as I will detail in sections 5.4 through 5.8.

The approach of this chapter illustrates that control and mapping may be considered at a very low-level: in the process of deciding upon the underlying sonic context for a given instrument. When one considers the type of sound they want to control they are, at least to some extent, imagining the space of possible transformations of this sound. Therefore, it is beneficial for one to consider the sort of transformations a given sound model affords, and the ways that control may be acted upon it. My primary interest, as I have outlined at various points in this writing, is the control of what I have called the grain element of sound in chapter 2, which includes timbral elements such as roughness and extends to general textural phenomena. As I've noted, such elements can arise from the interaction and modulation of stationary spectral components, as well as from stochastic elements of sound. Given this observation and general design criteria, I have focused on a family of sound models that parameterize the stationary and stochastic components using a spectral representation that is commonly based on an underlying Short-Time Fourier Transform (STFT) analysis [64]. Using this as a fundamental approach I have built a dynamic model of sound analysis and synthesis, focusing on a design that will simultaneously lead to interesting transformations of textural and noise-based sound features while allowing for control structures to be integrated into the sound dynamics. While many of the algorithms utilized for this system are not new in and of themselves, it is the resulting integration of several methods and my particular musical application which is novel. For example, the use of the Kalman filter for additive transformations was demonstrated in [195], but here I extend this by building on a recursive-exponential implementation, and exploiting a fast algorithm in order to process either additive data or the full underlying phase vocoder. Further, this model is augmented to allow for nonlinear adaptive control. As such I will present the derivation of the algorithms in the chapter rather than an appendix, as this is a fundamental part of the work itself. To help maintain perspective during the more mathematically-oriented moments, remember that the overarching form and trajectory for this chapter is as follows

- Section 5.2: provide relevant background on the chosen sound model and explain the underlying state-space representation. Present idiosyncratic effects that illustrate how this dynamical systems approach may be exploited for musical purposes.
- Section 5.3: derive an expression that allows us to extend the model from periodic sounds (DFT) to any given sound input (STFT).
- Section 5.4: extend this to the final form of the sound model, which uses a Kalman filter for spectral re-estimation of parameters under transformation.
- Section 5.5: derive an algorithm that allows for a significant reduction in computation time.
- Section 5.6: build an additive synthesis layer on top of this, allowing for more parametric control of perceptually and musically relevant sounds.
- Section 5.7: present the families of possible sound transformations that this model affords.
- Section 5.8: return to the notion of control, deriving a new form of the model in which control is embedded within the state description in a way that allows for nonlinear dynamics. Provide two canonical examples – one a basic building block and one more musically complex – that illustrate the power of the overall system.

Thus while there will be sections in which the details dominate the discussion, the reader can refer back to this list in order to keep in mind the primary goals of the chapter.

5.2 State-Space Analysis/Synthesis

5.2.1 The Phase Vocoder

The phase vocoder is a widely used tool for the analysis, transformation and synthesis of audio signals. It began as an attempt to efficiently code and transmit voice signals using filterbanks [47], was later represented by the STFT [51] and then began to find use in musical applications [49],[50]. The most common effects generated by the use of the phase vocoder are pitch shifting and time scaling, which are achieved through altering the time/frequency block increment size between the analysis and synthesis step and then interpolating. If the step increment for both analysis and synthesis is subjected to certain constraints based on the type of windowing function used in the STFT, then the input

signal is perfectly reconstructed upon re-synthesis. However the phase vocoder becomes musically interesting when the signal is distorted by transformations such as pitch/time scaling, cross-synthesis and others in which the amplitude and phase of each frequency bin are modified over time. As the representation itself is purely deterministic and able to capture the signal entirely, these distortions are externally applied to the spectral data in an intermediate (i.e. between analysis and synthesis) step. While this approach affords many interesting transformations and is the basis for much of the spectral processing used in computer music compositions, I have found that certain other interesting effects can be produced by embedding a stochastic representation within the phase vocoder itself. This approach represented my first use of a state-space representation (SSR) for sound processing, and so I will present the underlying SSR framework along with these effects as an illustration of the potential for noise-based sound processing.

5.2.2 State-Space Phase Vocoder

Generally speaking, rather than model a system in terms of its transfer-function representation one can express its actions on a stochastic process $y[n]$ by way of the two state-space equations

$$s[n + 1] = As[n] + w[n] \quad (5.1)$$

and

$$y[n] = Bs[n] + v[n] \quad (5.2)$$

where the sequence $s[n]$ is the state of the process at time n , and equation 5.1 represents the internal dynamics of the process as governed by dynamics matrix A . Equation 5.2 projects the state vector, which may be hidden, into a vector of observable output variables. Both w and v are assumed to be zero-mean, Gaussian white-noise processes. The first affects the progression of the state while the second is additive noise present in the output process x .

The creation of an SSR-based phase vocoder is possible by exploiting a recursive description of the Discrete Fourier Transform, as was presented in [196]. The essential

idea is that the complex exponentials of the DFT can be expressed as

$$e^{jn\theta} = e^{j\theta} e^{j(n-1)\theta} \quad (5.3)$$

for time n and frequency θ . Thus the DFT matrix and its inverse can be expressed as a first-order recursion, which from the above equations we can see is a necessity in order to work within this SSR framework. Therefore the DFT matrix may be represented by the state matrix A , and can be thought of as expressing the *process of the Fourier transform* in the way one might imagine a heterodyning filter to act over time in an analog filter-bank DFT implementation. The state s in this case is a vector representing the spectral frames acted on by the DFT matrix at each time step, while v can be thought of as an additive output noise similar to the residual found in the Spectral Modeling Synthesis (SMS) approach [64]. This basic representation forms the foundation for the control/sound models that are derived throughout this chapter.

Related work

A state-space approach to analysis/synthesis was presented in [197] in which the real and imaginary components of p sinusoidal partials, tracked over time, were represented in the state vector. The observation matrix summed across the real components of the partials, and the addition of observation noise generated a sinusoid+noise re-synthesis. The authors claim that this model represents a hybrid source-filter / sinusoidal model. The same underlying model was also used in [198], though the spectral components that were tracked did not necessarily represent partials. Similarly, a recursive state-space formulation is presented in [199] wherein the state is comprised of the real and imaginary components for N evenly spaced frequency bins. Thus, this implementation maintains all of the data from the phase vocoder while the aforementioned work directly tracks partials and so is a sinusoidal model. The motivation differs in [199] as well, with the goal being the interpolation of missing audio samples whereas the former two projects were concerned with building an analysis/synthesis scheme for audio transformations. The work I'll now present in this section is situated between these two in the sense that the motivation is towards musical transformations, yet I preserve the lower-level representation given by the complete Fourier spectral frames. However my state-space implementation differs from [199] in a reflection of the differing motivations: the desire to

track time-domain signals and interpolate missing values led to a stochastic representation as in equations 5.1 and 5.2. However, for musical effect I have decided to build uncertainty into the time-varying signal by adding noise to the dynamics matrix in order to perturb the structure of the *system itself* and to explore the complex couplings that result.

SSSPV Implementation

Even before adding a framework for estimation or control, this SSR approach to the phase vocoder can result in musically interesting effects simply by exploiting its explicit representation of spectral dynamics and a stochastic component. With this in mind I created my first implementation, referred to as the stochastic state-space phase vocoder (SSSPV). In this algorithm, the nature and size of the state varies depending on whether the analysis or synthesis step is being performed. For the analysis step, given an input block of real signal $x = \{x_1, \dots, x_N\}$, the state vector is initialized as

$$s = [x_1, 0, \dots, x_N, 0]^T \quad (5.4)$$

for the current block of N samples (the DFT window size). The state is re-initialized with a new input block at each signal boundary (each N samples), and during the state recursion s is propagated by the dynamics matrix

$$A = \mathbf{DIAG}(R(\theta_0), \dots, R(\theta_{N-1})) \quad (5.5)$$

where **DIAG** represents a block diagonal matrix and

$$R(\theta_k) = \begin{pmatrix} \cos(\frac{2\pi k}{N}) & \sin(\frac{2\pi k}{N}) \\ -\sin(\frac{2\pi k}{N}) & \cos(\frac{2\pi k}{N}) \end{pmatrix} \quad (5.6)$$

The observation matrix

$$B = \begin{pmatrix} 1 & 0 & \dots & 1 & 0 \\ 0 & 1 & \dots & 0 & 1 \end{pmatrix} \quad (5.7)$$

produces an output vector¹

$$\hat{s} = (s_{0,r}, s_{1,r}, s_{1,i}, \dots, s_{\frac{N}{2}-1,r}, s_{\frac{N}{2}-1,i}, s_{\frac{N}{2},r}) \quad (5.8)$$

¹The trivial imaginary values at $\theta_0, \theta_{\frac{N}{2}}$ are discarded.

which is comprised of the real and imaginary components of the spectrum for input block of signal x . We assume that x is real, and so only the first $\frac{N}{2}$ frequency bins are generated by the analysis state equations.

Now, the observed process \hat{s} becomes the state vector for the synthesis step, where the synthesis dynamics matrix is defined by

$$\hat{A} = \text{DIAG}(1, R^{-1}(\theta_1), \dots, R^{-1}(\theta_{\frac{N}{2}-1}), 1). \quad (5.9)$$

The new observation matrix

$$\hat{B} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & \dots & 1 & 0 & 1 \end{pmatrix} \quad (5.10)$$

produces output signal \hat{x} . In the absence of noise added to the state or observation equations for analysis and synthesis, this two-step recursion provides a perfect reconstruction. However, the addition of noise into the state equations at various points in the analysis/synthesis process and with different time-varying behaviors can introduce different textural qualities into the input sound that can then be controlled.

Introduction of Process Noise

This set of idiosyncratic and interesting effects begins by introducing noise into the state *matrix* rather than simply to the state vector as is typically done through process $w[n]$ of equation 5.1. This is achieved by reformulating the state equation as

$$s[n+1] = (A + W_n)s[n] \quad (5.11)$$

where W_n is an $N \times N$ time-varying matrix of Gaussian white-noise. In order to parameterize and specialize the type of effect and control possibilities, this matrix is decomposed as follows:

$$W_n = \alpha W_n^d + \beta W_n^r \quad (5.12)$$

where W_n^d contains non-zero values only along the block diagonal that corresponds to the non-zero values of the dynamics matrix A , while W_n^r provides the Gaussian values for the remaining upper and lower triangular parts of the matrix. α and β are free parameters that allows one to tune the contribution of these two different parts of the noise matrix.

The result of this addition is that the matrix W_n^d is added to the sinusoidal components of the dynamic rotation matrix, causing an uncorrelated and random fluctuation of amplitude and/or phase in each element of the state. The amplitude or phase of a given frequency can also be modified by converting the corresponding members of the state vector into polar form, acting on the appropriate values and then converting back to rectangular form before re-inserting them into the state equation. In this way one can e.g. introduce concurrent random modulations between partials that can induce jitter and lead to an influence of a given sound's texture [200].

Now, the matrix W_n^r causes a random fluctuation which behaves quite differently.

Random values added in this part of the matrix introduce a non-linear distortion, and noise can be added at specific matrix locations in order to introduce a coupling between two frequencies. This can be non-physical – such as if the frequency at bin i is coupled to bin j but j is not coupled to i – or it can maintain some physical coherence if frequencies remain coupled in a bi-directional manner. I have experimented with this and other process noise behaviors towards the end of creating musically interesting noise-based effects.

5.2.3 Example Effects

Different effects that influence the signal grain are observed depending on several factors, including where in the dynamics matrix noise was introduced, if and how it was propagated in time through the matrix, and whether it was added during the analysis or synthesis step. I will present a few musically interesting examples from this set of effects.

Noise in Analysis Step

When processing an input sinusoid at frequency f with added to the outer triangular regions of the matrix only – resulting in² $W_k = \beta W_k^r$ – a nearly white noise component is added to the entire signal, with a slight increase in energy at higher frequencies. In contrast, when noise is added to the matrix diagonal and $W_k = \alpha W_k^d$, a band of noise is introduced whose energy is concentrated around frequency f which falls off at higher frequencies. This difference is illustrated in figure 5-1.

²I use index k here to underscore the fact that the state recursion in the analysis step is a function of frequency rather than time.

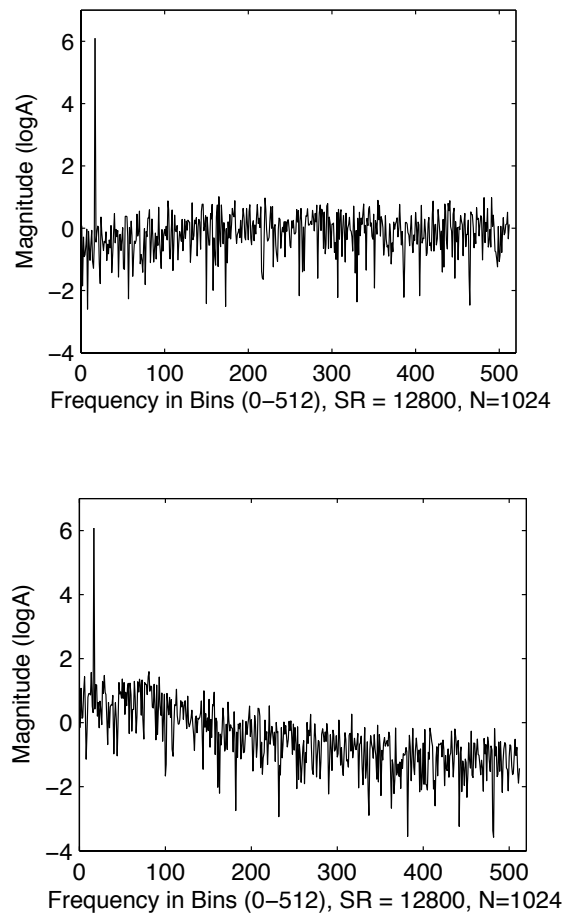


Fig. 5.1 Log FFT plot for stochastic state-space processed sinusoid. Gaussian noise added to analysis dynamics matrix in outer triangular portion (a) and to diagonal (b).

In order to synthesize more interesting grain-type qualities, noise values can be propagated through various parts of the matrix to create a time-varying effect. For example, it may be moved across a given column or row of the matrix. When noise was propagated down a single row or column for an input sinusoid, a beating noise with several small peaks was introduced. The rate of the beating can be controlled by the speed at which this scalar noise value is sent through the given row/column. This modulating behavior can be seen in figure 5-2(b).

While this time-varying single perturbation produces a more musical result, it is not physically accurate: the noise value causes an interaction between the frequency located at the given column where the propagation occurs and each other frequency bin at the instant that the noise is swept past it in the matrix. However, this is not truly a coupling between frequencies as it did not occur in both directions. Thus, to make the effect more physical, the same noise value must be passed down both column and row, so that at time t if the input noise value is added to matrix value $A(i, j)$, it is further added to the value at $A(j, i)$. This coupled time-varying effect creates a more sophisticated stochastic component to the sound - one that possesses “more texture” and somewhat resembles the sound of fire. For input sinusoid with frequency f , this effect is most prominent when it occurs at the column/row associated with the highest-energy frequency bin, namely $k = \frac{N*f}{F_s}$ where F_s is sampling frequency and N is the size of the input signal block. The difference between the “abstract” and physical roughness effects can be seen in figure 5-2. Beyond having more high frequency content, the coupled example of 5-2(a) possesses a spectral fine-structure that is present throughout the spectrum and which likely contributes to the overall textural quality.

Noise in Synthesis Step

It is important to remember that the state vector is not the same between analysis and synthesis steps. During analysis, the state is initialized with real and imaginary components of an input signal block of size N . At the synthesis stage, the state vector is initialized with the real and imaginary spectral values that are generated by the first $N/2$ iterations (assuming a real-valued input) of the recursive analysis process. Therefore, the addition of noise to the state matrix affects the dynamics of either the complex modulation or demodulation process associated with the DFT/iDFT and there is no

reason to assume that the addition of the Gaussian noise matrix would produce the same sonic result at each stage. Indeed, the addition of noise values in the synthesis state matrix — of coupled noise propagated down a column/row pair — produces a strong modulation effect not present in the previous examples. While the other examples produce fluctuations and a beating effect, this synthesis-step noise results in a quasi-periodic emergence of strong spectral peaks which modulate throughout the spectrum. For more discussion on this set of stochastic effects, the reader is directed to [201].

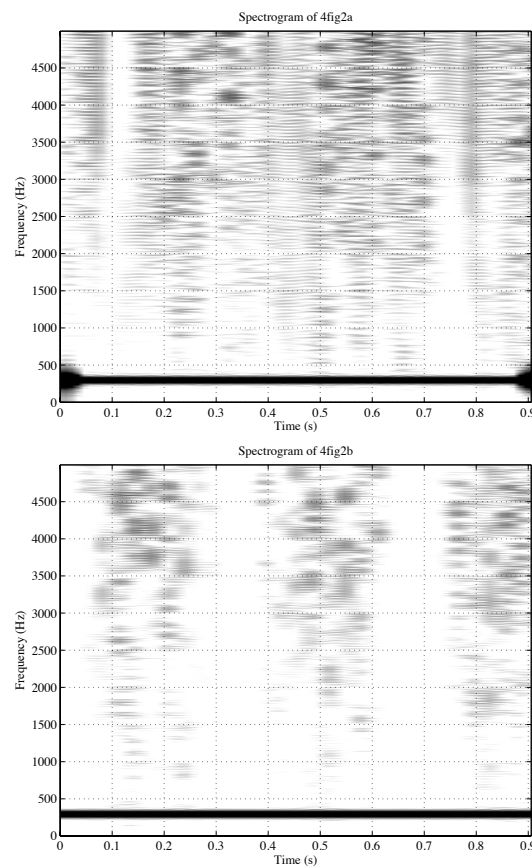


Fig. 5.2 Spectrogram of sinusoid affected by noise propagated through analysis state matrix at one column alone (a) and coupled between a column/row pair (b). Sampling frequency = 11.025 kHz.

5.2.4 Beyond SSSPV: From Effect to Transformation

The SSSPV can be considered as a phase vocoder in which a stochastic element has been built into the representation via a state-space framework. The focus with this implementation is on the fact that through the embedding of noise within the system representation itself (rather than as input to the state or observation equations) a sound can be re-synthesized with an added grain or roughness quality. In particular, musically interesting distortions can be introduced and controlled by altering the Gaussian noise matrix over time – including the position of noise values in the matrix, the α and β parameters and the mean and variance of the noise process.

Now, while this approach can provide a family of interesting effects, there are some limitations with this implementation. Primarily, this state-space representation diverges from a classic phase vocoder in that it is actually a recursive implementation of the DFT. In other words, it is only defined for a single analysis window, and the input signal must be re-introduced at the boundary between windows. This is not a problem for this set of noise-based effects, but it inhibits the implementation of deeper transformations based on time-stretching and pitch shifting across large time scales. The use of the state-space representation is interesting with the SSSPV in that it gives access to the internal dynamics of the analysis/synthesis process for control and processing as demonstrated by the example effects. However a deeper reason for using the state-space approach is that it allows for the use of tools such as the Kalman filter for parameter estimation. To this end, I returned to the “normal” state equations of 5.1 and 5.2, and using the same underlying spectral model as in SSSPV I have designed an STFT-based recursive implementation that functions as a “true” phase vocoder. This implementation is thus capable of deeper transformations beyond the idiosyncratic SSSPV effects, and as such it is what I have built the larger estimation and control framework upon.

5.3 Introduction of Recursive, Infinite Length Windows

Creating a state-space recursion based solely on equation 5.3 defines a DFT that implicitly assumes a periodic input signal whose fundamental period is defined by the N input samples. It is further an implicit rectangular windowing, and by re-introducing the input each N samples (as in SSSPV) there is no overlap in the analysis. In order to

extend this, I will illustrate how the STFT can also be expressed with a one-step recursion. In the process, this will introduce an infinite length window into the definition, which provides certain beneficial time-frequency properties. In particular I use a single-sided exponential window [202] $h[m]$ defined as

$$h[m] = e^{\lambda m} u_+[-m] \quad (5.13)$$

where $u_+[m]$ is the unit step function. The use of this window allows one to accurately detect the beginning of signals (due to the discontinuous front edge of the window) as well as to tune the influence of the past samples through changing the decay value λ . Following this, the STFT

$$X_{n,k} = \sum_{m=-\infty}^{\infty} x[m] h[m-n] e^{-j\omega_k m} \quad (5.14)$$

for input signal $x[m]$ becomes

$$X_{n,k} = \sum_{m=-\infty}^n x[m] e^{\lambda(m-n)} e^{-j\omega_k m} \quad (5.15)$$

The one-step recursion can be derived by looking at the value

$$X_{n+1,k} = \sum_{m=-\infty}^{n+1} x[m] e^{\lambda(m-(n+1))} e^{-j\omega_k m} \quad (5.16)$$

which expands to

$$X_{n+1,k} = \sum_{m=-\infty}^n x[m] [e^{\lambda(m-n)} e^{-\lambda}] e^{-j\omega_k m} + x[n+1] e^{-j\omega_k (n+1)} \quad (5.17)$$

and thus

$$X_{n+1,k} = e^{-\lambda} X_{n,k} + x[n+1] e^{-j\omega_k (n+1)} \quad (5.18)$$

Therefore the STFT at any given time – when using this window function – is a product of the exponentially-damped previous time-step and the Fourier Transform component from the current time-series value. In light of the state-space representation developed in the previous section, we can now re-write the state equation as

$$s[n + 1] = \hat{A}s[n] + D_n u[n + 1] + w[n] \quad (5.19)$$

where

$$\hat{A} = e^{-\lambda} A \quad (5.20)$$

$$D_n = (\bar{A})^n \quad (5.21)$$

with \bar{A} defined as the block diagonal from A extracted and collapsed in order to form a $2N \times 2$ matrix. Finally $u[n] = [x[n + 1] \ 0]^T$ is the time series at time $n + 1$ and functions as the “input vector” in terms of the state-space formalism.

Now, this implementation is an adaptive version of the STFT in which the infinite-length exponential window acts as a forgetting factor. This smoothing extends the previous recursive DFT into an STFT through a consideration of all past sample values. I will refer to this implementation as the recursive exponential STFT – or RESTFT – from this point forward. Note that this differs from a standard STFT in that the time-frequency resolution³ is not directly tied to N but rather only to frequency resolution [203]. Further, the bandwidth of the analysis window is a factor both of the overlap as well as the damping coefficient λ . These can therefore be chosen based on the transformation and synthesis requirements, and makes this a parametric representation.

Before moving onto example applications, however, there are further augmentations to be discussed that are geared towards refining the adaptive control of time-varying transformations, with particular focus on the noise part of the sound. To this end, I have built an adaptive framework on top of RESTFT using the Kalman filter, in order to parameterize the relative sine/noise quality of both attack and sustain such that one may re-estimate these trajectories under time-stretching and other time-varying transformations. Several algorithms were created in order to reach this goal, beginning with the creation of the initial Kalman framework.

5.4 Kalman Filter-Based Phase Vocoder

Two of the primary reasons for expressing the phase vocoder in state-space form were to access noise-based sound dynamics and because it allows for a hybrid between signal and

³Strictly speaking this resolution is tied to the window size, and here I am implicitly using a size N window and FFT.

physical models; while my use to this point has been towards abstract signal models, such as SSSPV and RESTFT, this will change when I introduce control structures inspired by physically-based interaction.

Taking full advantage of these qualities – including deeper sound transformations and control parameter estimation – requires the implementation of a Kalman filter (KF). The KF is an algorithm developed in order to optimally estimate the state of a linear dynamical system perturbed by Gaussian noise [204][205]. Many variants have been established in order to extend the tracking, estimation and prediction properties of this filter to nonlinear and non-Gaussian systems [206]. As I deal here with the STFT – which is a linear transform – then the standard form of the KF suffices in order to model the sound process. The result is a model that allows one to estimate the magnitude and phase of each bin for every time step, and to extract *both the state and observation noise* for individual control and processing. As is noted in [195], in which the author creates a similar Kalman-based additive model specialized to damped percussive sounds, the state and observation noise sources relate to the transient noise and sustain noise (respectively) of the underlying sound signal. In this research I focus on these noise values, in order to define novel transformations for them that adapt to varying time-stretching, such as one finds in the scrubbing of sound sources as was discussed in the control structure examples of the previous chapter.

Now, the Kalman filter is a recursive process that consists of a time update and a measurement update. The former first predicts future values of the state, and the latter then modifies this in order to provide an adjusted estimate for the current time step. More precisely, the *a priori* state estimate \hat{s}_n^- and the state covariance estimate P_n^- are conditioned on all prior values as

$$\hat{s}_n^- = \hat{A}\hat{s}_{n-1} + D_{n-1}u_{n-1} \quad (5.22)$$

$$P_n^- = \hat{A}P_{n-1}\hat{A}^T + Q \quad (5.23)$$

These are the *time update* equations. Recall that the underlying system is perturbed by process and observation noise w_n and v_n , which are zero-mean Gaussian white noise processes with variance Q and R respectively. The values \hat{s}_n and P_n are the *a posteriori* state and state covariance values, acquired from the following *measurement update* equations

$$K_n = P_n^- B^T (B P_n^- B^T + R)^{-1} \quad (5.24)$$

$$\hat{s}_n = \hat{s}_n^- + K_n(x_n - B\hat{s}_n^-) \quad (5.25)$$

$$P_n = (I - K_n B) P_n^- \quad (5.26)$$

A heuristic understanding of these equations would be to first consider the a priori state and covariance estimates as predictions based on all past values, with no noise disturbance; the updated estimates are defined by adjusting this initial estimate based on the influence of the *innovations sequence* $\epsilon = (x_n - B\hat{s}_n^-)$, which represents the difference between the observed value and the ideal noise-free observation. The degree of influence from the innovation is “tuned” by the so-called Kalman gain K_n .

It is suggested in [195] to extract the residual from the state-space model in order to drive the re-synthesis. I follow this general approach, though my extraction method differs as does the underlying representation (exponential STFT vs. damped sinusoidal model).

Thus the estimated process and observation noise sources are defined as

$$\hat{w}_n = \hat{s}_n - \hat{A}\hat{s}_{n-1} \quad (5.27)$$

$$\hat{v}_n = x_n - B\hat{s}_n \quad (5.28)$$

Thus we have an estimate of the state \hat{s}_n – providing instantaneous estimates of magnitude and phase values – which gives rise to estimates of the excitation noise \hat{w}_n and the additive output noise \hat{v}_n .

Initial Conditions and Model Parameters

The state-space equations and the derived Kalman time/measurement update equations are nearly sufficient for analysis of a given input audio signal. However, in order to provide the first a priori estimates, the model must be given initial conditions for the state and state covariance matrix. In the absence of any specific knowledge about the signal, we may safely define these such that

$$\hat{s}_0 = \mathbf{0}_{2Nx1} \quad (5.29)$$

$$P_0 = \sigma^2 \mathbf{I}_{2Nx2N} \quad (5.30)$$

The value $\sigma^2 \ll 1$ is the state covariance – it reflects the level of uncertainty in our initial state and governs the rate of convergence of the filter. If we were absolutely certain that the initial state was identically zero, we could let $\sigma^2 = 0$ as well. Since we do not in general know the state of the input at time zero, I therefore include σ^2 as a model parameter that must be tuned. The other parameters are the noise covariance for the state and process noise values, the damping coefficient λ for matrix \hat{A} and the state size N . The three noise covariance values influence the ability of the tracking of spectral data, changing the relative amount of energy that will be present in the residual signals \hat{w}_n and \hat{v}_n as compared to the state estimate \hat{s}_n . Meanwhile, the values λ and N together define the time/frequency resolution of the analysis as well as the computational power required (in the case of N). Therefore, while the kalman-based RESTFT will produce a *perfect reconstruction* of an input signal if there are no modifications to the spectral or residual data, this is a parametric approach in which the relative contribution of each time-series \hat{s}_n , \hat{w}_n and \hat{v}_n can be tuned by altering these model parameters.

Discussion

This KF-based RESTFT framework becomes musically interesting when used to re-estimate state and noise values after applying time/frequency transformations such as pitch shifting or time-stretching, or when one controls the relative level of each as well as the variance of the noise processes over time.

By adding a layer of control for such adaptive signal behavior, I am providing a bottom-up definition of mapping in which the general dynamic rules are described, rather than a static parametric mapping function. In this way, control and mapping are embedded in the signal definition, and happen *at the level of the sound transformation*. Before defining the layer for adaptive control and before suggesting possibilities for sound transformation, there are two more key extensions to the sound model that must be presented. As noted, while other Kalman-based spectral analysis/synthesis frameworks have begun with smaller state sizes (e.g. Additive partials), I begin from the entirety of

all phase vocoder spectral bins, as defined by the RESTFT. However the state size is a strong limiting factor for computation of the analysis, and so I have adapted an algorithm from the control theory literature in order to make high-quality transformations more feasible by allowing for more reasonable window (i.e. state matrix) sizes, and therefore higher frequency resolution.

5.5 Chandrasekhar Implementation

The computation of the state estimate in equation 5.25 is the essential step in which the Kalman gain – adapted by the updated state covariance – determines the amount of influence that the innovation sequence (i.e. the “novel” information) will have on the newly predicted state. This, as well as the Kalman gain of equation 5.24 may be factored differently, so that

$$\hat{K}_n = \hat{A}P_n^- B^T \quad (5.31)$$

$$\hat{R}_n = (BP_n^- B^T + R) \quad (5.32)$$

$$\hat{s}_n = \hat{s}_n^- + \hat{K}_n(\hat{R}_n)^{-1}\epsilon \quad (5.33)$$

The quantity \hat{R}_n is the variance matrix of the innovations sequence ϵ . Using this factorization, it was shown in [207] that the matrices \hat{K}_n and \hat{R}_n may be computed by utilizing intermediate operations that act on matrices having a substantially smaller rank – and so possibly much less computation time. These added operations are referred to as Chandrasekhar-type recursions. The RESTFT fits this case and so was re-written in this form. In particular, intermediate matrices Y_n and M_n arise which are defined as

$$Y_n = [\hat{A} - \hat{K}_n(\hat{R}_n)^{-1}B]Y_{n-1} \quad (5.34)$$

$$M_{n+1} = M_n + M_n Y_n^T B^T (\hat{R}_n)^{-1} B Y_n M_n \quad (5.35)$$

which then are used to redefine the Kalman equations as a recursion with

$$\hat{K}_{n+1} = \hat{K}_n + \hat{A}Y_nM_nY_n^TB^T \quad (5.36)$$

$$\hat{R}_{n+1} = \hat{R}_n + BY_nM_nY_n^TB^Tt \quad (5.37)$$

These new equations for Y_n , M_n , \hat{K}_n , and \hat{R}_n are of size $N \times \alpha$, $\alpha \times \alpha$, $N \times p$ and $p \times p$ respectively, where p is the output dimension and α is the rank of a matrix equation defined by the initial state covariance and Kalman gain matrices [207]. As the system observation is of the scalar audio value in our case, the matrix \hat{K}_n reduces to a column vector while \hat{R}_n is simply a scalar. Thus the inversion of this latter value in equation 5.35 can be replaced by a simple multiplicative inverse, further speeding up the calculations which is particularly time saving in the Matlab environment. Finally it can be shown – in this particular case of scalar observation – that the value $\alpha = 1$ as well, and so the overall speed for the calculation of the state estimate and related values is improved dramatically. Further, note that certain quantities such as $M_nY_n^TB^T$ are used several times, so that they can be stored in memory and only need be computed once.

Initial Conditions

The initial values for the matrices \hat{K}_n and \hat{R}_n are simply derived as

$$R_0 = R + BP_0B^T \quad (5.38)$$

$$K_0 = \hat{A}P_0B^T \quad (5.39)$$

However those for Y_n and M_n require further derivation. It is noted in [207] that in particular cases, the initial values Y_0 and M_0 assume a simple form. When the value \hat{s}_0 is known with a high degree of certainty – and one may safely assume that $\sigma^2 = 0$ – then the values reduce to

$$Y_0 = \mathbf{I}_{2N \times 2N} \quad (5.40)$$

$$M_0 = Q \quad (5.41)$$

At the same time, it is proven that if the state matrix \hat{A} is a stability matrix – meaning that its eigenvalues lie within the unit circle and the state sequence converges to a stationary process – then the initial conditions take another form. My sound model is a limit case of this, and so I include this as a possible scenario in the algorithm. In this case, the initial values become

$$Y_0 = \hat{A}P_nB \quad (5.42)$$

$$M_0 = (r + BP_nB^T)^{-1} \quad (5.43)$$

As these two scenarios are likely but not guaranteed outcomes for any given audio signal, I include them both and leave this as an option when conducting a given analysis. That said, the latter case has worked on all signals that have been processed thus far.

Now, this entire derivation has been based on the assumption that we have a stationary process, with \hat{A} constant over time. This is the case with our implementation based on the STFT. However, if we wish to extend this into an additive implementation then the Chandrasekhar recursions may not be used. Fortunately this is not a concern, as the need for this fast algorithm arises precisely in the case of having large state sizes brought on by the STFT. Therefore, we may switch to a different algorithm – and another model altogether – when building an additive version of this algorithm. In this way, fast implementations are used when they are needed, while the standard form still works in the additive case where state sizes are more manageable (e.g. N less than 100). I'll now present the additive model, and how it integrates with the aforementioned work to define the overall system architecture.

5.6 Additive Layer and Higher-Level Architecture

The advantages gained by having an adaptive framework based on the Kalman filter can be extended from the phase vocoder to a higher-level, additive framework as well. In this implementation, rather than tracking the state of N evenly-spaced frequency bins, I model L sinusoidal partials ($L \ll N$ in general) so that

$$x_n = \sum_{k=1}^L [a_{k,n} \cos(\sum_{r=1}^n \omega_{k,r} + \phi_k)] \quad (5.44)$$

In terms of the underlying state-space representation, rather than the time-invariant state matrix defined by equations 5.5 and 5.20, the system here is non-stationary with time-varying state matrix A_n where the frequency values to each block diagonal component are given from the L frequencies $\{\omega_{0,n}, \dots, \omega_{L-1,n}\}$. Therefore the matrix is of size $2L \times 2L$, where L is generally below 100. Note that the time resolution is therefore not bound implicitly to N samples in the same way as with the phase vocoder. Rather, the time-varying matrix A_n can be updated as much (i.e. each sample) or as little as desired. The classic method of smoothly varying between amplitude, frequency and phase values is to interpolate between the first linearly and the latter two with a cubic polynomial as discussed in [60]. This is achieved in this implementation by interpolating a set of values a priori and having the corresponding state matrices pre-computed in order to speed up processing.

There are two things that are important to note in adapting this implementation from an STFT to an additive representation. The first is that the system is time-varying, and so the Chandrasekhar-type recursions can not be used as with the RESTFT. This is not an issue here as the state matrix – the primary limiting factor in computation – is so much smaller. At the same time, note that the infinite-length exponential windows are still used in this analysis, preserving the time-frequency properties that exist with the RESTFT. The second point is that this analysis stage does not do any partial tracking itself. Rather, peaks of interest are given to the analysis system at any time-stamp deemed appropriate. Therefore, another analysis method may be used in order to provide high-quality partial tracking (e.g. [68]) and this information can be fed to the Kalman-based system in order to provide a complete set of data on the suggested location of each partial. The additive Kalman layer, then, computes a *re-analysis* which extracts specific magnitude and phase values at each partial while producing the residual in the process. As with the KF-RESTFT, the sound's transient noise is captured by the process \hat{w}_n while the additive noise is represented by \hat{v}_n . Therefore, rather than providing an improved partial-tracking algorithm, this additive implementation is geared towards providing a more refined analysis of the sine/noise energy decomposition. This serves the ultimate goal of improved noise-based transformations, in this case relative to a parametric additive framework.

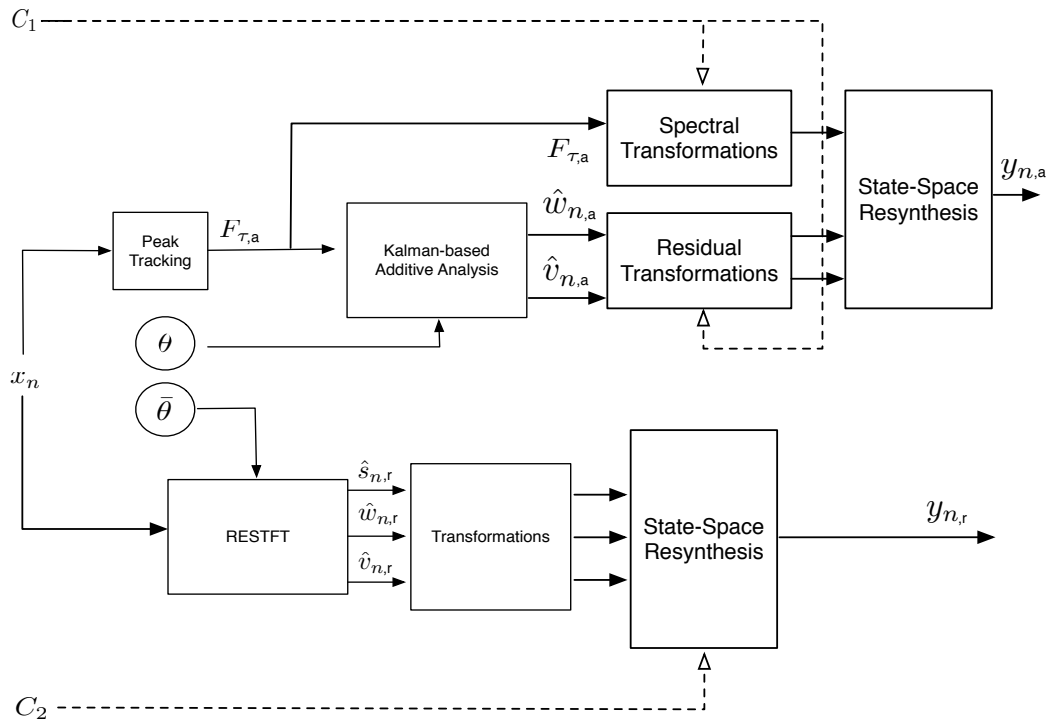


Fig. 5.3 Model Architecture for Kalman-based Additive and Recursive Exponential STFT Models, shown here in parallel. Amplitude or phase values may be extracted for processing during the respective transformation stages.

The overall sound processing architecture is depicted in figure 5.3, which shows the two analysis/synthesis schemes – KF-RESTFT and the Kalman-additive model – in parallel. It is not necessarily suggested that both schemes must be used in tandem, but rather this illustrates the manner in which they relate to one another in terms of model, input and control parameters. In each case, the input value x_n is sent into the analysis. In terms of the additive model, it is assumed that some other method is used for determining the list of peaks $F_\tau = \{\omega_{0,\tau}, \dots, \omega_{L-1,\tau}\}$ in the “peak tracking” step. Note that the peak tracking follows a slower time evolution, so that τ grows much more slowly than n . If we use a block-based analysis method, it may be that $\tau = N$, the block size for the RESTFT technique. Again, after this extraction the frequency trajectories must be interpolated so that they are known for each n , which is done prior to feeding them into the Kalman-based analysis. At this point, both for the additive and phase vocoder implementations, a set of model parameters must be fed to each. With the phase vocoder, these parameters are

$$\theta = \{\lambda, N, r, \sigma^2, q\} \quad (5.45)$$

where again the first two influence the time-frequency resolution and the latter three are the noise covariance values for the state, process and observation. For the additive model, the parameter set $\bar{\theta}$ varies slightly

$$\bar{\theta} = \{\lambda, L, H, r, \sigma^2, q\} \quad (5.46)$$

where L is the number of partials, and H is the hop size of the peak-tracking analysis. In general, these model parameters are not controlled online (and so have no inputs in this diagram), but they may potentially be changed in an adaptive fashion [208]. Once the Kalman-based additive or phase vocoder analyses are complete, they produce state estimates \hat{s}_n as well as residuals \hat{w}_n and \hat{v}_n . These are the values that are more likely controlled, or otherwise transformed before re-synthesis. Thus, any possible control input would map directly into these values. In the case of the additive model, the frequency values F_τ may be transformed separately, so that partials may be independently processed. After time/frequency transformations are applied to the spectral frames of the state or to the residual values, the entire process is re-synthesized using the state-space model and the equations given by 5.27 and 5.28.

The overall system presented to this point – as represented by this figure – has focused on the sound model and transformation possibilities. Control is simply shown as a possible input to affect the sound processes. As discussed in the introduction, there is no control embedded in the model except for that suggested by the sine/noise parameterization. That said, there are certain transformation possibilities that are made readily available by virtue of the model’s definition. Therefore I will present these different classes of transformation possibilities – focusing on processing the two noise time-series – which are themselves adaptive. This will establish what is possible in terms of sound processing and control, before closing the chapter with extensions to the system that embed a layer of adaptive control at the level of the sound model.

5.7 Sound Transformations

Depending on the type of input signal, these two models are fairly robust to changes made simply to the noise covariances. However these values do exhibit an influence when the signal in question is time stretched, and so become indirect tuning parameters in adding textural qualities to the signal. It is shown in [195] that in fact it is only the *ratio* of the state/observation noise covariances that affect the resultant sound. This ratio changes the relative influence of the state or observation noise process over the output, and so this must be tuned in conjunction with the transformations – whether applied primarily to the input (state) or output (observation) residual. My experience thus far has been that processing focused on the state noise vector present the most interesting texture-based transformations, as each value can be processed separately. At the same time, time-varying modifications of the partials (in the case of the additive model) can be achieved without time-stretching by focusing on the frequency trajectories and even the window decay value λ . Therefore, the transformations made possible with this approach can be broken down into three broad categories defined by

- Textural effects achieved through a combination of processing the two noise time-series and time-stretching.
- Spectral effects achieved through modifying the state matrix.
- Cross-synthesis by combining the noise time-series and state matrix values from different input signals.

As noted previously, the underlying sound model introduced in this chapter is a combination of a source-filter and an additive/phase vocoder approach. This fact is embedded in the above classification of transformations: the first can be considered a source-filter type of processing where the noise excitation is transformed and conditioned by the state matrix filter, while the second is primarily focused on additive transformations such as transposing all or certain partials over time by altering the decay parameter for the state matrix. This second class can also exhibit source-filter type processing if blocks of the state matrix itself are processed (rather than its frequency/decay parameters), thereby changing the filter response of the model. The third class combines both of these approaches: the excitation of one sound may be altered and crossed with another sound whose spectral content has likewise been processed by altering its state matrix parameters.

5.7.1 Textural Effects

For the first class of effects, while it is the textural quality of the signal that is being influenced from a perceptual point of view, the processing that these transformations have in common is time-stretching. In the classical algorithm employed for the phase vocoder, the hop size H is changed between analysis and synthesis by the desired time-stretch factor [42]. During resynthesis, for stretch factor α and input/output spectra X and Y the magnitudes have the relationship

$$|X_{k,rH_a}| = |Y_{k,rH_s}| \quad (5.47)$$

for each frequency bin k and hop increment r , where H_a and $H_s = \alpha H_a$ are analysis and synthesis hop sizes respectively. At the same time, one needs to change the phase between respective bins such that the instantaneous frequency (defined as the derivative of the phase) is preserved. In particular we must have

$$\Delta\phi_{k,rH_a}^x = \frac{H_s}{H_a} \Delta\phi_{k,rH_s}^y. \quad (5.48)$$

This is achieved by taking the phase increment between frames of the input, and adding this – multiplied by the proper stretch factor – to the phase-unwrapped version of the output phase for the target frame. Improvements can be had by “locking” the phases of

adjacent bins so that they are ideally out of phase with one another [54]. Finally, the overlap and add of the Fourier spectral frames provides an implicit interpolation of the intermediate phase information [51]. One can apply a similar process to the RESTFT by extracting the magnitude M_k and phase ϕ_k information for bin k from the state vector as

$$M_{k,n} = \sqrt{s_{k,n,r}^2 + s_{k,n,i}^2} \quad (5.49)$$

$$\phi_{k,n} = \tan^{-1}\left(\frac{s_{k,n,i}}{s_{k,n,r}}\right) \quad (5.50)$$

The steps to ensure continuity as described above are applied to this data, and the values converted back into a new state vector \hat{s}_n . One difference is that in the classic phase vocoder situation the phase must be interpolated implicitly through the overlapping and adding of IFFT data. In this model, I compute the time series itself at each *sample* rather than at hop increments, and so the information arises explicitly on the state vector.

Interpolation of this information occurs only after time-stretching, at which point the phase is calculated modulo 2π at each step to ensure proper evolution.⁴

Taking this approach, however, yields precisely the same sort of processing as the normal phase vocoder time-stretching algorithm. In order to achieve the sort of textural effects that influence a sound's grain quality, it is more effective to act on the *state matrix and noise residuals* independently of one another. Using the additive model in particular allows for interesting control of these two parameter sets. In doing so, one can change the time-varying behavior of the spectral model, the state noise and the observation noise separately. In terms of the state model, one can use a traditional re-sampling interpolation method in order to under or over-sample each frequency trajectory from $\{\omega_{1,n}, \dots, \omega_{L,n}\}$ by the appropriate stretch factor α . This produces the new block diagonal matrix A_α , and we may redefine the state matrix from equation 5.20 as

$$\hat{A}_\alpha = e^{-\frac{\lambda}{\alpha}} A_\alpha. \quad (5.51)$$

This ensures that the current time-altered signal value takes into account previous spectral data at the properly modified rate. However, one may treat the decay value itself as another control parameter in order to influence the temporal evolution of the

⁴Because of this different in magnitude/phase calculation, I will adopt M for the RESTFT (including additive-layer) magnitude and reserve $|X|$ for the standard phase vocoder.

spectrum, as I will illustrate in the next section.

Meanwhile, the noise values \hat{w} and \hat{v} are by definition outside of the model framework.

Therefore they cannot be subject to the above sort of interpolation of spectral parameters to achieve time scaling. In any case it is the time-domain nature of these signals – such as during transient regions – that they are of most interest in contributing the textural nature of the signal. For example, figures 5.4 through 5.6 depict the first 0.6 seconds of a piano note as well as the observation noise and the process noise values for the first seven frequency components (the maximum partials are 14 in this particular analysis). The majority of the state noise processes are focused in the transient region and are a product of the temporal behavior of the associated partials, while others have less energy and are more distributed in time. As such different processing can be applied to each state noise channel in order to influence the interaction between partials and thus the sound’s overall spectro-temporal evolution. In particular texturally-oriented sound transformations arise when this interaction causes modulations that involve the noise-part of the signal, to which end it is interesting to apply slightly different time stretching to each of these state residual processes.

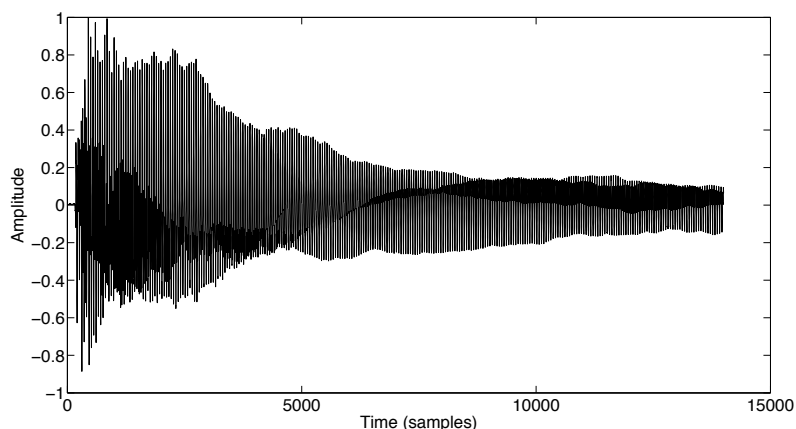


Fig. 5.4 Example piano note for KF-based analysis ($f_s = 22050$ Hz)

In order to create such transformations I utilize a set of classic methods from the family of Synchronized Overlap and Add (SOLA) algorithms, with the first such technique appearing in [209]. Several different variants exist – which each color the residual in different ways – and so I have employed several that can be used in parallel. While the

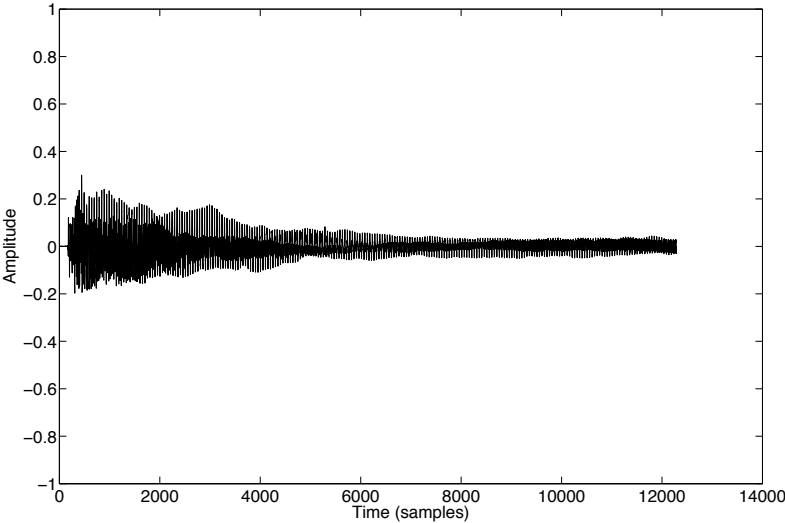


Fig. 5.5 Observation noise after analysis

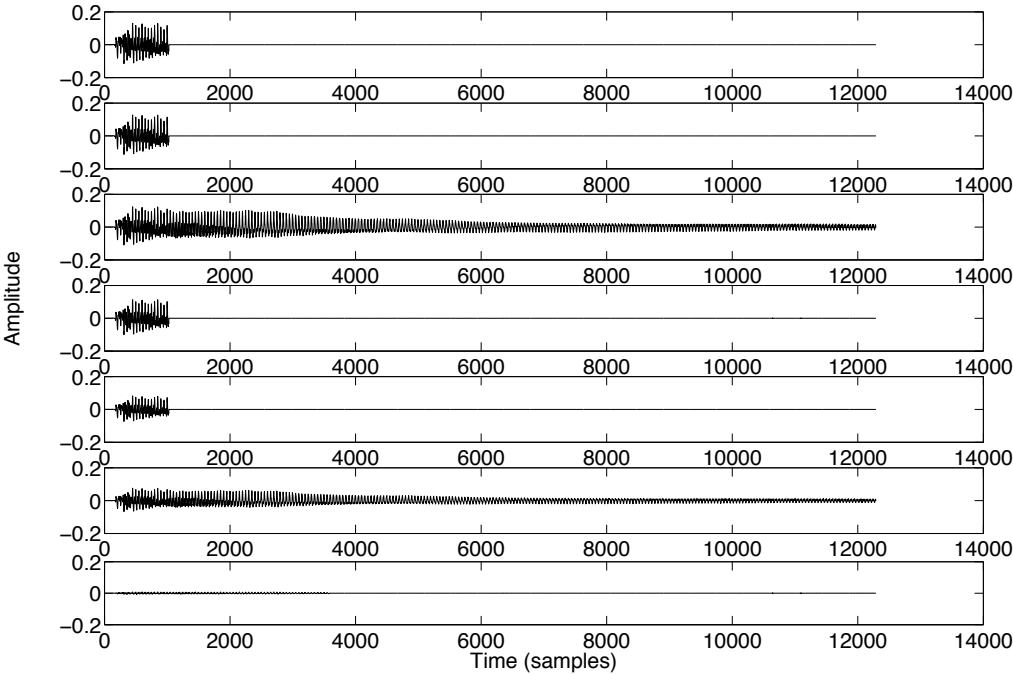


Fig. 5.6 Process noise related to first several partials after analysis

sound differs between techniques the general principle is the same: each is a block-based technique in which an input sound is segmented in blocks of length N having hop size H_a , and this hop is shifted by a time-stretch factor α so that output hop $H_s = \alpha H_a$. In this sense, the technique is similar to the phase vocoder. However, rather than ensuring the proper phase evolution this approach enforces temporal similarity by looking at the maximum value of the cross-correlation calculated within the overlapping-regions of the two signal-blocks. This provides a lag value r_{max} which can be used to adjust the hop size – either of S_s as in the original SOLA technique, or of S_a as with the more computationally-efficient SOLAFS approach [210] and the robust WSOLA [211] method. The latter two are the implementations that I have primarily built upon. Essentially, this set of techniques looks to define “periods” of the input signal, and to repeat or remove these as needed to achieve a given timeframe. In the case of noise signals, proper tuning of window parameters results in repeated material that is quite close to the notion of texture discussed previously – and that put forth by [91][92][93] – arising from signals that produce noisy yet quasi-periodic deviations within a globally stationary environment. For example, consider again the piano attack example, whose spectrogram is displayed in figure 5.7. There are well-defined partials, a quickly decaying noise region and a small transient region around 0.25 seconds. One approach to texture-oriented transformation⁵ is to apply the SOLAFs technique with small windows and overlap (256 and 128 samples) such that the stretch factor is a product of the relative transient/sustained nature of the process noise. The spectrogram for the resulting sound is shown in figure 5.8. Note that the noise energy centered around each partial has now been stretched fairly uniformly in time, while there are several more transient regions that are aligned in time. The increased noise around the partials makes sense as the process noise is the modulation of the partial’s temporal behavior, and so stretching this should prolong the narrowband noise modulation around each. As a different approach to the same general textural transformation type – resulting in different overall effect – the window size for the time-stretching can be made partial dependent. An interesting effect results if the stretch factor is slightly randomized with the general constraint that longer stretch times occur with process noise values associated to lower partials and the opposite for higher partials.

⁵To be clear: just as controlling spectral envelope is one of many ways to influence timbre, this type of noise-based modulation is one of a number of ways to influence the perceptually-related attribute of texture.

An example of this transformation is depicted in figure 5.9. Note that the higher partials were stretched much less than lower ones. Further, rather than temporally aligned transient regions, these are now offset and a modulation of the noise can be seen, particularly near the beginning of the first seven partials.

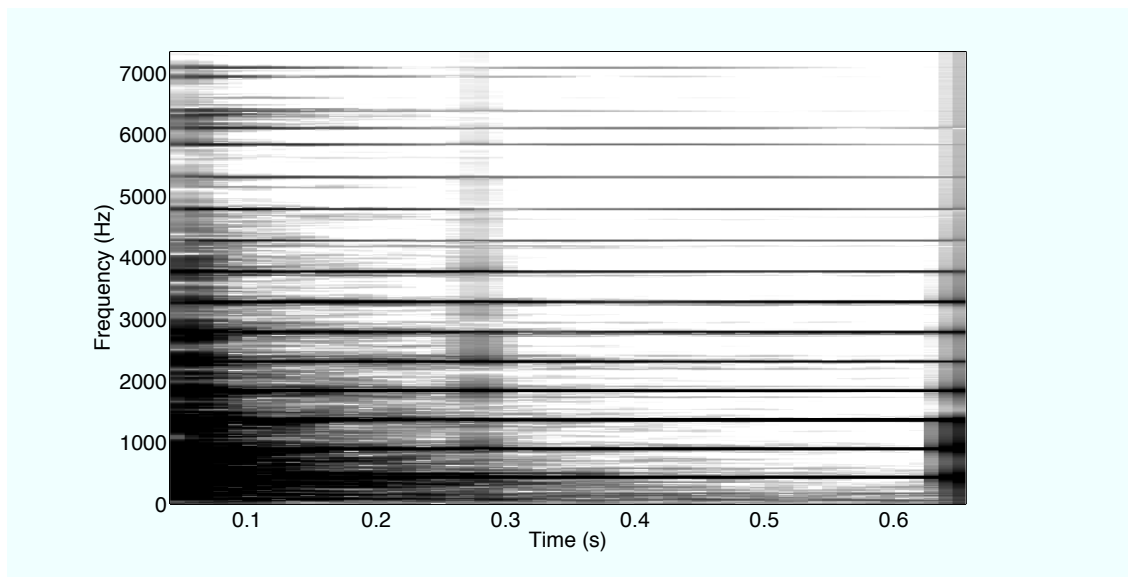


Fig. 5.7 Spectrogram of input piano attack

Various other grain-focused sonic gestural contours can be produced through such region-dependent processing. In these two examples I focused on processing that is based on the temporal support of the state residual and the given partial frequency. If one wanted to preserve the acoustic quality of an attack region, then all of the transient-based process noise values such as in figure 5.6 could be left unaltered while time-varying stretching and modulation could be applied to the sustain-oriented residuals. The SOLA-based parameters (window size, α , maximum shift size, overlap size, window type) themselves influence the resultant signal, as can be seen by comparing the modulatory nature of figure 5.9 to the lack of this in figure 5.8. The texture of the signal can be further altered by random modulations added to the state matrix, where jitter and shimmer type effects can be produced by adding noise to the matrix, as was discussed in section 5.2.2. In short, confluence of spectral-domain modifications to the state matrix and time-domain modifications to the residual signals yields a variety of sonic textures. The entire process for producing transformations from this class using this model consists

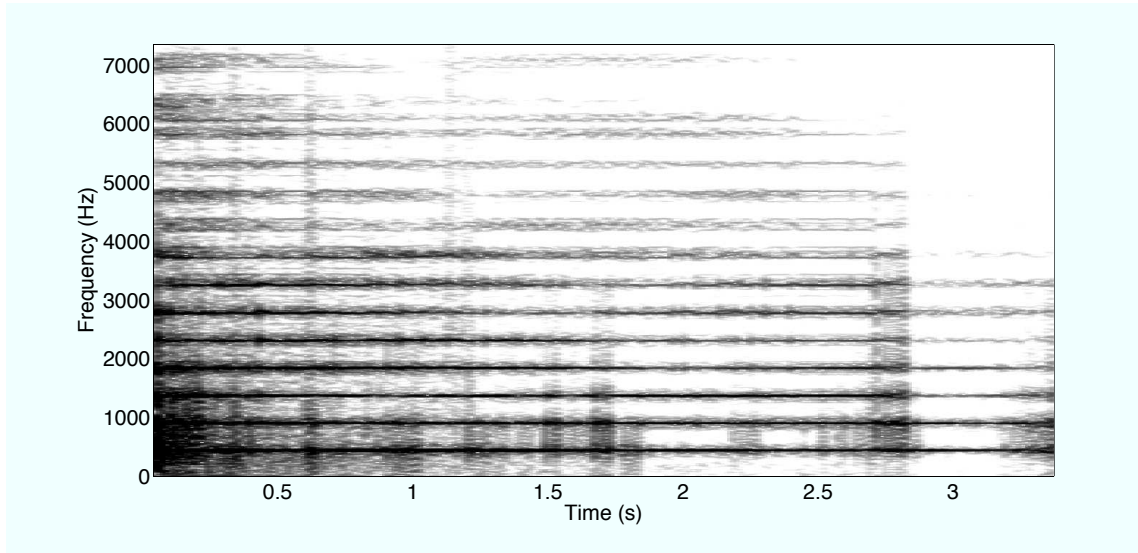


Fig. 5.8 Spectrogram of piano attack after SOLAF time-stretching of each state noise process, with stretch factor proportional to temporal support. Window size is 256 with overlap of 128.

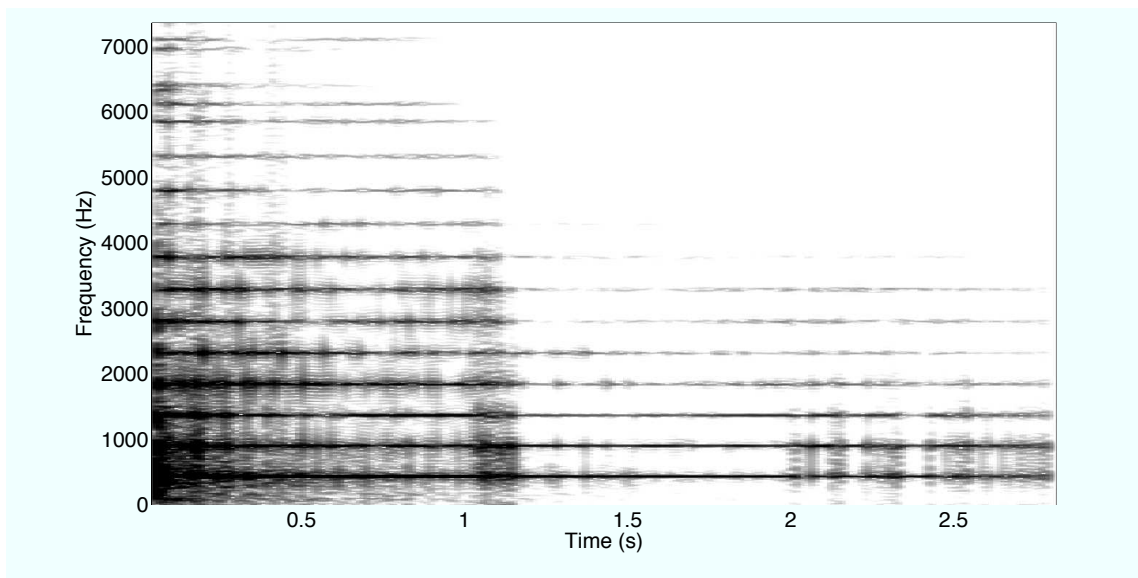


Fig. 5.9 Spectrogram of piano attack after SOLAF time-stretching of each state noise process, with stretch factor randomized with mean factor as well as window size proportional to frequency of related partial. Overlap is 50% in each case.

of

1. Resampling the extracted frequency trajectories.
2. Re-defining the state matrix by the addition of a single-sided window.
3. Optionally adding jitter and shimmer by adding Gaussian noise to the state matrix.
4. Time-stretching the noise processes \hat{w} and \hat{v} using a SOLA-based technique.
5. Dynamically changing their respective stretch factors (independently) as well as window size.
6. Re-synthesizing using the state-space model with transformed state matrix and noise signals.

5.7.2 Spectral Effects

The modification of the state matrix just described – by adding noise values to the block diagonal – could have been discussed here. However I included it in the previous section as the result is primarily a temporal modulation that is perceived more as influencing texture. Instead, there are other actions that can be more properly considered as spectral. The most basic of these is pitch shifting while preserving the time-scale. The classic methods for doing this are based on the block-based processing used for time-stretching discussed in the previous section [42]. One approach that may utilize the phase vocoder or SOLA technique is to time-stretch the signal, and then resample so that it is transposed upon playback at the original sampling rate. A second approach is to use two parallel delay lines with a cross-fade that will overlap and add their output, where each delay line is modulated by a sawtooth wave. The respective modulating waves are out of phase with one another by one-half of the length of the delay line, and the output cross-fades between the two at the end of each modulation cycle.

In the case of the Kalman based model, there are two approaches one may take to pitch shifting. Using the RESTFT, one may time-stretch and resample as with a standard phase vocoder. Alternatively, one may use the additive model and transpose by multiplying the frequency trajectories $\{\omega_{1,n}, \dots, \omega_{L,n}\}$ by a constant, and then re-calculating the state matrix \hat{A} . At the same time, one may introduce harmonicity/inharmonicity or exotic timbral effects by multiplying or adding an offset to each trajectory separately [32].

In the previous section I mentioned that the window decay itself can influence the temporal evolution of the spectrum. For example one can introduce different decay factors for different state values and modify these over time. As an example, I introduced scaling exponents between -0.3 and -0.98 to the resynthesis matrix from the time-stretching example of figure 5.8 presented in the previous section. The result is that higher partials and their associated noise-bands fall off more quickly as can be seen in figure 5.10, which results in a duller and more muted sound overall. While this control of the temporal evolution of individual partials subtly affected the overall timbre of the already-texturized signal, one may also alter the state matrix in order to impose a frequency-dependent delay that achieves more drastic timbral processing as well as an inherent time-stretching. While the random modulation of the diagonal as in section 5.2.2 influence texture, a more spectrally-focused effect results if alter the phase values between analysis and synthesis, leading to a dispersion effect [42]. For example, if we multiply each frequency

$$\omega_k = \frac{2\pi k}{N} \quad (5.52)$$

by the linear function of frequency

$$D_k = D \frac{k}{N} \quad (5.53)$$

then we introduce a linearly-varying, frequency-dependent time delay in the resynthesis. Applying this to the piano attack results in a sound that is smeared differently for each partial, changing the relative perceived influence of each over time and extending the sound overall. This transformation is depicted in figure 5.11.

5.7.3 Cross-Synthesis

Finally, one may combine the model parameters from two different systems to form a hybrid sound. We may rely on the state information for this process, use the noise processes, or a combination of these. In the first case, the process begins by extracting the magnitude and phase for each signal as in equations 5.49 and 5.50. At this point, one may choose to multiply, add or otherwise merge the magnitude values of the two signals. The phase information is much more sensitive, but successful hybrids can still result from adding the phase, preserving the phase of one sound or by combining the phases from the

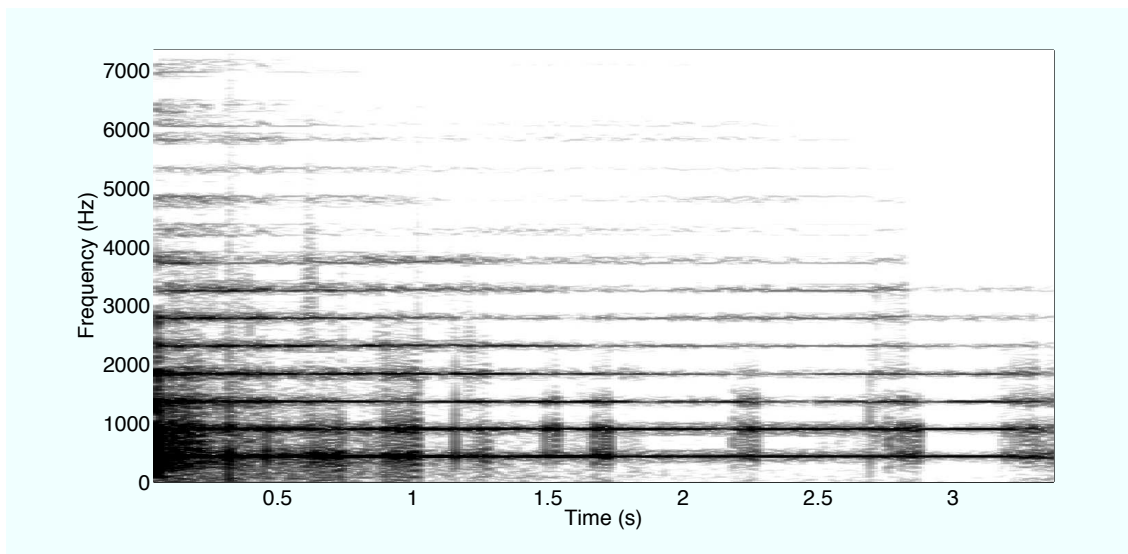


Fig. 5.10 Spectrogram showing different partial evolution for sound example from previous section, after altering state matrix window decay parameter.

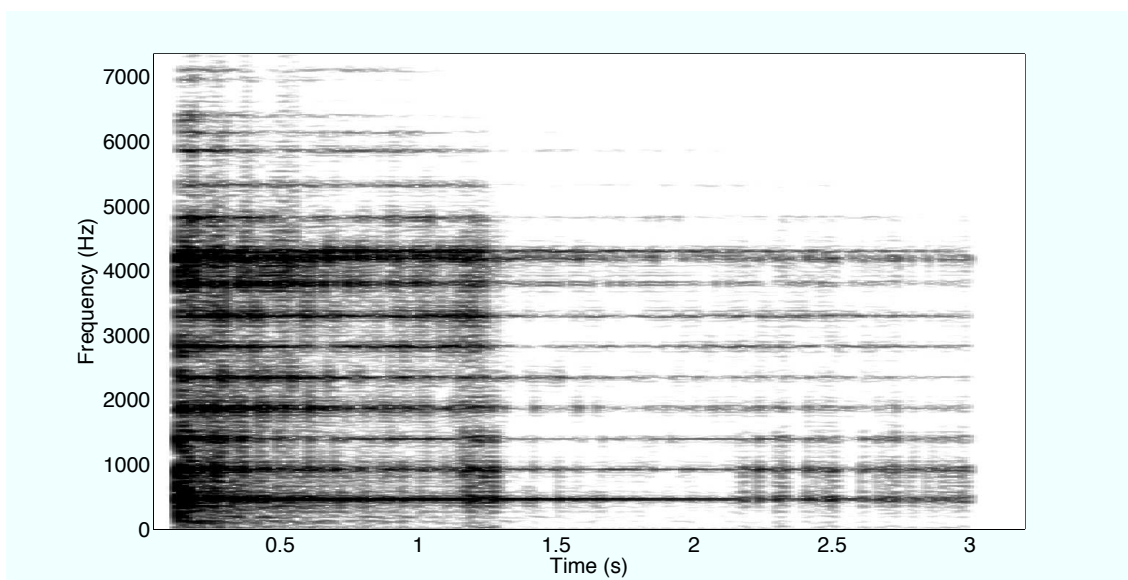


Fig. 5.11 Spectrogram showing a frequency-dependent delay and time smearing after altering state matrix frequency values.

two signals. A very classic approach would be to multiply the magnitude information and preserve the phase only from one signal (e.g. one with a well-defined pitch). In this case we may synthesize new state values as

$$s_{k,n,r} = (M_{k,n}^{(1)} M_{k,n}^{(2)}) \cos(\phi_{k,n}^{(1)}) \quad (5.54)$$

and

$$s_{k,n,i} = (M_{k,n}^{(1)} M_{k,n}^{(2)}) \sin(\phi_{k,n}^{(1)}) \quad (5.55)$$

Meanwhile, a source-filter type of cross-synthesis can be achieved if the process noise values \hat{w}_1 and \hat{w}_2 from two different sounds are exchanged. Alternatively, as suggested in [195], the process noise can be replaced with the residual from another source-filter technique such as LPC. However, the classic LPC technique results in a one-dimensional error signal, while the signal \hat{w} is multidimensional, having the same size as the state vector. I have experimented with multi-band LPC and cepstral techniques, in which each band is centered at the same frequency as the state vector. This provides for interesting musical effects in certain situations, but requires more thorough research into its potential. At the same time, the output noise process \hat{v} may be crossed between two sounds, resulting in a less drastic synthesis that sounds similar to an additive mixing between the sounds. Finally, a combination of these two methods can result if the state matrix data is crossed between two signals, while either of the two noise processes are crossed. In this way, one may choose to combine the overall spectral shape of one sound (magnitude) with the detailed pitch/frequency content of another (phase) while trading the textural nature of the excitation between the sounds (process noise). All of these methods, as with all cross-synthesis, are very contextual and depend on the musical interest and the nature of the input signals.

5.8 Adaptive Control of Sound Transformations

To this point in the chapter I have focused almost exclusively on developing a model for sound transformation. The resulting model is novel and interesting in its own right, allowing for various textural and spectral effects. However my purpose for re-considering classic techniques (phase vocoder, additive analysis/synthesis) in a new framework was not simply for new transformation possibilities, but rather to build the normally

“higher-level” control layer into the model itself, showing how mapping and control can be considered at the most atomic level. The use of the estimation framework afforded by the Kalman filter was chosen not only for residual estimation, but because it can allow for *consideration of control dynamics* during the analysis stage in parallel with sound dynamics. In this way, mapping becomes an expression of the way control dynamics co-vary with and influence sound dynamics, and vice-versa. Therefore, as the final piece in the iterative model-building process, I will illustrate an augmentation of the sound model to include control dynamics. I’ll first show an example based on a simple source-filter model to make the process clear and to illustrate the modeling of a time-domain ARMA process – a basic control system that can be used modularly to build more complex designs. After this I will return to a more musically-motivated example based on the RESTFT and physically-inspired control dynamics. This final example is not the end-all or intended to be *the* complete or defining musical performance system. Rather, it is one example system based on this new adaptive control structure-building framework, which is the main contribution of this chapter.

Recall that the RESTFT/additive model is based on a linear state-space equation, which expresses the temporal evolution of the STFT. While it is sufficient to use a linear model in order to describe the underlying analysis/synthesis system and certain sound transformations, as soon as one wishes to express musically-interesting control of the spectral state-evolution, the system becomes nonlinear very quickly. In this case, one must expand to a nonlinear state-space, and extend the modeling approach applied to RESTFT accordingly. This augmentation process begins with the modeling assumption that the state evolution and observation equations are governed by some nonlinear functions f and g such that

$$s[n + 1] = f(s[n], u[n], w[n]) \quad (5.56)$$

and

$$x[n] = g(s[n], u[n], v[n]) \quad (5.57)$$

where all input/output parameters maintain their meaning from the previous state-space equations. A direct application of the previous model is not possible, as the Kalman filter is defined only for a linear system. However, several nonlinear extensions to this technique have been developed, with one of the more popular variants being the so-called

Extended Kalman Filter (EKF) [205]. The EKF has been used in many applications for nonlinear state estimation and control, making it a standard for certain areas such as in design of navigation systems. The essential idea of this technique is to create a nominal linear trajectory \bar{s}_k in state-space around the true state trajectory. This is calculated from the a posteriori estimate \hat{s}_{k-1} at a previous time step and without input noise via

$$\bar{s}_k = f(\hat{s}_{k-1}, u_{k-1}, 0_{Nx1}) \quad (5.58)$$

and

$$\bar{y}_k = h(\bar{s}_k, u_k, 0). \quad (5.59)$$

Once this trajectory is calculated, a local linear estimate of the nonlinear state evolution is computed at each sample and the standard Kalman equations are applied to this linearized form. To achieve this, a Taylor series approximation of the nonlinear functions f and g are calculated as

$$s_k \approx \bar{s}_k + \bar{A}_k(s_{k-1} - \hat{s}_{k-1}) + \bar{W}_k w_{k-1} \quad (5.60)$$

$$y_k \approx \bar{y}_k + \bar{H}_k(s_k - \hat{s}_k) + \bar{V}_k v_{k-1} \quad (5.61)$$

where \bar{A} and \bar{H} are the Jacobian Matrices for functions f and g , respectively, taken with respect to the state input. \bar{W} and \bar{V} are similarly computed, with respect to the input noise processes. The Jacobian Matrix of a function is built from its partial derivatives, so that for example in the case of f with respect to the state input, each entry of the matrix is defined as

$$\bar{A}_{[i,j]} = \frac{\partial f_{[i]}}{\partial s_{[j]}}(\hat{s}_{k-1}, u_{k-1}, 0_{Nx1}) \quad (5.62)$$

where $f_{[i]}$ represents the i th place in vector-valued function f and $s_{[j]}$ is similarly the j th place in the state vector s .

Finally, with this local linearization at each instance – using the nominal state trajectory and Jacobian matrices – we may use the classic Kalman algorithm to produce state, residual and output estimates with appropriately-modified versions of equations 5.22 through 5.26. While the EKF is complex and fairly expensive computationally, it is used

regularly in real-time applications in which state or observation equations are nonlinear.

5.8.1 Application 1: Control of Modulated Source-Filter Model

In order to understand the paradigmatically different nature of this approach to control in comparison to the previous chapters, consider the following example system which starts from the desire to design a positional tablet-based control of a modulated source-filter model. The first step is to express a second-order IIR filter in a first-order recursion as a time-domain autoregressive (AR) process. The initial second-order expression begins with output $y[n]$ and filter coefficients $b_1[n], b_2[n]$ that relate to each other as

$$y[n] = b_1[n]y[n-1] + b_2[n]y[n-2] + a_0x[n] \quad (5.63)$$

where $x[n]$ is the input source signal and a_0 is an input gain. We may then express the state vector as

$$s[n] = \begin{pmatrix} x_1[n] \\ x_2[n] \\ x_3[n] \\ x_4[n] \end{pmatrix} = \begin{pmatrix} y[n] \\ y[n-1] \\ b_1[n] \\ b_2[n] \end{pmatrix}$$

which gives rise to the following state equation

$$s[n+1] = \begin{pmatrix} x_3[n+1]x_1[n] + x_4[n+1]x_2[n] \\ x_1[n] \\ x_3[n] \\ x_4[n] \end{pmatrix} + \begin{pmatrix} a_0x[n] \\ 0 \\ 0 \\ 0 \end{pmatrix} + w[n] \quad (5.65)$$

where again $w[n]$ is a white noise process. Note that computing the state at time n requires knowledge of the state itself, and so a priori values \hat{x}_3 and \hat{x}_4 are used in practice. While this equation already introduces nonlinearities due to the interaction between state values, it is more intuitive from a musical point of view to control the center frequency (f_c) and bandwidth (B_w) of the filter. Therefore, using knowledge of the relationship between these parameters for a second-order IIR filter, we can re-write the state equation in a more intuitive manner – though at the cost of introducing more nonlinearities. Further, a two-dimensional control input $u[n] = \{u_1[n], u_2[n]\}$ is introduced to influence

these two parameters, so that bandwidth changes linearly and the center frequency is sinusoidally modulated. To achieve this mapping we may include the control inside the state dynamics function. The state vector then becomes

$$s[n] = \begin{pmatrix} y[n] \\ y[n-1] \\ f_c[n] \\ B_w[n] \end{pmatrix} \quad (5.66)$$

while the new state equation is

$$s[n+1] = \begin{pmatrix} -\phi(\hat{x}_3[n+1], \hat{x}_4[n+1])x_1[n] + \psi(\hat{x}_3[n+1])x_2[n] \\ x_1[n] \\ x_3[n] + c_1[n] \\ x_4[n]\cos(2\pi u[n]n) \end{pmatrix} \quad (5.67)$$

where

$$-\phi(a, b) = -2e^{\frac{-a\pi}{f_s}} \cos\left(\frac{2\pi b}{f_s}\right) \quad (5.68)$$

and

$$\psi(a) = e^{\frac{-2a\pi}{f_s}} \quad (5.69)$$

for any real value a, b where f_s is the audio sampling rate. This new state space has several aspects that a traditional Kalman filter cannot handle: nonlinear state dynamics, use of the a priori state values in the state vector and the inclusion of control input in the state dynamics function itself. However, with the use of an EKF variant, we can estimate the state and control, driving the system as desired.

Now, with all of the complexity embedded in the state dynamics, the observation on this state is a simple projection of $x_1[n]$ onto output $z[n]$:

$$z[n] = [1 \ 0 \ 0 \ 0]s[n]. \quad (5.70)$$

For this example, the input “source” $x[n]$ is white noise, and so to align with our previous notation let $x[n] = w[n]$, while again the control input $u[n]$ comes from an external control source. In implementing this system I have used the 2D position data from a graphics tablet, though any continuous control could be used. Given this representation, the EKF

derivation begins with the Jacobian matrix for the state equation f which becomes

$$\bar{A}_n = \begin{pmatrix} -\phi(\hat{x}_3[n], \hat{x}_4[n]) & -\psi(\hat{x}_3[n]) & \frac{\partial f}{\partial x_3} & \frac{\partial f}{\partial x_4} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (5.71)$$

where

$$\frac{\partial f}{\partial x_3} = -2\hat{x}_1 e^{\frac{-\pi\hat{x}_3}{f_s}} \cos\left(\frac{2\pi\hat{x}_4}{f_s}\right) - \frac{2\pi\hat{x}_2}{f_s} e^{\frac{-2\pi\hat{x}_3}{f_s}} \quad (5.72)$$

and

$$\frac{\partial f}{\partial x_4} = \frac{4\pi}{f_s} e^{\frac{-\pi\hat{x}_3}{f_s}} \sin\left(\frac{2\pi\hat{x}_4}{f_s}\right) \quad (5.73)$$

After estimating the linearized state matrix, one can further estimate the state by the linear approximation as in 5.60, centered about the nominal state trajectory defined by 5.67. The matrices \bar{W} and \bar{V} are simply the identity multiplied by their initial noise covariances, and the matrix \bar{H} remains a simple projection. At this point, we can use the Kalman filter loop in order to estimate the state and output, given the control input values c_1 and c_2 . The a priori prediction at the next step is computed, in order to be used in the future calculation of the Jacobian matrix for the nonlinear state dynamics.

Therefore, in practice I use a one-step prediction in order to drive the nonlinear control dynamics, rather than directly affect the B_w and f_c parameters. In this sense the system implements predictive control.

5.8.2 Application 2: Control Dynamics for Partial/Residual Modification

We now return to the time-frequency models developed in this chapter in order to embed control dynamics within them. Note that the classic view of analysis/synthesis methods would consider control input to act on an intermediate transformation stage, between analysis and synthesis, as depicted in figure 5.3. However, if one were to consider control dynamics in the model during the analysis stage, then the process would be modeled as the outcome that occurs from the particular control input which is enacted at the time of analysis, such as in the previous example in which the process was considered as a frequency-modulated source-filter model. This difference must be considered when

planning for the transformation stage in the context of analysis-synthesis modeling. In the original Kalman implementation, note that the analysis stage was simply modeled as a linear STFT, and the parameters were modified or substituted afterwards. Instead, we can *augment* the initial model to include potentially nonlinear control dynamics, which are then estimated during the analysis stage.

Control of State-Space Additive Model

In order to explore an implementation of this instrument design methodology, return to the granular based scrubbing example as given in chapter 4, which made use of a graphics tablet controller. In particular, figure 4.26 of section 4.3.3 depicts a control structure in which the contact time with the tablet controls the initial position within the source file, speed influences the final playback time while X-Y position (among other parameters) modulates the timbral and textural qualities of the output. Dynamics, in this case, were introduced by extracting the velocity of control input, tuned by the RST mapping as well as the use of a leaky integrator. This was a modular construction in which control gestural dynamics were tuned by each block in cascade. We can extend this approach in a way that more closely considers the sound evolution, by combining the ARMA-EKF modeling approach from the previous example with the Kalman-based additive sound model. That is, the latter model is used for the sound analysis/modeling stage, and is augmented with temporal dynamics for the control analysis/transformation stage. Therefore this approach can properly be considered as *analysis-augmentation-transformation-synthesis*.

Now, the intent here is to build a system that is similar to the aforementioned scrubbing example in that the goal is to control timbral/textural characteristics through gestural dynamics that are built into the control structure. Rather than work on high-level granular parameters, however, we can work with the low-level spectro-temporal model as expressed by equation 5.19. In particular we can add a control for the separate levels of the sine vs. residual level such as

$$s[n+1] = \hat{A}\alpha s[n] + D_n u[n+1] + \beta w[n] \quad (5.74)$$

where α is a $1 \times N$ weighting for the partial values and β is similarly a $1 \times N$ weighting for the input residual value. The overall perceptual effect depends heavily on the choice of the distribution of the individual weight values. For example, the individual α weights

can target certain partials, be drawn from another spectral envelope to create cross-synthesis, or control the odd/even balance which can have a strong effect on the perceived timbre. Similarly, the relative balance of the partial/residual magnitude spectrum over time can strongly affect the dynamic matter properties of the sound. One must be careful in choosing the weight values when controlling the magnitude spectra so as not to influence the phase. For the k th and $k + 1$ th state values – which again define the real and imaginary components for partial number $\frac{k}{2}$ – we must require that $s_k = s_{k+1}$ for the phase to be preserved. Further, we can apply a given magnitude value m for the given partial by defining each of these values as $\frac{\sqrt{2}}{2}m$. Having said this, an interesting effect is produced simply when all values of the respective weight matrices are the same, particularly when control dynamics are added to these weights. I'll restrict the focus to this case for simplicity of presentation, and to focus on the control aspect of the model.

Implementing Control Gestural Dynamics

At this point we can compose the gestural dynamics to control the state parameters and residual process of the underlying sound model. Taking inspiration from the previous chapter's examples, I have found an interesting control gesture to arise from the use the velocity of X-Y values, conditioned by a leaky integrator, to guide certain aspects of a sound's resynthesis. This sort of gesture requires a sustained motion at a fairly high speed (which is a function of the integrator's response time) in order to maintain the influence. Therefore, oscillatory motion at regular speeds will result in a sustained value of the control output – which for a perceptually coherent mapping can be mapped into control of the stable part of the spectra, represented by the partial values of the state vector. For this example I will condition these dynamics on the x-position input, call it c_1 . In the scrubbing example of the previous chapter the differentiator and integrator were thought of as “black-boxes”, only considered in terms of their input/output effects. However, we can represent their dynamics as a recursive difference equation as well, with the velocity found simply by

$$\delta[n] = k_r(c_1[n] - c_1[n - 1]) \quad (5.75)$$

where k_r is the rate of the control signal. At this point, we can apply the leaky integrator to the velocity, which can be expressed as

$$L[n] = \frac{1}{k_r} \delta[n] + L[n-1] * 2^{\frac{-1}{k_r * \lambda}} \quad (5.76)$$

where λ is the response time of the integrator. Therefore, from a signal processing point of view, this system is a first-order filter acting on the velocity of control input c_1 . In order to design another expressive control gesture for y-position input c_2 that has considerably different dynamics, we can extend to a second order filter and add resonance to the system. Generally speaking we can condition this input by the equation

$$z[n] = b_1 z[n-1] + b_2 z[n-2] + b_0 c_2[n] \quad (5.77)$$

where z is the system output and b_1 and b_2 are general filter coefficients while b_0 is the input gain. The temporal behavior of this system depends heavily on these latter three coefficients. Further, to leverage the dynamics of this system we need to express it in a one-step recursion, which we can do by defining an intermediate equation as

$$z[n] = z[n-1] + \frac{1}{k_r} d[n] \quad (5.78)$$

with new intermediate variable d . Taking this further, define the general filter coefficients as

$$b_0 = \frac{a_1}{k_r^2}, b_1 = 2 - \frac{a_1 - a_2}{k_r^2}, b_2 = \frac{a_2}{k_r^2} - 1 \quad (5.79)$$

where again a_0 , a_1 and a_2 are tunable filter coefficients. Inserting equations 5.78 and 5.79 into the general second order equation results in a first order recursion of two variables defined as

$$d[n] = d[n-1] + \frac{1}{k_r} (a_1 (c_2[n] - z[n-1]) - a_2 d[n-1]) \quad (5.80)$$

Taken together, equations 5.78 and 5.80 constitute a state-space form of the general second order system with state variables d and z . Further this particular choice of b coefficients was made due the dynamics that it allows for: rather than building up of energy, the conditioned response here is such that fast input control changes result in an oscillation that converges to the given value (provided a coefficient values are subject to certain boundary constraints). From a perceptual control point of view, this behavior is an

interesting addition to the instrument design when used for control of the input residual gain. With such a mapping, sharp control actions result in a high excitation noise level, having amplitude modulation. Gradual movements have little or no perceptible modulations, depending on the filter coefficients. In the context of control dynamics, Menzies [212] has referred to this as a resonant-follower, which is certainly an appropriate description.

While these two dynamical control systems can be interpreted as signal processing tools as described above, they also possess a more intuitive physical interpretation. The leaky integrator is in fact a spring-damper system, with the response time parameter acting as a damping coefficient on the system. Similarly, the second order resonant filter expresses a mass-spring-damper (MSD) system, which can be checked by discretizing the physical equation and setting the mass to 1. Thus we can provide a physically-grounded expression by setting the coefficients to $a_1 = K_s$ and $a_2 = K_d$, which are spring and damping constants respectively. Further, the state variables z and d represent position and velocity, where this relationship can be seen by comparing equations 5.75 and 5.78. Therefore the perceptual effect of building up of energy in the former case or of modulating response in the later is in fact a product of designing the physics of a real system that has a resistive (spring-damper) and resonant (MSD) feel to it. In designing this physically-inspired interaction *in tandem with* the signal-focused STFT/additive sound model I am taking a hybrid approach that falls between signal and physical modeling.⁶ This is one of the great advantages and novelties to designs based on a state-space representation.

In order to augment the overall system to include these elements, however, we need to modify the expression a bit. If we kept equations 5.80 and 5.78 in their current form, the output would depend on an intermediate parameter at the same time-step and an a priori estimate would need to be used. This can be avoided by combining these equations so that

$$z[n] = z[n-1] + (1 - \frac{K_d}{k_r})d[n-1] + \frac{K_s}{k_r}(c_2[n] - z[n-1]) \quad (5.81)$$

This results in a single expression of the second-order control system in a first-order recursion.

⁶An example of a more physically-inspired sound model in state-space form can be found in [39]. Instead, the work presented here is a different form of hybrid in that it combines physically-modeled control elements with the signal-focused STFT.

Augmented State Space: Time-Domain Control, Time/Frequency Sound Model

With this, the mapping and control structure are fully defined. While I began from the control structure represented by figure 4.26, the final product is a slightly different mapping and new control dynamics, acting on an entirely different analysis/synthesis method. The new structure is depicted in figure 5.12.

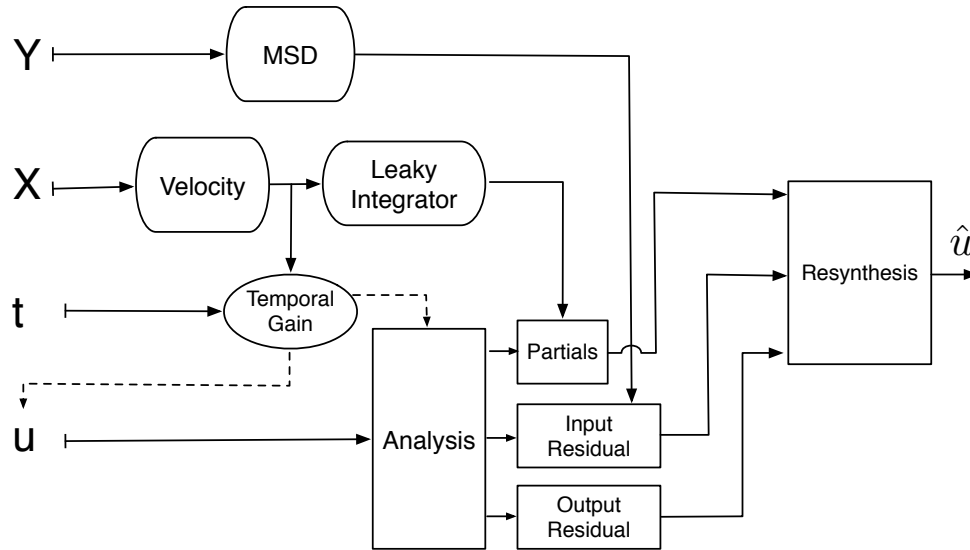


Fig. 5.12 The mapping structure for the given instrument design. X - Y tablet values feed the physically-inspired dynamics of the leaky integrator and mass-spring-damper systems, respectively. As with the similar example depicted in figure 4.26, time (of contact) is an implicit input that controls the “scrubbing”. This control is outside the model and so is not discussed in this chapter, but note that it may control either sound input u or the model parameters by way of some time-scaling algorithm such as a SOLA technique. Finally, the analyzed sound u is an input to the analysis process, which is then affected by the control dynamics before being resynthesized into output \hat{u} .

However, the system is not fully defined until we can integrate these new equations with the sound model’s state-space. While the essence of what we want to achieve may be expressed by

$$s[n+1] = \hat{A}L[n]s[n] + D_n u[n+1] + z[n]w[n] \quad (5.82)$$

This does not represent a full model of the dynamics, as the control structure must be

fully represented in the state equation. We can achieve this by augmenting the state vector to include the three control process variables. Let

$$s_c[n] = \begin{pmatrix} s[n] \\ L[n] \\ d[n] \\ z[n] \end{pmatrix} \quad (5.83)$$

which results in the new state equation defined by

$$s_c[n+1] = A_{c,n}f(s_c[n]) + D_{c,n}u_c[n] + z[n]w[n] \quad (5.84)$$

where

$$f(s_c[n]) = \begin{pmatrix} L[n-1]s[n-1] \\ L[n-1] * 2^{\frac{-1}{k_r * \lambda}} \\ d[n-1] - \frac{1}{k_r}(K_s z[n-1] + K_d d[n-1]) \\ z[n-1] + (1 - \frac{K_d}{k_r})d[n-1] - \frac{K_s}{k_r}z[n-1] \end{pmatrix} \quad (5.85)$$

Note that the scalar $L[n-1]$ is multiplied across all values of the state $s[n-1]$, so that the augmented state vector that results from this function is of size $N+3$ for state size N . Further $A_{c,n}$ is the block-diagonal matrix \hat{A}_n with a 3x3 identity matrix added to the diagonal, with proper zero padding of the first columns and rows of this new state matrix in order to make it well-defined.

Meanwhile, input vector $u_c[n]$ accounts for the current sound input value $u[n]$ as well as the control inputs, so that

$$u_c[n] = \begin{pmatrix} u[n] \\ c_1[n] \\ c_1[n-1] \\ c_2[n] \end{pmatrix} \quad (5.86)$$

Likewise $D_{c,n}$ extends the matrix D_n in order to account for the input value's contribution to the control dynamics. Thus the control matrix C is added to the block diagonal of $D_{c,n}$ where

$$C = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & \frac{K_s}{k_r} \\ 0 & 0 & \frac{K_d}{k_r} \end{pmatrix} \quad (5.87)$$

with zero padding of the first columns and rows of $D_{c,n}$ in the same manner as applied to $A_{c,n}$. Similarly $u_c[n]$ is the input value $u[n]$ padded with zeros, thereby finalizing this new form of the process equation. The observation is essentially the same as in equation 5.2 except that it acts on the augmented state vector s_c , and the new observation matrix is an augmented form of B defined as

$$B_c = \begin{pmatrix} B & 0 & 0 \\ \vdots & 1 & 0 \\ \vdots & 0 & 1 \end{pmatrix} \quad (5.88)$$

with zero-padding prepended to the bottom two rows. This new matrix is needed to account for the two control values, which are added to the observation that still includes the audio output $y[n]$. Thus the system can be thought of as *an observation of input/output values, leading to an estimate of sound/control dynamics which in turn are synthesized to produce the final audio output.*

EKF for Control/Sound Integration

Note that this control-augmented dynamics model is now nonlinear due to the interaction between state variables. Because of this, we once again must use the EKF structure in order to take advantage of the underlying Kalman framework for estimating the dynamics. Again we must linearize about a nominal state trajectory

$$\bar{s}_c[n] = A_{c,n-1}f(\hat{s}_c[n-1]) + D_{c,n-1}u_c[n-1] \quad (5.89)$$

As with the previous example we calculate the nominal state based on the a posteriori estimate of the last time step; however in this case we have control dynamics that are taken into account in making the state prediction. Given this value we can linearize the state trajectory in a similar fashion to the expression of 5.60. The primary difference is that we must take the control into account for the state noise process, as only the control dynamics for the partials are taken into account in the derivation of the nominal state

trajectory. Thus we arrive at

$$s_c[n] \approx \bar{s}_c[n] + \bar{A}_n(s_{n-1} - \hat{s}_{n-1}) + \bar{W}_n w[n] \quad (5.90)$$

where the two Jacobians are found to be

$$\bar{A}_n = \begin{pmatrix} \hat{L}[n-1] & \hat{s}[n-1] & 0 & 0 \\ 0 & 2^{\frac{-1}{\lambda k_r}} & 0 & 0 \\ 0 & 0 & 1 + K_d & 0 \\ 0 & 0 & 1 - \frac{K_d}{k_r} & 1 - \frac{K_s}{k_r} \end{pmatrix} \quad (5.91)$$

for the process linearization and

$$\bar{W}_n = \begin{pmatrix} \hat{z}[n-1] & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (5.92)$$

for the linearization of the input residual control matrix⁷As with the AMRA-KF example, using the newly-linearized approximation from 5.90 and the already-linear observation equation, we can use the standard Kalman loop – in this case as it was defined for analysis of the additive sound model.

At this point – after taking into consideration the nonlinear interaction between control and sound dynamics – we have arrived at a complete instrument design that takes into account gestural dynamics. The layout for the entire instrumental system is depicted in figure 5.13. This particular system – in terms of the chosen mapping and control dynamics – is not what is most important. Rather, this example serves to show the process of *designing an instrument that builds control and sound dynamics together in an estimation framework for control/sound analysis/synthesis*. After designing the dynamics, state space and extended Kalman filtering, the system’s runtime behavior is as follows:

- Sample the control and sound input simultaneously.
- Drive the respective control/sound state equations with these values.

⁷In practice, the value $\hat{L}[n]$ from \bar{A}_n is actually an $N \times N$ matrix comprised of copies of $\hat{L}[n]$ where N is the number of partials, and the same for the \hat{z} entry of the latter matrix. I express it in this simplified form for clarity of notation, without any loss of generality.

- Estimate the intermediate control/sound state variables using the EKF based on nonlinear control/sound interaction model.
- Use the estimated state values to drive this nonlinear model, thereby re-synthesizing new sound output $y[n]$.

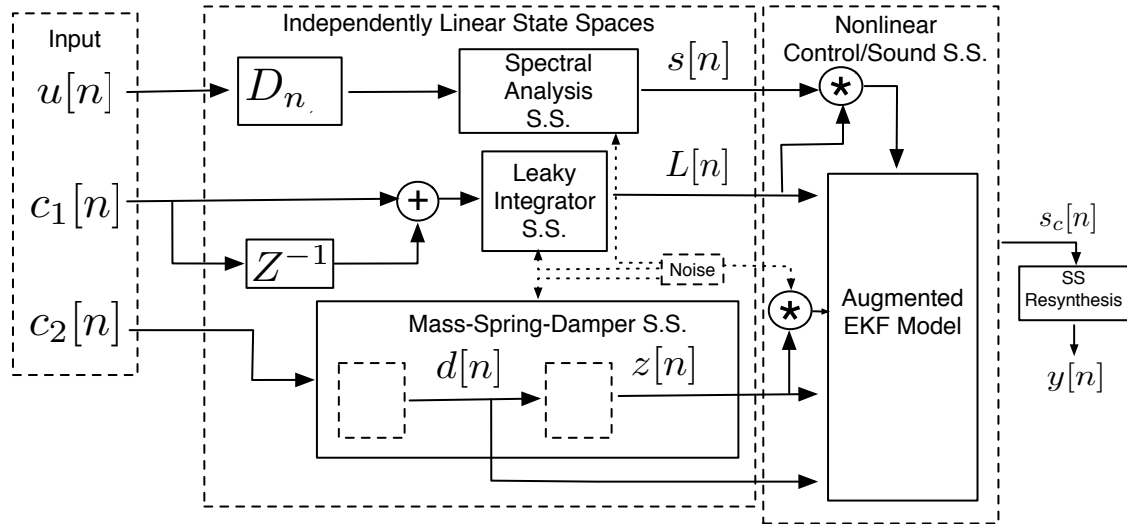


Fig. 5.13 Nonlinear state-space system for controlling time/frequency model. Analyzed input sound and control values are observed, while linear dynamical systems for control/sound are combined into nonlinear control-sound dynamics model, augmented by an EKF. This dynamics model predicts the control/sound state value, and this is re-synthesized to produce new sound output.

Many other control structures and mappings are possible using this underlying STFT/additive sound model and ARMA-based model for control gestures. The main contribution of this chapter is a canonical way to build such instruments using a combination of these two modeling approaches, the state-space representation for dynamics, the Kalman filter for estimation and the EKF for nonlinear control. The combined use of these tools has led to both a set of novel algorithms as well as a working methodology for embedding control-sound dynamics *at the sound level* of an instrumental system, so that control structures can be built from the bottom-up in a temporally-focused way.

5.9 Chapter 5 Summary

In this chapter, the issues of mapping and control structure design were approached from a completely different angle than in the previous two. Rather than constraining a set of possible values (i.e. parameters) and optimizing these to produce dynamics, the actual sound-control interaction dynamic was of primary focus and taken as the point of departure. The key idea is to build not only control possibilities but also control/sound dynamics in at the low level of the sound processing. This approach necessarily means considering musical aesthetics, mapping and gestural dynamics a priori as these are built explicitly. As such this chapter differed from 3 and 4 in that I did not present a unified theory and set of tools, followed by a variety of example applications. Rather, the examples here are themselves the presentation of the overall framework that is the main contribution of this chapter. Thus there were much fewer examples, with the key point being their formalization and the process of their construction.

This process began with a time/frequency model that was based on a state-space representation. After developing simple textural effects by way of the SSSPV implementation, the Kalman filter was introduced first as a tool for re-estimating spectral (magnitude, phase) dynamics as well as the state/observation residuals, leading to more interesting sound transformation possibilities than those afforded by the SSSPV. A novel algorithm was derived for the phase vocoder based on the well-known Chandrasekhar-type recursion in order to make the calculations more stable and efficient. After outlining the family of sound transformations that this method affords, the model was extended to nonlinear processes by use of the Extended Kalman filter. This derivation was created with the knowledge that most musically interesting control-sound dynamics are indeed nonlinear. Finally, control dynamics were embedded in the model by way of two systems, with the first being a simple source-filter model that illustrates an ARMA-type model implementation. Then, a much more involved second example showed how such a time-domain approach may be combined with the time/frequency sound model, illustrating both a more musically-inspired instrument design and the hybrid physical/signal model nature of this approach. While one could certainly take a more black-box approach reminiscent of chapter 4 to such an instrument by conditioning control parameters with leaky integrators and MSD-type processing before mapping these into an intermediate transformation phase, this would only loosely describe control/sound

dynamics as in the fabric examples. Rather this integrated approach allows for an expression of how these should co-evolve, with an estimation and adaption layer to ensure that this occurs. This system is therefore inherently different than a modular combination of control conditioning and analysis/synthesis.

The notions of mapping and control structuring in this chapter are thus implicitly extended to include this description of coupled control-sound dynamics, making the temporality of mapping and control explicit in the process. In this way chapter 5 is a combination of the phenomenological sound-first approach of chapter 2 and the gestural conditioning of the previous chapter. From a design point of view (and practically speaking), in using the framework developed in this chapter the actual output may adhere closely or loosely to the control input, depending on the noise statistics among other tuning parameters. Therefore, while in the geometric control examples of the previous chapter one could tune the exactness vs. approximateness of an interpolator, here one can tune the degree of adherence to the dynamical systems model or add additional control values into the state vector. While it is beyond the scope of this presentation, some future directions include a more robust tuning by adaptively estimating the noise statistics at the same time as the state dynamics [208] or using the control dynamics residual itself as an interesting signal to be sonified or mapped.

Chapter 6

Conclusion

6.1 Mapping, Gestural Dynamics and Musical Control Context

This dissertation is a synthesis of ideas centered around one primary goal: the design of real-time control of sound processing and digital musical instruments from the “software side”. Several approaches were taken that, on a surface level, can be seen to be distinct projects. However these approaches are in fact puzzle pieces that complement one another, leading to a design methodology that is greater than the sum of their parts. The key ideas explored were mapping and control structuring, gestural dynamics and designing an instrument for musical intent and context.

Mapping was an integral area of exploration in this research, both in terms of the theoretical and conceptual work and the practical applications that resulted. Throughout the discussion, concepts were addressed in regards to how they did or did not fit within the classic framework of mapping as a parameter association. The work of chapter 2 presented a design methodology that grounded mapping in a more perceptually and temporally oriented way, suggesting the notion of an *inverse mapping* approach in which the dynamic sonic response of an instrument could gauge the type of perceived mapping that resulted. In order to quantitatively measure this *sonic gesture* output, I developed a quantitative signal processing framework that was designed based on electroacoustic music theory. Mapping was then the explicit focus of chapter 3, where I furthered its discussion through a presentation on mapping theory that provides a stronger conceptual grounding for future discussion. I examined this concept from different points of view – systems, functional, perceptual – and further differentiated the parameter from warping

aspect of mapping (how vs. what). The functional view on mapping was examined more deeply in a mathematical context, providing another grounding force. Beyond its relevance in providing theoretical and conceptual rigor, this work allowed me to examine existing uses of mapping in the literature, and extrapolate to a first-order *mapping design space* that hopefully will generate further discussion. Applying this design space, I constructed the LoM toolbox of mapping functions that provide a significant addition to existing software for holistic, geometric-oriented mapping design. In order to link this functional view back to a perceptual one, I examined the influence of two mapping's geometric structure on perceived sound quality, controllability and ability to mimic and define different sonic gestural contours. Apart from the studies of Hunt et al [5], which focused on parameter mapping complexity, these studies provide the first (to my knowledge) attempt at examining the *perceptual control structure* defined by a mapping structure and control/synthesis pairing. While more testing must be run to arrive at definitive results, the magnitude of effect of the interactions illustrates that there is an influence of mapping structure on perceived sound quality/controllability, and that these influence performance as a product of sonic control context (i.e. what features are being traced). This provides a firm foundation for further study in this area. Finally, chapter 5 examined mapping from the point of view of a constraint on a dynamical system, expressing the co-evolution of parameters. This work shifted the focus from modeling parameter space to modeling the temporal response of action-to-sound coupling in a way that combined physically-modeled dynamics with abstract time/frequency signal models. In considering the perceptual nature of mapping, the concomitant issue of gestural dynamics was examined in detail. This was at the heart of chapter 2, where the notion of linking control and sonic gestures was introduced. The former highlighted the focus on the gestural nature *of the control data itself* and not of the initial physical gesture. The sound aspect drew on electroacoustic music theory and the concept of *dynamic morphology* to explain the gestural element of sound output from an instrumental system. Considering gestural dynamics illustrated the disparity between the different approaches to mapping and control design. In order to design a given dynamic response in the multi-parametric context presented in chapter 3, the control structure designs of chapter 4 illustrated the need to define modular mapping layers wherein the *mapping itself was controlled*, which illustrated the advantage of parameterizing the mapping layer. The fact that this approach is tied in to the multi-parametric paradigm of instrument design was made clear

through the fabric examples that showed instances in which the conditioning of control signals goes beyond this realm and becomes a *gestural conditioning*. This notion of defining control/sound temporal response in the design process was formalized in a much deeper way in chapter 5. Here the focus was on explicitly modeling the control dynamics *and* sound dynamics so that one did not simply feed the other in a decoupled fashion. Finally, all of the work presented was grounded in a musical control context: what type of sound features are relevant to control, and what sort of interaction should be used to achieve this. Chapter 2 applied electroacoustic music theoretic rigor from the Schaefferian tradition in order to define the sound-focused philosophy that would underlie the musical intention of all the work presented. The conceptual notion of balancing gesture and texture were used in order to give definition to the problem of designing control of different morphological facets of sound in a coherent way. These facets were described, with a particular focus on the *mass* and *grain* aspects of the signal. The analysis framework of this chapter was designed to explain the dynamics qualities of these morphological features, and the user studies of chapter 4 (particularly the last) illustrated the way that mapping or sound synthesis can influence one's ability to control these sound qualities. It was apparent that such timbral and textural-related sound features are driven in an indirect way, and the control structure designs of chapter 4 illustrated this through indirect adaptation of the mapping layer in order to achieve a given sonic gestural response. In order to more explicitly control spectral (timbral) and quasi-stationary temporal deviations (textural) that these instruments could afford, chapter 5 showed that deeper control warrants working on the lower-level of the sound model. The classic phase vocoder and additive approaches were re-parameterized by using a Kalman filter to estimate the spectra and process/observation noise sequences, thereby providing control "handles" for timbral and textural elements of sound. Going beyond the analysis-transformation-synthesis paradigm, the key result was the augmentation with nonlinear control dynamics, thus allowing for *modeling of the control/sound interaction* in a way that was customized to the musical context of controlling timbre and texture.

6.2 Future Work

This dissertation can be seen as three complementary approaches to control of sound processing that each display their own balance of music theory and practice,

mathematics, signal processing, HCI design and social science. The first axis was a theoretical construction of an electroacoustic music framework for instrument design and related analysis implementation from chapter 2, the second was a formalized approach to mapping and related applications (toolbox, user studies, control designs) of chapters 3 and 4, while the last was the creation of a time-frequency model that allows for expression and estimation of control and sound dynamics with specialization to timbral and textural sound transformations. While each project takes a unique path to the same ultimate goal, there does exist certain interactions between these approaches as can be seen in the application of the sonic gestural analysis to the user studies of chapter 4 as well as the control structure designs from this same chapter. There are many other such interactions that are beyond the scope of this work, but which are very much relevant to future plans of combining top-down and bottom-up control design. This includes building a sonic gestural analysis layer into the additive/RESTFT models in order to define adaptive control that reacts to sonic gestural output. Building a higher-level control structure (as from chapter 4) on top of such a model and using N-step prediction of the sound process (rather than 1-step, as in the current EKF-based model) would then allow for a more refined feedback control that could parametrically alter the mapping layer as a product of predicted sonic gestural output. This vector of potential research points to a more adaptive and flexible instrumental control system design, and this dissertation provides a firm grounding for the future of this work.



Appendix A

Subject Background for User Studies of Section 4.2

A.1 Questionnaire given to subjects

1. How many years of musical training do you have? Consider this to be learning about the theory of music (in one of its established forms), through playing or listening, as taught by a professional instructor.
2. Do you currently play an instrument. If so, which and for how long? Please focus on years that you were actively engaged in playing.
3. Do you create Electroacoustic (EA) style music? Consider this to be music that is primarily focused on sound (timbre, texture) rather than pitch or rhythm structures (these being secondary), and that requires the use of electronic technology (recording, processing or synthesis) for its creation. If so, for how many years are/were you active in this?
4. Do you listen to EA music (defined as above) often, and on a regular basis?
5. Are you a drawing or sketching artist? If so, for how many years are/were you active at this?
6. Have you used a Wacom tablet prior to the experiment? If so, how often? (either the one from the study or another - please differentiate as needed).

A.2 Participant Information Form

 Schulich School of Music of McGill University 555 Sherbrooke Street West Montreal, Quebec H3A 1E3	 Schulich School of Music École de musique Schulich École de Musique Schulich de l'Université McGill 555, rue Sherbrooke Ouest Montréal (Québec) H3A 1E3	☎ (514) 398-4535 📠 (514) 398-8061 www.mcgill.ca/music
---	---	---

Information for Prospective Participants

Dear Participant,

This experiment aims to evaluate the suitability of mapping strategies for specific tasks in a digital music performance system. The experiment will involve your attempting a number of musical tasks with several different mapping strategies and with or without visual feedback. You will be asked to evaluate different musical performance systems, giving a subjective rating within certain guidelines. Your participation will require 0.5 to 1.5 hours. You will be compensated \$5 for every 30 minutes for your time (minimum \$5).

The aim of this test is not to evaluate your individual performance, but to gather findings of the role of mapping in musical performance, with a focus on electroacoustic-style music. Your performance will be assigned a number and will be grouped with those of other participants and later analyzed according to specific variables. Your name will not be disclosed at any time.

You may discontinue participation in this study at any point during the process. If you need a break, kindly inform the test administrator.

If you would like to learn about the results of the study, please contact Doug Van Nort at 398-4535 (ext. 00271). We thank you for your interest in this study.

Sincerely,

 Doug Van Nort, PhD Candidate, Music Technology Area, Department of Music Research.

Study Supervisor: Prof. Marcelo M. Wanderley,
 Music Technology Area, Department of Music Research. Tel: 398-4535 (ext. 00917)

Participant Consent Form

My participation in the Study "Evaluation of Mapping Strategies for Digital Music Performance" is voluntary. I understand that I may discontinue participation at any point during the experiment and that my name will not be disclosed at any time during the analysis or the dissemination of findings.

Participant's name _____

Signature _____ Date _____

Fig. A.1 Form that each subject in the perceptual studies signed.

Appendix B

Rating and Performance Data for User Studies of Section 4.2

B.1 Subjective Response Data for Experiment 1 of Section 4.2

MI Modal-GC	SI Modal-GC	MI Gran	SI Gran	MI Modal-AD	SI Modal-AD	EAC	EAL	PI	MT
0.01570	0.0000	0.45669	0.78740	1.0000	0.0000			<i>x</i>	<i>x</i>
0.60630	0.031496	0.39370	0.94488	0.18110	0.40157				<i>x</i>
0.65354	0.56693	0.75591	0.36220	0.16535	0.20472	<i>x</i>	<i>x</i>	<i>x</i>	
0.36220	0.70866	0.48031	0.27559	0.49606	0.62992		<i>x</i>	<i>x</i>	
0.59055	0.26772	0.58268	0.81890	0.58268	0.51969	<i>x</i>	<i>x</i>		
0.59843	0.21260	0.023622	0.93701	0.37795	0.21260		<i>x</i>		
0.81102	0.86614	0.31496	0.62992	0.70866	0.90551	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>
0.85039	0.49606	0.74803	0.14173	0.62992	0.33858			<i>x</i>	<i>x</i>
0.60630	0.65354	0.38583	0.27559	0.62992	0.76378	<i>x</i>	<i>x</i>		
0.16535	0.62205	0.78740	0.30709	0.36220	0.90551		<i>x</i>	<i>x</i>	<i>x</i>
0.73228	0.44094	0.88189	0.95276	0.74803	0.57480			<i>x</i>	<i>x</i>
0.92126	0.72441	0.70866	0.43307	0.79528	0.52756		<i>x</i>	<i>x</i>	<i>x</i>
0.25197	0.52756	0.45669	0.0000	0.67717	1.0000		<i>x</i>	<i>x</i>	<i>x</i>
0.11811	0.0000	0.11811	0.92126	0.77165	0.70079		<i>x</i>	<i>x</i>	
0.46457	0.070866	0.42520	0.73228	0.85827	0.8268			<i>x</i>	

Table B.1 Sound preference rating values and background: whether EA composer (EAC), EA listener (EAL), plays instrument (PI) or is musically trained (MT). Each row represents a single subject.

MI Modal-GC	SI Modal-GC	MI Gran	SI Gran	MI Modal-AD	SI Modal-AD	EAC	EAL	PI	MT
0.0000	0.22047	0.44882	0.79528	1.0000	0.44882			<i>x</i>	<i>x</i>
0.85827	0.74803	0.48031	0.27559	0.76378	0.48031				<i>x</i>
0.63780	0.36220	0.23622	0.63780	0.59055	0.57480	<i>x</i>	<i>x</i>	<i>x</i>	
0.29921	0.65354	0.45669	0.15748	0.28346	0.52756		<i>x</i>	<i>x</i>	
0.44094	0.22835	0.70866	0.79528	0.62205	1.0000	<i>x</i>	<i>x</i>		
0.20472	0.61417	0.96063	0.51181	0.18898	0.60630		<i>x</i>		
0.66142	0.66929	0.82677	0.96850	0.62992	0.79528	<i>x</i>	<i>x</i>	<i>x</i>	<i>x</i>
0.74016	0.57480	0.81890	0.69291	0.64567	0.38583			<i>x</i>	<i>x</i>
0.24409	0.70079	0.97638	0.60630	0.039370	0.63780	<i>x</i>	<i>x</i>		
0.39370	0.37795	0.77953	0.55118	0.96850	0.38583		<i>x</i>	<i>x</i>	<i>x</i>
0.44094	0.80315	0.55906	0.95276	0.34646	0.68504			<i>x</i>	<i>x</i>
0.66929	0.77953	0.85827	0.75591	0.56693	0.57480		<i>x</i>	<i>x</i>	<i>x</i>
0.14173	0.34646	0.66929	0.61417	0.0000	1.0000		<i>x</i>	<i>x</i>	<i>x</i>
0.02360	0.0000	0.32283	0.86614	0.92913	0.0000		<i>x</i>	<i>x</i>	
0.54331	0.59055	0.29134	0.63780	0.88189	0.87402			<i>x</i>	

Table B.2 Controllability rating values and background: whether EA composer (EAC), EA listener (EAL), plays instrument (PI) or is musically trained (MT). Each row represents a single subject.

B.2 Mean Performance Values for Each Subgroup of Experiment 2, Section 4.2

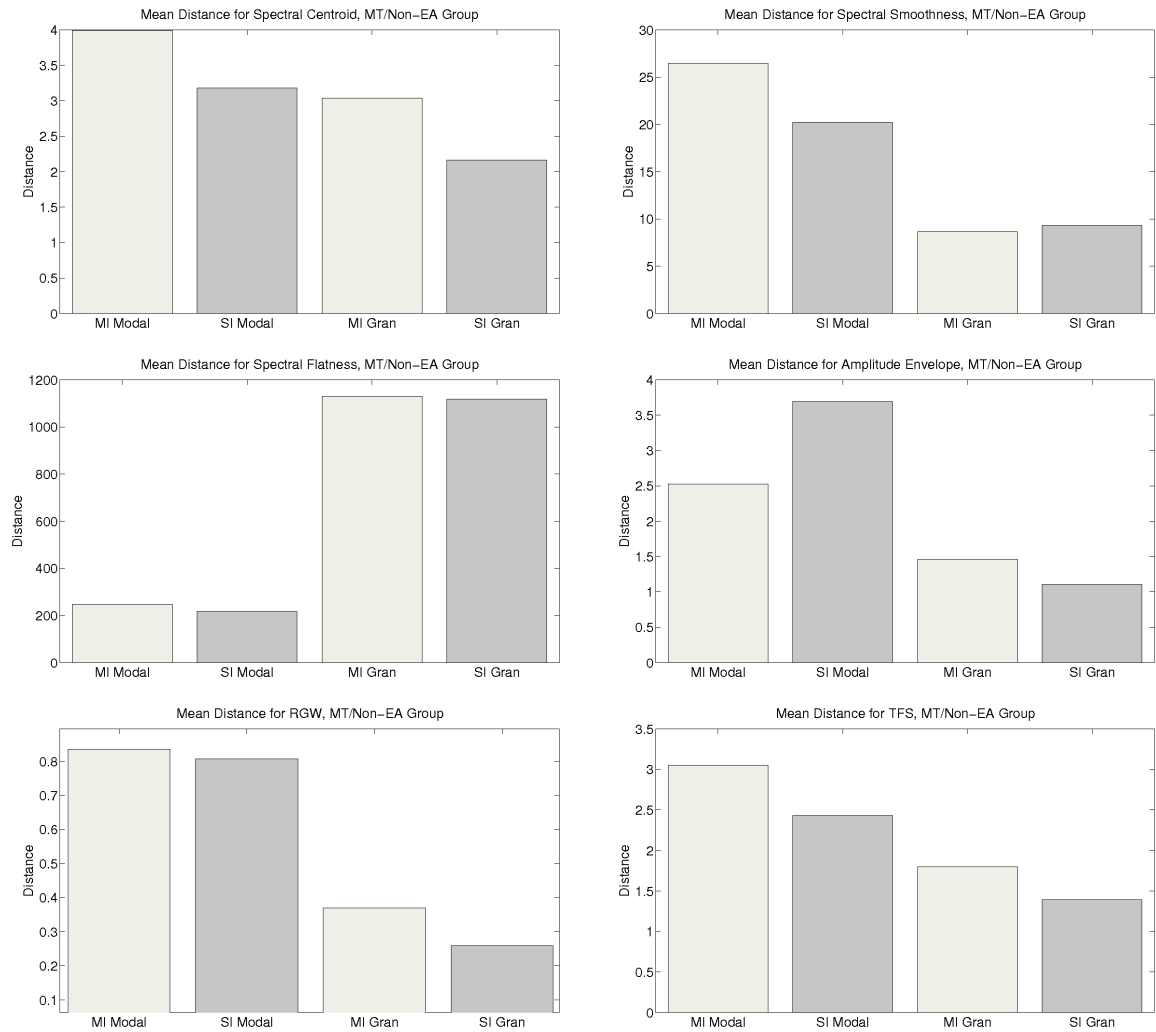


Fig. B.1 Mean Distances for (a) Spectral Centroid (b) Spectral Smoothness (c) Spectral Flatness (d) Amplitude Envelope (e) Relative Grain Weight and (f) Temporal Fine Structure from the Musically Trained/ No EA Experience Group.

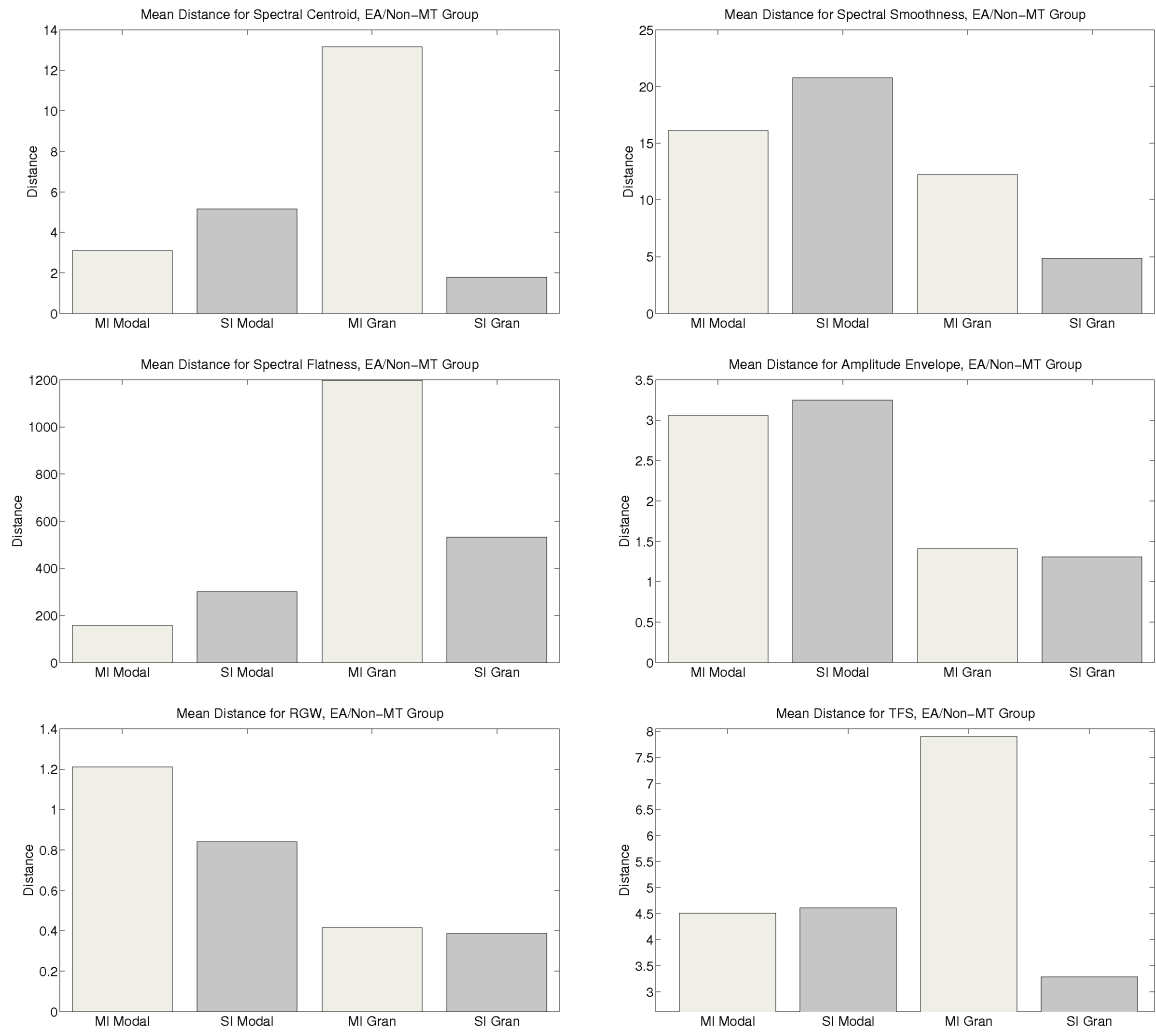


Fig. B.2 Mean Distances for (a) Spectral Centroid (b) Spectral Smoothness (c) Spectral Flatness (d) Amplitude Envelope (e) Relative Grain Weight and (f) Temporal Fine Structure from the EA Experience/ No Musical Training Group.

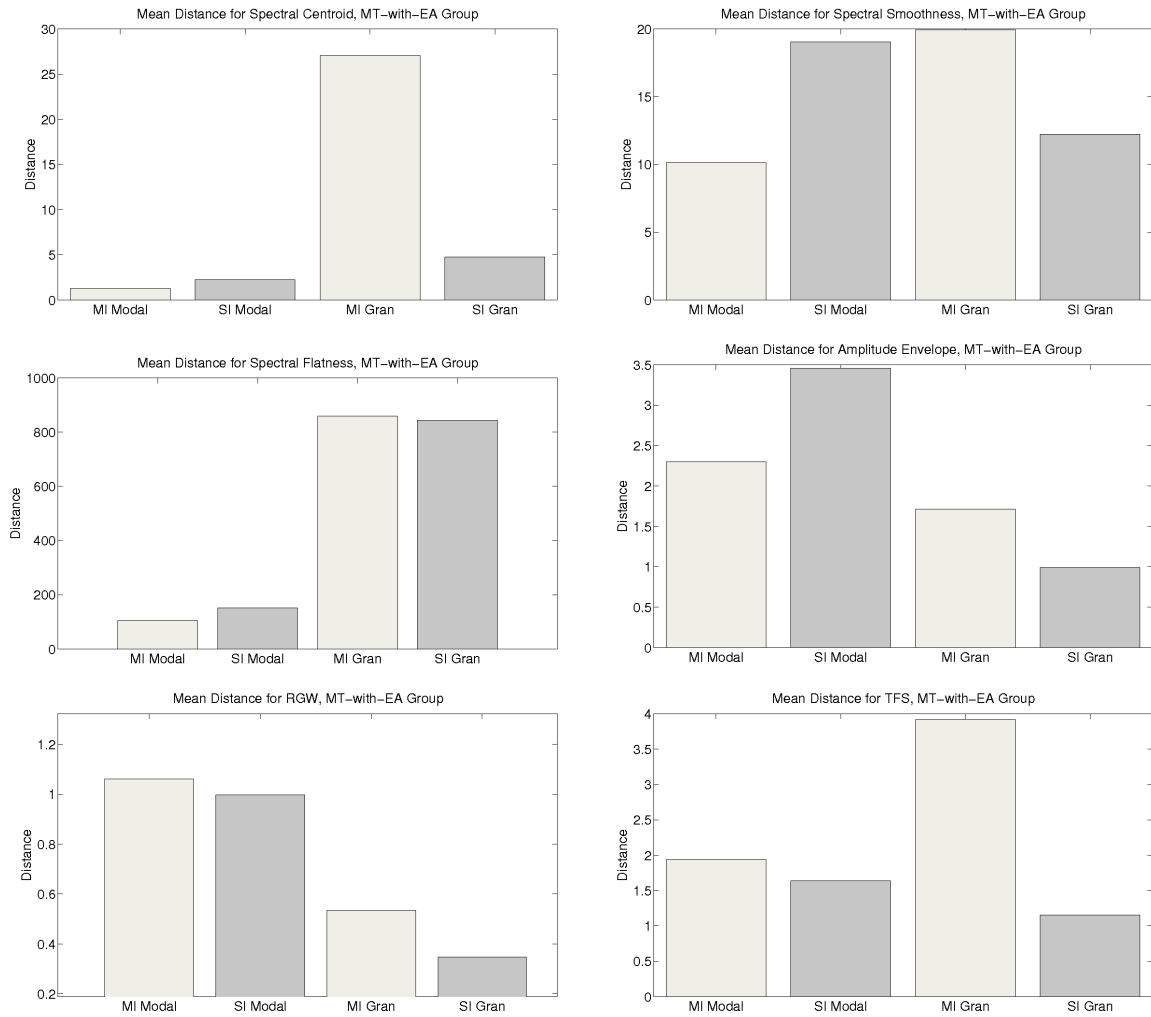


Fig. B.3 Mean Distances for (a) Spectral Centroid (b) Spectral Smoothness (c) Spectral Flatness (d) Amplitude Envelope (e) Relative Grain Weight and (f) Temporal Fine Structure from the Musically Trained with EA Experience Group.

Appendix C

Correlation Analysis for Blanket Instrument

C.1 Spatio-Temporal Correlation

Consider the raw control input stream as a single time-varying vector $X[n] = \{x_1[n], x_2[n], \dots, x_N[n]\}$, wherein each dimension represents a different point on the sensor grid. From this information, I extract features based on cross-correlation across spatial channels as well as temporal autocorrelation at given points along the cloth surface. Considering $X[n]$ as a wide-sense stationary stochastic process [80], we can express the generalized spatio-temporal correlation sequence as

$$R_n[k, i, j] = E(x_i[n]x_j[n - k]) \quad (\text{C.1})$$

giving us an expression of the dependence between variables across space and time. From this generalized statistical framework we must build a proper real-world estimate, working with a potentially intractable amount of data. The former problem is dealt with by looking at time-smoothed as well as instantaneous estimates of the data streams, which results in two respective approaches:

$$\hat{R}_n[k, i, j] = x_i[n]x_j[n - k] \quad (\text{C.2})$$

and

$$\tilde{R}_n[k, i, j] = \sum_{l=n-L}^n x_i[l]x_j[l - k] \quad (\text{C.3})$$

where L is the window size of observation for the input streams of control data. The problem of reducing the amount of data from which to extract meaningful gestural features is dealt with by consideration of the manner in which one might interact with the Blanket instrument, including some observations on the set of gestural actions that it affords and elicits.

C.2 Fabric-Based Interaction and Resulting Feature Extraction

While I did not want to strongly enforce cognitive models or schemas such as one would have in a classic instrumental performance context, there are certain modes of interaction that I considered to be indicators of intentional movement, including periodic motion and general wave-like or repeated movements. I will briefly mention three approaches to feature extraction that were designed to respond to such Blanket gestures.

C.2.1 Multi-dimensional, Area-Based Correlates

Certain areas of the cloth surface have particular importance simply due to the shape and installation of the Blanket. From these regions of interest I extracted the multi-dimensional cross-correlation. Topologically speaking, this relation does not have to constrain itself in a rectilinear fashion to the underlying sensor grid. Defining other correlative structures allows one to discover other natural and organic gestures, as these often do not arise in perfect orthogonality to the Blanket surface dimensions. For example, the interaction between boundary and center is of importance to the Blanket, as these two represent perceptual limits of the surface as well as natural points of interaction for individual (waving of the blanket) as well as collective play (“covering” of an inner participant, sending gestural waves back and forth). The fundamental difference between this approach to feature extraction and that of grid-based column correlation was depicted in figure 4-32. Note that this extraction differs from that of $\hat{R}_n[k, i, j]$ and $\tilde{R}_n[k, i, j]$ in that I consider an entire area of fabric space as a single entity. In this case, interaction between two columns of the $M \times M$ grid becomes¹

$$\bar{R}_n = \sum_{i=0}^{M-1} x_i[n]x_{i+M}[n] \quad (\text{C.4})$$

¹Again, the max $M=5$, and this presentation is for the sake of generality.

so that we are taking the inner product of two columns, in this example the first and the k th. In the case of the interaction between boundary and center, I take M copies of the center point and treat this as a repeated vector, then taking the product with the boundary in question as above.

C.2.2 Instantaneous Spatial Correlates

The above approach examines the instantaneous cross-correlation of two areas of the Blanket surface. At a different level of granularity, I also examine the momentary correlation between a collection of different Blanket locations. This can give insights into concurrent motion and the phase relationship between different points along the Blanket surface. In terms of the underlying Blanket gestures, this relates to both concurrent motion from group participation as well as waves that result from oscillations of single or multi-users. From a software design perspective, this amounted to constructing a matrix that allowed me to “tap” the Blanket surface at various locations, providing a time-varying function expressing the point-wise correlation (to augment the area correlation of above) as well as giving information about the directionality of movement. This implementation leverages the data structuring and processing of the FTM package [152], within the Max/MSP environment.

C.2.3 Concurrent Autocorrelation Analyses

As a third approach to system design and gestural analysis, I extracted autocorrelation sequences from each spatial location along the Blanket surface. While the spatial correlates provide information about directionality and the “gestural shape” or contour of the fabric, this approach provides cues for the direction and regularity of motion of the Blanket. The most fundamental information provided is the degree of periodicity, presumed in this work to be a strong measure of human intention. This use of the autocorrelation sequence is in parallel to the classical use for fundamental frequency detection in music analysis [40]. In particular I utilized a window-adaptive normalized autocorrelation function that specialized to gestures in the 0.25 to 5 Hz range.

References

- [1] C. Goudeseune, “Interpolated Mappings for Musical Instruments,” *Organised Sound*, vol. 7, no. 2, pp. 85–96, 2002.
- [2] M. M. Wanderley and P. Depalle, “Gestural control of sound synthesis,” *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, 2004.
- [3] J. Rován, M. Wanderley, S. Dubnov, and P. Depalle, “Instrumental Gestural Mapping Strategies as Expressive Determinants in Computer Music Performance,” in *Proc. of Kansei, The Technology of Emotion Workshop*, pp. 68–73, 1997.
- [4] A. Hunt and R. Kirk, “Mapping Strategies for Musical Performance,” in *Trends in Gestural Control of Music* (M. M. Wanderley and M. Battier, eds.), pp. 231–258, IRCAM – Centre Pompidou, 2000.
- [5] A. Hunt, M. M. Wanderley, and M. Paradis, “The Importance of Parameter Mapping in Electronic Instrument Design,” *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.
- [6] M. M. Wanderley, *Performer-Instrument Interaction: Applications to Gestural Control of Sound Synthesis*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2001.
- [7] A. Hunt and M. Wanderley, “Mapping Performance Parameters to Synthesis Engines,” *Organised Sound*, vol. 7, no. 2, pp. 97–108, 2002.
- [8] S. Fels and G. Hinton, “Glove Tak II: An Adaptive Gesture-to-Formant Interface,” in *Proc. of the Conference on Human Factors in Computing Systems (CHI '95)*, pp. 456–463, 1995.

-
- [9] M. Lee and D. Wessel, "Connectionist models for real-time control of synthesis and compositional algorithms," in *Proc. of 1992 International Computer Music Conference (ICMC 92)*, pp. 277–280, 1992.
 - [10] P. Modler, "Neural networks for mapping gestures to sound synthesis," in *Trends in Gestural Control of Music* (M. M. Wanderley and M. Battier, eds.), pp. 301–313, IRCAM – Centre Pompidou, 2000.
 - [11] J. Mandelis and P. Husbands, "Musical interaction with artificial life forms: Sound synthesis and performance mappings," *Contemporary Music Review*, vol. 22, no. 3, pp. 69–77, 2003.
 - [12] F. Bevilaqua, R. Muller, and N. Schnell, "MnM: a Max/MSP Mapping Toolbox," in *Proc. of 2005 Conference on New Interfaces for Musical Expression (NIME 05)*, pp. 85–88, 2005.
 - [13] M. Wanderley, N. Schnell, and J. Rován, "Escher – Modeling and Performing Composed Instruments in Real Time," in *Proc. 1998 IEEE International Conference on Systems, Man and Cybernetics*, pp. 1040–1044, 1998.
 - [14] D. Arfib, J. Couturier, L. Kessous, and V. Verfaillie, "Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces," *Organised Sound*, vol. 7, no. 2, pp. 127–144, 2002.
 - [15] R. Plomp, "Timbre as a multidimensional attribute of complex tones," in *Frequency analysis and periodicity detection in hearing* (R. Plomp and G. Smoorenburg, eds.), pp. 397–414, Leiden A W Sijtho, 1970.
 - [16] J. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.
 - [17] J. Grey and J. Gordon, "Perceptual effects of spectral modifications on musical timbres," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.
 - [18] D. Wessel, "Timbre Space as a Musical Control Structure," *Computer Music Journal*, vol. 3, no. 2, pp. 45–52, 1979.

- [19] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Soete, and J. Krimphoff, "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes," *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.
- [20] I. Bowler, A. Purvis, N. Bailey, and P. Manning, "On Mapping N Articulation onto M Synthesiser-Control Parameters," in *Proc. of 1990 International Computer Music Conference (ICMC 90)*, pp. 181–184, 1990.
- [21] I. Choi, R. Bargar, and C. Goudeseune, "A Manifold Interface for a High Dimensional Control Space," in *Proc. of 1995 International Computer Music Conference (ICMC 95)*, pp. 181–184, 1995.
- [22] G. Garnett and C. Goudeseune, "Performance Factors in Control of High Dimensional Spaces," in *Proc. of the 1999 International Computer Music Conference (ICMC 99)*, pp. 268–271, 1999.
- [23] A. Momeni and D. Wessel, "Characterizing and Controlling Musical Material Intuitively with Geometric Models," in *Proc. of the 2003 conference on New interfaces for Musical Expression (NIME 03)*, pp. 54–62, 2003.
- [24] R. Bencina, "The Metasurface: Applying Natural Neighbor Interpolation to Two-to-Many Mappings," in *Proc. of 2005 Conference on New Interfaces for Musical Expression (NIME 05)*, pp. 101–104, 2005.
- [25] J. Chadabe, *Electric sound: the past and promise of electronic music*. Upper Saddle River: Prentice Hall, 1997.
- [26] J. O. Smith, "Virtual acoustic musical instruments: Review and update," *Journal of New Music Research*, vol. 33, no. 3, pp. 283–304, 2004.
- [27] C. Chafe, "A Short History of Digital Sound Synthesis by Composers in the U.S.A." <http://www-ccrma.stanford.edu/cc/lyon/historyFinal.pdf>, *Last Accessed June 11, 2009*.
- [28] G. Young, *The Sackbut Blues: Hugh Le Caine, Pioneer in Electronic Music*. Ottawa: National Museum of Science and Technology, 1997.

- [29] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *Journal of the Audio Engineering Society*, vol. 27, no. 3, pp. 134–140, 1979.
- [30] J. Flanagan, "Computers that talk and listen: Man-machine communication by voice," *Proceedings of the IEEE*, vol. 64, pp. 405–415, April 1976.
- [31] K. Cascone, ed., *The Laptop and Electronic Music*, vol. 22. Contemporary Music Review, 2003.
- [32] T. Wishart, *Audible Design*. Orpheus the Pantomime Ltd., 1994.
- [33] P. Guillemain, R. Kronland-Martinet, and S. Ystad, "Physical modelling based on the analysis of real sounds," *Proceedings of the Institute of Acoustics*, vol. 19, no. 5, pp. 445–450, 1997.
- [34] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proceedings of the 4th International Congress on Acoustics*, pp. 1–4, 1962.
- [35] P. R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. PhD thesis, Stanford University, 1990.
- [36] K. Karplus and A. Strong, "Digital synthesis of plucked-string and drum timbres," *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983.
- [37] J. M. Adrien, "The missing link: Modal synthesis," in *Representations of Musical Signals* (G. D. Poli, A. Piccialli, and C. Roads, eds.), pp. 269–297, MIT Press, 1991.
- [38] X. Rodet, D. Matignon, and P. Depalle, "State space models for wind-instrument synthesis," in *Proc. of 1992 International Computer Music Conference (ICMC 92)*, vol. 277–280, 1992.
- [39] P. Depalle and S. Tassart, "State space sound synthesis and a state space synthesiser builder," in *Proceedings of the 1995 International Computer Music Conference (ICMC 95)*, 1995.

-
- [40] C. Cadoz, A. Luciani, and J. Florens, "Cordis-anima: A modeling and simulation system for sound and image synthesis-the general formalism," *Computer Music Journal*, vol. 17, no. 1, pp. 19–29, 1993.
- [41] C. Cadoz, "The Physical Model as Metaphor for Musical Creation. pico.. TERA, a Piece Entirely Generated by a Physical Model," in *Proc. of the 2002 International Computer Music Conference (ICMC 02)*, pp. 87–89, 2002.
- [42] U. Zölzer., ed., *Dafx: Digital Audio Effects*. New York, NY, USA: John Wiley & Sons, Inc., 2002.
- [43] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, April 1975.
- [44] S. Haykin, *Adaptive Filter Theory*. Prentice-Hall, fourth ed., 2002.
- [45] P. Lansky, "Compositional Application of Linear Predictive Coding," in *Current Directions in Computer Music Research* (M. Mathews and J. Pierce, eds.), Cambridge, MA: MIT Press, 1989.
- [46] D. Schwarz and X. Rodet, "Spectral envelope estimation and representation for sound analysis-synthesis," in *Proceedings of the 1999 International Computer Music Conference (ICMC 99)*, pp. 351–354, 1999.
- [47] J. L. Flanagan, D. I. S. Meinhart, R. M. Golden, and M. M. Sondhi, "Phase vocoder," *The Journal of the Acoustical Society of America*, vol. 38, no. 5, pp. 939–940, 1965.
- [48] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.
- [49] J. A. Moorer, "The use of the phase vocoder in computer music applications," *Journal of the Audio Engineering Society*, vol. 26, no. 1-2, pp. 42–45, 1978.
- [50] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

- [51] J. B. Allen and L. R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis,” *Proceedings of the IEEE*, vol. 65, pp. 1558–1564, 1977.
- [52] J. Laroche and M. Dolson, “Phase-vocoder: about this phasiness business,” in *1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 97)*, Oct 1997.
- [53] A. Röbel, “A new approach to transient processing in the phase vocoder,” in *Proc. of the 2003 International Conference on Digital Audio Effects (DAFx 03)*, 2003.
- [54] M. S. Puckette, “Phase-locked vocoder,” in *Proceedings of the 1995 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 95)*, pp. 222–225, 1995.
- [55] J. Laroche and M. Dolson, “Improved phase vocoder time-scale modification of audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 323–332, May 1999.
- [56] T. Karrer, E. Lee, and J. Borchers, “PhaVoRIT: A Phase Vocoder for Real-Time Interactive Time-Stretching,” in *Proc. of the 2006 International Computer Music Conference (ICMC 06)*, pp. 708–715, 2006.
- [57] N. J. Bailey, A. Purvis, I. W. Bowler, and P. D. Manning, “Applications of the phase vocoder in the control of real-time electronic musical instruments,” *Journal of New Music Research*, vol. 22, no. 3, pp. 259–275, 1993.
- [58] L. Polansky and T. Erbe, “Spectral mutation in soundhack,” *Computer Music Journal*, vol. 20, no. 1, pp. 92–101, 1996.
- [59] J.-C. Risset, “Computer music experiments, 1964-...,” *Computer Music Journal*, vol. 9, pp. 11–18, 1985.
- [60] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

- [61] K. Fitz, L. Haken, and P. Christensen, “Transient preservation under transformation in an additive sound model,” in *In Proceedings of the 2000 International Computer Music Conference (ICMC 00)*, 2000.
- [62] J. R. Beltràn and F. Beltràn, “Additive synthesis based on the continuous wavelet transform: A sinusoidal plus transient model,” in *Proc. of 2003 International Conference on Digital Audio Effects (DAFx 03)*, 2003.
- [63] D. Van Nort, “Noise/music and representations systems,” *Organised Sound*, vol. 11, no. 2, pp. 173–178, 2006.
- [64] X. Serra and J. O. Smith, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, vol. 14, no. 4, pp. 14–24, 1990.
- [65] T. S. Verma and T. H. Y. Meng, “Extending spectral modeling synthesis with transient modeling synthesis,” *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.
- [66] K. Fitz, L. Haken, and P. Christensen, “A New Algorithm for Bandwidth Association in Bandwidth-Enhanced Additive Sound Modeling,” in *Proc. of the 2000 International Computer Music Conference (ICMC 00)*, 2000.
- [67] M. Goodwin, “Residual modeling in music analysis-synthesis,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 96)*, vol. 2, pp. 1005–1008, May 1996.
- [68] P. Depalle, G. Garcia, and X. Rodet, “Tracking of partials for additive sound synthesis using hidden markov models,” in *Proceedings of the 1993 International Conference on Acoustics, Speech and Signal Processing (ICASSP 93)*, vol. 1, pp. 225–228, 1993.
- [69] M. Wright, J. Beauchamp, K. Fitz, X. Rodet, A. Röbel, X. Serra, and G. Wakefield, “Analysis/synthesis comparison,” *Organised Sound*, vol. 5, no. 3, pp. 173–189, 2000.
- [70] E. Métois, “Musical gestures and audio effects processing,” in *Proc. of 1998 International Conference on Digital Audio Effects (DAFx 98)*, 1998.

-
- [71] V. Verfaillie, M. Wanderley, and P. Depalle, “Mapping strategies for gestural and adaptive control of digital audio effects,” *Journal of New Music Research*, vol. 35, no. 1, pp. 71–93, 2006.
- [72] C. Traube, P. Depalle, and M. M. Wanderley, “Indirect acquisition of instrumental gesture based on signal, physical and perceptual information,” in *Proc. of 2003 Conference on New Interfaces for Musical Expression (NIME 03)*, 2003.
- [73] E. Wold, *Nonlinear Parameter Estimation of Acoustic Models*. PhD thesis, UC-Berkeley, 1987.
- [74] I. Choi, “Gestural primitives and the context for computational processing in an interactive performance system,” in *Trends In Gestural Control of Music* (M. M. Wanderley and M. Battier, eds.), IRCAM – Centre Pompidou, 2000.
- [75] R. S. Hatten, *Interpreting musical gestures, topics, and tropes: Mozart, Beethoven, Schubert*. Indiana University Press, 2004.
- [76] F. Delalande, “La gestique de gould: Éléments pour une sémiologie du geste musicale,” in *Glenn Gould Pluriel* (G. Guertin, ed.), pp. 85–111, Louise Courteau, 1988.
- [77] J. Chadabe, “Electronic music and life,” *Organised Sound*, vol. 9, no. 2, pp. 3–6, 2004.
- [78] C. Cadoz and M. M. Wanderley, “Gesture-Music,” in *Trends in Gestural Control of Music* (M. M. Wanderley and M. Battier, eds.), pp. 315–334, IRCAM – Centre Pompidou, 2000.
- [79] R. I. Godoy, E. Haga, and A. Refsum-Jensenius, *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop (GW 2005)*, ch. Playing “Air Instruments”: Mimicry of Sound-producing Gestures by Novices and Experts, pp. 256–267. Springer-Verlag, 2006.
- [80] D. Van Nort and M. Wanderley, “Control Strategies for Navigation of Complex Sonic Spaces,” in *Proc. of the 2007 International Conference on New Interfaces for Musical Expression (NIME 07)*, pp. 379–382, 2007.

-
- [81] R. I. Godoy, E. Haga, and A. Refsum-Jensenius, "Exploring music-related gestures by sound-tracing - a preliminary study," in *2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems*, (Leeds, UK), May 9-10 2006.
 - [82] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1-36, 1985.
 - [83] R. I. Godoy, "Motor-mimetic music cognition," *Leonardo*, vol. 36, no. 4, pp. 317-319, 2003.
 - [84] R. I. Godoy, "Gestural-Sonorous Objects: embodied extensions of Schaeffer's Conceptual Apparatus," *Organised Sound*, vol. 11, no. 2, pp. 149-157, 2006.
 - [85] P. Schaeffer, *Traité des Objets Musicaux: Essai Interdisciplines*. Paris: Éditions du Seuil, 1966.
 - [86] P. Schaeffer, *Solfège de l'Objet Sonore*. Paris: INA/GRM, 1998.
 - [87] M. Chion, *Guide des Objets Sonores*. Buchet/Chastel/INA-GRM, 1983.
 - [88] J. Ricard, *Towards Computational Morphological Description of Sound*. PhD thesis, Universitat Pompeu Fabra, September 2004.
 - [89] L. Thoresen, "Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology," *Organised Sound*, vol. 12, no. 2, pp. 129-141, 2007.
 - [90] V. Verfaille, C. Guastavino, and P. Depalle, "Perceptual evaluation of vibrato models," in *Proceedings of the 2005 Conference on Interdisciplinary Musicology (CIM 05)*, 2005.
 - [91] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Synthesizing sound textures through wavelet tree learning," *IEEE Computer Graphics and Applications*, vol. 22, no. 4, pp. 38-48, 2002.
 - [92] G. Strobl, "Parametric sound texture generator," Master's thesis, Technische Universität Graz, 2007.

- [93] N. St-Arnaud and K. Popat, "Analysis and synthesis of sound textures," in *Computational Auditory Scene Analysis* (D. F. Rosenthal and H. Okuno, eds.), pp. 293–308, Lawrence Erlbaum Association, 1998.
- [94] M. Athineos and D. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP '03)*, pp. 648–651, 2003.
- [95] D. Van Nort, "Texture modeling: Signal models and compositional approaches," in *Proceedings of the 2007 Conference of the Society for Music Perception and Cognition (SMPC 07)*, 2007.
- [96] J. M. Hajda, R. A. Kendall, E. C. Carterette, and M. L. Harschberger, *The Psychology of Music*, ch. Methodological issues in timbre research, pp. 253–306. AP Press, 2nd ed., 1999.
- [97] D. Smalley, "Defining Timbre - Refining Timbre," *Contemporary Music Review*, vol. 10, no. 2, pp. 35–48, 1994.
- [98] S. McAdams, "Psychological constraints on form-bearing dimensions in music," *Contemporary Music Review*, vol. 4, no. 1, pp. 181–198, 1989.
- [99] A. Schoenberg, *Harmonielehre*. Vienna: Universal Edition, 1922.
- [100] L. Teodorescu-Ciocanea, "Timbre versus spectralism," *Contemporary Music Review*, vol. 22, no. 1, pp. 87–104, 2003.
- [101] W. Gaver, "The sonicfinder: An interface that uses auditory icons," *Human Computer Interaction*, vol. 4, no. 1, pp. 67–94, 1989.
- [102] W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," *Journal of Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.
- [103] W. Gaver, "How do we hear in the world? explorations in ecological acoustics," *Journal of Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.
- [104] B. Gygi, *Factors in the Identification of Environmental Sound*. PhD thesis, Indiana University, 2001.

-
- [105] S. Lakatos, S. McAdams, and R. Caussé, “The representation of auditory source characteristics: simple geometric form,” *Perception and Psychophysics*, vol. 59, no. 8, pp. 1180–1190, 1997.
- [106] B. Giordano and S. McAdams, “Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates,” *Journal of the Acoustical Society of America*, vol. 119, pp. 1171–1181, February 2006.
- [107] D. Cohen and S. Dubnov, “Gestalt phenomena in musical texture,” in *Music, Gestalt and Computing: Studies in Cognitive and Systematic Musicology* (M. Lehman, ed.), pp. 386–405, Springer-Verlag, 1997.
- [108] D. Smalley, “Spectromorphology: Explaining sound-shapes,” *Organised Sound*, vol. 2, no. 2, pp. 107–126, 1997.
- [109] S. Emmerson, ed., *The Language of Electroacoustic Music*. Palgrave Macmillan, 1986.
- [110] N. Schnell and M. Battier, “Introducing composed instruments, technical and musicological implications,” in *Proc. of 2002 Conference on New Interfaces for Musical Expression (NIME 02)*, 2002.
- [111] J. Ricard and P. Herrera, “Using morphological description for generic sound retrieval,” in *Proc. of 2003 International Symposium on Music Information Retrieval (ISMIR 03)*, 2003.
- [112] G. Peeters, S. McAdams, and P. Herrera, “Instrument sound description in the context of mpeg-7,” in *Proc. of 2000 International Computer Music Conference (ICMC 00)*, 2000.
- [113] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the Institute of Phonetic Sciences 17*, pp. 97–110, 1993.
- [114] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the cuidado project,” tech. rep., IRCAM, 2004.

- [115] E. Terhardt, "Frequency analysis and periodicity detection in the sensation of roughness and periodicity pitch," in *Frequency analysis and periodicity detection in hearing* (R. Plomp and G. Smoorenburg, eds.), pp. 278–287, Leiden A. W. Sijtho, 1970.
- [116] D. Pressnitzer and S. McAdams, "An effect of the coherence between envelopes across frequency regions on the perception of roughness," in *Psychophysics, Physiology and Models of Hearing*, pp. 105–108, London: World Scientific, 1999.
- [117] P. Daniel and R. Weber, "Psychoacoustical roughness: Implementation of an optimized model," *Acustica*, vol. 83, pp. 113–123, 1997.
- [118] S. McAdams, "Perspectives on the contribution of timbre to musical structure," *Computer Music Journal*, vol. 23, pp. 85–102, 1999.
- [119] J. Opolko, F. Wapnick, *McGill University Master Samples [Compact Disc]*. Montreal, Quebec: McGill University, 1987.
- [120] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [121] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, 2004.
- [122] A. J. McDonald, A. J. G. Baumgaertner, G. J. Fraser, S. E. George, and S. Marsh, "Empirical mode decomposition of the atmospheric wave field," *Annales Geophysicae*, vol. 25, pp. 375–384, 2007.
- [123] P. Le, E. Ambikairajah, and V. Sethu, "Speech enhancement based on empirical mode decomposition," in *Signal Processing, Pattern Recognition, and Applications 2008*, 2008.
- [124] J. Nunes, S. Guyot, and E. Delechelle, "Texture analysis based on local analysis of the bidimensional empirical mode decomposition," *Machine Vision and Applications*, vol. 16, no. 3, pp. 177–188, 2005.

- [125] B. Behm and J. Parker, "Creating audio textures by samples: Tiling and stretching," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 04)*, vol. 4, pp. 317–320, 2004.
- [126] G. Rilling, P. Flandrin, and P. Goncalves, "On empirical mode decomposition and its algorithms," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP 03)*, 2003.
- [127] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proceedings - Royal Society. Mathematical, physical and engineering sciences*, vol. 460, no. 2046, pp. 1597–1611, 2004.
- [128] P. Heydarian and J. D. Reiss, "Extraction of long-term rhythmic structures using the empirical mode decomposition," in *122nd AES Convention*, May 5-8 2007.
- [129] P. Vassilakis, "Auditory roughness as means of musical expression," *Selected Reports in Ethnomusicology (Perspectives in Systematic Musicology)*, vol. 12, pp. 119–144, 2005.
- [130] D. Cabrera, S. Ferguson, and E. Schubert, "Psysound3: software for acoustical and psychoacoustical analysis of sound recordings," in *Proceedings of the 13th International Conference on Auditory Display*, pp. 356–363, 2007.
- [131] I. Xenakis, *Concret PH, from Xenakis: Electronic Music [Compact Disc]*. Electronic Music Foundation EMF003CD, 1958 (original recording date).
- [132] A. Boudraa, J. Cexus, and Z. Saidi, "EMD-based signal noise reduction," *International Journal of Signal Processing*, 2004.
- [133] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette, "Vibrato: Detection, estimation, extraction, modification," in *Proc. of 1998 International Conference on Digital Audio Effects (DAFx 98)*, 1998.
- [134] T. Wishart, *On Sonic Art*. Harwood Academic Publishers, 1996.
- [135] R. Erickson, *Sound Structure in Music*. Berkeley: University of California Press, 1975.

-
- [136] T. Wishart, *Red Bird [Compact Disc]*. Electronic Music Foundation EMF 022CD, 1980 (original recording date).
- [137] T. Wishart, *Tongues of Fire, from Voiceprints [Compact Disc]*. Electronic Music Foundation EMF029CD, 1995 (original recording date).
- [138] D. Terrugi, “Technology and musique concrète: the technical developments of the groupe de recherches musicales and their implication in musical composition,” *Organised Sound*, vol. 12, no. 3, pp. 213–231, 2007.
- [139] M. Wright, *The Shape Of an Instant: Measuring and Modelling Perceptual Attack Time with Probability Density Functions*. PhD thesis, Stanford University, 2008.
- [140] D. Levitin, S. McAdams, and R. L. Adams, “Control parameters for musical instruments: a foundation for new mappings of gesture to sound,” *Organised Sound*, vol. 7, no. 2, pp. 171–189, 2002.
- [141] V. Verfaillie, U. Zölzer, and D. Arfib, “Adaptive digital audio effects (A-DAFx) : A new class of sound transformations,” *IEEE transactions on audio, speech and language processing*, vol. 14, no. 5, pp. 1817–1831, 2006.
- [142] E. R. Miranda and M. Wanderley, *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Inc., 2006.
- [143] D. Van Nort, M. Wanderley, and P. Depalle, “On the choice of mappings based on geometric properties,” in *Proc. of 2004 Conference on New Interfaces for Musical Expression (NIME 04)*, pp. 87–91, 2004.
- [144] D. Wessel and M. Wright, “Problems and Prospects for Intimate Musical Control of Computers,” in *Proceedings of 2001 Conference on New Interfaces for Musical Expression (NIME 01)*, 2001.
- [145] J. Chadabe, “The limitations of mapping as a structural descriptive in electronic music,” in *Proceedings of 2002 Conference on New Interfaces for Musical Expression (NIME 02)*, 2002.
- [146] P. Doornbusch, “Composers’ view on mapping in algorithmic composition,” *Organised Sound*, vol. 7, no. 2, pp. 145–156, 2002.

-
- [147] M. Babbitt, "Twelve-tone rhythmic structure and the electronic medium," *Perspectives of New Music*, vol. 1, no. 1, pp. 49–79, 1962.
- [148] A. Milne, W. Sethares, and J. Plamondon, "Isomorphic controllers and dynamic tuning: Invariant fingering over a tuning continuum," *Computer Music Journal*, vol. 31, no. 4, pp. 15–32, 2007.
- [149] D. Van Nort and M. Wanderley, "Exploring the Effect of Mapping Trajectories on Musical Performance," in *Proc. 2006 Sound and Music Computing Conference (SMC 06)*, pp. 19–24, 2006.
- [150] S. Axler, *Linear Algebra Done Right*. Springer, 1995.
- [151] E. Métois, *Musical Sound Information*. PhD thesis, Massachusetts Institute of Technology, 1996.
- [152] H. Hülsen, *Self-Organising Locally Interpolating Maps in Control Engineering*. Dr.-Ing., Universität Oldenburg, 2007.
- [153] R. Jacob, L. Sibert, D. McFarlane, and M. Preston Mullen, Jr., "Integrality and separability of input devices," *ACM Transactions on Computer-Human Interaction*, vol. 1, no. 1, pp. 3–26, 1994.
- [154] U. Axen, *Topological Analysis Using Morse Theory and Auditory Display*. PhD thesis, University of Illinois at Urbana Champaign, 1998.
- [155] T. Kohonen, *Self-Organizing Maps*. Springer-Verlag, 2001.
- [156] C. Goudeseune, *Composing with Parameters for Synthetic Instruments*. PhD thesis, University of Illinois at Urbana Champaign, 2001.
- [157] A. Momeni and C. Henry, "Dynamic independent mapping layers for concurrent control of audio and video synthesis," *Computer Music Journal*, vol. 30, no. 1, pp. 49–66, 2006.
- [158] B. Schoner, C. Cooper, C. Douglas, and N. Gershenfeld, "Data-Driven Modeling of Acoustical Instruments," *Journal of New Music Research*, vol. 28, no. 2, pp. 81–89, 1999.

-
- [159] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Muller, "FTM - Complex Data Structures for Max," in *Proc. of 2005 International Computer Music Conference (ICMC 05)*, 2005.
- [160] R. Bencina, "Audiomulch." <http://www.audiomulch.com>, *Last Accessed June 11, 2009*.
- [161] R. Sibson, "A brief description of natural neighbor interpolation," in *Interpreting Multivariate Data*, pp. 21–36, John Wiley, 1981.
- [162] C. DeTar, "Color blobs." <http://tirl.org/software/colorblobs/>, *Last Accessed June 11, 2009*.
- [163] G. Farin, *Nurbs: From Projective Geometry to Practical Use*. Natick, Massachusetts: A.K. Peters, Ltd., 1999.
- [164] S. Omohundro, "The Delaunay Triangulation and Function Learning," *Intl. Computer Science Institute Tech Report*, pp. 1–10, 1990.
- [165] A. Talmi and G. Gilat, "Method for smooth approximation of data," *Journal of Computational Physics*, vol. 23, pp. 93–123, 1977.
- [166] J. Duchon, "Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces," *R.A.I.R.O. Anal. Num.*, vol. 10, pp. 5–12, 1976.
- [167] L. Mitas and H. Mitasova, "General variational approach to the interpolation problem," *Comp. Math. Applications*, vol. 16, pp. 983–992, 1988.
- [168] H. Mitasova and L. Mitas, "Interpolation by Regularized Spline with Tension: I. Theory and Implementation," *Mathematical Geology*, vol. 25, no. 6, pp. 641–655, 1993.
- [169] H. Mitasova and L. Mitas, "Interpolation by Regularized Spline with Tension: II. Application to Terrain Modeling and Surface Geometry Analysis," *Mathematical Geology*, vol. 25, no. 6, pp. 657–669, 1993.
- [170] D. Van Nort and M. Wanderley, "The LoM Mapping Toolbox for Max/MSP/Jitter," in *Proc. of the 2006 International Computer Music Conference (ICMC 06)*, pp. 397–400, 2006.

-
- [171] C. Goudeseune, "Simplicial interpolation." <http://zx81.isl.uiuc.edu/interpolation/>, *Last Accessed June 11, 2009*.
- [172] W. Buxton, "The Haptic Channel," in *Readings in Human-Computer Interaction*, pp. 357–365, Morgan Kaufmann Publishers, 1987.
- [173] S. Card, J. Mackinlay, and G. Robertson, "A Morphological Analysis of the Design Space of Input Devices," *ACM Transactions On Information Systems*, vol. 9, no. 2, pp. 99–122, 1991.
- [174] R. Vertegaal, T. Ungvary, and M. Kieslinger, "Towards a Musician's Cockpit: Transducers, Feedback and Musical Function," in *Proc. of 1996 International Computer Music Conference (ICMC 96)*, pp. 181–184, 1996.
- [175] M. Wanderley, J. Viollet, F. Isart, and X. Rodet, "On the Choice of Transducer Technologies for Specific Musical Functions," in *Proc. of the 2000 International Computer Music Conference (ICMC 00)*, pp. 244–247, 2000.
- [176] N. Orio, N. Schnell, and M. Wanderley, "Input Devices for Musical Expression: Borrowing Tools from HCI," in *Proc. of 2001 Conference on New Interfaces for Musical Expression (NIME 01)*, 2001.
- [177] T. Jehan, A. Freed, and R. Dudas, "Musical Applications of New Filter Extensions to Max/MSP," in *Proceedings of the 1999 International Computer Music Conference (ICMC 99)*, 1999.
- [178] D. Van Nort, "Noise/nature shift : balancing attention towards noise optimization," Master's thesis, Rensselaer Polytechnic Institute, 2003.
- [179] J. Ryan, "Effort and expression: Some notes on instrument design at steim," in *Proceedings of the 1992 International Computer Music Conference (ICMC 92)*, 1992.
- [180] S. Olejnik and J. Algina, "Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs," *Psychological Methods*, vol. 8, no. 4, pp. 434–447, 2003.

-
- [181] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 2nd ed., 1988.
- [182] M. Barthet, R. Kronland-Martinet, and S. Ystad, “Consistency of Timbre Patterns in Expressive Music Performance,” in *Proc. of 2006 International Conference on Digital Audio Effects (DAFx 06)*., pp. 19–24, 2006.
- [183] M. Barthet, R. Kronland-Martinet, and S. Ystad, “Improving musical expressiveness by time-varying brightness shaping,” in *Computer Music Modeling and Retrieval. Sense of Sounds: 4th International Symposium, CMMR 2007, Copenhagen, Denmark, August 27-31, 2007. Revised Papers*, 2008.
- [184] M. Wright, D. Wessel, and A. Freed, “New Musical Control Structures from Standard Gestural Controllers,” in *Proc. of the 1997 International Computer Music Conference (ICMC 97)*, pp. 87–89, 1997.
- [185] C. Roads, *Microsound*. MIT Press, 2001.
- [186] B. Maubrey, “Audio jackets and other electroacoustic clothes,” *Leonardo*, vol. 28, no. 2, pp. 93–97, 1995.
- [187] G. Weinberg, M. Orth, and P. Russo, “The embroidered musical ball: a squeezable instrument for expressive performance,” in *CHI ’00: CHI ’00 extended abstracts on Human factors in computing systems*, (New York, NY, USA), pp. 283–284, ACM Press, 2000.
- [188] D. Van Nort, D. Gauthier, S. Xin Wei, and M. Wanderley, “Extraction of Gestural Meaning from a Fabric-Based Instrument,” in *Proc. of the 2007 International Computer Music Conference (ICMC 07)*, pp. 441–444, 2007.
- [189] X.S. Labs. <http://www.xslabs.net>, *Last Accessed August 25, 2009*.
- [190] Arduino. <http://www.arduino.cc/>. *Accessed August 25, 2009*.
- [191] S. Xin Wei, M. M. Wanderley, F. Abtan, D. Birnbaum, D. Gauthier, R. Koehly, E. Sinyor, and D. Van Nort, “WYSIWYG: Wearable Sounds, Gestural Instruments.” *Submitted for Publication*.

-
- [192] D. Birnbaum, F. Abtan, S. Xin Wei, and M. Wanderley, "Mapping and dimensionality of a cloth-based sound instrument," in *Proc. 2007 Sound and Music Computing Conference (SMC 07)*, pp. 386–389, 2007.
- [193] C. Bascou and L. Pottier, "GMU, A Flexible Granular Synthesis Environment in Max/MSP," in *Proceedings of the Sound and Music Computing Conference 2005*, 2005.
- [194] T. Kailath, *Linear Systems*. Prentice-Hall, 1980.
- [195] H. Thornburg, *Detection and Modeling of Transient Audio Signals with Prior Information*. PhD thesis, Stanford University, 2005.
- [196] G. H. Hostetter, "Recursive discrete fourier transformation," *IEEE transactions on Audio, Speech and Signal Processing*, vol. 28, no. 2, pp. 184–190, 1980.
- [197] H. D. Thornburg and R. J. Leistikow, "Analysis and resynthesis of quasi-harmonic sounds: An iterative filterbank approach," in *Proc. of 2003 International Conference on Digital Audio Effects (DAFx 03)*, 2003.
- [198] Y. Qi, T. P. Minka, and R. W. Picard, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *Proceedings of the 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP 02)*, 2002.
- [199] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *13th European Sig. Proc. Conf.*, (Antalya, Turkey), 2005.
- [200] S. Dubnov, N. Tishby, and D. Cohen, "Influence of frequency modulating jitter on higher order moments of sound residual with applications to synthesis and classification," in *Proc. of 1996 International Computer Music Conference (ICMC 96)*, pp. 378–385, 1996.
- [201] D. Van Nort and P. Depalle, "A Stochastic State-Space Phase Vocoder for Synthesis of Roughness," in *Proc. of 2006 International Conference on Digital Audio Effects (DAFx 06)*, pp. 177–180, 2006.

-
- [202] S. Tomaszic, “On short-time fourier transform with single-sided exponential window,” *Comp. Math. Applications*, vol. 55, no. 2, pp. 141–148, 1996.
- [203] S. Tassart, “Infinite length windows for short-time fourier transform,” in *Proceedings of the 1998 International Computer Music Conference (ICMC 98)*, 1998.
- [204] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transaction of the ASME – Journal of Basic Engineering*, pp. 35–45, 1960.
- [205] G. Welch and G. Bishop, “An Introduction to the Kalman Filter,” Tech. Rep. TR 95-041, University of North Carolina, Department of Computer Science, 1995.
- [206] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice*. Englewood Cliffs: Prentice Hall, 1993.
- [207] M. Morf, G. S. Sidhu, and T. Kailath, “Some new algorithms for recursive estimation in constant, linear, discrete-time systems,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 315–323, 1974.
- [208] A. Moghaddamjoo and R. L. Kirlin, “Robust Adaptive Kalman Filtering with Unknown Inputs,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 8, 1989.
- [209] S. Roucos and A. Wilgus, “High Quality Time-Scale Modification of Speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 85)*, pp. 236–239, 1985.
- [210] D. Hejna and B. R. Musicus, “The SOLAFS Time-Scale Modification Algorithm,” tech. rep., BBN, 1991.
- [211] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 93)*, pp. 554–557, 1993.
- [212] D. Menzies, “Composing instrument control dynamics,” *Organised Sound*, vol. 7, no. 3, pp. 255–266, 2002.

-
- [213] A. de Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.