

Audio Segmentation, Classification and Visualization

A thesis submitted to
Auckland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Jessie Xin Zhang

2009

School of Computing and Mathematical Sciences

Primary Supervisor: Dr. Jacqueline Whalley

Table of Contents

Attestation of Authorship.....	xii
Acknowledgements.....	xiii
Abstract.....	xiv
Preface: Organization of this thesis.....	xv
Chapter 1 Introduction	1
1.1 Motivation and Research Objectives.....	1
1.2 Challenges in Audio Visualization.....	5
1.3 Audio Visualization Literature Review.....	7
1.3.1 Audio Visualization Introduction	7
1.3.2 The Visualization of Audio Properties	9
1.3.3 Auditory-Visual Associations.....	11
1.3.4 Visualizing the Structure of an Audio File	12
1.3.5 Music Visualization.....	13
1.3.6 Speech Visualization	15
1.3.7 The Visualization of Audio Files using Images	16
1.3.8 Audio Database Visualization	17
1.4 Summary	20
Chapter 2 The Audio Visualization System Framework	22
2.1 Development Methodology.....	22
2.2 The Audio Visualization System Framework	23
2.3 The Development Environment	25
Chapter 3 Audio Segmentation.....	26
3.1 The Audio Segmentation Module	26
3.2 Literature Review	29
3.3 A Novel Two-Phase Audio Segmentation Method for General Audio Files.....	34
3.3.1 A Framework for the Two-Phase Audio Segmentation Method.....	34
3.3.2 Phase One – Silence Detection	35
3.3.3 Phase Two – Edge Detection within Self-similarity Maps.....	38
3.4 Evaluation Method	47
3.5 Experiments and Analysis	49
3.6 Chapter Summary.....	61
Chapter 4 Audio Classification.....	62
4.1 Introduction to the Classification Module.....	62

4.2 Literature Review	64
4.3 Training and Learning for the Classification Module	71
4.4 General Classification Method	74
4.4.1 NFL Method in Audio Classification	74
4.4.2 Cross-Validation for the Classifier and Feature Set Selection	75
4.4.3 Feature Set and Accuracy Experiments	76
4.5 Introducing a New Class Detection Method	79
4.5.1 Overview	80
4.5.2 Parameters Used in New Class Detection	81
4.5.3 Evaluation of the "New Class Detection" Method	83
4.5.4 The <i>Uncertain</i> Criterion: "New Class Detection" Experiments	86
4.6 Hierarchical Classification	89
4.7 Classification of Mixed Sound Audio Files	92
4.8 Chapter Summary	97
Chapter 5 Visualization Approach 1- Time Mosaics	98
5.1 The Framework for Time Mosaic Generation	99
5.2 Literature Review: Image Processing	101
5.3 Audio/Image Database for Testing	105
5.4 Image Tile Generation	107
5.5 Interpretable Audio Features	108
5.6 Audio-Visual Feature Mapping	109
5.7 Image Tile Generation	111
5.8 Mosaic Time-Line Generation	114
5.9 Template Image Selection	120
5.10 Experimental Results	124
5.11 Limitations	128
5.12 Other Usages for Time Mosaics	129
5.13 Summary	132
Chapter 6 Visualization Approach 2-Video Texture	133
6.1 Limitations of Image Mosaics	134
6.2 Video Texture Literature Review	137
6.3 Video Texture Generation	141
6.3.1 The Video Texture Module Framework	141
6.3.2 Video Texture Generation Based on Random Transitions	144
6.3.3 Adaptive Video Texture Generation Method Based on Audio Matching	149

6.3.4 Template Video Limitations	163
6.3.5 Video Texture Clip Duration	167
6.3.6 Mapping Audio Features	167
6.4 Blended Video Texture Mosaic Generation	168
6.4.1 Parallel and Sequential Video Texture Mosaic Generation Methods.....	169
6.4.2 Blended Video Texture Mosaic for Hierarchical Databases	174
6.5 Summary	178
Chapter 7 Summary and Future Work	180
7.1 Summary	180
7.2 Limitations.....	183
7.3 Future work	185
7.3.1 Audio Segmentation and Classification.....	185
7.3.2 Audio Visualization	187
7.4 User Studies.....	190
References	192
Appendix A Audio feature extraction	207
Appendix B Related publications by the author.....	214

List of Figures

Figure 1-1: Audio database query.....	3
Figure 1-2: An example of a time mosaic audio visualization.	4
Figure 1-3: Relevant research areas of audio visualization.	5
Figure 1-4: The first audio visualization device – a Phonautograph.	8
Figure 1-5: An Oscilloscope (L) and a spectrogram (R) for a cat's meow.	8
Figure 1-6: <i>TimbreGrams</i> for speech and classical music [8].	9
Figure 1-7: Visualization results for single notes on various instruments in phase space [3].....	10
Figure 1-8: Visualization of two songs using chromatic graphs (in terms of colouring) [9].	14
Figure 1-9: (L) Graphic matching in speech learning [32]; (R) Clock visualization for conversation [33].	15
Figure 1-10: Sound visualizations: (L) from method in [38]; (R) from method in [11]	17
Figure 1-11: (L) Music collection visualization result by "Islands of Music" [43]; (R) Content-based exploration of music archives [44].	18
Figure 1-12: Resultant visualization of audio file query results from system in [47].	19
Figure 1-13: Browser of selected audio files to show their distances (differences) in [47].	19
Figure 1-14: Visualizing audio database using the Sonic Browser [48].	20
Figure 1-15: Audio database structure visualization results (L: in [48]; R: in [49]).	20
Figure 2-1: Framework of audio visualization system.....	23
Figure 2-2: Processes of visualizing an audio input.	24
Figure 3-1: An overview of the segmentation process.....	27
Figure 3-2: Categories for existing segmentation methods.	29
Figure 3-3: Framework for the novel two-phase audio segmentation method.	34
Figure 3-4: Wave shape of audio file "Hands clapping" and its SP, RMS; (Inset) Enlarged wave shape of sound 7 in "Hands clapping".....	35
Figure 3-5: Correct audio segmentation result for a mixed sound audio file.	37
Figure 3-6: Top: (L) Similarity map image; (R) Edge detection of similarity map image; Bottom: Segmentation result based on similarity map image.	39

Figure 3-7: Result of segmenting the audio file (Figure 3-5) using Euclidean-d similarity segmentation method.....	40
Figure 3-8: Segmentation result for long and short signals.	41
Figure 3-9: The comparison of the similarity map with and without adding silence frames..	42
Figure 3-10: (L) The comparison of $d_{\cosine}(a,b)$ and $d_{Angle}(a,b)$. (R) The comparison for distances d_{log} , d_{Sine} , and d_{Linear}	46
Figure 3-11: Visualization of distance $d_{Chen}(a,b)$ and $d_{Linear-Chen1}(a,b)$	47
Figure 3-12: Comparison of different frame sizes (16ms and 32ms).	52
Figure 3-13: Comparison of different audio feature sets.	55
Figure 3-14: Segmentation experiment results using 5 different methods.	58
Figure 3-15: The starting audio file amplitude vs. frames (16ms).....	59
Figure 3-16: Segmentation results for audio shape in Figure 3-15 using our two-phase method.	60
Figure 4-1: Scheme of classification module.....	63
Figure 4-2: Framework for the classification module.....	72
Figure 4-3: Training/Learning process for feature set selection.	72
Figure 4-4: Training/Learning process for parameter selection and automatic threshold determination in new class detection.	73
Figure 4-5: Generalization of two feature points x_1 and x_2 to the feature line $\overline{x_1x_2}$	75
Figure 4-6: Parameters $NFLd$, $RCCd$ and $NCCd$ used in "new class detection" experiments.....	81
Figure 4-7: Parameter $Fnum$ used in "new class detection" experiments.....	81
Figure 4-8: Expanded $Fnum$	82
Figure 4-9: LOFO and LOCO experiments.	83
Figure 4-10: Audio file classification using the <i>uncertain</i> criterion.....	87
Figure 4-11: <i>VisualData</i> ontology overview.....	90
Figure 4-12: An example of an audio file that is incorrectly classified.	93
Figure 4-13: Incorrectly classified middle segment that is not identified as anomalous.	96
Figure 5-1: Framework for the time mosaic generation.....	99
Figure 5-2: Image tiles placed according to their corresponding time sequences.	100
Figure 5-3: Blended image mosaic generated with image tiles of the same size.....	100
Figure 5-4: The mapping matrix of chromatic synthesis [174].	102
Figure 5-5: Legend for visualized features in the audio visualization system.....	110

Figure 5-6: Algorithm to represent the audio noise-to-signal ratio.	113
Figure 5-7: Result of image tiles with width of each image used to represent its corresponding audio clip duration.	115
Figure 5-8: Background texture generation.	116
Figure 5-9: Steps for background texture image generation.	118
Figure 5-10: Generation of time mosaic images for the audio in Figure 5-2.	120
Figure 5-11: Bird-dog-cat example with alternate template images.	121
Figure 5-12: Bird-dog-cat example with alternate template images.	121
Figure 5-13: Hierarchical structure of the <i>VisualData</i> for template image selection.	123
Figure 5-14: Bird-dog-cat example with filtering.	124
Figure 5-15: Time mosaic image for an audio file different from that in Figure 5-14.	125
Figure 5-16: Frog-bee-cow example.	126
Figure 5-17: An audio file containing sounds from class "bee", "cow" and "frog"	126
Figure 5-18: An audio file containing duck and rooster sounds.	126
Figure 5-19: A time mosaic of five images representing five different sounds.	127
Figure 5-20: A time mosaic for an audio file containing three distinct instrument sounds (oboe, trombone and cello).	127
Figure 5-21: A time mosaic for a music audio file containing a violin and oboe sounds.	127
Figure 5-22: Incorrect segmentation and classification produce an incorrect result.	128
Figure 5-23: Images representing audio files in the "dog" class from the <i>VisualData</i> database.	130
Figure 5-24: Filtered bird-dog-cat example with different mapping relationships.	131
Figure 6-1: Image mosaic generation when the durations of the component audio clips are different;	135
Figure 6-2: Difference between audio clips visualized by time mosaic module and video texture module.	142
Figure 6-3: Video textures generation method based on random transition.	144
Figure 6-4: Cross-fading method from Frame A to Frame B.	147
Figure 6-5: Pseudo code for video texture synthesis based on random transitions and three functions.	149
Figure 6-6: Framework for adaptive video textures generation method.	151

Figure 6-7: Example of a template processing result.....	152
Figure 6-8: An example input audio file (dog barking).....	153
Figure 6-9: Frame images from the template video.....	155
Figure 6-10: Resultant frames by Frame-based likelihood method.....	155
Figure 6-11: 3D feature vector curve similarity distance calculation.....	158
Figure 6-12: Similarity distance curve of feature vector curve similarity.....	159
Figure 6-13: Absolute length for audio feature vectors in template and target audio piece.....	160
Figure 6-14: An example of visualizing audio with video textures.....	161
Figure 6-15: Single sound input result.....	162
Figure 6-16: Structures of input videos which are suitable for video textures generation.....	164
Figure 6-17: Structures of video frames which are not suitable for video textures generation.....	165
Figure 6-18: Random structure of input video.....	166
Figure 6-19: Combination of video texture components.....	169
Figure 6-20: A frame generated from three individual frames using Poisson image editing.....	170
Figure 6-21: Parallel blended video mosaics results with audio features;.....	172
Figure 6-22: Sequential blended video mosaics results with audio features;.....	173
Figure 6-23: Combination of video texture components for a hierarchical database.....	175
Figure 6-24: Visualization using a hierarchically structured database and video components.....	176
Figure 6-25: A parallel blended video texture mosaic result for sounds from two parent classes.....	177
Figure 7-1: Three sounds misrepresented as two images tiles.....	184
Figure 7-2: Overlapping audio signals producing an incorrect result.....	184
Figure 7-3: Fold-line division for the uncertain criterion.....	186
Figure 7-4: Audio visualization: (L) with wave shape illustration; (R) using curve shapes.....	187
Figure 7-5: 3D audio visualization: (L) cylinder mapping; (R) Z-values mapped to feature.....	188
Figure 7-6: Using image elements to represent audio clips.....	189
Figure 7-7: Element mosaic time-lines result.....	189

Figure 7-8: Schematic diagram of the speech visualization..... 190

List of Tables

Table 3-1: Existing equations for similarity between two vectors.....	44
Table 3-2: New methods for calculating the similarity between two vectors.....	45
Table 3-3: The 410 audio files and 16 classes in the <i>MuscleFish</i> audio database.	48
Table 3-4: % average accuracy of segmentation methods using different frame sizes.....	51
Table 3-5: Average accuracy of segmentation methods using different normalization approaches.	52
Table 3-6: Accuracy of the different segmentation methods by audio file group.	53
Table 3-7: The three new feature sets tested.....	55
Table 3-8: Percentage average accuracy using four different feature sets and seven different vector distance calculations in the two-phase method.....	55
Table 3-9: The percentage average accuracy of segmentation using the <i>VisualData</i> database.	57
Table 4-1: Number of correct classifications for 32 audio feature sets using <i>MuscleFish</i> database (n = 410).	77
Table 4-2: Number of correct classifications for 32 audio feature sets using <i>VisualData</i> database (n = 611).	78
Table 4-3: Number of files incorrectly classified for the LOFO and LOCO experiments and the <i>MuscleFish</i> database using various parameter sets.	85
Table 4-4: Number of files incorrectly classified using various parameter sets with the <i>VisualData</i> database.	86
Table 4-5: Results for LOFO and LOCO experiments using the <i>uncertain</i> criterion and the <i>MuscleFish</i> database (<i>NFLd</i> + <i>NCCd</i>).	88
Table 4-6: Results for LOFO and LOCO experiments using the <i>uncertain</i> criterion and the <i>VisualData</i> database (<i>NFLd</i> + <i>NCCd</i>).	89
Table 4-7: Number of incorrectly classified files for LOFO, using parent level classes (n = 611).	91
Table 4-8: Number of incorrectly classified files for LOFO, using child level classes.	91
Table 4-9: Total number of correctly classified files after two passes.	91
Table 4-10: Classification accuracy for mixed sound audio files.	93

Table 5-1: Comparison of <i>MuscleFish</i> and <i>VisualData</i> Ontological Structures.	106
Table 5-2: Accuracies for the segmentation and classification for <i>MuscleFish</i> and <i>VisualData</i>	106

Attestation of Authorship

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted to the award of any other degree or diploma of a university or other institution of higher learning."

Signed: _____

Acknowledgements

I am very grateful to the many people who encouraged me to begin, continue and finish my study at AUT. First and foremost, I am deeply grateful to my supervisors Dr. Jacqueline Whalley and Dr. Stephen Brooks, for their invaluable suggestions, support, and guidance throughout my period of research. The work presented in this thesis owes much to their enthusiasm and careful guidance.

Dr. Jacqueline Whalley has provided timely guidance and support since the initial stages of this research. Her supervision covered every detail of my study. She offered considerable extra effort as English is my second language. This thesis would not have been possible without her constant support, valuable discussions and detailed reviews. I appreciated all her effort and valuable time.

Thanks to Dr. Stephen Brooks for providing the initial inspiration for the research topic and for the opportunities he provided for me to undertake some of the research under his supervision at Dalhousie University. He has provided valuable guidance and direction at various stages of this research and has generously provided partial financial support towards my attendance at conferences.

David Whalley, I am deeply grateful for the considerable thought and time you put into proofreading my thesis. Thank you for being so patient in correcting any mistakes I've made. I would also like to acknowledge my gratitude to Dr. Qun Song for valuable discussions about the development of the system in this thesis. My thanks go to Felix Suwen Wang and Waseem Ahmad for spending much time reading my work.

Thanks to AUT and SERL for funding my study and providing financial support towards the presentation of papers at overseas conferences. I am grateful to the students in the PhD lab for providing a wonderful community in which to learn and conduct research.

Finally, I would like to add personal thanks to my mother Xuejun Li for her endless encouragement and support, and to my husband Peiliang Zhang for his love and support. Thank you, Mum, you have given me the strength to keep going and to get where I am today.

Abstract

This thesis presents a new approach to the visualization of audio files that simultaneously illustrates general audio properties and the component sounds that comprise a given input file. New audio segmentation and classification methods are reported that outperform existing methods. In order to visualize audio files, the audio is segmented (separated into component sounds) and then classified in order to select matching archetypal images or video that represent each audio segment and are used as templates for the visualization. Each segment's template image or video is then subjected to image processing filters that are driven by audio features. One visualization method reported represents heterogeneous audio files as a seamless image mosaic along a time axis where each component image in the mosaic maps directly to a discovered component sound. The second visualization method, video texture mosaics, builds on the ideas developed in time mosaics. A novel adaptive video texture generation method was created by using acoustic similarity detection to produce a resultant video texture that more accurately represents an audio file. Compared with existing visualization methods such as oscilloscopes and spectrograms, both approaches yield more accessible illustrations of audio files and are more suitable for casual and non expert users.

Preface

Organization of this Thesis

This thesis presents research in the field of content-based audio visualization. The research reported in this thesis is cross disciplinary so the in-depth review of literature of particular relevance to each specific aspect of the research is presented in the applicable chapter.

Chapter 1 introduces the reader to the existing literature in the field of general audio visualization. The motivation for this research is discussed and the research objectives are introduced.

Chapter 2 begins with a description of the methodology used. Then the reader is introduced to the framework of our audio visualization system and the development environment used.

Chapter 3 introduces relevant literature in the field of audio segmentation. Subsequently the development and evaluation of a novel 2-phase audio segmentation method is detailed.

Chapter 4 covers the current literature on audio classification and other relevant classification methods. The development and evaluation of an accurate, adaptive classification method with new class detection is described.

Chapters 5 and 6 describe two alternate and complementary methods for visualizing audio; using images and video textures. The chapters also review the relevant literature on image processing and video texture generation respectively.

Chapter 7 analyzes the audio visualization system and proposes potential avenues for future research and development.