http://cnmat.berkeley.edu/publications/exploration-real-time-visualizations-musical-timbre

An Exploration of Real-Time Visualizations of Musical Timbre

Kai Siedenburg
kai.siedenburg@gmx.de
CNMAT - The Center for New Music and Audio Technologies
University of California, Berkeley

August 13, 2009

Abstract

This study explores real-time visualizations of timbre-related audio features using the music programming environment *Max/MSP/Jitter*. Following the approach of exploratory data-analysis, we present a palette of different visualization techniques, each one providing a different perspective on the feature-data. Additionally, we introduce a simple notion of timbral similarity, which can be used in real-time performance situations. The visualizations are further used to inform the control of audio effects by feature trajectories. Finally, we present a brief analysis of a section of Gerard Grisey's *Modulations*, which aims to combine score oriented methods with acoustical analysis of recordings.

1 Introduction

"The role of timbre has extended to that of central subject of the music. Then, paradoxically, the very notion of timbre, this catchall, multidimensional attribute with a poorly defined identity, gets blurred, diffuse and vanishes into the music itself." - J.-C. Risset & D. Wessel [1]

Timbre - most often defined by what it is not - is central to 20th and 21st century music. Ironically, the more the phenomenon of timbre is understood by research lately, the more complex timbral structures are composed and the more problematic gets its notion as a clearly defined musical parameter. In today's music, the relevance of timbre reaches far beyond sound objects which have a clearly defined pitch. As Tristan Murail notes, *"the new materials that offer themselves to the composer [...] are often complex sounds, intermediate sounds, hybrids, sounds that possess new dimensions (transitions, development over time), sounds that are neither harmonic complexes nor timbres but something between the two."* The *"Revolution of Complex Sounds"* [2] - starting in the 1950s in the electronic studios and firstly becoming obvious in instrumental music with Ligeti and during the beginning of the french spectral movement - challenges the multidimensionality of timbre, the difference between timbre and harmony and the possibilities of its manipulation.
Research on the complex acoustical phenomenon of timbre is naturally divided into different fields, each one taking a different perspective: sound synthesis, physical modeling, timbre-perception and cognition, timbre-based music information retrieval, composition. However, all fields more or less depend on, and contribute to the understanding of timbre-perception, since timbral information in the aether is only relevant in respect to its perception. An example for this is the *"Analysis by Synthesis"* paradigm, which has lately been reversed by work on Feature Modulation Synthesis into *"Synthesis by Analysis"* [3]. These are approaches to close the apparent gaps between the field-specific perspectives on timbre.

This project follows a similar hands-on approach: *timbre analysis by visualization, musical control by analysis.*. The paper focusses on the key points of the project: audio feature based timbre-modeling, -visualization and -manipulation.[1]

[1]The notion of timbre refers in this context to single instruments as well as to complex textures.

The *analysis by visualization*-approach is inspired by the methodology of exploratory data analysis in statistics: exploring underlying structures of high dimensional data by graphical representation. This serves a two-fold purpose: gaining experience of the sound-dependent behavior and correlations of the visualized audio features, as well as contributing to the understanding of physical properties of certain sounds, as reflected by the respective audio features. In the work done so far, a palette of different real-time feature visualizations was programed in the music programming environment *Max/MSP/Jitter*. In the following, we mention briefly previous research done in this field and explain the main parts of the project and its main result, the *Max*-program. Section three is an example of using the program in music analysis. Section four outlines promising areas for future research in this field.

2 Structure of the Program

For the sake of efficiency, the program was divided into its main functional parts. It consists of four communicating subprograms (patches), namely *Feature Extraction*, *Visualization*, *Similarity*, and *Audio Control* which can be run separately. The latter three patches receive and process the multivariate feature-time-series from the initial extraction patch. For all visualizations holds that the selection of the respective feature is variable, as well as its scaling. This means that the patch could be used for any kind of multivariate time-series visualization.

2.1 Feature Extraction

Starting with a powerful set of audio features is naturally a prerequisite for successful Music Information Retrieval (MIR). Many studies in MIR, dealing with timbre-classification, segmentation and music-similarity, start with a set of about 10-20 Mel frequency cepstral coefficients (MFCC). McKinney and Breebart [4] demonstrate the performance of auditory-filter-based features and their fluctuations in different bands. Jensen [5] uses a gaussian-smoothed, bark-based representation of the spectrum for extracting timbral information. Moerchen et al. propose nonlinear transformations to unskew the feature distributions of low level features [6], a method that is adapted in this study.[2]

[2]Interestingly, their evaluation of a large set of features shows that psycho-acoustically weighted features correlate highly with their unweighted counterparts, a finding that con-

Pachet and Roy suggest automatically generated 'analytical' features for supervised classification tasks [7]. However, for the purpose of direct visualization, it is necessary to use features which are 'meaningful', i.e. which refer to a more or less intuitive property of the spectrum.
The *Feature Extraction* patch extracts the multivariate series of features from *MSP* - audio in real-time and sends it to the other patches. Additionally, the time series can be written into a text file, which can be the starting point for data-analysis in programs such as *MATLAB*. Three different kinds of features are extracted. For most low-level features, we use the real-time sound description library **zsa.descriptors** by Emmanuel Jourdan and Mikhail Malt [8]. We use its features *(spectral) -centroid, -spread, -skewness, -kurtosis, -flux, rolloff* (default 95%), *-slope, -decrease*. Additionally, second order features, *bark-flux, bark-flux-flux, centroid-flux, peaks above threshold, (bark based) spectral crest factor, centroid-fund-ratio*, are computed, using the zsa.descriptors data. For higher level perceptual features, such as *Loudness, Noisiness, Roughness and Fund(amental frequency)* we use CNMAT's **Analyser** -object. The default settings of the FFT are a window size of 512 samples with 50% overlap.

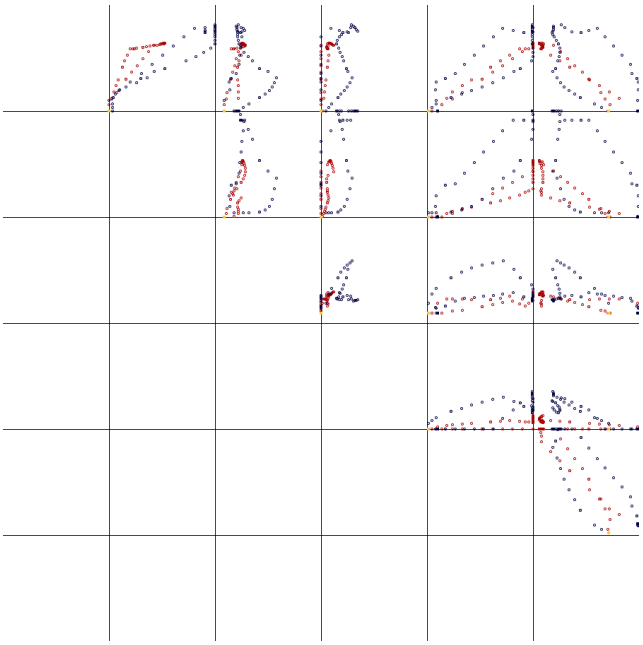
2.2 Visualization

Besides being of interest in other fields, feature visualization can be a valuable tool for music analysis. In music-analysis, dealing with sound itself is difficult, since its traditional methods are mainly based on the representation of music by a score. This is especially true for electronic music, where often no scores exists at all. Jean-Claude Risset notes: *"The lack of an objective representation makes it difficult to study these works."*[9] Visual representations based on the trajectories of audio features of recordings are an attempt to generate "objective" representations of pieces. This is what Stephen Malloch suggests in [10], using trajectories of perceptually oriented features to inform the analysis of a critical section of Ligeti's *Atmospheres*. Unfortunately, his set of descriptors is small and his visualizations rather unappealing.

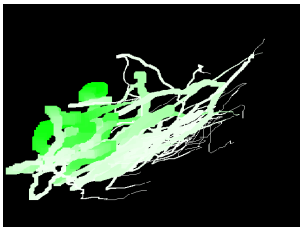
In the patch *Visualization*, several real-time tools are developed, each providing a different perspective onto the data. [3]

tradicts the tenor of [4].

[3]The following graphics are of course more valuable in the form of videos, tied to the music that generated them. Compare the short video-documentation of the project.



(a) Scatterplot of basson (red) and clarinet (blue) - "Grey-tones", used in (11) and (12). Features: centroid, spread, skewness, flux, bark-flux, centroid-Flux



(b) First minute of Berio's *Sequenza 3 for women's voice* Features: x-axis: spread, y-axis: centroid, green-intensity: bark-flux, width: loudness

Figure 1: *Scatter Plots* and *Lineto* visualizations

Slider simply draws the magnitude of six features over time, using *Max* standard objects. A histogram for each feature gives a more global point of view of the feature's distribution.

Scatter Plots provides the scatter plots of all pairs of a 6-d feature-space. Time is encoded by changing color of the dots. It is highly suited to explore correlations between features.

Histos is a compact view on the histograms of the six features selected in Scatter Plots. By accumulating the distribution of features over time, it relates to the way in which texture distances are computed in the *Recognition*-patch.

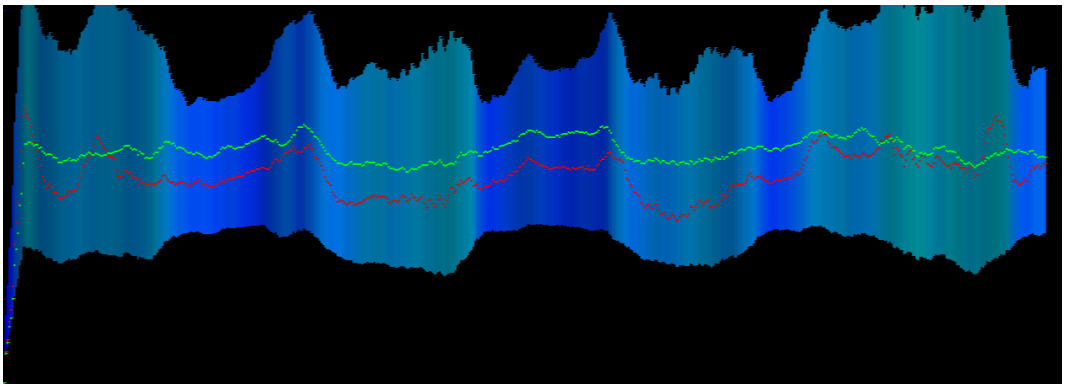
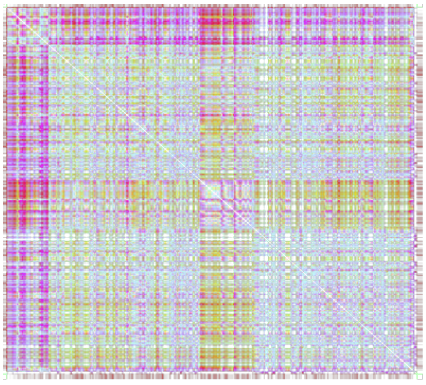


Figure 2: Klangfarbenmelodie?! Schoenberg: *"Fuenf Orchesterstuecke"* op.16,no.3, *"Farben"* - first three measures. Lower bound: centroid, width: spread, blue-intensity: bark-flux, green-intensity: flux, red-dots: skewness, green-dots: rolloff.

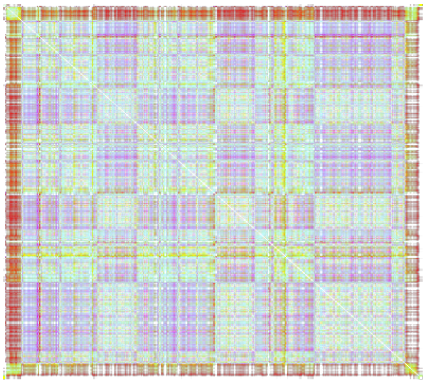
Lineto represents a 4-d feature-space by encoding 2 dimensions in x-y, a third dimension in the intensity of the line's color and the fourth dimension in the width of the line. It seems to be especially suited for the context of feature driven control and the capturing of gestures.

Stock Market simply plots the magnitude of four features over time. It can give an overview of long term properties and has more temporal resolution and flexibility than possible with *multislider*-based *Slider*.

River is inspired by the metaphor of a "river of sound". The approach was to represent timbre and textures not only by dots and lines, but also to let it shape the surface-texture of an object. In this high-dimensional "timbre-gram", a notion coined by Cook [13], two features control the upper and lower bound of the "river". Two features contribute to intensity of blue and green, respectively, and the relative position of green and red dots is shaped by two other features.This representation connects many features in one visual entity - the "river". It metaphorically tries to stay close to music-perception as the perception of one entity with several layers.
Consider the beginning of Schoenberg's famous *Fuenf Orchesterstcke* op.16. no.3. Whether or not a realization of his vision of an *Klangfarbenmelodie*, it could be seen as one of the first pieces to pose analytical problems, difficult to



(a) Billie Holiday's "All of Me". Turquoise = first four moments, magenta = loudness, yellow = flux. Smoothing = ca. 175ms.



(b) Shakira's "Whenever, Wherever". Turquoise = first four moments, magenta = loudness, yellow = flux. Smoothing = ca. 175ms.

Figure 3: *SimMatrix* visualizations.

resolve with tradional pitch oriented analysis. In the first three measures, the pitch content stays constant, only its orchestration shifts back and forth subtly. MacCallum and Einbond [14] compare the different orchestrations by their values in an auditory model of roughness. They measure a higher roughness value in the second orchestration, corresponding to the bright brass instruments as english horn and trumpet. Using *River*, it becomes obvious that the different sound of the second orchestration is related to a higher spectral centroid and a lower spread. Zones of stationary sound exhibit a low Bark-Flux value, whereas the transitions from one orchestration to the other have naturally higher flux. The *River*-representation can thus serve as a starting point for an acoustically oriented study of orchestration.

Similarity matrices are naturally suited for exploring temporal (a-)periodicity. A large body of research in automatic beat detection and music-segmentation has dealt with similarity matrices (for a thorough review of their role in MIR, see e.g. [15]). Jensen compares segmentation of popular music with audio features based on timbre (i.e. spectral) , rhythmic and harmonic information. He concludes that segmentation using timbral features performs best and presents dissimilarity matrices for the songs "Whenever, Wherever" sung by Shakira

and Billie Holiday's "All of Me". [5] [4]
SimMatrix plots a 3-d similarity matrix, by color-coding and superimposing similarity matrices (based on euclidean distances) of three 4-d feature vectors. The intensity of colors corresponds to the distances between feature vectors over time. New is the superposition of different matrices in real-time, which forms one complex higher-dimensional, less reductive representation of feature-time-series. Depending on the purpose of a visualization, complexity can be good or bad. However, this project strives for representations of timbre which take its multi-dimensionality into account.[5]

2.3 Similarity

Consider Figure 4-5, the plot of features of eight musical textures, drastically differing in genre. The features' oscillation patterns of each texture seem to be chaotic and hard to describe. Nevertheless, each texture seems to be at least distinguished in its features' mean values, its deviations and oscillation frequencies. This naive observation motivates the measurement of timbral similarity of textures by the comparison of their statistical distributions over long term time-windows. In the MIR-community, most research in music-similarity is devoted to global non-real-time music browsing and collection management, therefore using higher level statistical models like Gaussian mixtures, hidden Markov models or clustering. In contrast to that, this program applies low-level ways of calculating timbre-distances which can be used in real-time. The patch *Similarity* starts with a 10-d feature vector, and calculates its first four moments over a time window of variable length. The moments are non-linearly scaled and collected in a 40-d vector. A measure of similarity is then computed by taking the euclidean distance of the 40-d moments vector at time t and a reference point at time t_0.
Employing a naive notion of recognition, a distance below a variable threshold triggers the event of "recognition". The results of this real-time recognition-mechanism are surprisingly good, compared to the simplicity of its underlying mathematical model. Consider figure 6, the plot of the similarities of four reference textures, compared to eight textures overall. Note that a uniform threshold for recognition could be set in this collection of textures. Texture

[4] The method presented here achieves similar visual results on a global scale. Locally, the real-time 3-d method seems to be more differentiated.

[5] As a wise man used to say: "Things should be made as simple as possible, but not any simpler."

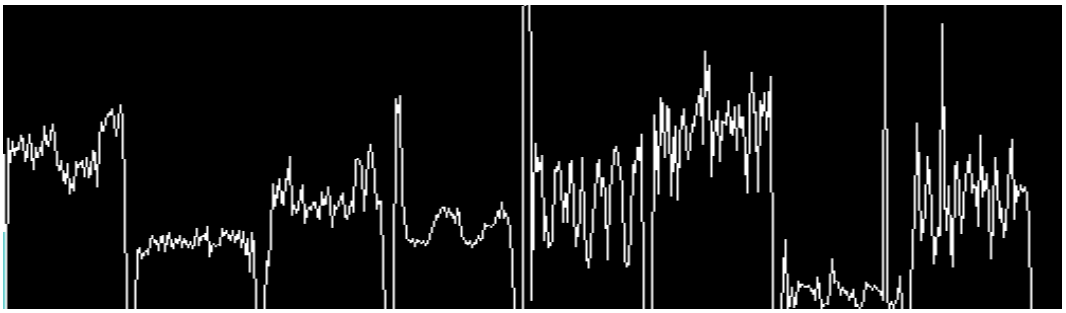


Figure 4: Feature trajectory over time of eight musical textures, 10 sec each, centroid = white, moving average length = ca. 30ms.

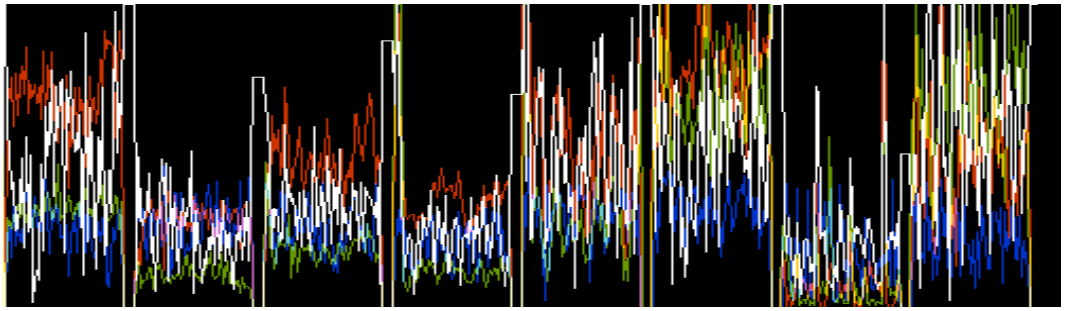


Figure 5: Set of feature trajectories over time of the same eight musical textures, 10 sec each, Red = Centroid, Green = Spread, Blue = Flux, White = Centroid-Flux, no smoothing.

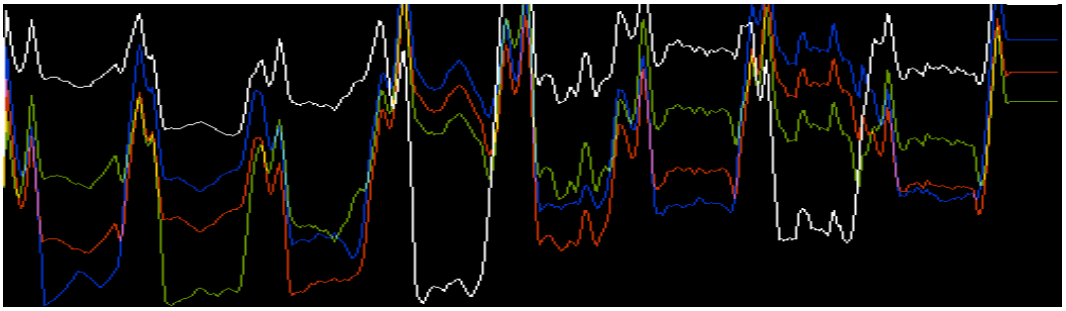


Figure 6: *Similarity*-values of the same textures. The higher the value on the y-axis, the more dissimilar are the textures from each other. The first four textures were set as references for blue, green, red, white, respectively.

1 and 3 seem to be similar in the plot and indeed, both are dense settings of strings. The video documentation of this project shows examples for the use of the recognition function in performance.

2.4 Audio Control

This part of the program is devoted to the application of audio feature extraction in performance contexts. It is basically an communication platform mapping the feature series or similarities onto the control of audio effects. It receives the scaled values of the Lineto-visualizations which can serve as an orientation for players, using feature driven audio effects. Control parameters of the effects are then shaped by the analysis of the incoming audio material itself (compare [16]). This method has an interesting dynamic behavior of the effect as a result, something which can lead to a more natural fusion of effect and original sound. The patch also receives the distance values and the recognition cues of the *Similarity* patch. The final mapping is of course up to the individual user of this patch.[6]

3 Feature Visualization and Music Analysis - Timbral Convergence in Gerard Grisey's *Modulations*

The last section (no.44 - end) of Grisey's *Modulations for 33 Instruments* is a musical process from harmonicity to inharmonicity, realized as a crescendo in ten steps. From a global point of view, this section features two different evolving sound objects: a continuous string section, build on the Les Espaces Acoustiques E, and developing chords in the wind instruments. These chords again divide into high chords, mainly played by the woodwinds and succeeded by deeper chords, mainly set for the brass[7]. The following section applies the notion of "convergence" to describe the musical development of those three

[6] Our first demonstration mapped e.g. the spectral centroid onto the oscillation-frequencies of a filter, loudness onto a (linear) frequency-shift and the event of "recognition" triggered an audio sample.

[7] Grisey uses for these chords the spectral model of a trombone sound which makes up the first high chord at the very beginning of the process. The mirrored set of intervals leads to the low chord, which he considers as the first chord's "subharmonic" pendant.

elements during the process. The aim is to develop a fertile interaction of score-analysis and auditory-feature-based visualizations.

The strings start in the first section in a unison on the cycles-E (in their respective range), only the first four violins play the fifth a in an flageolet. The dynamics start at *pppp* and are stricly increasing in each chord, leading to a fff in the last crescendo. During each of the nine following chords with their successors, inharmonic tones are added to the first E. The hammond organ comes in at the forth chord, with a cluster of a deep E, D-sharp, C-sharp to which in every following step one ascending tone is added. All the changes in pitch from the first unison E grow monotonously from zero to the range of a minor third.

The pitch-range of the high and low chords in the winds become more and more similar during the whole process, as intended by the model Grisey uses. The instrumentation changes: in the last three chords, the dichotomy in the setting of brass and woodwinds becomes more and more fuzzy. This process of convergence is supported by the temporal development of the overall attack time of the chords and the duration of the crescendos.

The addition of the chords in the winds seems to make the perceived difference between each continuous string section even smaller. Grisey's writing plays at this point with the limits of perception: listening to the ten sections in direct succession with the harmonic and subharmonic chords "cut out", the difference between them is clearly perceivable. Nonetheless, while listening to the section as a whole, attention is rather attracted by the chords, which let each string part appear very similar to its predecessor. Grisey uses the chords as a tool to mask the change in the strings to a certain extend and let them appear truly continuos. At some point though, listeners are confronted with dissonance in its very physicality without having been able to notice any process of change. The change is captured by a visualization, nevertheless. This case highlights an interesting divergence of the audio feature based representation of sound and the listening experience, caused by a subtle composition of change in repeating musical material.

Consider figure 7, the *River* - visualization of the process.[8] It gives an overview of the process and confirms that the dichotomy of continuos string

[8]In this paper the following recording was used for the audio feature extraction: Gerard Grisey: Les Espaces Acoustiques, Kairos, Vienna 2005. This section is the beginning of *Transitoires* on this recording.

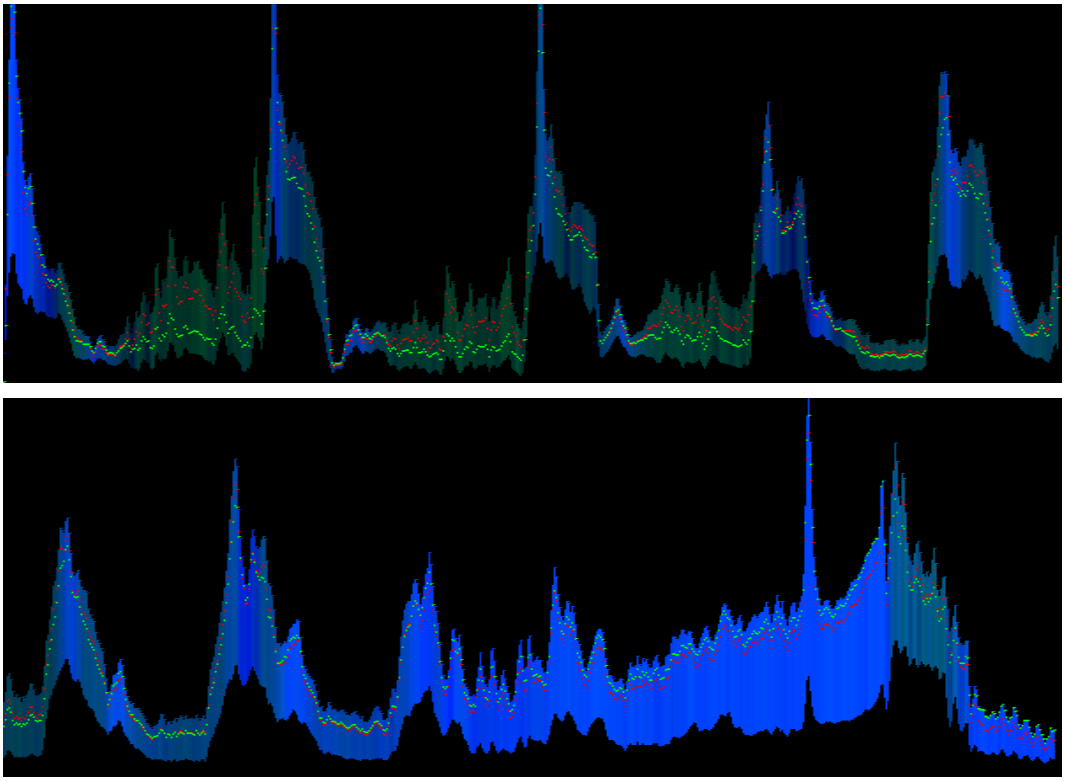


Figure 7: *River* visualization of the process in *Modulations*. Lower bound = centroid, width = spread, intensity of blue = bark-based-flux, intensity of green = flux, relative position of red dots = noisiness, rel.pos. green dots = loudness. Running average = ca. 580ms

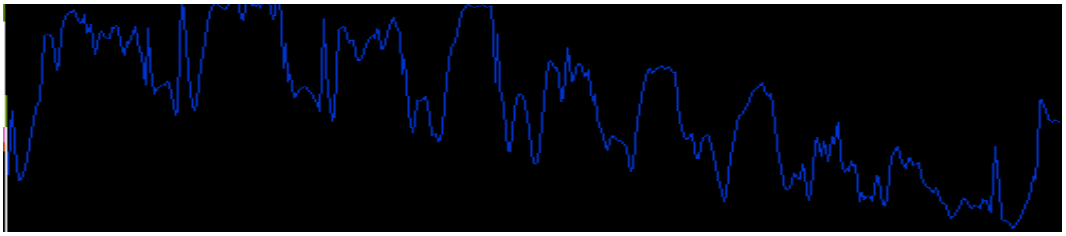


Figure 8: *Similarity* - values for the process with the reference point set at the last crescendo.

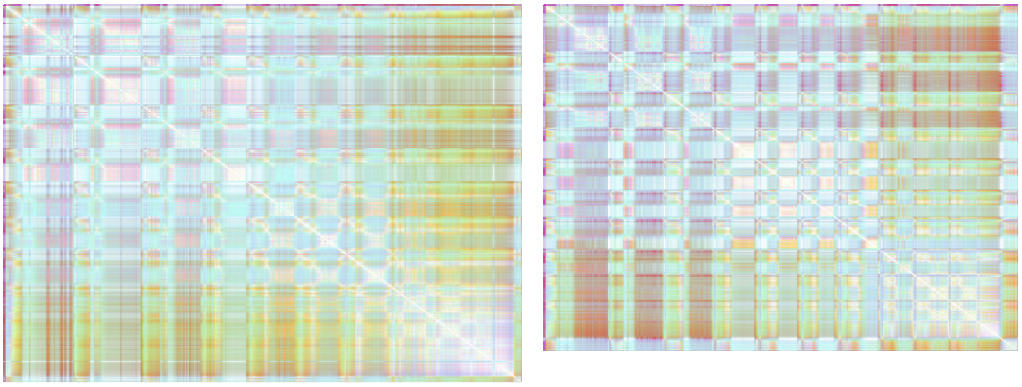


Figure 9: *SimMatrix* - visualization of the process. Spectral envelope encoded as turquoise, spectro-temporal cues as yellow, loudness as red. Smoothing = ca. 58ms

section and peaked chord in the winds is converging acoustically. This can even be made more explicit by considering the *Similarity* - values of the whole section with the last crescendo set as the reference (see figure 8). The "timbral distance" in respect to the last crescendo is shrinking to zero in an overall trend. Nonetheless, the difference of the distance between string section and chords is also diminishing, which supports the above thesis of mutual timbral convergence.

For the discussed section, similarity matrix representations highlight the repetitive but developing form of the process. Consider representations of two different recordings of the process in figure 9. These large scale visualizations can work like a global map for comparing recordings. It is clearly visible that the underlying recording is highly similar. Bright squares on the main diagonal correspond to the self-similar, stationary nature of the continuous string sections. The spectral envelope shows a pattern reflecting the dichotomy of continuos string section and chords in the winds. The increasing intensity of red in the right hand corner signifies a drastic change in loudness, compared to the beginning. The increasing portion of yellow encodes a similar structure for flux. The end of the process is visible by an increased distance to the early past, but smaller distance to the beginning of the process. However, the second recording is obviously faster. Also, it seems to be segmented into at least three self-similar sub-sections, whereas the first recording seems to develop more continuously. Interestingly, both representations differs in many details,

that might correspond to fine perceptual cues, which enable listeners finally to discriminate between the distinct "sound" of different recordings.

Both, the score-analysis and the visualizations of feature trajectories and similarities point out that this process can be seen as as a convergence of three distinct objects: the high and low chords in the winds and the string section. During the process, each one of them is continuously transformed in a direction of mutual convergence, so that they naturally appear as one mass of sound in the last crescendo. In reference to this evolving structure of converging sound objects, it might be illustrative to use an abstract conception of counterpoint and speak about a "counterpoint of complex sounds". Such a counterpoint may not exclusively be visualizable by points and lines, but maybe by points, lines, colors, shadows and textures.

4 Conclusion and Future Work

This project explores audio-feature-based real-time visualizations, timbre similarity and audio-control. The presented dynamic representations of multivariate time-series of audio features proved to be a powerful tool for acoustical and musicological analysis. Additionally, it could be used for educational purposes, e.g. as a "listening aid" in new music. The implemented notion of timbral similarity has produces promising results and will soon be integrated in my own (improvisation-oriented) performance setup. This project has not yet touched the big problem of systematically evaluating the performance of features and timbre-similarity measures, something connected with the general methodological problem of missing ground truth in this field of MIR. Currently, it seems preferable to rely in this point on the results of more exhaustive studies.
Each part of the presented project could be taken to another level in future work: Firstly, the choice of the feature set has to be re-adressed, as it seems necessary to extend the palette of features. For the purpose of visualization, higher-level perceptual features, like those McKinney and Breebart [9][9] suggested, appear to be highly suited and will be implemented. The task of finding a representation for a set of MFCCs is another point to be considered. Since sets of MFCCs ideally include more than 10 coefficients, finding a complete representation will involve some creativity. To the authors knowledge, these

[9]Sharpness, Roughness, Loudness and their respective modulation energies in different frequency bands.

kind of descriptors do not yet exist as real-time Max/MSP abstractions, and will therefore have to be programmed from scratch.
Secondly, in the realm of visualization, the question of how to optimize the preprocessing, i.e. smoothing/filtering of the data has to be posed. So far, only a moving averages were used, but other filters will probably improve the transparency of the data.
There is great potential in using the openGL - environment in *Jitter* for 3-d real-time representations. Using open GL will improve the visual qualities and the intuitiveness of the representations, while conveniently operating on the GPU. The aim is to letting sound dynamically shape three dimensional objects and their surface textures.
Thirdly, the field of timbre-recognition and timbre-similarities turns out to be highly interesting, both conceptually and musically, and it is closely related to general questions of feature based texture-modeling. Higher level statistics and classifiers, as e.g. GMM or HMM will be evaluated for real-time use. Methods of spectral analysis of the feature time-series are another idea which will be tested. For the task of recognition, the inclusion of neural network based learning schemes, as e.g. deep belief networks, seems promising.
The audio-control part, is the least developed so far. However, this field could evolve in respect to realizing more concrete musical ideas, perhaps in cooperation with an interested composer. In a similar sense, music-video interaction via audio-features could be explored artistically, and some of the implemented visualizations might serve as a starting point for this endeavor.

References

1 J.-C. Risset, D. Wessel: *Exploration of timbre by analysis and synthesis* in The Psychology of Music, 2. Edition, Diana Deutsch (Edt.), Academic Press, San Diego 1999, p. 151

2 T. Murail: *The Revolution of Complex Sounds* in Contemporary Music Review 24 (2/3), 2005, pp.123-124

3 T.H. Park, J. Biguenet, Z. Li, R. Conner, S. Travis: *Feature Modulation Synthesis*, Proceedings of the International Computer Music Conference 2007, Copenhagen, Denmark, August 2007

4 M. McKinney, J. Breebart: *Features for Audio and Music Classification*,

Proceedings of the 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA, 2003

5 K. Jensen: *Multiple Scale Music Segmentation Using Rhythm, Timbre, and Harmony*, EURASIP Journal on Advances in Signal Processing, Volume 2007

6 F. Morchen, A. Ultsch, M. Thies, I. Lohken, M. Nocker, C. Stamm, N. Efthymiou, M. Kuemmmerer: *MusicMiner: Visualizing timbre distances of music as topographical maps*, Report of the Data Bionics Research Group, Philipps-University Marburg, 2005

7 F. Pachet, P. Roy: *Analytical Features: A Knowledge-Based Approach to Audio Feature Generation*, EURASIP Journal on Audio, Speech, and Music Processing, Volume 2009

8 E. Jourdan, M. Malt: *Zsa.Descriptors: A library for real-time descriptors analysis*, Proceedings of 5th Sound and Music Computing Conference Berlin, 2008

9 J.-C. Risset: Foreword to *Electroacoustic Music - Analytical Perspectives*, T. Licata (Edt.), Library of Congress Cataloging-in Publication Data, 2002

10 S. Malloch: *Timbre and Technology: An Analytical Partnership*, Contemporary Music Review, Vol. 19, Part 2, pp. 155 -172, 2000

11 J. Grey: *An exploration of musical timbre*, Center for Computer Research in Music and Acoustics, Department of Music, Report No. STAN-M-2, Stanford University, 1975

12 D. Wessel: Timbre Space as a Musical Control Structure in Computer Music Journal 3(2), 1979

13 P. Cook: *Computer Music* in Springer Handbook of Acoustics, T. D. Rossing (Edt.), New York, 2007

14 J. MacCallum and A. Einbond: *Real-Time Analysis of Sensory Dissonance* in Computer Music Modeling and Retrieval. Sense of Sounds, 4th International Symposium, CMMR 2007, Copenhagen, Springer, New York, 2008

15 R. B. Dannenberg, M. Goto: *Music Structure Analysis from Acoustic Signals* in Handbook of Signal Processing in Acoustics, Vol. 1, Springer 2008

16 V. Verfaille, U. Zoelzer, D. Arfib: *Adaptive Digital Audio Effects: A New Class of Sound Transformations* IEEE Transactions on audio, speech, and language processing, Vol. 14, No. 5, September 2006