

# ON AUTOMATED ANNOTATION OF ACOUSMATIC MUSIC

Volkmar Klien<sup>1,2</sup>, Thomas Grill<sup>1,2</sup>, and Arthur Flexer<sup>1</sup>

<sup>1</sup>Austrian Research Institute for Artificial Intelligence, Freyung 6/6, A-1010 Vienna, Austria

<sup>2</sup>University for Music and Performing Arts Vienna, Anton v. Webern Platz 1, A-1030 Vienna, Austria

## ABSTRACT

This paper presents an inquiry concerning the feasibility of using existing methods from the field of Music Information Retrieval (MIR) for automated annotation of acousmatic music. Thorough discussion and appraisal of the meaning and role of annotation in this context leads to our conclusion that: (i) full automation is not possible due to the lack of a ‘ground truth’ and the absence of semantic comprehension on the side of the computer, (ii) MIR can nevertheless play a valuable role by providing human annotators with tools for interactive annotation. We present two possible approaches to interactive annotation applied to compositions of acousmatic music, namely John Chowning’s *Turenas* and Denis Smalley’s *Wind Chimes*. We also discuss the potential impact of such semi-automatic annotation on the theoretical consideration and practice of acousmatic music.

## 1 INTRODUCTION

This paper is about automated annotation of acousmatic music through application of Music Information Retrieval (MIR) methods. Following Smalley’s definition (Smalley, 1997) we refer to electronic (or *electroacoustic*) music as *acousmatic* if it is not traditionally note based, does not rely heavily on the listeners’ understanding of anecdotal content or their ability to recognize the sounds’ physical origins. Even though this holds true for a fair number of ‘tape’ compositions in a musical tradition often referred to as ‘acousmatic’, our use of the term includes organized sound mediated over loudspeakers attributed to all kinds of traditions and practices in the sonic arts.

Since the majority of acousmatic music is made with electronic technology, its sounds drawn from nature and studio, synthesized or prerecorded and often extensively processed and altered, annotation of acousmatic music is both more challenging and less developed than analysis of instrumental or vocal music. There exist no ‘pre-segmented’ discrete units equivalent to a note, there is no score and no universally established or widely used system for analysis. Although musicology has developed various sets of tools for

analysis of acousmatic music, the tediousness of manual annotation has prevented the application of these theories to a larger body of music also limiting how fully those ideas are put to the test. Since Music Information Retrieval (MIR) has developed a rich repertoire of algorithms for analysis of music, including methods that can be used for automatic annotation, it is our objective to bring together musicological theories of acousmatic music and MIR methods with the aim of exploring the possible automation of annotation.

We will focus on an existing framework for analysis of acousmatic music: the theory of spectromorphology developed by New Zealand composer Denis Smalley (1997). This theory provides descriptive tools based on aural perception and concerns itself with identifying ‘carriers of meaning’, i.e. structural and sonic entities in acousmatic music. It covers aspects from very low-level sound properties to highly abstract concepts of form and interdependencies of musical matter. The contribution of this paper is in answering the following research questions:

- Can Music Information Retrieval be employed to identify carriers of meaning within works of acousmatic music?
- Which parts of the analysis have to be conducted manually and which parts can be fully automatic?
- Which level of abstraction/complexity can be reached?

To address these questions, this paper is divided into two sections, the first concerned with theoretical considerations about the meaning and role of annotation in general (section 2) and the second with the presentation of two examples of possible approaches to semi-automatic annotation of acousmatic music (section 3). Section 2 discusses the role of annotation and score in both traditional and acousmatic music. We give a critical evaluation of the subjective nature of annotation and score especially in the context of acousmatic music and discuss possible problems and benefits of automating annotation. This more general argumentation is then focused on using MIR methods in the context of spectromorphologic annotations. At the end of section 2, we describe our proposed solution to resolve the apparent contradiction concerning the subjective nature of annotations of

acousmatic music versus the quasi-objective nature of any fully automatic approach towards annotation: providing human annotators with software tools for interactive annotation. In section 3 we present two approaches that allow semi-automatic interactive annotation of acousmatic music. One is based on self-similarity matrices and produces a structural overview of complete pieces of music while the other uses a dimensionality reduction and clustering technique to visualize representative sound groupings. Both approaches are applied to two prominent works of acousmatic music thereby sketching possible analysis strategies. Section 4 concludes our paper with a critical discussion of what has been achieved and by answering the above-raised three questions central to our research endeavor.

## 2 ON ANNOTATION

### 2.1 Analyzing and annotating acousmatic music

Setting out to analyze an individual acousmatic work it cannot be taken for granted what the adequate, most relevant analytical parameters should be. Hence one of the main analytical objectives is to unearth those aspects central to the organization of sound in a given composition. Whereas in instrumental, scored music the analyst is able more often than not to focus confidently on pitch and rhythm, in electroacoustic composition the structurally most salient matter may be the spatial aspects, gestural development or energy trajectories over time.

Throughout the centuries musicology has developed a rich tool-set of methods for the analysis of instrumental and vocal music bound to the existence of a notated score. Traditional approaches to music theory leave us rather helpless in any attempt to analytically describe electronic sound that goes beyond simply modeling traditional musical instruments. A key reason for this lies in the fact that acousmatic music does not restrict itself to sonic material that, by convention, has been defined as musical. There are no predefined building blocks or basic musical objects (represented by notes) and no pre-defined or obvious syntax guiding the arrangement of these musical units across time and frequency grids. Given the high output of electronic music, it is surprising that the emphasis of contemporary musical analysis remains with the various genres of instrumental rather than electronic music. In recent years a number of publications (e.g. Simoni, 2006; Licata, 2002) presented collections of analyses of acousmatic compositions. Although practically all of them were created using some form of computational representation of sound (see Adams, 2006 for a review), the vast majority relied on purely manual annotation of a composition's sound. A number of tools for computer-aided annotation have been developed: *Clam Music Annotator*<sup>1</sup> (Universitat Pompeu Fabra), *ASAnnotation*<sup>2</sup> (IRCAM), *Acousmographe*<sup>3</sup> (INA-GRM), *Sonic Visualiser*<sup>4</sup> (Centre for Digital Music at Queen Mary, University of London) and *iAnalyse*<sup>5</sup> by Pierre Couprie. These are software packages allowing for manual annotation of acousmatic sound with different levels of support by integrated digital signal processing tools (e.g. for transient or pitch detection). The software framework SQEMA (Park et al., 2010, Tulane University, New Orleans) specifically aims at integrating methods from Music Information Retrieval into a more quantitative analysis of acousmatic music.

<sup>1</sup>[http://clam-project.org/wiki/Music\\_Annotator](http://clam-project.org/wiki/Music_Annotator),

*tion*<sup>2</sup> (IRCAM), *Acousmographe*<sup>3</sup> (INA-GRM), *Sonic Visualiser*<sup>4</sup> (Centre for Digital Music at Queen Mary, University of London) and *iAnalyse*<sup>5</sup> by Pierre Couprie. These are software packages allowing for manual annotation of acousmatic sound with different levels of support by integrated digital signal processing tools (e.g. for transient or pitch detection). The software framework SQEMA (Park et al., 2010, Tulane University, New Orleans) specifically aims at integrating methods from Music Information Retrieval into a more quantitative analysis of acousmatic music.

### 2.2 The analytical score

It would be difficult to overestimate the role traditional scores played and continue to play in the development of both Western art music composition and music theory, the score not uncommonly being viewed as the 'true location' of the musical work (cf. Dahlhaus, 1965; Goodman, 1981, see also Delalande, 2007). Leaving aside the details of that discussion it will suffice to say that the role of the annotation score will be fundamentally different (cf. Gayou, 2006). The aim of our research for facilitating automated annotation of acousmatic sound (and hence the production of 'scores') is not to finally mend a perceived shortcoming but to detect and describe perceptually relevant musical materials, their interrelation and evolution in time.

The classical musical score, although based on analysis of pitch and metric rhythm is also shaped distinctly by the necessities of instrumental sound production, such scores in turn influencing the sound production itself. In acousmatic music with its multitude of production techniques production scores are no longer an imperative and (in practice) hardly ever exist in monolithic form. Hence the annotated listening score in acousmatic music is completely de-coupled from sound production.

### 2.3 Manual annotation and potential benefits of automating annotation

Manual annotation of acousmatic music is extremely time consuming (Hirst, 2006), a fact that has prevented broader application and recognition of already existing theoretical frameworks. As has been widely discussed in musicology it would be plainly wrong to presume the existence of one single 'most-correct' analysis (and annotation is always the result of analytical decisions) of a given piece, to which we intend to apply computational methods in order to achieve a

retrieved 2011-02-23

<sup>2</sup><http://recherche.ircam.fr/equipes/analyse-synthese/ASAnnotation/>, retrieved 2011-02-23

<sup>3</sup><http://www.ina-entreprise.com/entreprise/activites/recherches-musicales/acousmographe.html>, retrieved 2011-02-23

<sup>4</sup><http://www.sonicvisualiser.org/>, retrieved 2011-02-23

<sup>5</sup><http://logiciels.pierrecouprie.fr/>, retrieved 2011-02-23

closer approximation. Automation simply automates things; it does not make them ‘more objective’.

Any attempt to formalize and automate a task necessarily places that effort and concepts relating to it under added scrutiny. In the case of acousmatic music this will mean testing the ability of the chosen conceptual toolset to provide clear and unambiguous descriptions of sound independent of personal communication with its added channels of bodily gesture and vocal mimicry of sonic behavior.

Automated annotation will provide the musicologist with an un-emphatic view of the sonic material, against which his or her own listening experience can be measured and vice versa. This process in itself can provide insight regarding the workings of the annotation algorithms, the analyzed composition and indeed the analyst’s own listening behavior. We envisage automated or semi-automated annotation to break new ground in musical analysis by significantly accelerating the process of annotation as well as stabilizing the analysis’ parameters and results.

Even though individual analyses might legitimately follow rather different strategies, automated annotation will allow for an underlying accumulative process of collecting data on the musical works analyzed. Hence automated annotation of acousmatic music will help make acousmatic music research a more data-rich endeavor, which in itself has to be seen as a desideratum (cf. Clarke and Cook, 2004).

## 2.4 Artistic practice and automated annotation

Analysis, not only in its choices of conceptual tools, but also in its individual reading of music is a creative act in itself and as such has always played a role in musicians’ individual approaches to music. We envision automated annotation of electronic sound as constituting a further step towards enabling analysis to take on new roles in the creative process of electronic music making.

Frisius sketches out the implications machine-aided notation of the listening score will have on the role reception plays in musical production. “Music of all kinds, in the context of the related listening experience, will be described no longer via its abstract visual score, but rather in its concrete sounding image. All that is audible thus turns into the potential objects of musical analysis. [...] The relationship between musical reception and production will change fundamentally, as soon as music has become analyzable to a point enabling the analyst not only to describe, but also to experimentally alter that music.”<sup>6</sup> (Frisius, 2002)

## 2.5 Methodologies of acousmatic music analysis and MIR

MIR is concerned with engineering solutions to the problem of how to automatically analyze, annotate, store and

retrieve music. As outlined by Wiggins (2009), MIR tends to treat music, which is undoubtedly a cultural artifact and the human annotation thereof as the ‘ground truths’ of its scientific endeavor. For any quantitative approach to music analysis it is essential not to lose sight of the fact that, even without proposing or agreeing on one specific definition of music, music is a communal, cultural construct and a dynamic process in its social context rather than a natural object in its physical environment. Although genre classifications of, say, ‘mambo’ or ‘heavy metal’ do in themselves seem stable categories, they are so only to a certain point.

In the context of acousmatic music practice no comparable established categories which one could refer to during annotating are to be found. Rather there are several approaches, such as those proposed by Schaeffer (1952), Smalley (1997), Roy (2003), Clarke (2010a) and Hirst (2006) giving rise to the impression that there are nearly as many methodologies suggested as there are full analyses published. This ‘lack of consistency in nomenclature’ has already been noted in the research community (Park et al., 2010). Given, for example, that machine learning<sup>7</sup> techniques are deployed to learn from databases of human annotation, this lack of a coherent methodology for annotating acousmatic music is a difficult position from which to start.

This state of affairs cannot be interpreted as a shortcoming of the analytic methods at hand or a lack of analytical practice focusing on acousmatic music only, but has to be seen as rooted in the nature of acousmatic music itself. Despite the fact that in the realm of academic music making the acousmatic might present itself as a rather settled and well-defined art form, it is the inquisitive nature of acousmatic music itself that is the basis of many of the difficulties one faces when attempting to automate annotation thereof. This is of course only the case, if one’s ideal annotation is the description of the musical work in unambiguous categorical terms, apparently the goal for a fair share of the MIR research directed at managing databases of music and sound.

It is acousmatic music’s investigating and testing of established musical categories, listening habits and frames of reference that lies at the heart of many difficulties encountered by the human annotator (and hence also the intended machine annotator). The questions: ‘Is a specific pitch audible or not?’ or: ‘When exactly does this one stream of sonic material turn into two distinct ones?’ might stand not only at the beginning of the analytical approach, but also at the inception of the work itself. This exploration and deconstruction of established musical parameters and categories in compositional practice lends itself less easily to a systems approach to annotation and conceptualization of music

<sup>6</sup> translation: Volkmar Klien

<sup>7</sup> Machine learning, a class of technologies widely used in MIR, is a subfield of artificial intelligence and is concerned with the design and development of algorithms and techniques that allow computers to learn from data (e.g. the relationship between audio representations of music and semantic descriptions).

than others might.

Many of the techniques widely applied in MIR involve attaching categoric labels to the concrete musical work or sound in question. It is these categoric grids of reference, implicitly seen as underlying all musical practice, that reveal themselves as inapplicable to acousmatic music. Acousmatic music is an artistic practice constantly renegotiating its situatedness alongside established traditional musical grids and their connected music-theoretical edifices of rule sets, which are often interpreted as underlying the concrete work's structuring.

Compositional practice in acousmatic music at least in its more innovative incarnations needs to be interpreted as an ongoing artistic examination of established musical practice, theoretical approaches, production norms and listening habits. This approach will need to be reflected in any analytical elucidation of a work, which MIR technologies were simply not designed to achieve. Hence there exists not only the 'semantic gap' between audio representations of music and their semantic description, or excessive layers of indexing that might get in the way of successful automation of annotation, but also the fact that acousmatic music rarely evolves along lattices transcending individual compositions.

It is not only from a technical point of view that one has to distance oneself from a perceived ideal of creating one annotation machine unearthing or modeling ground truths about musical works at hand. As will be shown below, this does not mean that automating annotation cannot play a valuable role in acousmatic music analysis.

## 2.6 Manual annotation in the context of MIR

In MIR research manual annotation of the audio signal is of crucial importance for the development of algorithms allowing computational systems to connect the purely technical representations of the audio signal to first person descriptions thereof: its human intentionality (cf. Lesaffre et al., 2004). Research into acousmatic music can draw on ample experience concerning methodology and practicalities of manual annotation even though most of MIR research primarily concerns itself with more mainstream forms of music.

While our research into methods for automating annotation of acousmatic music finds a multitude of MIR methods on which to build, manual annotation of acousmatic music presents rather specific challenges. Firstly, there is no existing corpus of annotated compositions to draw from. Secondly there exist limitations inherent to manual annotation of acousmatic sound in terms of accuracy with regard to time, pitch and timbre. In dense musical passages (perhaps with various spectromorphologies overlapping, or quickly moving 'clouds' of short sonic events) exact manual annotation in time becomes a sheer impossibility. While it might be easily possible to aurally identify a quiet click stream

'behind' the musical 'foreground' it can be completely impossible to annotate these events and check the correctness of these annotations in time, even for expert annotators.

This systemic lack of reliable and exact annotation data presents a fundamental difficulty for any MIR algorithm development. In our search for reliable testing grounds we settled upon Chowning's composition *Turenas* (Chowning, 1988) as one of our first objects of interest. Our choice was not only motivated by the piece's relative sonic clarity and homogeneity, but also – the composition being a product of pure synthesis – by the fact that there exists a related production score (in Music IV score file format), which was made available to us by John Chowning. This does not mean that our intended methods would be reliant on production scores to function (the ones proposed in chapters 3.4 and 3.5 do not directly reference the production score themselves), but it does provide anecdotal evidence of what convenient things scores really are for musicology. The existence of several published analyses and annotations thereof (Zelli, 2001; Pottier, 2005; Jure, 2004a,b) also factored into our choice of *Turenas*.

As an example of the problems faced in manual annotation see figure 1. Below the waveform pane each dot represents the onset of a short sound event with the exact timing provided by the original production score. These events grow ever denser (to maximum of approx. 40 onsets per second) until they finally form one extended grainy sound. Pottier (2005), in his detailed analysis of the piece refers to the sound morphologies in this section as 'grainy lines'. Though it might well be sufficient in the context of a musical analysis, for our goals outlined above, this description remains far too general.

## 2.7 Spectromorphology

A number of approaches have been made to categorize and describe the origin and nature of the carriers of meaning in acousmatic music and to establish an appropriate framework for consideration of the conceptual. Two larger frameworks have become influential, both building on the basis of Pierre Schaeffer's works. One is the system of *Unités Sémiotiques Temporelles* (UST, Delalande et al., 1996) developed at the Laboratoire Musique et Informatique de Marseille (MIM). The other, which we are focusing on<sup>8</sup>, is the theory of *spectromorphology* (SM) developed by Denis Smalley.

SM is the theory of temporal unfolding and development of sound spectra and it provides "tools for describing and analyzing listening experience" (Smalley, 1997). Given that SM has been influential and widely received<sup>9</sup> we in the following examine the extent to which it could be possible to

<sup>8</sup> The main reasons being the first author's long-standing familiarity with the concepts of SM and also (having been published in English) its greater general accessibility and influence on artistic practice.

<sup>9</sup> For a critical review of SM's influence see Normandeau (2010)

base an automated annotation system on SM's methods of describing acousmatic sound.

It is spectromorphology's embodied approach to musical gesture and its conceptual flexibility in the description of structural levels and dynamic attribution of function that allowed it to resonate with numerous practitioners and theoreticians in the sonic arts. These dynamics – as will be discussed here – can simultaneously be seen as the central difficulty in any attempt to systematize and formalize SM.

### 2.7.1 *The system of spectromorphology?*

SM is not a static sound ontology for the description and classification of sound objects, but rather a set of conceptual tools, a vocabulary for describing the listening experience and its dynamics. It does not propose fixed functional, structural levels or static hierarchies but describes dynamic attributions of functions within the context of the listening experience, directed primarily at *intrinsic* relations within the acousmatic work.

Discussing the evolution of SM Smalley writes: “My elaborations of ‘motion and growth processes’, of ‘behavior’ and of ‘structural functions’, were conceived of as relational frameworks for considering the musical context, and I now think that they are best understood as metaphorical mappings which might simultaneously embody intrinsic and extrinsic views”<sup>10</sup> (Smalley, 1999, quoted and translated in Atkinson, 2007). As Smalley acknowledges, ‘intrinsic to music’ reveals itself as a less well-defined notion as soon as the various modes of perception are recognized as being heavily integrated. Approaching this issue from semiotics rather than an ecological approach to music perception, Atkinson (2007) reaches a similar conclusion.

Although many of SM's terms for describing sonic behavior have proven to be very helpful in addressing acousmatic music in human dialogue (which is exactly what they set out to do) these terms cannot be interpreted as strict perceptual categories of sonic behavior. Given the intended conceptual overlap between ‘neighboring’ categories (e.g. of *drifting* and *floating*), generating sufficiently unambiguous annotated test and training sets of such categories for the development of MIR algorithms seems to be problematic.

Marsden, speaking in the context of computational approaches to traditional music analyses, rightly states that “the concepts of traditional music analysis are not entirely systematic and precise (their power is precisely in their allowance for expert knowledge and experience), so those who apply computers to analytical problems must redefine concepts for the impersonal, digital domain. What does it mean, for example, for a passage of music to be in a particular key [...]?” (Marsden, 2009). If trying to formulate the

<sup>10</sup> In this SM displays similarities to ecological approaches to listening in the context of everyday sound (Neuhoff, 2004) as well as music (Clarke, 2005).

concept of tonality, which on first sight appears to be such an unambiguous notion, turns out to be fraught with particular difficulties, the problems of formalizing SM's terms and concepts for describing sound shapes can only be hinted at.

In this SM faces similar problems to those of any linguistic/symbolic representation of musical sound faces: namely, establishing the exact extent to which agreement between different listeners can be reached, and similarly the extent to which this agreement is determined by cultural aspects. This is one of the reasons behind recent sub-symbolic approaches to musical gesture that emerged in the area of embodied music cognition, musical performance and embodied approaches to new interfaces for musical expression (e.g. recent work by Godøy, 2006; Godøy and Leman, 2009 and Leman, 2007).

In a recent commentary on the theory of spectromorphology and its impact over the years Smalley (2010) clarifies his intentions as follows: “I can suggest viewpoints, methods of approach, concepts, schemas, provide a selection of tools, sometimes invent ordered and systematic frames of reference, and at other times provide frameworks for speculating and imagining, but I cannot tie them up in a neat little bundle primed conveniently for analytical or compositional action.” In the same article he further states: “My spectromorphology, [...], was not formulated to facilitate any *systematic* analytical method. It is, rather, a collection of tools for describing sound shapes, structures, and relationships, and for thinking about certain semiotic aspects – potentially analysis of a kind.”

### 2.7.2 *Problems of automating spectromorphological description of the listening experience*

Setting aside all the potential problems for automating annotation arising from this open approach to describing musical behavior, implementation difficulties for such annotation remain, even when dealing with aspects of spectromorphology, which, like the classifications of onsets, continuants, and terminations, do systematically describe specific sonic behavior.

The standard MIR approach towards classification is to divide a set of annotated examples into training and test sets. The training set is used to learn models summing up the relation between the examples and their annotations. The test set is used to evaluate the success in linking examples to their annotations in a fair way.

Even in straight-forward musical situations of, for instance, tempo estimation or genre classification of popular music the quality of these training and test sets is one of the central concerns. It is currently unclear how such sets for SM descriptors can be accumulated, especially given the fact that in real life situations (i.e. acousmatic compositions) hardly any one single spectromorphology appears ‘solo’.

It might be promising to ask several acousmatic com-

posers to produce their individual quasi-prototypical sonic examples corresponding to SM's descriptors. This would result in an artificial test set, which, probably even more than linguistic annotation, would need to be seen as a product of its times and might have sounded rather different in the 1970s than it would have in the 1990s.

## 2.8 Discussion

To sum up our arguments concerning the meaning and role of annotation in the context of acousmatic music, it has to be said that any attempt at *full automation* of annotation is rather problematic. Even annotations of traditional music need to be seen as communal, cultural constructs in their social context rather than objective 'ground truths'. It is ever more so in the case of acousmatic music, having an inquisitive nature and constant exploration and deconstruction of established musical parameters at its very heart. It would therefore be plainly wrong to presume the existence of one single 'most-correct' analysis putting the quasi-objective nature of any fully automatic approach at odds with a more subjective reality. This situation seems to be the main reason for the systemic lack of reliable and precise annotation data which further hampers all efforts towards full automation of annotation.

Our proposed solution to resolving this seeming futility of any fully automatic approach towards annotation is to provide human annotators (analysts) with software tools for *interactive annotation*. For automated annotation to play a potentially meaningful role in the process of acousmatic music analysis the annotation algorithms do not need to 'understand' the music signal's meaning for the human listener. This becomes possible in an interactive setting, where the analytical decisions are taken by the person in front of the computer, who also interprets the computer's output. The automatic annotation's output can then feed back into the analytical process leading to refinements in the machine annotator's parameter setting. Our presentation in section 3 of two practical approaches towards interactive annotation will make clearer the nature of these analytical analyst decisions: the choice of audio descriptors to investigate, the algorithms employed, tuning of parameters governing resolution and scope of the analysis.

Most of the published analyses of full acousmatic pieces known to us devote considerable effort to identifying groups of different sonic materials. This identification of similarity/unity, depending on the specifics of the compositional work at hand, may be based on rather different aspects. Clarke (2010b), in his interactive analysis of Smalley's *Wind Chimes* (Smalley, 2004) distinguishes in his sound taxonomy's labeling between different types of events; for example, between 'metal', 'ceramic', and similar categories of sounds. Pottier (2005), in his listing of sound morphologies (see table 1) found in Chowning's *Turenas* (Chowning,

1988), distinguishes, amongst others, between percussive sines of short duration and long granular sines.

This process of classifying the individual sonic materials into groups of similar behavior and timbre can hardly in itself (i.e. without any further interpretation) be seen as a valuable contribution to music analysis as such, but as a preparatory step it is an essential part of the process. Given the current state of technology, the creation of such an overview of a composition's evolution in time seems to be the most useful role that an interactive software tool assisting the expert human annotator can play.

## 3 TWO EXAMPLES OF AUTOMATED ANNOTATION

In the following we present two different practical approaches to automating aspects of manual annotation of acousmatic music. The first employs *self similarity matrices* for the generation of an overview of a composition on a higher temporal scale. The other relies on *dimensionality reduction and clustering* techniques allowing for an interactive approach to visualization and identification of representative groupings of sound materials based on their timbral and structural similarity.

Both of these experimental approaches are based on existing and established methodologies in MIR and are presented here in the context of recreating in a partially automated manner certain aspects of two published analyses of prominent works of the acousmatic music repertoire:

- Pottier's analysis of John Chowning's *Turenas* (Pottier, 2005)
- Clarke's interactive analysis (Clarke, 2010a) of Denis Smalley's *Wind Chimes*

The methods outlined here take a composition's digital audio signal as their point of departure aiming (as will be outlined in the following) at helping to unearth aspects of human listening experience. This process, in both of the scenarios presented is one of interaction between the human analyst and their interpretation of the algorithm's output. While it may well be that in this 'dialogical' process between individual listening and parameter tuning leads to new personal insights into a compositions' structuring and related compositional processes, our aim in the following two exemplaric analyses is to show that there is a valuable role automated annotation can play in the analysis of electroacoustic music, regardless of the fact that in itself it might not be able to unveil information about individual works that would principally remain beyond the scope of manual annotation.

The fact that for *Turenas* there exists a production score is helpful here only for verifying the analytical results against

an existing description, reliable in terms of temporal resolution, which to produce manually would be fraught with intricate problems (see 2.6). As will be shown in the following, our aim was not to unearth underlying numerical, ‘structural’ relationships through a study of the production score, as has been done – for example – in analyses by Meneghini (2007) and Dahan (2007) of Chowning’s *Stria*.

For Smalley’s *Wind Chimes*, for which no detailed uniform production score comparable to the one for *Turenas* exists, we were able to refer back to the detailed analytical scores of Hirst (2006) and Clarke (2010a).

These two rather different kinds of ‘scores’, one a production score for technical sound synthesis, the others listening scores we use to evaluate our attempted linking of signal based analysis to the human listening experience with the help of perceptually relevant analytical methods. This we believe is possible only in highly interactive settings. As will be detailed below both of the processes presented are highly dependent on the individual tuning of parameters in the algorithms involved. Manual parameter tuning is not generally seen as a desirable method in MIR research, which focuses rather on designing algorithms that allow for calculations to be performed over large databases with the specific goal of rendering manual parameter adjustments by human experts in response to the individual musical work or sound unnecessary. In the case of acousmatic music with all its fluid concepts and its lack of reference to stable lattices and systematic grids<sup>11</sup>, this tuning of parameters is not only necessary in response to the differences in the individual signals to which the algorithm is applied, but also and most importantly in accordance with the analyst’s analytical intentions.

The two examples of automated annotation presented here are thus to be seen as case studies of potential contributions to processes in acousmatic music analysis – rather than stable analytical methods with their algorithmic toolsets that would only need to be applied to other musical works without further need of adaptation.

### 3.1 Turenas

As one of the first pieces produced entirely in the digital domain (it was produced in 1972 in the laboratories of Stanford University using the Music IV<sup>12</sup> and SCORE (Smith, 1972)) John Chowning’s composition *Turenas* is one of the most prominent pieces in the history of early digital music.

The following short structural overview of the piece draws heavily on the analysis by Pottier (2005), which is published in book form, but also accompanied by an interactive, multimedia presentation of the analysis’ results on-

line<sup>13</sup>. For reasons of computational limitations at the time of its creation the production score to the piece is divided into seven different score files. The piece itself, according to Pottier, who in the process of analyzing *Turenas* was able to refer back to the composer for feedback, is structured into four sections, forming a relatively symmetrical structure<sup>14</sup>.

The introductory and closing sections make use of similar sonic materials (see figure 2). Similar elements are to be found in the introduction and the closing section of the piece: “Crystalline percussion, spatialized lines, bells and rich sustained sound. They precede or, respectively, conclude two melodic canons, both constructed in quite similar ways. Between the two canons, two melodic and polyphonic parts can be observed: the first is composed of sustained sounds, originating from an acceleration and from deformations of the canon’s notes; the second, longer than the other, is composed of diverse percussive sounds of various timbres and leads to a final crescendo, the climax of the piece.”<sup>15</sup> (Pottier, 2005)

From a morphological point of view, Pottier identifies ten different groups of sonic morphologies, which he orders according to envelope, duration, harmonics, spectrum, register and modulation (see table 1).

### 3.2 Wind Chimes

Denis Smalley’s *Wind Chimes* (Smalley, 2004), a piece commissioned by the South Bank Centre in London, was premiered in 1987. The composition is mainly produced from recorded sounds. Clarke (2010a) describes the composition’s source sounds: “As the title might imply, the most significant single source is a set of ceramic wind chimes the composer discovered on a visit to his native New Zealand in 1985. Other sources extend the sound palette in various directions with different materials giving them different qualities. Metallic chimes from Japan and resonant metal bars contrast with a bass drum and sounds from the interior of a piano (both string sounds and sounds from hitting the body of the piano). Some sounds were also imported from earlier compositions, *Piano Threads* and *Vortex*, including frequency modulation sounds.”

The reasons for our choice of *Wind Chimes* as the second composition to test our preliminary methods for automated analysis against are two-fold. Firstly, produced with entirely different technologies, compositional methods and sound palettes it helps to ensure that the methods presented here do work not only in the overly specific context of one piece in its individual spectral shaping. Sec-

<sup>13</sup> [http://www.ina-entreprise.com/sites/ina/medias/upload/grm/portraits-polychromes/extraits/chowning/analyse/turenas/turenas\\_app.html](http://www.ina-entreprise.com/sites/ina/medias/upload/grm/portraits-polychromes/extraits/chowning/analyse/turenas/turenas_app.html), retrieved 2011-02-23

<sup>14</sup> For an alternative (drawing rather different conclusions about the piece’s structure) refer to Luis Jure’s analysis (Jure, 2004a)

<sup>15</sup> translation: Thomas Grill

<sup>11</sup> See Wishart (1996) for a detailed discussion of this.

<sup>12</sup> <http://en.wikipedia.org/wiki/Music4>, retrieved 2011-02-23

only, even though no production score with exact timing information exists for *Wind Chimes*, the detailed analyses of Hirst (2006), and Clarke (2010a) allow for testing the algorithm’s results against reliable annotations by human experts. Clarke identifies ten sections in *Wind Chimes* which are listed in table 2.

### 3.3 Preliminaries

In chapter 3.4 we demonstrate how self-similarity matrices can be utilized for generating structural overviews of full compositions or individual sections thereof. In chapter 3.5 we outline how different groups of sonic material within the individual compositions (in Pottier’s analysis ‘morphologies’, in Clarke’s analysis ‘gestural and textural elements’ in his ‘aural taxonomy’) can be identified automatically once the analyst has identified and isolated individual prototypes of these morphological groups.

It is undeniably the case that the analytical methods presented in the following are highly selective in their scope. We are concentrating on only two aspects from the many presented by Pottier’s and Clarke’s analyses since it is not our intention to present another full analysis but to rather sketch possible new methods for analytical practice.

In the case of *Turenas* (a 4 channel composition) and *Wind Chimes* any analysis trying to present a full, if always individual account of the piece, would have to address spatial aspects, pitch relationships and rhythmic developments. The algorithms presented here work on monophonic signals only and are hence oblivious to any spatial developments, just as they will (in the case of *Turenas*) never be aware of the canonical structures present, to mention just two of the most obvious omissions.

### 3.4 Structural overview

The task of automated structural discovery in acousmatic music is comparable to the same task performed on a mixture of environmental sounds. This is because the sound materials used in acousmatic music (as opposed to typical classical or pop music) are generally not restricted to traditional acoustic or amplified instruments, voice or specific electronic synthesis algorithms. Materials found in acousmatic music can rather originate from any conceivable sound source that is also found in everyday, non-musical sonic environments. Sound materials used in acousmatic compositions are often of a pre-structured nature, because of having been recorded (sampled) – hence the internal complexity of such recorded material is folded with the complexity of its deployment in a composition. Consequently, the separation of individual components is often impossible, or at least dependent on a specific focus (or thresholding) applied to the process of structure analysis. In this sense, computer-based annotation of acousmatic music is essentially comparable

to tasks in general *Computational Auditory Scene Analysis* (CASA) (Bregman, 1994) for everyday sonic environments.

#### 3.4.1 Related work

In order to apprehend the contents of a sound stream, so-called *audio descriptors* (see e.g. Herrera et al., 1999b,a) are used to describe significant individual features of the analyzed audio data and also to reduce the amount of data for subsequent analysis. Many different descriptors have been introduced in recent years at various levels of complexity. Audio descriptors are usually computed on *frames*, that is, short audio segments of constant length, typically connected to the calculation of a spectral representation using the discrete Fourier transform (DFT). A number of tools, e.g. IRCAM’s CataRT (Schwarz, 2004), have been developed to allow the exploration of *descriptor spaces*, i.e. multi-dimensional spaces spanned by different audio descriptors.

For the visualization and identification of perceptually significant changes in audio, Foote (1999) first proposed the use of *self-similarity matrices*. A self-similarity matrix is a two-dimensional matrix consisting of the distances between chosen audio descriptors for all pairs of frames in the audio under analysis. The calculation of distances between audio features, or combinations thereof, is done by the use of appropriate *distance measures*. A great variety of such measures is in common use, most notably the Euclidean distance, cosine similarity, Kullback-Leibler divergence, and many more – the actual choice being dependent on the nature of the employed audio features in question.

The work on self-similarity matrices relies for the most part on timbre-based audio descriptors for modeling the audio content, mostly by calculating chromagrams for score-based music or Mel Frequency Cepstrum Coefficients (MFCCs) for more general sounds (e.g. Cooper et al., 2006; Ong, 2007; Chai, 2005). MFCCs are a perceptually meaningful and spectrally smoothed representation of the frequency spectrum of audio signals and are now a standard technique for modeling spectral similarity in music analysis (see e.g. Logan, 2000).

#### 3.4.2 Results

Figure 3 shows a self-similarity matrix of *Turenas*, using MFCCs and the following cosine distance measure as proposed by Foote (2000)<sup>16</sup>:

$$d_{\cos}(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad (1)$$

The calculation of the MFCCs was performed using Pampalk’s MA toolbox<sup>17</sup>, with an FFT size of 1024 samples (23.2 ms), a hop size of 512 samples (11.6 ms) and

<sup>16</sup> Using standard Euclidean distance ( $L^2$  norm) with normalized arguments yields very similar results for all the use cases presented in this paper.

<sup>17</sup> <http://www.pampalk.at/ma>, retrieved 2011-02-23

full spectral resolution with 24 MFCCs. The MFCC vectors have been resampled by calculating the mean over segments of one second length, resulting in 600 segments for *Turenas*' 10 minute duration. The self-similarity matrix can be read in the following way: The brighter a region in the matrix is, the more similar the two audio segments at the points in time corresponding to the region's row and column coordinates will be. The top-left to bottom-right diagonal represents the timeline, and all points in a row left of the diagonal refer to regions appearing earlier in the piece, and those to the right of the diagonal at later times. Looking at figure 3, a block structure can be vaguely discerned with boundaries at ca. 0:30, 2:00, 5:00, and 7:45 minutes (i.e. 30, 120, 300, 465 segments). Comparing this to the structure identified by Pottier (see figure 2), the latter three points in time correspond to the transitions between the main sections I, II, III and IV.

The contrast between adjacent sections of the piece which contain different sound materials, as represented by the self-similarity matrix, can be improved by adding temporal features to the existing timbral ones. The focus of analysis thus shifts from purely timbre-based to more general structure-based criteria. Based on ideas developed by Fröhwhirth and Rauber (2001), Pampalk devised audio descriptors termed *rhythm patterns* (Pampalk, 2001; Pampalk et al., 2002), capable of describing rhythmic characteristics of 'songs', i.e. whole pieces of music. These rhythm patterns, often more generally referred to as *fluctuation patterns* (FPs), describe the amplitude modulation of the loudness per frequency band (see figure 4 for an overview of the feature extraction process). Grill (2010) applied this method to the characterization of textural sounds demonstrating that FPs are capable of efficiently modeling micro-rhythmic temporal structure (i.e. amplitude fluctuations perceived as granularity or roughness) using segments of about one to five seconds duration. FPs are represented by a two-dimensional matrix of considerable size (typically 24 frequency bands  $\times$  30 modulation frequencies) for each analyzed segment. To reduce feature dimensionality we do not use the full FP matrices but the means along both axes, yielding timbral and fluctuation information. This reduces the amount of data to typically 24 + 30 elements per segment. Experiments identified an equally weighted combination of those timbral and fluctuation coefficients to be most significant. Figure 5 shows a self-similarity matrix using timbre and fluctuation coefficients equally weighted on one-second segments, again with a cosine distance measure. In comparison to figure 3 the block structure becomes more distinct, also with less variance within individual blocks.

The automatic detection of block boundaries has been demonstrated by Foote (2000), using a so-called *novelty detection function* with subsequent retrieval of prominent peaks therein. Such detection functions are used throughout various methods of audio segmentation, in most cases

related to *onset detection*. Since in our case we are more interested in the discovery of contiguous regions sharing sonic characteristics than in the detection of events, an appropriate novelty detection function must be based on longer durations than just individual segments. This can be accomplished by row-wise convolution of the self-similarity matrix  $M$  with half-Gaussian kernels, once with the left half ( $G_{\sigma}^{-}$ ) and once with the right half ( $G_{\sigma}^{+}$ ), and then evaluating the distances (again with a cosine distance measure) between those two convolutions:

$$n(t) = 1 - d_{\cos}((M \star G_{\sigma}^{-})(t), (M \star G_{\sigma}^{+})(t)) \quad (2)$$

with the discretely sampled half-Gaussians

$$G_{\sigma}^{\pm}(x) = H(\pm x) \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma} \quad (3)$$

and  $H$  representing the Heaviside step function. The resulting novelty function  $n(t)$  for column index  $t$  (equation 2) yields values in the range 0 (audio features left and right of  $t$  are identical) to 1 (audio features left and right of  $t$  are totally dissimilar) with maxima revealing the block boundaries of the self-similarity matrix. The novelty function  $n(t)$  for *Turenas* is shown in figure 6 together with Pottier's manually annotated blocks. In order to limit the number of identified peaks a threshold can be used, with useful values in the range of approximately  $-2.7$  to  $-2.0$  on  $\log_{10}$ -scale, as determined empirically. Local maxima above a threshold of  $-2.5$  are also shown in figure 6. The other adjustable parameter involved is the width  $\sigma$  of the half-Gaussians used in the convolution. It describes the temporal extent of the two matrix parts weighted by the convolutions with the half-Gaussian functions and then compared with the distance function: one reaching from  $t$  to earlier frames, the other reaching from  $t$  to later frames. Values for  $\sigma$  between 2 and 10 (corresponding to 2 to 10 seconds) result in amounts of detail useful for our annotation purposes. For the examples in this paper a value for  $\sigma$  of 5 seconds was used for segmentation on the timescale of entire pieces and lower values for parts thereof.

The boundaries resulting from the peaks in the novelty function can be used to perform high-level segmentation of the audio data, yielding regions of similar audio descriptor characteristics. Figure 7 shows such a segmentation with the partitioning, as manually annotated by Pottier, superimposed. The self-similarity matrix in figure 7 is computed by: (i) segmentation in time according to peaks of the novelty function, (ii) averaging the audio descriptors within each segment, (iii) calculating the cosine distances between all those segment averages (equation 1). As with the self-similarity matrices in figures 3 and 5, the brighter a region is, the more similar are the two segments identified by the respective row and column positions. For example, figure 7 suggests that section II-A should exhibit some resemblance

to section IV-A, which is indeed the case, due to the use of material 6b, a sustained drone-like sound. Sections II-B, III-A and III-B in the middle part on the other hand resemble each other due to the persistent use of rather percussive material, being particularly dry and impulsive within the two sections of part III. The plot also shows that sections I and IV-B are both composed of a few considerably differing subsections, related to the different materials identified by Pottier.

The detailedness of the high-level segmentation is dependent on the number of peaks appearing in the novelty function, as explicated above. Figure 8 shows the results of slightly different thresholding. The proper value will certainly always be the result of some experimentation although values in the narrow range between  $-2.5$  and  $-2.3$  haven proven to robustly yield meaningful results. The tuning of this value could also be made dependent on some alternative pre-defined setting for segment density or average segment length.

As a second example figure 9 demonstrates the same analysis procedure applied to Denis Smalley’s composition *Wind Chimes*. Again a segmentation based on a novelty function (this time with a slightly lower threshold of  $-2.5$ ) was calculated clearly exposing section boundaries that closely align with those identified by Clarke (2010a). Switching to a higher time resolution by using a smaller filter width  $\sigma$  lets the subsection boundaries and the use of various sound materials emerge in the self-similarity matrix. This is demonstrated in figure 10, depicting a more detailed automatic analysis of the first main section of *Wind Chimes*. Therein, a limitation of the current implementation is revealed, concerning the rather coarse time resolution of at least a few seconds depending on the temporal support of the FP analysis and the width of the subsequent Gaussian smoothing filter. Hence, due to the fact that it is only about 1.5 seconds long, subsection 1B remains beyond the proposed method’s scope. The other subsection boundaries correspond well to Clarke’s manual annotation.

### 3.5 Interactive detection of representative sound groupings

Besides segmentation, the detection and localization of representative or prototypical sounds, e.g. appearing repeatedly within a piece or representing its building parts is of particular interest. On the basis of self-similarity matrices this is already somewhat possible by bringing into relation the bright bits in the graphs, indicating mutual similarity. However, a more direct and visually clearer way of relating temporal and sound-based groupings to one another would be desirable.

#### 3.5.1 Related work

One application of visualizing the similarities of musical data has been demonstrated by Pampalk in his project *Islands of Music* (Pampalk, 2001). Therein a set of musical pieces (termed ‘songs’) have been arranged on a two-dimensional map, exhibiting *islands* (i.e. clusters) of shared characteristics, as modeled by fluctuation patterns used as audio descriptors. Similar approaches towards visualization of music databases can be found in the literature (see e.g. Rauber and Frühwirth, 2001; Gasser and Flexer, 2009). The main challenge of such a visualization task is to translate high-dimensional data (i.e. audio features) to a low-dimensional (in most cases two-dimensional) representation preserving significant structure inherent in the data. Once the data has been visualized, structure and clusters can be identified and denoted by simply looking at the two-dimensional representation and using one’s own subjective judgement. This has been referred to as ‘clustering via visualization’ (Flexer, 2001). Among the many different existing visualization techniques (such as principal component analysis (PCA), multi-dimensional scaling (MDS), self-organizing feature maps (SOM), a survey is given by Ferreira de Oliveira and Levkowitz (2003)) a method termed ‘t-Distributed Stochastic Neighbor Embedding’ (*t-SNE*, van der Maaten and Hinton, 2008; van der Maaten, 2009) has of late received much attention. T-SNE performs particularly well in preserving both the local and global structure, revealing the presence of clusters within the data.

#### 3.5.2 Results

Figure 11 (left graph) shows the entirety of one-second long snippets of *Turenas* (overlapped with a hop size of 0.2 seconds), modeled by an equally-weighted combination of timbre and fluctuation information (as in section 3.4) and mapped to two dimensions by use of the *t-SNE* method. The data points have been color-coded according to their temporal position in the piece. Dark points correspond to positions early in the piece, light ones to positions later on<sup>18</sup>.

As a result of the *t-SNE* algorithm the formation of clusters can be observed on the two-dimensional map. Data points in close proximity on the map correspond to similarly sounding segments. Continuous developments in sound are reflected by line-shaped agglomerations (also supported by the overlapping of the segments), while varying (e.g. stochastic) sounds form more irregular arrangements in the map. Distinct clusters can be associated with representative or even *prototypical* sounds or morphologies of a piece. Many of the clusters in figure 11 (left graph) are uniformly colored, corresponding to sound groupings that appear at

<sup>18</sup> The gray scale has only been chosen for restrictions of print, a rainbow-colored scale would make things even clearer.

unique temporal positions. Other clusters exhibit mixed colors, for sounds that appear alike at multiple positions.

An interactive Max/MSP<sup>19</sup>-based application has been created to allow browsing through the sounds in the map. Sound spaces can be directly experienced and listened to with the computer mouse traveling from sound object to sound object (circles in figure 11, left graph). The combination of location and color for agglomerated data points facilitates the association of morphologies as identified by Pottier (see figure 2) to clusters within the map. Figure 11 (right graph) explicates this association to the sound morphologies. The ellipses have been manually positioned according to the observed clusters in the left graph of the figure. Most of the clusters correspond very well to Pottier's identified morphologies, others represent combinations of those sounds, and some are miscellaneous sounds, like pure or noisy silence, spurious attacks and other mixed sounds.

With the clustering method at hand the identification of sound groupings is still restricted to morphologies appearing 'solo'. Simultaneous sounds cannot be separated, although, to some extent, contributions of sound groupings (as modeled by the audio descriptors) over the duration of a piece can be analyzed, as shown in figure 12. Four quite distinct morphologies, namely 1, 6b, 7e and 8c as identified by Pottier, have been chosen from the whole set (influenced by the criterium of solo-ness), models using the already established audio descriptors generated, and their respective similarities to all the instances in the piece calculated. The zones highlighted in gray in the figure depict the locations where the individual sound prototypes have been retrieved. The four different curves show the similarities to those prototypes throughout the whole piece. In the bar on top of the graphs predominant similarities have been annotated. This analysis allows inference about sound characteristics (percussive, drone-like, grainy, etc.), represented by those prototypical sounds, or likewise, by external exemplary sounds not contained in the piece itself. For instance, the attributed similarity to material 6b between seconds 25 and 80 is due to a predominant drone-like nature, caused by material 3a according to Pottier. Materials 3a and 6a are alike in many respects. In the same manner, the morphological prototype 7e represents the dry percussive nature of sounds between seconds 300 and 450, including materials 7a through 7e, while 8c stands for the succession of reverberant events between seconds 210 and 300, involving materials 8a through 8d.

For the second example, *Wind Chimes*, the same attempt of clustering the occurring sounds was carried out, as shown in figure 13. In this case the association of distinct clusters on the two-dimensional map to the annotation by Clarke (in his 'Aural Paradigmatic Chart') proves to be more difficult, as very rarely do the identified prototypes appear isolated in the piece. They rather function as fundamental components

in varying forms merging to composite sound evolutions. The simultaneous occurrence of those gestural and textural constituents as annotated by Clarke inhibits their discovery by use of the methods presented here. Nevertheless, the formation of location-color clusters in the map can be observed. Their relationship to certain sections in the piece is obvious, which is what has been manually annotated in figure 13 (right graph). In most cases the individual ellipses can be associated to unique sections, in some cases sections share sonic characteristics and consequently fall into the same cluster.

### 3.6 Discussion

As already indicated in the text above the presented methods for semi-automatic discovery and annotation of structural components in acousmatic music have a number of limitations. One unavoidable issue is the inevitability of user-adjustable parameters. Some of those parameters are less influential, like all those inherent in the audio feature extraction process. Others, like FP analysis length, Gaussian filter width or the novelty threshold are directly connected to a chosen *analysis focus* in regard to temporal resolution or segmentation sensitivity. Adequate values for those parameters are also dependent on the type of sonic material present in a specific piece of music. Therefore the analyst must be aware of the existence and impact of those parameter settings and actively work with them to achieve the desired results.

Other limitations are due to the specific technological methods or implementations used for the analysis. For instance, the chosen FP-based audio descriptors merely incorporate characteristics of timbre and amplitude fluctuations. Hence, any morphological developments and the presence of singular events can not be explicitly modeled. Nevertheless those structures could leave distinct indirect traces in what may in fact be inapt models. The limited minimal time resolution of FPs, amounting to about one second, could be alleviated by using Constant-Q-based (Brown, 1991) instead of FFT-based temporal processing, effectively making the time support dependent on the fluctuation frequency and improving overall temporal resolution.

A fundamental drawback of the techniques for structure discovery presented here is the inability to detect sound groupings that only ever appear in combination with others, but never standing alone. Many of the sound primitives annotated by Clarke for *Wind Chimes* are of that nature. The technologies researched in the literature tackling the superposition of prototypical sounds, most notably perceptually weighted Non-Negative Matrix Factorization (NMF) (Virtanen, 2007; Kirbiz and Günsel, 2010) or other forms of mixture models, are not yet capable of handling real-life situations, involving a larger number of sound prototypes appearing in varied forms.

<sup>19</sup> <http://cycling74.com/products/maxmsp/jitter>, retrieved 2011-02-23

All the technical implementations for methods presented in section 3 are in a prototypical state with command-line interfaces only. Actual annotation practice of acousmatic pieces would gain tremendously from the existence of appropriate user interfaces and ad-hoc re-calculation upon changing of parameters – then indeed simplifying and accelerating the cumbersome procedures of manual analytical work. First steps towards such integrated systems have already been reported in the literature (Park et al., 2010).

## 4 CONCLUSION

This paper presented an inquiry concerning the feasibility of using Music Information Retrieval (MIR) methods for automated annotation of acousmatic music. As outlined in the introduction (section 1), this inquiry allows us to answer three central questions concerning automated annotation.

The first question concerns the applicability of MIR methods to identify carriers of meaning within works of acousmatic music. The answer to this question is positive: MIR methods can be used for annotation of carriers of meaning but only in an interactive setting. Only the integration of a human analyst into the workflow allows us to sidestep the seeming impasse that the lack of a ‘ground truth’ in annotation of acousmatic music presents. A human analyst in front of the computer, taking all the analytical decisions and also interpreting the output of a repertoire of algorithms is able to compensate for the lack of semantic comprehension on the side of the computer. This has been discussed in greater detail in section 2.8.

The second question concerns the level of automation achievable. Answering our first question, we already established that full automatic annotation is not possible and that a good deal of the process would have to be conducted manually. This is due to the above mentioned lack of a ‘ground truth’ but also due to the present state-of-the-art in MIR and signal processing. Many of the decisions concerning algorithm parameters and threshold settings cannot reliably be automated but require human intervention and decisions as to what the focus of analysis should be. In section 3 we presented two possible routes of exploration into the structural and sonic nature of acousmatic compositions: one provides a structural overview of complete pieces of music while the other allows identification and clustering of representative sound groupings at a more detailed level. Although we did apply these two approaches to two prominent works of acousmatic music we like to emphasize again that it was not our intention to present yet two more full analyses. We only presented interactive workflows that are feasible in relation to the current state of technology and that do provide us with insightful results.

The third question concerns the level of abstraction and complexity that can be reached in automating annotation of acousmatic music. This question cannot be answered eas-

ily since it is dependent on the highly varying complexity of acousmatic music. While some compositions might be quite simple and clear in their design and hence lend themselves well to comprehensive abstraction, the analyses of others do indeed present formidable challenges. It is important though to understand however that complexity in this context cannot be interpreted as a problem limited only to the engineering side of things. A composition’s complexity, be it based on the frequency of changes in spectrum or the ambiguities thereof, increases not only the challenges in signal processing, but also in terms of manual annotation. The more complex a composition is, the less uniform the results of human annotation will be. Since the human annotations are used as ‘ground truth’ for evaluation of algorithms for automated annotation, this introduces a certain blurring in the overall system.

The two examples of automated annotation provided in section 3 are limited to classifying sonic materials into groups of similar timbral and structural attributes. As discussed in section 2.8, this is an essential initial step of any analysis. One of the main current technological challenges to making improvements in this direction is the multitude and diversity of temporally overlapping sound sources present in most acousmatic compositions. Possible routes for future work also concerning other important dimensions of sound analysis (e.g. morphological aspects) have been discussed in section 3.6.

As a final comment we would like to re-iterate that the work presented here of course does not conclusively answer all questions concerning the application of MIR methods to automated annotation of acousmatic music. But we are confident that we were able to clarify at least two major points: (i) MIR can play a valuable role in analysis of acousmatic music, (ii) MIR methods will not be able to replace human annotators but will aid experts by providing interactive tools.

## Acknowledgements

This work has been supported by the Austrian Research Fund FWF (project number P21247), Uni:vision2 (University of Music and Performing Arts Vienna) and the Vienna Science and Technology Fund WWTF (Audiominer, project number MA09-024).

## References

- Adams, N. (2006). *Analytical Methods of Electroacoustic Music*, chapter Visualization of Musical Signals, pages 13–28. Routledge.
- Atkinson, S. (2007). Interpretation and musical signification in acousmatic listening. *Organised Sound*, 12(2):113–122.

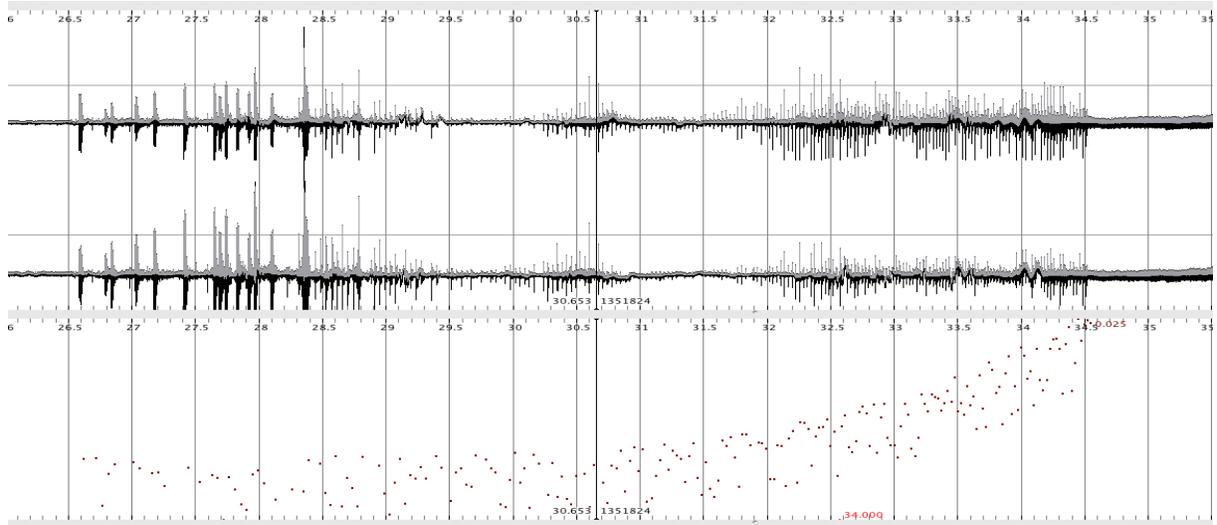
- Bregman, A. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, Cambridge, MA.
- Brown, J. C. (1991). Calculation of a constant  $q$  spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434.
- Chai, W. (2005). *Automated analysis of musical structure*. PhD thesis, Massachusetts Institute of Technology.
- Chowning, J. (1988). Turenas – Stria – Phoné – Sabelithe. WERGO, WER 2012-50.
- Clarke, E. F. (2005). *Ways of Listening. An Ecological Approach to the Perception of Musical Meaning*. Oxford University Press.
- Clarke, E. F. and Cook, N. (2004). *Introduction: What is Empirical Musicology?*, pages 3–14. Oxford University Press.
- Clarke, M. (2010a). Wind chimes, an interactive aural analysis. In *Denis Smalley, Polychrome Portraits*. INA-GRM.
- Clarke, M. (2010b). Wind chimes, an interactive aural analysis. <http://www.inagram.com/polychromes/smalley/clarke>, retrieved 2011-03-03.
- Cooper, M., Foote, J., Pampalk, E., and Tzanetakis, G. (2006). Visualization in audio-based music information retrieval. *Comput. Music J.*, 30(2):42–62.
- Dahan, K. (2007). Surface tensions: Dynamics of Stria. *Computer Music Journal*, 3(3):65–74.
- Dahlhaus, C. (1965). *Notation neuer Musik*, chapter Notenschrift heute, pages 9–34. Darmstädter Beiträge zur Neuen Musik. Schott.
- Delalande, F. (2007). The technological era of ‘sound’: a challenge for musicology and a new range of social practices. *Organised Sound*, 12(3):251–258.
- Delalande, F., Formosa, M., Frémiot, M., Gobin, P., Malbosc, P., Mandelbrojt, J., and Pedler, E. (1996). *Les Unités Sémiotiques Temporelles: Éléments nouveaux d’analyse musicale*. Éditions MIM - Documents Musurgia, Marseille.
- Ferreira de Oliveira, M. and Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378–394.
- Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis*, 5(5):373–384.
- Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 77–80.
- Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME’00)*, volume 1, pages 452–455, New York, NY, USA.
- Frisius, R. (2002). *Forum Analyse: Medienspezifische Analyse – Analyse medienspezifischer Musik; Einführung*, volume Konzert-Klangkunst-Computer, Wandel der musikalischen Wirklichkeit, pages 170–171. Schott.
- Frühwirth, M. and Rauber, A. (2001). Self-organizing maps for content-based music clustering. In *Proceedings of the 12th Italian Workshop on Neural Nets (WIRN01), Vietri sul Mare*. Springer.
- Gasser, M. and Flexer, A. (2009). FM4 soundpark: Audio-based music recommendation in everyday use. In *Proceedings of the 6th Sound and Music Computing Conference*, Porto, Portugal.
- Gayou, É. (2006). Analysing and transcribing electroacoustic music: the experience of the Portraits polychromes of GRM. *Organised Sound*, 11(2):125–129.
- Godøy, R. I. (2006). Gestural-sonorous objects: embodied extensions of Schaeffer’s conceptual apparatus. *Organised Sound*, 11(2):149–157.
- Godøy, R. I. and Leman, M., editors (2009). *Musical Gestures: Sound, Movement, and Meaning*. Routledge.
- Goodman, N. (1976/1981). *Languages of Art; An Approach to a Theory of Symbols*. Harvester Press.
- Grill, T. (2010). Re-texturing the sonic environment. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, AM ’10, pages 6:1–6:7, New York, NY, USA. ACM.
- Herrera, P., Serra, X., and Peeters, G. (1999a). Audio descriptors and descriptor schemes in the context of MPEG-7. In *Proceedings of ICMC*.
- Herrera, P., Serra, X., and Peeters, G. (1999b). A proposal for the description of audio in the context of MPEG-7. MPEG.
- Hirst, D. (2006). *The development of a cognitive framework for the analysis of acousmatic music*. PhD thesis, University of Melbourne.
- Jure, L. (2004a). Escuchando Turenas de John Chowning. <http://www.eumus.edu.uy/revista/nro1/jure.html>, retrieved 2011-03-03.

- Jure, L. (2004b). Resintetizando Turenas. [http://www.eumus.edu.uy/docentes/jure/pubs/resintetizando\\_turenas/](http://www.eumus.edu.uy/docentes/jure/pubs/resintetizando_turenas/), retrieved 2011-03-03.
- Kirbiz, S. and Günsel, B. (2010). A perceptually enhanced blind single-channel audio source separation by non-negative matrix factorization. In *Proceedings of 18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark.
- Leman, M. (2007). *Embodied Music Cognition and Mediation Technology*. MIT Press.
- Lesaffre, M., Leman, M., de Baets, B., and Martens, J.-P. (2004). Methodological considerations concerning manual annotation of musical audio in function of algorithm development. In *Proceedings of the International Conference on Music Information Retrieval*.
- Licata, T. (2002). *Electroacoustic Music: Analytical Perspectives*, volume 63 of *Contributions to the Study of Music and Dance*, chapter Introduction, pages xxi – xxiv. Greenwood.
- Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Conference on Music Information Retrieval*, Plymouth, Massachusetts.
- Marsden, A. (2009). "What was the question?": Music analysis and the computer. In *Modern Methods for Musicology*, pages 137–147. Ashgate Pub Ltd.
- Meneghini, M. (2007). An Analysis of the Compositional Techniques in John Chowning's *Stria*. *Computer Music Journal*, 31(3):26–37.
- Neuhoff, J. G., editor (2004). *Ecological Psychoacoustics*. Elsevier Academic Press.
- Normandeau, R. (2010). Spectromorphology of Denis Smalley. In Gayou, É., editor, *Denis Smalley, Portraits Polychromes*, pages 79–88. INA-GRM.
- Ong, B. S. (2007). *Structural Analysis and Segmentation of Music Signals*. PhD thesis, University Pompeu Fabra, Barcelona, Spain.
- Pampalk, E. (2001). Islands of music: Analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, Vienna, Austria.
- Pampalk, E., Rauber, A., and Merkl, D. (2002). Content-based organization and visualization of music archives. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 570–579, New York, NY, USA. ACM.
- Park, T. H., Hyman, D., Leonard, P., and Wu, W. (2010). Systematic and quantitative electro-acoustic music analysis (sqema). In *Proceedings of ICMC2010*, pages 199–206.
- Pottier, L. (2005). *John Chowning, Portraits polychromes*, chapter Turenas, analyse. Éditions Michel de Maule.
- Rauber, A. and Frühwirth, M. (2001). Automatically analyzing and organizing music archives. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 402–414, London, UK. Springer-Verlag.
- Roy, S. (2003). *L'analyse des musiques électroacoustiques; Modèles et propositions*. L'Harmattan.
- Schaeffer, P. (1952). *À la recherche d'une musique concrète*. Éditions du Seuil, Paris, France.
- Schwarz, D. (2004). *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Ircam - Centre Pompidou, Paris, France.
- Simoni, M. (2006). *Analytical Methods of Electroacoustic Music*, chapter Acknowledgments, page vii. Routledge.
- Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised Sound*, 2(2):107–26.
- Smalley, D. (1999). Établissement de cadres relationnels pour l'analyse de la musique postschaefferienne. In Dufour, D. and Thomas, J.-C., editors, *Oùir; entendre, écouter, comprendre après Schaeffer.*, pages 177–213. INA-GRM/Buchet-Chastel.
- Smalley, D. (2004). Wind chimes. On: Impacts Intérieurs, Montreal: empreintes DIGITALes, IMED 0409.
- Smalley, D. (2010). Spectromorphology in 2010. In Gayou, É., editor, *Denis Smalley, Polychrome Portraits*, pages 89–101. Ina-GRM.
- Smith, L. (1972). SCORE – a musician's approach to computer music. *J. Audio Eng. Soc.*, 20(1):7–14.
- van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AI-STATS)*, volume 5, pages 384–391. JMLR W&CP.
- van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.
- Virtanen, T. (2007). Monaural sound source separation by perceptually weighted non-negative matrix factorization. Technical report, Tampere University of Technology, Institute of Signal Processing.

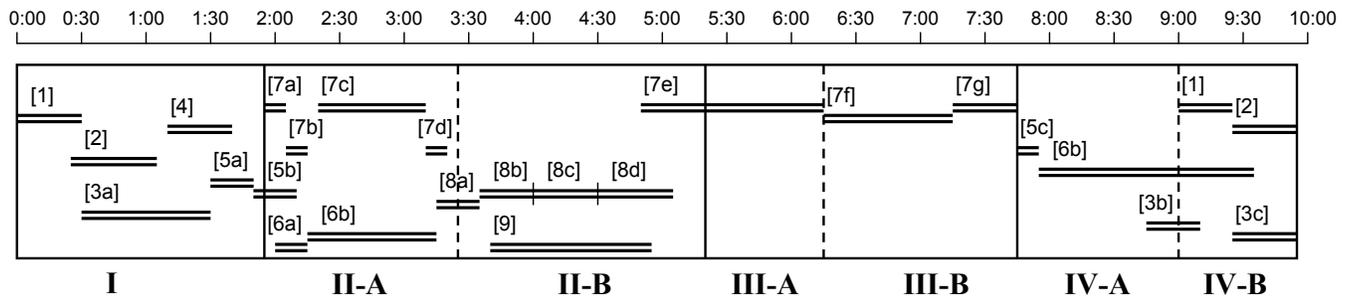
Wiggins, G. (2009). Semantic Gap?? Schemantic Schmap!!  
Methodological Considerations in the Scientific Study of  
Music. In *2009 11th IEEE International Symposium on  
Multimedia*, pages 477–482. IEEE.

Wishart, T. (1985/1996). *On Sonic Art*. Harwood, Amster-  
dam.

Zelli, B. (2001). *Reale und virtuelle Räume in der Comput-  
ermusik*. PhD thesis, Technische Universität Berlin.



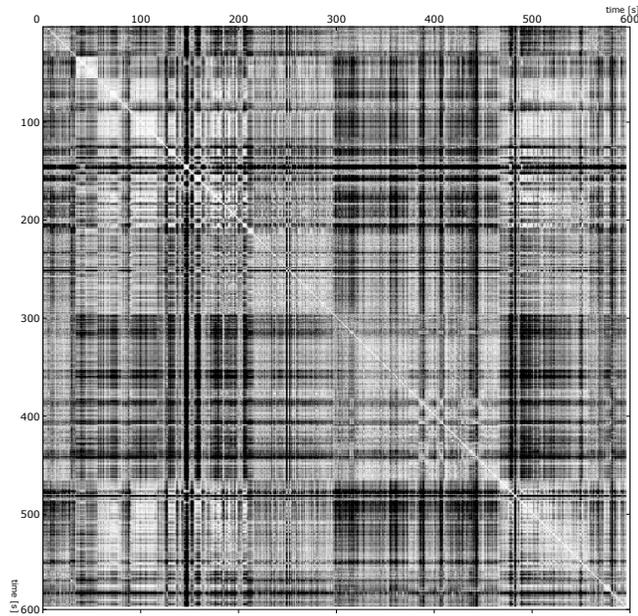
**Figure 1.** *Turenas*, seconds 00:26 – 00:35, stereo waveform (top two curves) and event onsets below (dots at bottom), as taken from original production score. Time is depicted on the x-axis. Please note the ever denser onsets (dots) towards the end of this excerpt which are almost impossible to annotate manually (see section 2.6 for detail).



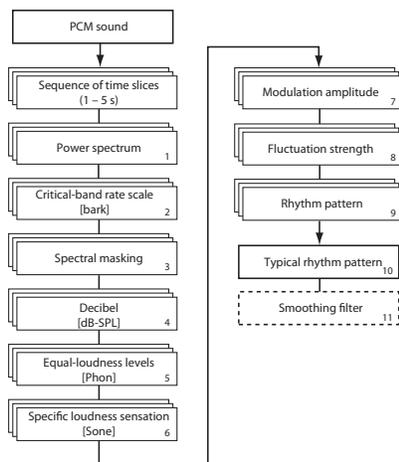
**Figure 2.** *Turenas*, structural overview according to Pottier (2005), time depicted on x-axis. Numbers correspond to morphologies given in table 1.

| sound | envelope   | duration | harmonics  | spectrum    | register   | modulation   |
|-------|------------|----------|------------|-------------|------------|--------------|
| 1     | percussive | short    | sine       | none        | high       | mod          |
| 2a    | granular   | long     | sine       | none        | high       | granular     |
| 2b    | granular   | long     | simple     | little      | high       | granular     |
| 3a    | dampened   | long     | harmonic   | little      | middle     | beating      |
| 3b    | dampened   | long     | harmonic   | rich        | low        | beating      |
| 4     | percussive | short    | varying    | varying     | high       | –            |
| 5     | percussive | long     | inharmonic | rich        | middle/low | medium/heavy |
| 6     | dampened   | long     | nasal      | medium      | medium     | beating      |
| 7a    | percussive | short    | inharmonic | noisy       | wide       | –            |
| 7b    | percussive | short    | harmonic   | medium      | wide       | –            |
| 8a    | dampened   | medium   | harmonic   | medium      | medium     | –            |
| 8b    | dampened   | medium   | inharmonic | medium      | medium     | –            |
| 9     | dampened   | medium   | inharmonic | rather rich | low        | beating      |
| 10    | dampened   | long     | inharmonic | hollow      | low        | heavy        |

**Table 1.** Listing of sound morphologies found in Chowning’s *Turenas* according to Pottier (2005).



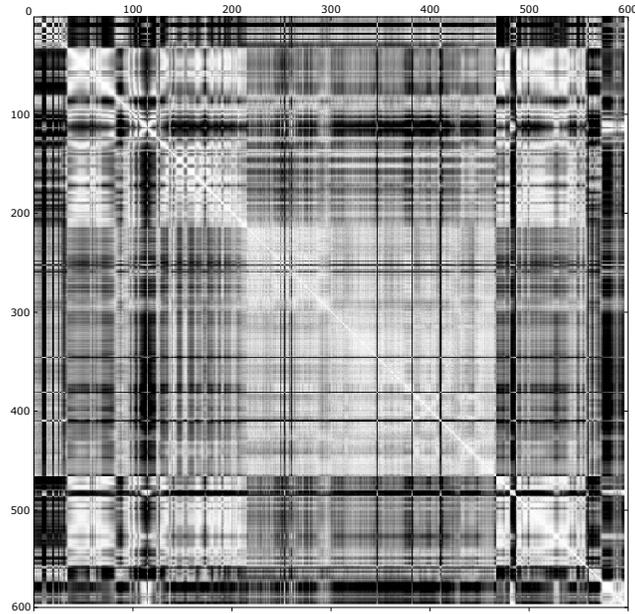
**Figure 3.** *Turenas*, self similarity matrix of 600 one-second segments modeled by timbral features (24 MFCCs).



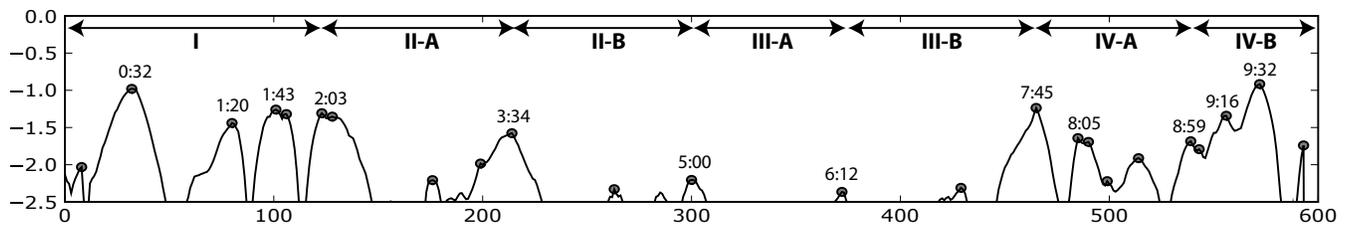
**Figure 4.** Overview of the feature extraction process for the fluctuation pattern model (adapted from Pampalk et al., 2002).

|     |           |
|-----|-----------|
| 1   | 0         |
| 2   | 0'45,5''  |
| 3   | 1'25,7''  |
| 4   | 3'33,9''  |
| 5   | 4'38,5''  |
| 6   | 6'28,5''  |
| 7   | 7'11,8''  |
| 8   | 10'50,8'' |
| 9   | 12'52,5'' |
| end | 15'16,5'' |

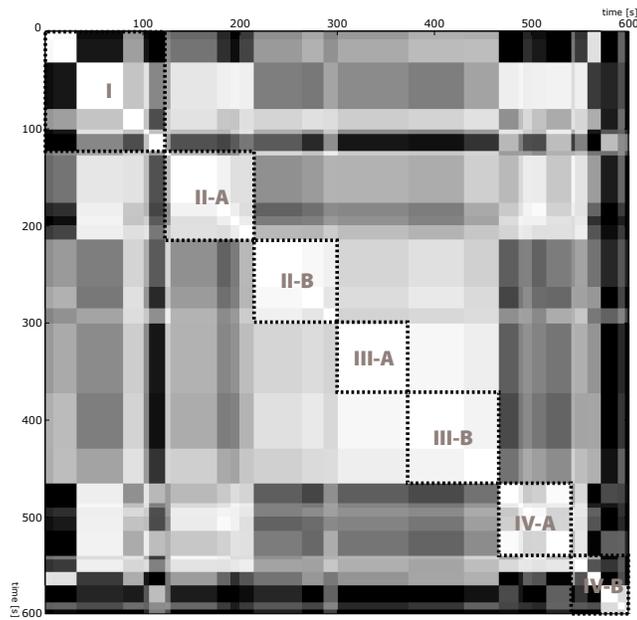
**Table 2.** *Wind Chimes*, begin times of sections (2004 recording) according to Clarke (2010a).



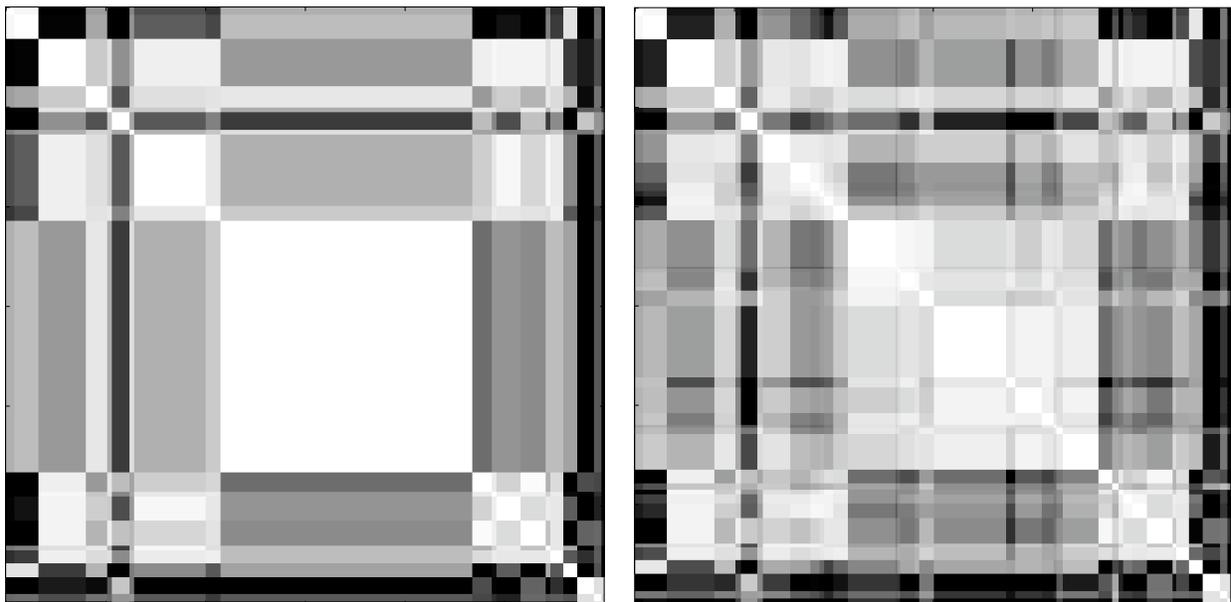
**Figure 5.** *Turenas*, self similarity matrix of 600 one-second segments modeled by an equally-weighted mixture of timbre and fluctuation coefficients.



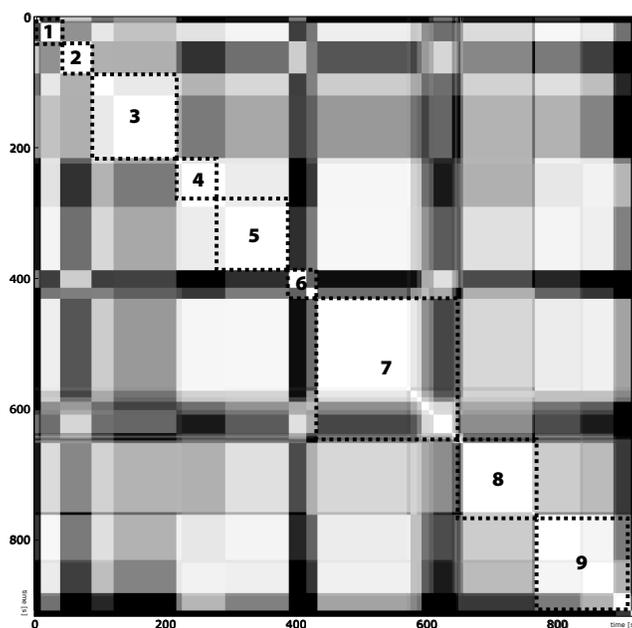
**Figure 6.** *Turenas*, novelty function ( $\log_{10}$ -scaled, y-axis) used for high-level segmentation (calculated from the self-similarity matrix in figure 5) in comparison to Pottier's manually annotated blocks, time depicted on x-axis. Local maxima of the novelty function above a threshold of  $-2.5$  are indicated as circles plus time information in minutes and seconds.



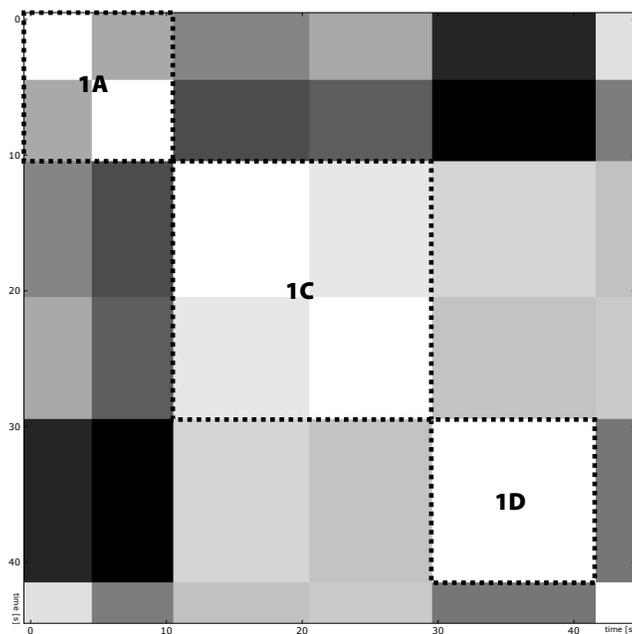
**Figure 7.** *Turenas*, self similarity matrix of automatically segmented high-level blocks in comparison to Pottier’s manually annotated blocks. Novelty threshold has been set to  $-2.4$ .



**Figure 8.** *Turenas*, self similarity matrices of automatically segmented high-level blocks, with novelty thresholds for segmentation set to values lower and higher than in figure 7 ( $-2.2$  on the left-hand side,  $-2.6$  on the right-hand side).

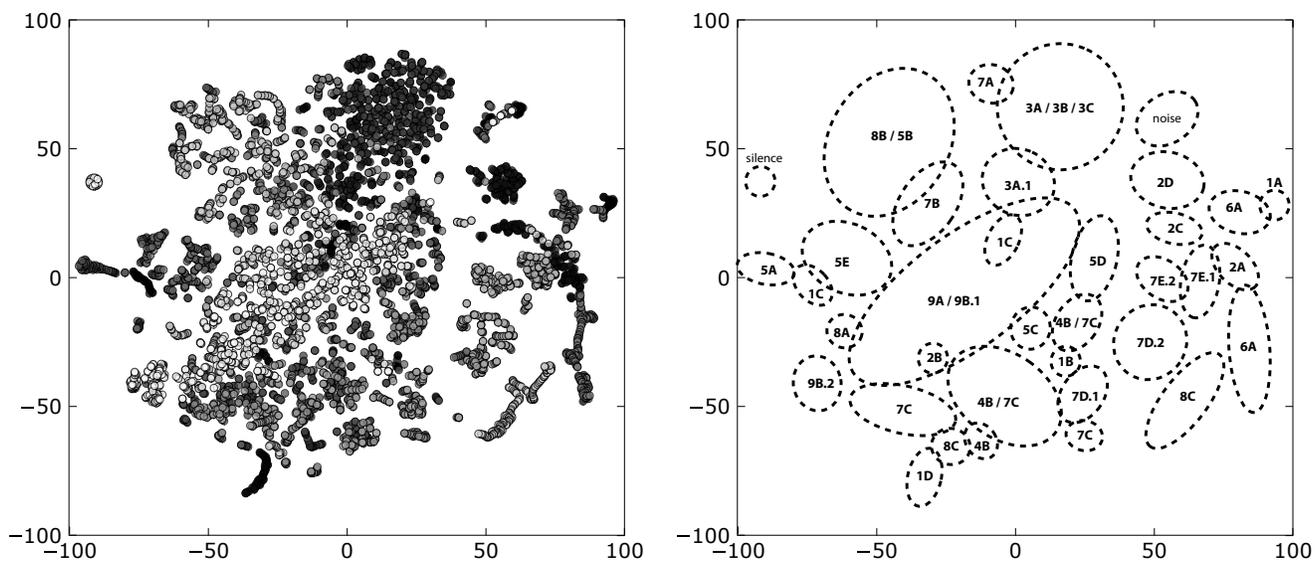


**Figure 9.** *Wind Chimes*, self similarity matrix of automatically segmented high-level blocks in comparison to Clarke’s manually annotated blocks (see Clarke, 2010a). Novelty threshold has been set to  $-2.5$ .



**Figure 10.** *Wind Chimes*, self similarity matrix of the first 45 seconds, covering the entire first main section. In order to reveal subsections, the width  $\sigma$  of the Gaussian filter kernel has been set to 2 seconds, novelty threshold stays at  $-2.5$ .





**Figure 13.** *Wind Chimes*, two-dimensional map of sounds in the piece. Left graph: mapping achieved with t-SNE visualization method, each circle represents one second of sound, proximity on map signifies similarity of sound, temporal information coded from dark (beginning) to light (end of *Wind Chimes*). Right graph: manual annotation of map taking into account sections identified by Clarke (2010a).