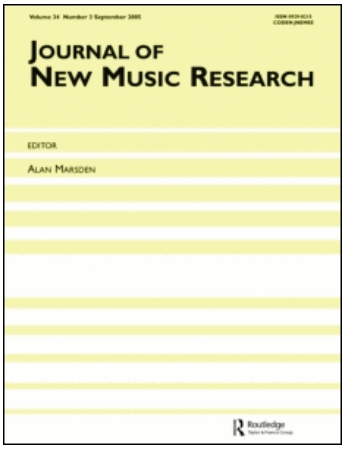

Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713817838

Automatic Harmonic Description of Musical Signals Using Schema-based Chord Decomposition

Francesco Carreras; Marc Leman; Micheline Lesaffre

Online publication date: 09 August 2010

To cite this Article Carreras, Francesco , Leman, Marc and Lesaffre, Micheline(1999) 'Automatic Harmonic Description of Musical Signals Using Schema-based Chord Decomposition', Journal of New Music Research, 28: 4, 310 — 333
To link to this Article: DOI: 10.1076/0929-8215(199912)28:04;1-O;FT310
URL: http://dx.doi.org/10.1076/0929-8215(199912)28:04;1-O;FT310

Journal of New Music Research, 28 (1999), No. 4, pp. 310–333 0929-8215/99/2804-310$15.00

Automatic Harmonic Description of Musical Signals Using Schema-based Chord Decomposition*

Francesco Carreras[1], Marc Leman[2] and Micheline Lesaffre[2]

[1]Multimedia Systems Department, CNUCE/CNR, Pisa, Italy
[2]IPEM — Department of Musicology, University of Ghent, Ghent, Belgium

ABSTRACT

This paper presents a model for the harmonic description of musical signals using schema-based chord decomposition. The model is based (i) on a bottom-up processing of musical signals into auditory images of different kinds and (ii) on a schema-based top-down processing of the images which produces chord decomposition. The schema is carried by a neural network and its content is derived from the self-organization of a pre-selected set of chords in all keys. The system architecture of this model is described and its performance is analyzed in detail through several excerpts of piano pieces. An evaluation procedure has been developed to compare the results with the ones derived using a music theoretical approach to chord decomposition.

INTRODUCTION

Automatic transcription of polyphonic music is known to be a hard problem. Its goal is to transcribe an audio signal into a score, much like a spoken voice is translated into a written text. Music, however, is often made up of many voices, "speaking" or playing all at the same time. The recognition of different auditory objects in synchrony or as streams adds, therefore, complexity to the problem.

Several researchers have tackled the problem in the past. Two approaches may be distinguished: a sonological approach, adopted since the beginning of the seventies, and an auditory-based approach that, following the progress in auditory modeling and speech recognition, was introduced in the early 1990s.

The sonological approach is based on frequency and time analysis techniques. The first attempts for computer transcription of sounded music started in the 1970s. They were limited to monophonic music (Askenfelt, 1976; Sundberg and Tjernlund, 1970). Moorer (1975) developed a more general method for two-part music. The approach was constrained to harmonic sounds and primes; octaves and duodecimes (octave + fifth) were not allowed. By applying Artificial Intelligence techniques, knowledge about the musical style was used to improve the process of automatic transcription (Chafe et al., 1982; Foster et al., 1982). However, good performance was not generally achievable other than for the musical style chosen. The adoption of improved techniques derived from digital signal processing, such as asynchrony onset detection (Schloss, 1985) and peak frequency extraction (Katayose and Inokuchi, 1989), contributed to more precise note identifica-

*Sound examples & color images are available in the JNMR Electronic Appendix (EA) which can be found on the WWW at http://www.swets.nl/jnmr/jnmr.html

Correspondence: Francesco Carreras, Multimedia Systems Department, CNUCE/CNR, via S. Maria 36, I-56126 Pisa, Italy. Tel.: +39-050-593224. E-mail: Francesco.Carreras@cnuce.cnr.it

tion and voice separation. Recent work by Casajus-Quiros and Fernandez-Cid (1994) focused on sophisticated pattern-matching and peak detection.

The approach based on auditory analysis holds that automatic recognition should rely on the capabilities of human information processing. Progress in the physiology of the auditory system[1] has recently led to a number of computational models that simulate the peripheral auditory information processing (e.g., Balliello et al., 1998; Cosi et al., 1994; Meddis and Hewitt, 1991; Van Immerseel and Martens, 1992). Some authors have applied auditory models to the problem of automatic transcription. Hanappe (1994) used a schema-based approach in which an auditory model was connected to a neural network. Martin (1996) uses an auditory model as a front end and a blackboard framework containing (among other things) information about the analyzed music.

From the viewpoint of signal processing, however, a distinction between sonological and auditory-based approaches cannot always be strictly maintained. A time-frequency analysis technique such as Wavelet transform displays some similarity with the acoustical front end of auditory models in that it is based on a logarithmic spacing of (constant Q) bandpass filters in the frequency domain (Solbach et al., 1998). Most of the auditory models, on the other hand, implement the hydromechanics of the cochlea in terms of (quasi constant Q) band-pass filtering and a subsequent signal to neural transformation.

The auditory approach, however, implies that the goal of automatic transcription is not to reproduce the exact original score, but to describe the music the way it is perceived. The basic level at which features are extracted is not the musical signal, but the primary auditory image thereof, i.e., the signal as filtered by the human ear and encoded by the neurons that innervate the ear. Hence, automatic transcription, based on auditory information processing, is necessarily constrained in that aspects of pitch perception and onset detection are dependent on constrained feature extraction processes of the auditory system, such as masking and virtual pitch perception. The perceived features, however, are believed to be the ones that are most needed for exploiting the results of automatic transcription in a practical application (e.g., musical data mining applied to recorded music). In that sense, the recent focus on auditory modeling introduced a diversion from the original (somewhat idealized) idea of fully automated transcription.

Automatic transcription, therefore, is not only a hard problem, but it is also an ill-defined problem. It is indeed hard to believe that an automated transcription of a recorded version of Mahler's "Das Lied von der Erde" would reproduce the original score for the human voice and any of the instruments that play in the orchestra. But perhaps a piano reduction can be aimed at? Or maybe other, more general, content can be extracted that contains sufficient information to be useful in a practical application? We have, therefore, developed our approach in line with the pragmatic view in which automated transcription is an aspect of musical content extraction. Rather than automatic *transcription*, we prefer to call our approach automatic *description*. We subscribe the idea that:

1. A certain degree of automatic description may already be sufficient for certain purposes. For example, the specification of a sequence of chord labels rather than the specification of the exact pitches, may provide already sufficient information for retrieving a musical sequence from an audio database. In order for such a retrieval to be successful, it would then be sufficient to automatically describe the music into chord labels (which is assumed to be an easier task than transcribing the sound into exact pitches). The same argument holds for higher level harmonic analysis.
2. Global information, in general, is easier to extract than detailed information. In agreement with the Gestalt-based approach (Leman, 1997), this statement implies that different degrees of description are possible, useful, and even wishful in view of the way humans deal with musical content. On this basis, Hanappe (1994) explored the possibilities of chord recognition using the schema approach for tonality analysis developed in Leman (1994). Camurri

[1]See Zenner (1994) for an overview.

and Leman described a system in which recognition of chords and tone centers are combined within a logical reasoning system in order to track harmonic patterns such as cadences (Camurri and Leman, 1997). These studies demonstrated the feasibility and usefulness of recognizing objects at different levels of abstraction.

Automatic description is interpreted as part of a more general approach to musical content extraction. What is described are not necessarily pitches, but aspects of information that can be of use in practical applications. In what follows, we base automatic description on harmonic content extraction. It includes the description of chord fundamentals and chord labels from which pitch names (= pitch classes) can be deduced. A section of the paper is also devoted to the description of an objective evaluation procedure for analyzing the results of the transcription.

THE THEORETICAL FRAMEWORK

The model basically contains two parts:

1. A bottom-up part in which musical signals are processed into *auditory images* of different kinds. This part is based on the knowledge of the physiology of the auditory system.
2. A top-down part in which images are analyzed. This part is based on psychology and music theory.

A central aspect of the top-down part is based on the idea that given chords (represented in our model as *chord images*)[2] can be decomposed into (known) sub-chords (or *sub-chord images*). The idea is similar to the decomposition of a signal in terms of a (finite) set of basic signals, such as in a (Discrete) Fourier Transformation or Wavelet transformation. Decomposition can also be conceived of in terms of a $1 \rightarrow n$ $(n \geq 1)$ projection of a chord $c \in C$ onto sub-chords $s \in S \subset C$. All sub-chords span a space that contains a large subset of the set of all possible chords.[3]

The chord decomposition theory put forward by Balsach (1997) makes reference to the auditory phenomenon of virtual pitch, i.e., the fact that a collection of partials tends to be completed, provided the partials fit as harmonics of a lower pitch (or fundamental). The approach can be linked to our auditory-based bottom-up part which also does virtual pitch extraction. We adopted part of Balsach's theory as the basis of our decomposition model.

Figure 1 represents the harmonic decomposition model which relates the harmonic spectrum of the note C to the chord C-E-G-Bb. The chord notes are represented as prime harmonics of a note. Each prime harmonic generates its own harmonic multiples as shown in each staff. The harmonics generated by the prime harmonics belong to the harmonic spectrum of the note, which can also be conceived of as the fundamental (top staff) of the chord composed by the prime harmonics. The perception of a chord works on two levels. First, the perception of the prime harmonics (chord notes) can be recalled from a number of its harmonics, in compliance with the phenomenon known as "the perception of the missing fundamental". Second, the perception of the fundamental harmonic can be recalled from its prime harmonics.

[2]A distinction is made between a chord as a music theoretical concept and the representation of a chord in the auditory-based system. In the latter, we speak about chord images. Images are obtained by physiologically constrained feature extraction processes, which operate on sound waves. Hence, they reflect certain properties of the sound environment within a representational system. Technically speaking, images are represented as ordered arrays of numbers (= vectors). From a conceptual point of view, images are carried by neurons, whose activation is described in terms of the values of the vector components.

[3]The idea of chord decomposition has attracted the attention of many researchers through the centuries. Since the time of the ancient Greeks, relationships between prime harmonics and the ratios of integers associated with them have played an important role in philosophy and music theory. The "gradus suavitatis" (or pleasantness of intervals) proposed by L. Euler (1707–1783) for the calculation of the degree of consonance is based on a decomposition of natural numbers into a product of powers of different primes. Tanguiane (1993) proposes a decomposition method for chord identification based on correlational analysis of the spectra of the sounds that make up the chords. A drawback of the latter approach is that it does not take into account the constraints of the auditory system nor the spectral complexity of real sounds.

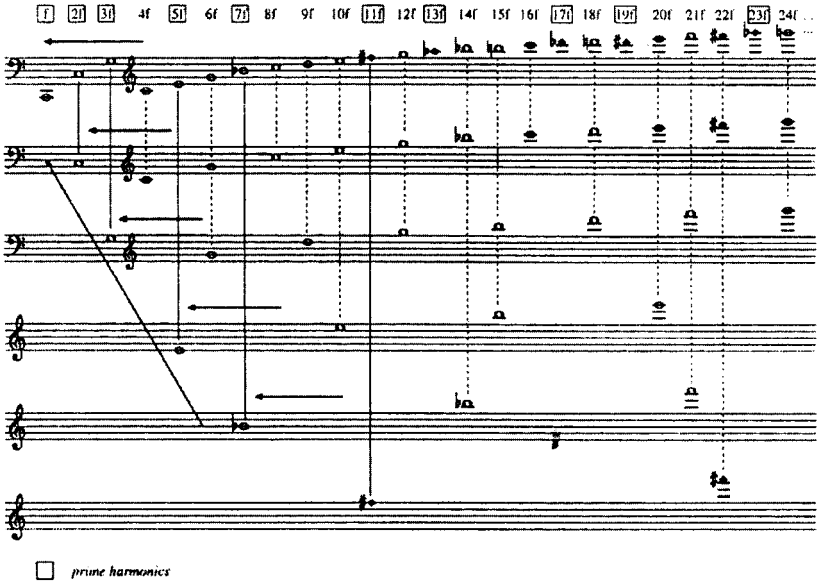

Fig. 1. Harmonics of a tone (Balsach 1997).

Taking into account combinations of the first five prime harmonics of a tone, Balsach constructs eleven sub-chords for one fundamental. The sub-chords support the perception of the virtual pitch with different degrees of strength. They are ranked from 1 to 11 according to the supporting power for the fundamental (Fig. 2).

The first sub-chord, containing the notes C-C′, has a low support for the fundamental C, because the notes can be considered as harmonics of the fundamental F (or/and Ab, or/and other fundamentals). The sub-chord C-E-G-Bb, on the other hand, gives the highest support for the fundamental C, because C is the only fundamental that has these tones as its harmonics. The sub-chords are thus ranked according to the ranking shown in Figure 2. A similar set of eleven sub-chords can now be constructed for each note of the chromatic scale in an octave. This yields a set of 132 (= 12 fundamentals * 11 sub-chords) sub-chords that span the space of $S \subset C$. The chords are labeled according to the fundamental in combination with the numbering in Figure 2. For example, C9 is the sub-chord composed of the notes C-E-G, F#4 is the sub-chord composed of the notes F#-C#-E.[4] A given chord, such as C-E-A, is then decomposed into the sub-chords A2, C8, and F5, giving support to the fundamentals C, A, and F, although with different strength. This is the part of Balsach's theory that we adopted as the basis of our decomposition model.

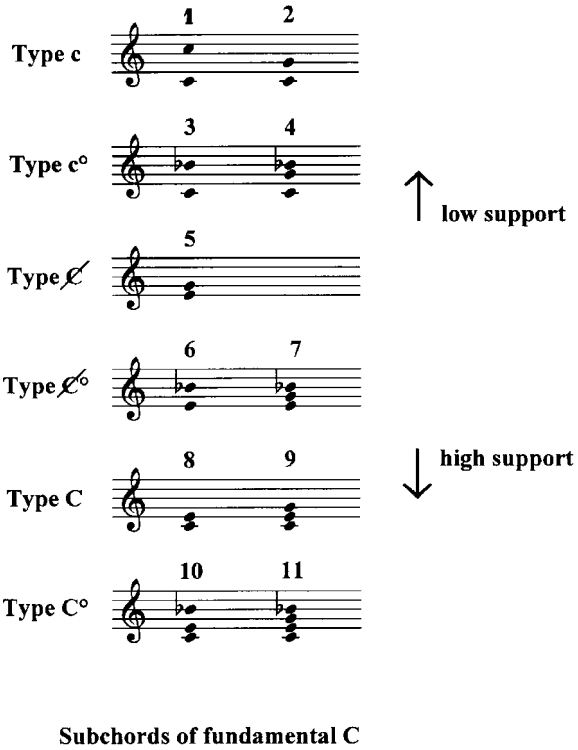

Fig. 2. Sub-chords in support of the fundamental C.

[4]Enharmonic notation is used.

In order to be able to use Balsach's music theoretical proposal in a computational analysis, we define the notion of *optimal decomposition* (to be used for validation first, and for later evaluation of the schema-based decomposition of musical signals). Optimal decomposition is based on the criteria of (i) maximum support for the fundamental pitch and (ii) minimum set (i.e., the minimum number of sub-chords). The algorithm can be summarized in three steps:

1. For each note of the chord, find all possible sub-chords (out of the set of 132 sub-chords) which contain that note;
2. If different sub-chords support the same fundamental, select the one with the maximum support;
3. Merge all the selected sub-chords and discard redundant ones so that a minimum set is obtained.

Similar configurations can belong to different keys. The following example illustrates the decomposition of E-G-A# into sub-chords:

E-G-A# is (optimally) decomposed into the sub-chords C7 F#6

-E-G-A# belongs to C7
-E-G belongs to C5
-G-A# belongs to D#5
-E-A# belongs to C6 and F#6
-E belongs to E1
-G belongs to G1
-A# belongs to A#1

Other examples of optimal chord decomposition are:

C-E-F#-G# is (optimally) decomposed into the sub-chords G#10 E8 C8

D#-F-G#-A is (optimally) decomposed into the sub-chords F10 C#5

The optimal decomposition algorithm, however, is not unique in that more than one set of sub-chords may decompose a given chord.

Optimal decomposition is part of a music theoretical approach that does not take into account real musical signals but only written notation. We apply the theory to the decomposition of *chord images* that are obtained from the processing of real sounds by an auditory model. This is worked out in terms of a schema-based architecture, as developed in Leman (1995), where an auditory model is connected with a two-dimensional neural network. The network can be viewed as a topological memory structure that develops a *schema (= structured knowledge)* after training. It is used here both as a container (= carrier or representation) for the images that represent the 132 sub-chords and as a projection operator for the decomposition during recognition.

The decomposition of a given chord image is thus achieved by a projection of the chord image onto a topological map. The mapping activates different regions on the schema from which we then infer the precise nature of the decomposition in terms of sub-chord labels. The latter provides the transcription of the musical signal. The evaluation of the schema-based chord decomposition model is done in reference to the optimal decomposition of the score.

In the next section, we give a description of our approach in terms of the system architecture. The following two sections are devoted to a detailed description of the several parts that make up the model. After that follows a description of the evaluation procedure. The final section provides some examples of the system performance.

THE SYSTEM ARCHITECTURE

The functional diagram of the system architecture is represented in Figure 3. The musical signal is sampled and then processed by an auditory model

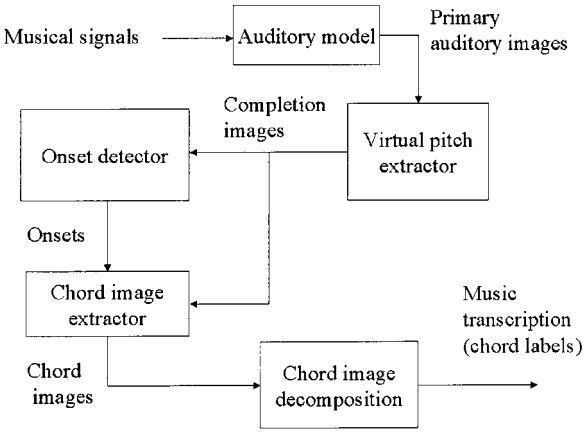

Fig. 3. Architecture of the model.

that produces the so-called *auditory nerve images*, i.e., representations of the neural activation as the neural firing rate code in a number of auditory nerves.

A virtual pitch extraction module transforms the auditory images into *completion images*. In order to detect and isolate the musical events, an onset detection module processes the completion images and outputs *chord images* via a chord image extractor module. A chord image decomposition module then has the task of analyzing, decomposing, and labeling (in terms of sub-chords) the chord images. The extracted harmonic content comprises chord fundamentals and chord labels (from which chord notes are deduced).

The Auditory Model

The auditory model of Van Immerseel and Martens (1992) is used as a front end. The model takes as input a digitized sound and produces an output that represents the auditory nerve patterns of twenty auditory channels at 1 Bark distance (= 1 critical band) from each other.

The musical signal is sampled as a mono 16-bit signal at a rate of 20.000 Hz and the amplitude is in the range $-1/+1$. The sampled signal *s(n)* is fed into an outer and middle ear filter and is then transformed into neural excitation (rate-code) patterns. These rate-code patterns represent the probability of neural firing during intervals of 0.4 ms. The sampling rate of the auditory nerve patterns is 2500 sa/sec. The transformation takes into account the asymmetric band-pass filtering of the cochlea and performs a transduction from a (filtered) sound signal to neural excitation. The transduction part simulates the half-wave rectification, short-time adaptation, and dynamic compression performed by the hair cells. The output at each time instance (= 0.4 ms interval) over all twenty channels is called an *auditory nerve image*.

The Virtual Pitch Extractor

The virtual pitch extractor takes the auditory nerve images as input and performs a periodicity analysis along each of the twenty channels. This part of the system acts as a bridge between the front end and the decomposition part of the model. The output images are called virtual pitch images or *completion images* (Leman, 1994).

The periodicity analysis is based on a short time auto-correlation of the auditory nerve images, i.e., the rate-codes, in each of the 20 channels. The auto-correlation analysis is described by the following equations:

$$R_i(n) = \sum_{k=0}^{F-1-n} e_i(k)e_i(k+n)\ W(k) \qquad \forall i = 1 \ldots 20$$

$$w(k) = e^{\frac{k}{D}}$$

$$R(n) = \sum_{i=1}^{20} R_i(n)$$

Where:

- R_i is the auto-correlation analysis of the rate-code signal e_i (output of the i-th channel).
- n is the time lag of the auto-correlation ($0 \leq n \leq F-1$).
- F is the time window of the auto-correlation (with F=80 samples or 32 ms).
- R is the summary auto-correlation pattern.
- An exponential function w is used to decrease the effect of the samples as the time in the window increases. Here, D = 40 samples which amounts to a decay of 16 ms.

The auto-correlation patterns R_i (for $i = 1, \ldots, 20$) are produced at steps of 4 ms. The summary auto-correlation pattern R is obtained by taking the sum of the auto-correlation patterns R_i over the 20 channels. At steps of 4 ms, this means that we have 250 summary auto-correlation patterns (R) per second.

Finally, the 80 components of the vectors R are reduced to 56 components (components from 0 to 4 and over 60 are deleted). This is done in order to focus on the region where the frequency representation[5] has an optimal resolution. Such a vector or summary auto-correlation pattern is now called a *completion image*. The term *completion* refers to the fact that the auto-correlation analysis completes a harmonic pattern in terms of its sub-harmonics.

[5]Alternatively, a log-lag representation could be used.

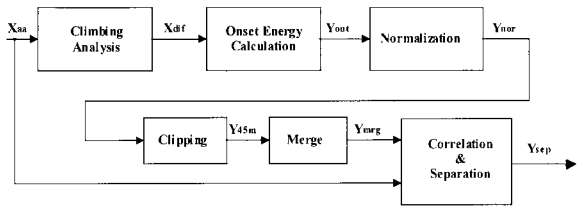

Fig. 4. The onset detection process.

The Onset Detector

The onset detector takes as input the completion images and produces a list of values containing onset information of detected events. This module is a slightly adapted version of the model described in Moelants and Rampazzo (1997). The list values indicate (i) the time of the onset, (ii) the time of the start of the stable regime of the event, and (iii) the onset energy.

Complete or partial masking effects, as well as intensities too low to be perceived, will influence the number of distinct events this system will perceive. Both pitch and intensity are considered here in order to recognize and discard performance elements such as vibrato, glissando, and dynamic openings. Figure 4 gives a general scheme of the steps involved in the onset detection process. In order to be able to extract stable chord images, the end of the attack part of the event is estimated (Climbing Analysis). The detection of the harmonic content is then based on the stable part (Onset Energy Calculation). Different criteria are used to discard and merge detected events (Clipping and Merge). The onset detector can be fine-tuned (Correlation and Separation). In general, it is preferred to have more onsets recognized than to have missed onsets.

The Chord Image Extractor

The chord image extractor takes as input two types of information: the completion images (= one image every 4 ms) and the list of onset events. A *chord image* is then defined as the weighted mean[6] completion image, calculated by taking into account all the completion images starting from the stable regime of one event to the onset of the next event, and normalized to an Eucledian unitary norm. There are as many chord images as detected onset events.

The Chord Image Decomposition

The decomposition module takes as input the chord images and maps these onto a schema (= a neural network trained with completion images that represent sub-chords). The schema responses correspond to labeled areas (see below); hence, the projection defines a *transcription of a chord in terms of its sub-chord labels*.

At this point, a distinction must be made between the actual decomposition on the one hand, and the development of the schema itself through a training procedure on the other. In the next two sections, we first describe how the schema is constructed and then give a description of the details of the decomposition process.

CONSTRUCTION OF THE SCHEMA

The schema is to be conceived of as a neural network-based memory representation for 132 prototype sub-chords. The schema is constructed by means of a training procedure described in the *Appendix* (Carreras and Leman, 1996). The neural network that carries the schema is a two-dimensional Self-Organizing Map (Kohonen, 1995) composed of a grid of 100 x 100 neurons on a torus surface. During the learning process, the network is trained with a set of images representing the sub-chords. After training, sub-chord images are used to attach labels to the most reacting neuron in the area activated (calculated as the centroid of the neurons activated). The trained neural network, called *schema*, can afterwards be used for giving a label to patterns that are in general different from the ones of the training set. In that way, a given chord image can thus be decomposed into labeled sub-chords. The generalization power of the SOM is exploited here.

The construction of the schema is an important aspect of the model. Three different candidate sets for the 132 prototype sub-chords have been considered for the training process, called the Shepard set, the one-octave set, and the four-octave set.

- The Shepard set: A first set of sub-chords was built using Shepard tones. A Shepard tone is

[6]The weight is computed with a negative exponential function such that the last image weights only 2% of the first one.

composed of sinusoidal frequencies at an octave distance from each other. The spectral envelope of the Shepard tones has a bell-shaped form, which favors the spectral energy between 500 Hz and 1000 Hz. Shepard tones thus reduce the perceptual range of pitches to one octave such that a chord and all its inversions produce the same perceptual effect. Shepard prototypes have been used in a number of modeling studies (e.g., Leman, 1995). In the present case, the projection of chord images (taken from real sounds of a piano keyboard) onto the schema (trained with the images of Shepard sub-chords) did produce satisfying results for the recognition of the chord fundamental. However, unsatisfying results were obtained for the chord notes. A test with a set of all possible tetrachord images (= 495 different images) produced results that are 10% below the ones obtained with the other prototype sets used.

- The one-octave set: A second set of 132 sub-chords was produced using an electronic keyboard with piano timbre. The chords were constructed such as to keep the note range within one octave. The tetrachord test gave good results, but, when the schema was tested with chords that contained tones outside the octave range, too many incorrect transcriptions were obtained.
- The four-octave set: To overcome the above-mentioned difficulty, we decided to expand the one octave range to four octaves. This was achieved by replicating the notes of the chord in one octave simultaneously over four octaves (from 2 to 5, with a3 = 440 Hz). The resulting schema was trained with images representing 132 sub-chords obtained from piano sounds that span chords over four octaves.

In the sequel, some more details are given of how the set of images that represent (Shepard, one-octave, and four-octave) *sub-chords* have been composed.

Sounds of individual notes were recorded during 200 ms (=4000 samples). After processing these sounds with the auditory model, we obtained completion images for each note. The completion images for a *sub-chord* were then built by taking the sum of the note completion images for all the notes that make up the sub-chord.[7] In the case of the Shepard tones, the sounds of the individual notes were built with a computer program. In the one-octave case, the notes span only one octave, while in the four-octave case, each note was played over four octaves. In all three cases, the amplitudes of all sounds were modeled with a piano sound envelope using the following modulating function:

$$m(k) = \begin{cases} \frac{1-e^{\frac{k}{400}}}{1-e} & \text{for k} < 400 \\ \frac{1-e^{\frac{4000-k}{3600}}}{1-e} & \text{otherwise} \end{cases}$$

The variable k ranges between 0 and 3999 (i.e., the number of samples) producing an amplitude $m(k)$ between –1 and 1. The amplitude modulated sounds were then processed by the auditory model and the virtual pitch extractor, resulting in 50 completion images for each note. The first 10 images were discarded on the assumption that the harmonic information during the attack phase was not yet complete. The last 40 completion images were retained. A unitary Euclidean norm was used to smooth the effects of possible differences in amplitude. The whole training set for the 132 sub-chords was therefore composed of 5280 completion images (= 40 completion images per sub-chord). The neural network was then trained with these 5280 completion images. After training, the network was first tesselated and then labeled.

Tesselation

After training, each of the 5280 completion images is presented to the network and the coordinates (bx, by) of the neuron with the minimum distance (i.e., the maximum similarity or maximum convergence) to the image are found according to the formula:

$$conv(i) = \min_{(A,B)\in Gb} || \, synaps(A,B) - image(i) ||$$

[7]Note images, as well as the images that represent sub-chords, are vectors of 56 components. Their sum is obtained by adding the corresponding components of all of the note images.

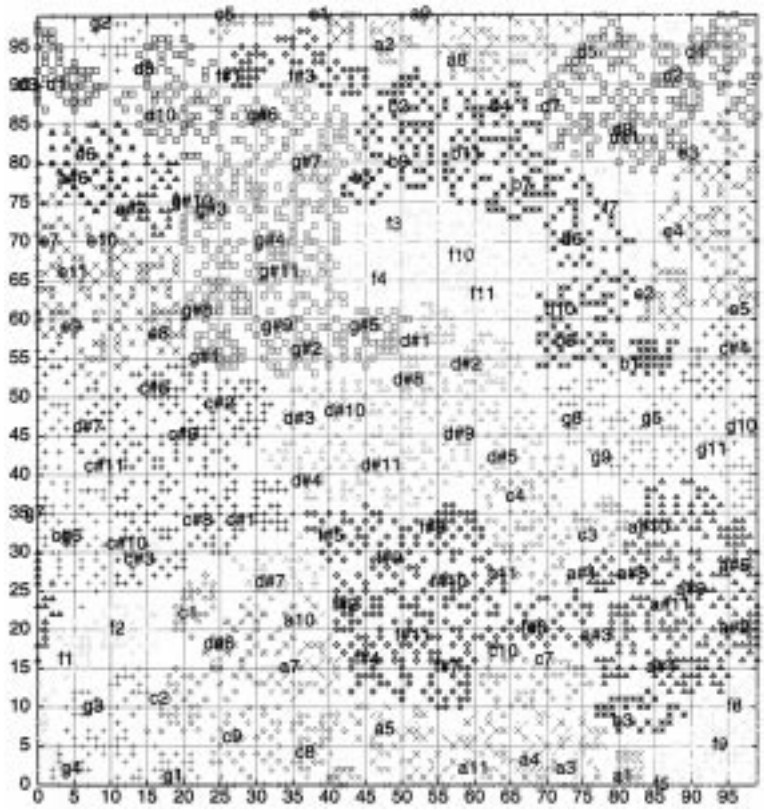

Fig. 5. Map trained with Shepard tones. The map represents a neural grid of 100×100 neurons on a torus structure such that the array of neurons at the top of the map connects to the array of neurons at the bottom of the map. Similarly, the array at the left connects to the array at the right. The different marks show the way in which the map is tesselated into sub-chords that support different fundamentals. (A color map is available in the Electronic Appendix of this journal).

$$best(i) = (b_x, b_y) = \arg \min_{(A,B) \in Gb} conv(i)$$

where the use of A and B means that all possible distances between two neurons on a torus surface (Gb) are considered. *Conv(i)* is the neuron and *best(i)* are its coordinates on the torus.

It is useful to visualize the neural response to the 5280 different completion images on a neural grid. On that grid, all of the completion images that represent sub-chords that belong to one fundamental (i.e., 40 images * 11 sub-chords per fundamental = 440 images) can be given the same color. Accordingly, a different color was assigned to each of the 12 distinct sub-chord sets of the total training set. A tessellation of the grid appears in which regions of the same color represent response areas for the same fundamental.

Labeling

The label for each of the 11 sub-chords for each fundamental is calculated by finding the centroid neuron for the 40 completion images of that sub-chord. This is calculated as the weighted mean of the 40 completion images on the map. The weight is defined as inversely proportional to the similarity value:

$$w_i = \begin{cases} \frac{10^{-6}}{conv(i)} & \text{if } conv(i) \geq 10^{-6} \\ 2 & \text{otherwise} \end{cases}$$

The labeled tessellation map for the case of the Shepard tones appears as in Figure 5.

The map shows a well-structured topology in which each fundamental is assigned one area (sometimes more) of contiguous responding neu-

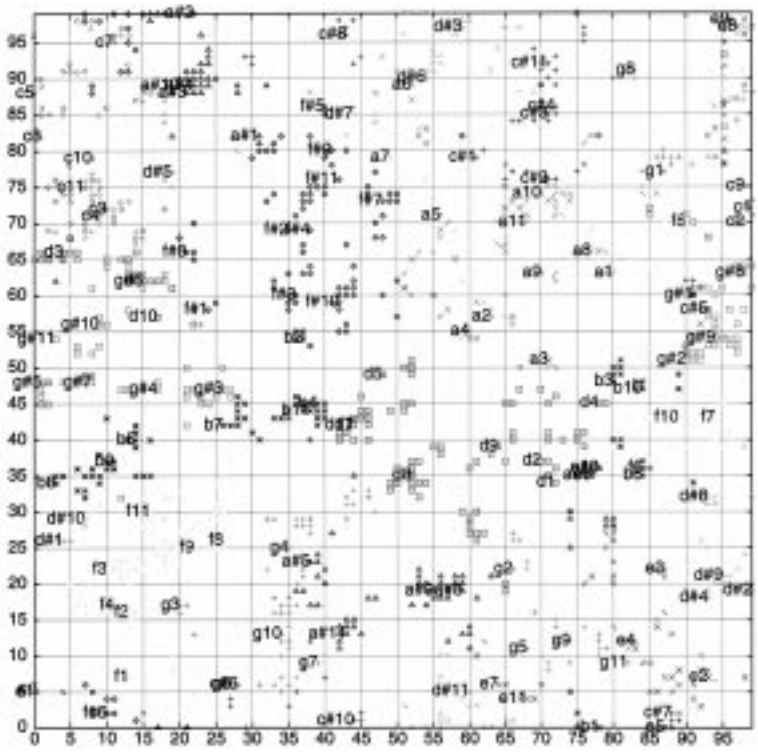

Fig. 6. Map trained with the one-octave set. (A color map is available in the Electronic Appendix)

rons. Chords of type 6 (see Fig. 2) (representing the interval of an augmented fourth) appear in pairs because the same note configuration, with different labels, belongs to two different fundamentals (separated at an interval of an augmented fourth). In Figure 5, the responding neurons are regularly distributed over the grid.

In the case of the one-octave set, neurons are often responding to more than one completion image. As a consequence, the map shows more compressed response areas and larger no-response areas (Fig. 6). Figure 7 shows the map with four-octave tones.

SCHEMA-BASED DECOMPOSITION OF CHORD IMAGES

Chord image mapping is defined as a projection of *chord images* onto a *schema* (= a trained neural network) holding images of 132 sub-chords. When dealing with real sampled music, a sequence of chord images is extracted with the modules described in the previous section (see Fig. 3).

Grey maps

A qualitative evaluation of a chord image decomposition can now be obtained with a visual representation tool designed for this purpose. A gray map is created where each point has a gray value *gray(A,B)* proportional to the response of the corresponding neuron in a scale from 0 (maximum darkness and highest response) to 255. The response of a neuron is calculated as the Euclidean distance between its associated 56-component vector *synaps(A,B)* and the vector of the *chord_image*:

$$ans(\mathrm{A},\mathrm{B}) = ||synaps(\mathrm{A},\mathrm{B}) - chord\_image||$$

After having found the highest (*max_a*) and lowest (*min_a*) value of *ans(A,B)*, the gray level of a point is normalized as:

$$gray(\mathrm{A},\mathrm{B}) = \mathrm{round}\left(225 * \frac{ans(\mathrm{A},\mathrm{B}) - min\_a}{max\_a - min\_a}\right)$$

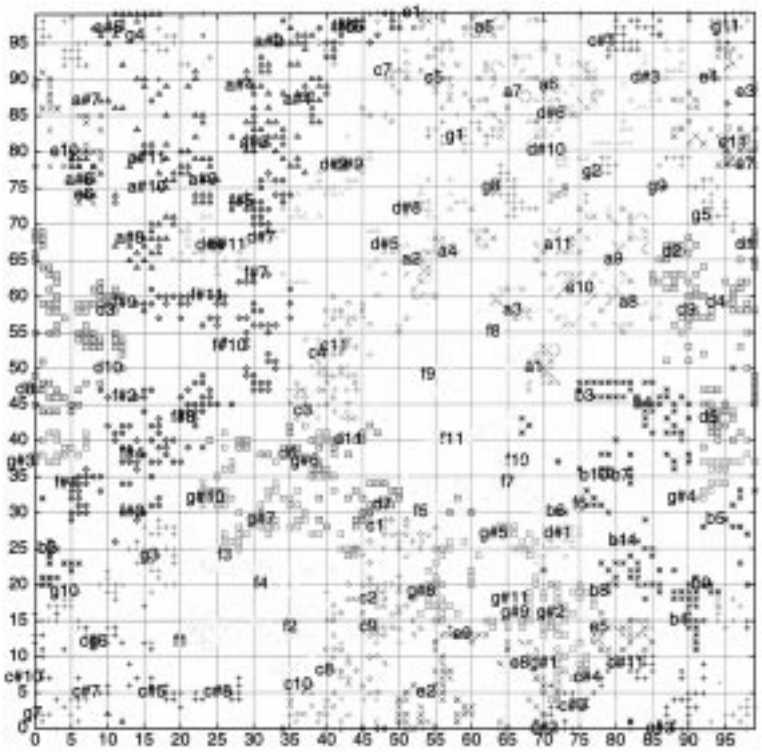

Fig. 7. Map trained with four-octave tones. (A color map is available in the Electronic Appendix)

In Figure 8, the representation of the decomposition of the chord C-D#-E-G is shown by the dark areas that correspond to the chord labels of C9 (C-E-G) and D#8 (D#-G). The transcription thus obtained corresponds to the optimal decomposition (see the section on Theoretical framework). In this example, the input chord was not part of the training set of the network. The example illustrates the generalization power of the schema-based approach.

Labeling

Schema responses to chord images have to be translated into labels. Given a chord image, the list of 40 best-matcher neurons is taken into consideration for each of the 132 sub-chords.[8] The gray value for each neuron is compared with a predefined threshold and, if the value is lower than the threshold, a counter *fit(cc)* is incremented as well as a cumulative value *cum(cc)*.

A fitness measure *measure(cc)* is then defined as:

$$measure(cc) = \begin{cases} \frac{fit(cc)^2}{cum(cc)} & \text{if } fit(cc) \geq minfit \\ 0 & \text{otherwise} \end{cases}$$

where the parameter *minfit* is 10 if the threshold is <40 and 2 otherwise. The resulting values give an indication of the network responses in terms of the sub-chords. These values are ordered per fundamental and normalized to 100. The following criteria are then used for selecting (i) the chord fundamental(s), (ii) sub-chords that define the decomposition, and (iii) chord notes.

1. The chord fundamental is calculated by ranking the sub-chords according to the fundamental they support (Fig. 2). Then, the measured values of the different sub-chords are summed up, weighted according to the hierarchical ranking (from 1 to 11) of the sub-chords, as defined in Figure 2. This gives a list of 12 chord fundamentals, with associated total weight values, out of which the highest is selected.[9]

[8]Recall that during training, each sub-chord was represented by 40 completion images.

[9]Although Balsach considers more than one fundamental, we select only one in this stage of the research.

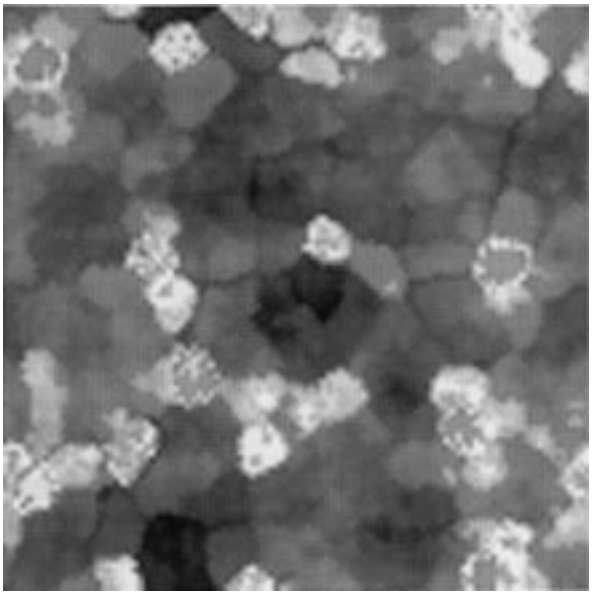

Fig. 8. Decomposition of a chord image in terms of sub-chords contained in a schema. Black regions indicate high activation. The network is a 100x100 neuron grid on a torus topology (cf., Fig. 5, Fig. 6, and Fig. 7).

2. The sub-chords are ranked according to the weighting values calculated by *measure(cc)*. The highest valued sub-chord (= 100%) is then selected, together with the second highest sub-chord, provided that the weight of the latter is greater than 25%.[10]
3. Out of these sub-chords, the note names are extracted by a simple merge.

EVALUATION PROCEDURE

The results of our model can be evaluated in many different ways. It can be done in view of perceptual results by asking listeners to compare a (MIDI)-played version of the transcription with the original version or by adopting a more pragmatic evaluation in terms of the practical application one has in mind. The method employed here goes back to the theoretical basis from which we started in that we evaluate the transcribed results in terms of the optimal decomposition theory. This entails an objective measure, which can afterwards be qualitatively discussed in terms of perceptual, pragmatic, and performance issues.

Taking for granted a successful detection of the onsets, the transcription yields three types of output for each chord image: (a) a fundamental, (b) a set of sub-chords, and (c) a set of note names:

- Since we take into account only one fundamental, it can be either correct or wrong.
- The set of sub-chords may contain chords that are correct or wrong, or not found. The idea that the found sub-chords should be of the same type as the sub-chords obtained by optimal decomposition cannot be used as a criterion, because the optimal decomposition is not unique. Therefore, a quantitative evaluation should be based on note names only.
- The set of note names may contain note labels that are either correct or wrong, or not found.

EXAMPLES

The first four measures of Träumerei (Kinderszenen op.15) in F by R. Schumann, in Figure 9, have been automatically described. Two versions have been used in this analysis:

- version 1: Daniel Barenboim piano, Resonance 4311672 GR
- version 2: Alfred Brendel piano, rec.1980 Philips 434732-8

The versions are compared in the following table by putting, for each event, the results of version 1 on top of the results of version 2.

Table 1 gives an overview of the results obtained.

The first column (1) reports the chord event number. Each event has the extension .1 and .2, which respectively refers to version 1 and version 2. The second column (2) is the onset time. The third column (3) contains the notes of the chord event as written in the score. The fourth column (4) shows the optimal decomposition in sub-chords; the list contains one or more possible sub-chord configurations for the fundamentals that are represented in the original chord. Note that this set is not necessarily unique (hence, its use in the

[10]This choice has been made on an experimental basis.

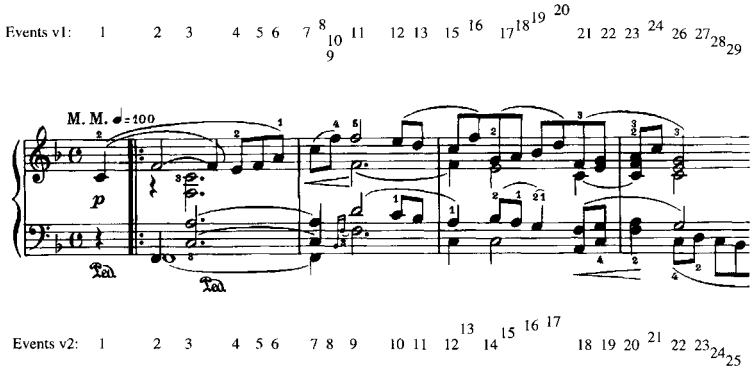

Fig. 9. First four measures of Träumerei by R. Schumann.

Table 1.

1	2	3	4	5	6	7	8	9	10
1.1	0.596	c	c1	C+	c	c1	1	0	0
1.2	0.912			C+		c1	1	0	0
2.1	1.776	f f	f1	F+	f	f1	1	0	0
2.2	2.096			F+	f	f1	1	0	0
3.1	2.896	f c a c f	f9	F+	c f a	f2 f9	3	0	0
3.2	3.128				c f a	f2 f9	3	0	0
4.1	4.956	f c a c e	f9 c8	E–	e	e1	1	3	0
4.2	4.744			E–	e	e1	1	3	0
5.1	5.600	f c a c f	f9	F+	f	f1	1	2	0
5.2	5.152			F+	f	f1	1	2	0
6.1	6.060	f c a c a	f9	F+	c f a	f8 f9	3	0	0
6.2	5.556			F+	f a	f8	2	0	0
7.1	6.528	f c a c	f9	F+	c f a	f9 f8	3	0	0
7.2	5.964			F+	c f f# a	f2 d7	3	0	1
8.1	7.180	f c a f	f9	F+	c f a	f9 f8	1	0	0
8.2	6.404			F+	c f a	f9 f8	1	0	0
9.1	7.316	a#	a#1	A#+	f g# a#	a#2 a#4	1	0	0
10.1	7.544	f	f1	A#+	f a#	a#2 f1	1	0	0
11.1	7.876	f d f f	a#5 f1	A#+	d f	a#5	2	0	0
9.2	7.140			A#+	d f	a#5	2	0	0
12.1	10.168	f c f e	c8 f2	C–	c	c1	1	2	0
10.2	9.436			C–	c	c1	1	2	0
13.1	10.732	fa# fd	a#9	A#+	d f a#	a#9 a#8	3	0	0
11.2	99.944			A#+	d f a#	a#9 a#8	3	0	0

(*Continued*)

Table 1. Continued.

15.1	11.300	c a f c	f9	F+	c f# a	f5 f7	2	1	1
12.2	10.496			C–	c f# a	f5 f7	2	1	1
16.1	11.724	c a f f	f9	F+	c f a	f9 f1	3	0	0
13.2	11.032			F+	c f# a	f5 d7	2	1	1
17.1	12.292	c a# e g	c11	C+	e g	c5	2	1	0
14.2	11.448			C+	e g a#	c5 d#5	3	1	0
18.1	12.788	c a e a	c8 f5	A–	a	a1	1	2	0
15.2	11.888			A–	a	a1	1	2	0
19.1	12.860	c g e a#	c11	A–	a	a1	0	4	1
16.2	12.300			D#–	g g# a#	d#5 a#3	2	2	1
20.1	13.916	c g e d	c9 d3 e3	G–	d f g	g2 g4	1	2	0
17.2	12.780			A#–	d f a#	a#8 a#9	1	2	2
21.1	14.552	a f c f	f9	F+	f	f1	1	2	0
18.2	13.220			F+	c f	f2	2	1	0
22.1	15.044	c g c e g	c9	C+	e g	c5	2	1	0
19.2	13.716			C+	e g	c5	2	1	0
23.1	15.576	f a c f a	f9	F+	f a	a1 f8	2	1	0
20.2	14.252			F+	c f a	f9 f8	3	0	0
24.1	16.072	f a c f c	f9	F+	c f a	f9 f8	3	0	0
21.2	14.812			F+	c f	f2	2	1	0
26.1	16.648	c g c e g	c9	C+	c e g	c5 c2	3	0	0
22.2	15.324			C+	c e g	c9 c8	3	0	0
27.1	17.104	d g c e g	c9 d3 e3	A–	d f g	g4 g2	2	2	1
23.2	15.800			A–	d f g g#	g4 a#7	2	2	2
28.1	17.672	c g c e g	c9	C+	c e g	c9 c2	3	0	0
24.2	16.268			C+	c e g	c9 c1	3	0	0
29.1	18.016	a# g c e g	c11	G–	f g a#	d#5 g3	2	2	1
25.2	16.736			A#–	f g a#	f#7 c10	3	1	1

evaluation should be handled with care). The fifth column (5) contains the found fundamental (the sign "+" indicates a correct estimation, "–" an incorrect estimation). The sixth column (6) gives the chord notes as found by the automatic transcription process (obtained by merging the sub-chords of column 7). The seventh column (7) shows the highest valued sub-chord and (a possible) second highest sub-chord with a weight greater than 25% as found through decomposition of the chord images. The eighth (8), ninth (9), and tenth (10) columns, respectively, evaluate the chord notes from the sixth column: the eighth column indicates how many correct notes have been found, the ninth column indicates how many notes were not detected, and the last column indicates the number of mistakes made in the transcription.

The results, in percentage, of the automatic description are:

Version 1:
1) — correct fundamentals 81.5%
2) — correct transcribed notes versus all notes in the score 66.6%
   — notes not recognized by the transcription process 32.0%
   — mistakes 6.4%

Version 2:
1) — correct fundamentals 76.0%
2) — correct transcribed notes versus all notes in the score 65%
   — notes not recognized by the transcription process 29.9%
   — mistakes 11.8%

In the following analysis, we describe the first four measures of the piano sonata K331 in A by W.A. Mozart (see also Fig. 10). In this analysis, two versions have been used:

- version 1: Alicia de Larrocha piano, Decca 4178177-2
- version 2: Paul Badura-Skoda forte piano (Johan Schantz 1790), rec.1989 Astrée E8683

The versions are compared in Table 2 by putting, for each event, the results of version 1 on top of the results of version 2.

The results, in percentage, of the automatic description are:

Version 1:
1) — correct fundamentals 84.3%
2) — correct transcribed notes versus all notes in the score 77.7%
   — notes not recognized by the transcription process 22.3%
   — mistakes 14.8%

Version 2:
1) — correct fundamentals 68.4%
2) — correct transcribed notes versus all notes in the score 79.6%
   — notes not recognized by the transcription process 20.4%
   — mistakes 7.4%

In the next analysis, we describe the first four measures of the Prelude 8 in e flat (WTK TI) by J.S. Bach (see also Fig. 11). We started from a CD-recorded performance played by Kenneth Gilbert, harpsichord (Jan Couchet, Antwerpen, 1671), Archiv produktion 413 439-2, rec. 1983.
Table 3 gives an overview of the results obtained.
The results, in percentage, of the automatic description are:
1) — correct fundamentals 82.1%
2) — correct transcribed notes versus all notes in the score 77.9%
   — notes not recognized by the transcription process 22.0%
   — mistakes 18.2%

DISCUSSION

To test a fragment of Träumerei by Schumann, we started from two performances on piano, version 1 by D. Barenboim and version 2 by A. Brendel. The second version is a slightly quicker performance (version 1 has a duration of 17.42 sec, version 2 has a duration of 15.82 sec). Due to performance characteristics, like the use of the pedal and the existence of very short onsets, the two versions turn out to have a different number of onsets (27 in version 1 and 25 in version 2). In the first version, it was sometimes difficult to check the correctness of the onsets. In the present analysis of version 1, onsets that were first included such as events 14 and 25 were later removed after careful checking. In version 2, there are no separate onsets for the decoration notes A# and F in the second measure (in version 1, these notes are represented in events 9.1 and 10.1). A first list with onset information of detected events shows that the system perceives an event at onset time 6.648 seconds. The intensity for this event was too low to be perceived through aural observation and has therefore been removed.

If we compare the different versions for the first measure, there is a remarkable similarity in so far that the detected fundamentals, notes, and sub-chords are identical with the exception of only one note C at event 6 that was not detected in version 2. By checking the sub-chords of that event, we see that F9 has much more weight in version 1 (in version 1, the weight for F9 is 72.9 as opposed to version 2 where the weight for F9 is 23.9) whereas the weights for the second sub-chords of the events 1 to 5 turn out to be approximately equal. The simi-

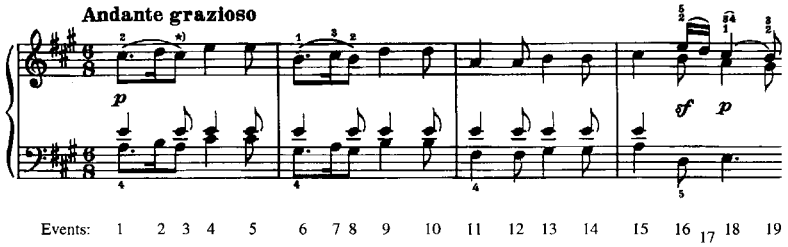

Fig. 10. First four measures of piano sonata KV331 by W.A. Mozart.

Table 2.

1	2	3	4	5	6	7	8	9	10
1.1	0.976	a e c#	a9	A+	c# a	a8 c#1	2	1	0
1.2	0.684			A+	c# e a	a5 a9	3	0	0
2.1	1.712	b d e	g5 e4	G–	d g b	g5 g9	2	1	1
2.2	1.136			G–	g# b	b1 e5	1	2	1
3.1	1.936	a e c#	a9	A+	c# e a	a8 a9	3	0	0
3.2	1.524			A+	c# e a	a9 a8	3	0	0
4.1	2.508	c# e e	a5 e1	A+	c# e	e1 a5	2	0	0
4.2	2.008			E–	e	e1	1	1	0
5.1	3.312	c# e e	e9	A+	c# e f# a#	a5 f#11	2	0	2
5.2	2.804			E–	e	e1	1	1	0
6.1	3.792	g# e b	a9	E+	g# b	e5 b1	2	1	0
6.2	3.300			E+	e g# b	e8 a9	3	0	0
7.1	4.508	a c# e	a9	A+	c# e a	a8 a9	3	0	0
7.2	3.880			A+	c# e a	a9 a8	3	0	0
8.1	4.744	g# e b	e9	E+	g# b	e5 b1	2	1	0
8.2	4.088			E+	e g# b	e8 e9	3	0	0
9.1	5.252	b e d	g5 e4	E+	e g# b	e8 e9	2	1	1
9.2	4.544			E+	e b	e2 e1	2	1	0
10.1	6.166	b e d	g5 g4	E+	d e g#	e3 e10	2	1	1
10.2	5.360			E+	d e	e2 e1	2	1	0
11.1	6.616	f# e a	a2d5	D–	f# a	d5	2	1	0
11.2	5.840			A–	e g a	a2 a4	2	1	1
12.1	7.496	f# e a	a2 d5	A–	a	a1	1	2	0
12.2	6.636			A–	e g a	a2 a4	2	1	1

(*Continued*.)

Table 2. Continued.

13.1	8.008	g# e b	e9	E+	g# b	b1 e5	2	1	0
13.2	7.120			E+	e g#	e8 g#1	2	1	0
14.1	8.992	g# e b	e9	E+	g# b	e5 b1	2	1	0
14.2	7.888			E+	e g# b	e8 e9	3	0	0
15.1	9.432	a e c#	a9	A+	c# e a	a8 a9	3	0	0
15.2	8.392			A+	c# e a	a5 a9	3	0	0
16.1	10.372	d b e	g5 e4	E+	d e a b	b3 e4	3	0	1
16.2	9.220			E+	d e b	g5 e4	3	0	0
17.1	10.472	d b d	d1 g5	E+	deg#b	e4 e11	2	0	2
17.2	9.436			A#–	d	d1	1	1	0
18.1	10.908	e a c#	a9	A+	c# e a	a9 a5	3	0	0
18.2	9.720			A+	c# e g a	a9 a11	3	0	1
19.1	12.056	e g# b	e9	E+	g#b	e5 b1	2	1	0
19.2	10.656			E+	g#b	e5	2	1	0

larity continues in the next measures where the detected sub-chords, if not identical, nevertheless are related. The detected notes correspond quite well and the mistakes made in the transcription are located on the same event level. The percentage of mistakes (added notes) is higher in version 2. In version 1, three from the five (mistakes) added notes (60%) are due to the second sub-chord. Two added notes come from the first sub-chord. In version 2, eight of the nine added notes come from the second sub-chord. One added note (88.8%) comes from the first sub-chord. This shows that the inclusion of the second sub-chord contributes to the mistakes.

Some chords can be completed by taking the fundamental into account. For example, in events 17.1 and 14.2, C is not recognized as a note but as a fundamental of the chord.

In several cases, the system added a note that is half a tone higher than the correct one. This added semitone comes from the second sub-chord. In events 7.2, 15.1, 12.2, and 13.2, F# is added to the chord F-A-C. F# forms part of the sub-chord D7.

The results, in percentage, of the automatic description concerning the misrepresented sub-chords are equivocal. In the analysis, we look upon similar sub-chords with different weight as correct estimations. This leads up to results that are more

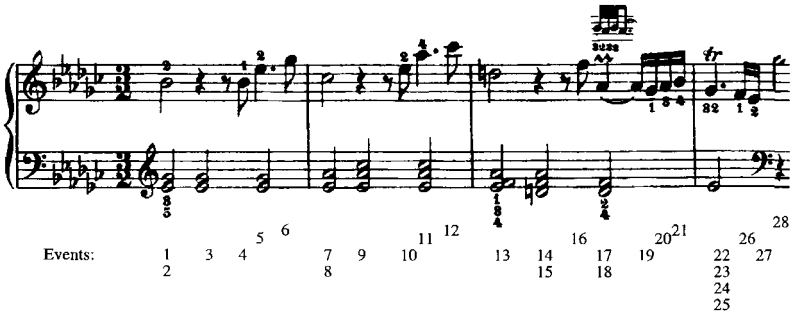

Fig. 11. First four measures of Prelude 8 by J.S. Bach.

Table 3.

1	2	3	4	5	6	7	8	9	10
1.	0.320	d# f# a#	b5 f#8	D#+	d# f#	d#1 b5	2	1	0
2.	0.464	d# f# a#	b5 f#8	D#+	d# g a#	d#9 d#2	2	1	1
3.	1.500	d# f#	b5 f#1	B+	d# f#	b5	2	0	0
4.	2.428	d# f# a#	b5 f#8	A#–	a#	a#1	1	2	0
5.	2.564	d# f# d#	b5 f#1	B+	d# f#	b5	2	0	0
6.	3.636	d# f# f#	b5 f#1	F–	d# f# a#	b5 f#8	2	0	1
7.	3.756	d# g# b#	b8 e5	D#–	d# f#	d#1 b5	1	2	1
8.	3.928	d# g# b#	b8 e5	B+	d# f# b	b8 b9	2	1	1
9.	4.864	d# g# b#	b8 e5	B+	d# g# b	b8 g#2	3	0	0
10.	5.844	d# g# b# d#	b8 e5	B+	d# f# b	b8 b9	2	1	1
11.	5.976	d# g# b# g#	b8 e5	G#+	d# g# b	g#2 e5	3	0	0
12.	7.000	d# g# b# b#	b8 e5	B+	g# b	b1 e5	2	1	0
13.	7.168	d# f# g# d#	a#7 f3	A#+	d f g# a#	a#9 a#11	3	1	1
14.	8.348	d# f# g#	a#7 e6	A#+	d f a#	a#5 a#11	2	0	1
15.	8.440	d# f# g#	a#7 e6	A#+	d f g#	c#5 a#7	3	0	0
16.	9.272	d# f# g# f#	a#7 e6	A#+	d f	a#5 f1	2	1	0
17.	9.476	d f (g# a#)*	a#7 e6	A#+	d f a#	a#5 a#9	3	0	0
18.	9.632	d f (g# a#)*	a#9	A#+	d f g# a#	a#9 a#11	4	0	0
19.	10.304	d f f#	d8 a#5	F#–	f# a#	f#8 f#1	1	2	1
20.	10.440	d f g#	a#7 e6	G#–	f f# g#	g#3 c#5	2	1	1
21.	10.556	d f a#	a#9	A#+	g# a#	a#1 a#3	1	2	1
22.	10.720	d# (f# g#)*	b5 g#4	G#+	d# g#	g#2 d#1	3	0	0
23.	10.940	d# (f# g#)*	b5 f#1	G#+	d# f# g#	g#2 g#4	3	0	0
24.	11.084	d# (f# g#)*	b5 f#1	G#+	d# f# g#	g#4 g#2	3	0	0
25.	11.328	d# (f# g#)*	b5 f#1	G#+	d# f# g#	b5 g#3	3	0	0
26.	11.452	d# f	d#1 f3	F+	d# f g	f3 g3	2	0	1
27.	11.516	d# d#	d#1	D#+	d# f f#	f3 b5	1	0	2
28.	11.888	f#	f#1	B+	d# f#	b5	1	0	1

*( )*Trill.*

or less alike for both versions, namely 22.2% misrepresented sub-chords in version 1 and 20.0% in version 2. These results cannot be taken too literally, because the optimal decomposition is not unique (see section Theoretical framework).

By comparing the sub-chords found through decomposition of the chord images and the sub-chord list of the optimal decomposition, many of the misrepresented sub-chords appear at events that reproduce no mistakes but where several notes of the chord were not detected by the automatic description process. In events 4.1 and 4.2, only the note E was recognized from the chord F-C-A-E. The optimal decomposition is F9-C8, whereas E1 was found through decomposition of the chord image. In events 18.1 and 15.2, only the note A was recognized from the chord C-A-E. The optimal decomposition is C8-F5, whereas A1 was found through decomposition of the chord image.

Sometimes the decomposition of the chord images generates sub-chords that contain notes that are half a tone higher or lower than the correct one. This may be due to a coarse resolution of the chord images.

To test a fragment of the Piano Sonata K331 by Mozart, we selected two performances on instruments with different timbre qualities, namely piano on version 1 and forte piano on version 2. For version 2, we have to bring into account that the schema was trained with piano sounds. The analysis of

both versions leads up to the same amount of events. Version 2 is played slightly quicker (version 1 has a duration of 11.08 sec, version 2 has a duration of 9.97 sec) and has higher onset energy, especially for events 8 (version 1 produces an onset energy of 35.42%, version 2 produces an onset energy of 100.00%) and 14 (version 1 produces an onset energy of 3.09%, version 2 produces an onset energy of 86.83%).

In the forte piano version, there are half as many mistakes (4 instead of 8) than in the piano version and three of the four mistakes appear at other events than the mistakes in version 1.

For events 11.2 and 12.2, the chord notes detected are E-G-A instead of F#-E-A. The note G is found instead of F# (rising semitone). G forms part of the second sub-chord A4 and fits in chord F#-A-C#-E-G.

For event 18, the chord notes detected are C#-E-G-A instead of E-A-G#. The note G is found instead of G# (falling semitone) and G forms part of the second sub-chord A11.

Four times there is a succession of identical chords with different duration: a quarter note in a strong metrical position followed by an eighth note in a weaker metrical position. The decomposition produces various but allied sub-chords. In version 2, there is a great similarity and there are more sub-chords in common between the succeeding chords than in version 1. Table 4 summarizes these results.

The chord A-E-C# in event 3 is a repetition of the chord in event 1 and should have the same results. In event 1.1, only the notes A and C# were found and the note E was not detected (sub-chords A8-C#1).

Table 4.

Chord notes	(score) Version 1		sub-chords Version 2	
C#-E-E	4.1	E1-A5	4.2	E1-F#6
	5.1	A5-F#11	5.2	E1-A5
B-E-D	9.1	E8-E9	9.2	E2-E1
	10.1	E3-E10	10.2	E2-E1
F#-E-A	11.1	D5-B7	11.2	A2-A4
	12.1	A1-D4	12.2	A2-A4
G#-E-B	13.1	B1-E5	13.2	E8-G#9
	14.1	E5-B1	14.2	E8-E9

Roughly all mistakes emanate from the second sub-chords. In version 1, seven of the eight notes that were added to the chords (87.5%) came from the second sub-chords. Two of the seven added notes are also included in the first sub-chord. Two added notes that come from the first sub-chord can be explained by making reference to the chord context. For example, in event 9.1 the notes E-G#-B were detected instead of B-E-D: G# forms part of sub-chords E8 and E9, but also fits in the chord E-G#-B-D. In event 17.1, the notes D-E-G#-B were detected instead of D-B. The note E forms part of the sub-chords E4 and E11, G# forms part of the second sub-chord E11, and E and G# here again fit in the chord E-G#-B-D. If we take the fact in consideration that some added notes fit in a chord, the global percentage of mistakes for version 1 can be reduced to 7.4% and for version 2 to 5.5%. In version 2, all added notes (100%) come from the second sub-chord.

Some chords can be completed by taking the fundamental into account. In version 1, for events 6, 8, 13, 14, and 19, E is not recognized as a note but as the fundamental of the chord.

The comparison of the sub-chords found through decomposition of the chord images and through optimal decomposition for this analysis results in 0.5% misrepresented sub-chords in version 1 and 0.0% misrepresented sub-chords in version 2. In spite of the fact that these results, which are very much alike in both versions, seem to be promising, we have to bring into account the no-uniqueness of the optimal decomposition (see above).

To test the Prelude 8 in e flat by J.S. Bach, we used a CD-recorded harpsichord performance. The harpsichord was tuned with pitch A=392 Hz, so that the results had to be transposed one tone higher. Here, we have to reckon with the fact that the schema was trained with piano sounds and that notes played on a harpsichord generate high harmonics. As the performer does not always play all the notes from a chord simultaneously and also plays appogiatura, we sometimes get two onsets for one chord:
the chord D#-F-A# is represented in events 1 and 2 (onset interval of 0.144 sec)

the chord D#-G#-B is represented in events 7 and 8 (onset interval of 0.172 sec)
the chord D-F-G# is represented in events 14 and 15 (onset interval of 0.092 sec)

The trill in the score also produces several events. The events 17 and 18 are the result of the trill which starts on the upper note (A#-G#). Therefore, we do not consider it as a mistake that the chord D-F-G# from event 17 leads up to the detection of D-F-G#-A#. Events 22 to 25 are the outcome of a trill on the upper note (F#-G#).

By comparing the sub-chords obtained by decomposition of the chord-images with those obtained by optimal decomposition, we get 28.6% misrepresented sub-chords.

Eleven of fourteen (78.6%) added notes originate from the second sub-chord. Three of those eleven added notes are also included in the first sub-chord. Three of fourteen added notes only originate from the first sub-chord. Some added notes fit within the tonal composition of the chord. In event 6, the note A# fits in the chord D#-F#-A#

If we take these considerations into account, the global percentage of mistakes can be reduced to 11.68%. The previous results, as well as results from the analysis of a few other pieces, are summarized in Appendix B.

SUMMARY

1. The pieces analyzed pose several problems for automatic description:
- performance characteristics show particular problems. The use of the pedal in Schumann's Träumerei and the performance with many arpeggio's in Bach's Prelude 8 cause problems for the onset detector. Pedals fuse the tones and may make the correct recognition much more difficult.
- the forte piano and harpsichord are instruments with different timbre qualities, although their timbre approaches that of the piano. The analyses of Mozart's Sonata K331 and Bach's Prelude 8 show that the automatic description model is able to deal with different timbres.
- the model also manages to deal with the use of different styles. The selection of compositions by Bach, Mozart, and Schumann covers an important part of the classical repertoire.

2. Our evaluation procedure follows a rather strict schema. Some of the mistakes can be explained by:
- making reference to the chord context: some added notes that fit within the tonal composition of the chord are not that problematic from the point of view of perception.
- in a number of cases, chords can be completed by taking into account the fundamental. Therefore, it is not a bad idea to include the fundamental as an added chord name.
- the onset detection may have problems with the recognition of very short onsets. The pitch of very short tones is difficult to detect given the frame of 32ms needed for pitch analysis (see the section System architecture). Alternative methods, however, are currently being explored.
- occasional mismatch in the sub-chords of the decomposed original chord is found for notes of the sub-chords that result half a tone higher or lower than the correct one. This is a problem that can partly be attributed to the way in which the images are constructed.
- the system often detects the harmonic of a chord note rather than the note itself. Since this can be more problematic, this phenomenon needs further study.
- a high percentage of mistakes emanates from the second sub-chord. In practical applications, a trade-off can be stipulated between less notes, but more correct ones, or more notes with more mistakes.

CONCLUSION

The model of schema-based chord decomposition has been based on an auditory analysis of sound signals. The model presents a new method for the description of harmonic content based on chord decomposition in terms of sub-chords. Automatic description of musical examples, taken from CD recordings, shows that the method at present performs results up to a degree of accuracy that can be useful in certain applications, ranging from music analysis (e.g., functional harmonics analysis, tonality analysis), data mining, or mappings from

sound to score. The description has a focus on chord fundamentals, sub-chords, and chord notes.

Automatic description is not only hard and ill-defined, but also difficult to evaluate. In this paper, we therefore worked out an evaluation method in terms of a comparison between schema-based chord decomposition (starting from sounds) and a so-called optimal chord decomposition (starting from the score). This quantitative analysis has been completed with a qualitative analysis which takes into account particular features of the tone system as well as performance characteristics. The qualitative analysis in general may explain some of the mistakes in that, for example, notes may belong to the tonality of a chord, although they are not written in the score.

The results obtained with the schema-based chord decomposition method are promising in that the system is able to deal with a number of different performance characteristics, musical styles, and even different timbres. Thus far, however, the performance of the model has been tested on a limited amount of musical examples. Future work will concentrate on an even broader range of performances, styles, and timbres.

ACKNOWLEDGEMENTS

The authors wish to thank D. Petrolino for help in developing the model.

REFERENCES

Askenfelt, A. (1976). Automatic notation of played music. In *Report STL-QPSR 1/1976* (pp. 1—11). Stockholm: Royal Institute of Technology.

Balliello, S., De Poli, G., and Nobili, R. (1998). The colour of music: spectral characterisation of music sounds filtered by a cochlear model. *Journal of New Music Research*, 27(3), 325—358.

Balsach, L. (1997). Application of virtual pitch theory in music analysis. *Journal of New Music Research*, 26(3), 244—265.

Camurri, A., and Leman, M. (1997). AI-based music signal applications — a hybrid approach. In C. Roads, S. Travis Pope, A Piccialli, and G. De Poli (Eds.), *Musical Signal Processing* (pp. 349—381). Lisse: Swets & Zeitlinger.

Carreras, F., and Leman, M. (1996). Distributed parallel architectures for the simulation of cognitive models in a realistic environment. In E. D'Hollander, G. Joubert, F. Peters, and D. Trystram (Eds.), *Parallel Computing: State-of-the-Art Perspective* (pp. 585—588). Amsterdam: Elsevier.

Casajus-Quiros, F.J., and Fernandez-Cid, P. (1994). Real-time, loose-harmonic matching fundamental frequency estimation for musical signals. In *Proceedings of the ICASSP94* (pp.211—214).

Chafe, C., Mont-Reynaud, B., and Rush, L. (1982). Toward an intelligent editor of digital audio: recognition of musical constructs. *Computer Music Journal*, 6(1), 30—41.

Cosi, P., De Poli, G., and Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, 23(1), 71—98.

Ellis, D.P.W. (1996). *Prediction-driven computational auditory scene analysis*. Unpublished Ph.D thesis, MIT, Cambridge, MA.

Foster, S., Schloss, W.A., and Rockmore, A.J. (1982). Toward an intelligent editor of digital audio: signal processing methods. *Computer Music Journal*, 6(1), 42—51.

Hanappe, P. (1994). *Het herkennen van akkoorden in een muzikaal signaal*. Unpublished Masters' Thesis, University of Ghent, Ghent.

Katayose, H., and Inokuchi, S. (1989). The Kansei music system. *Computer Music Journal*, 13(4), 72—77.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin/Heidelberg: SpringerVerlag.

Leman, M. (1994). Schema-based tone center recognition of musical signals. *Journal of New Music Research,* 23(2), 169—204.

Leman, M. (1995). *Music and Schema Theory: Cognitive Foundations of Systematic Musicology*. Berlin/Heidelberg: Springer-Verlag.

Leman, M. (Ed.) (1997). *Music, Gestalt, and Computing – Studies in Cognitive and Systematic Musicology*. Berlin/Heidelberg: Springer-Verlag.

Martin, K.D. (1996). *A Blackboard System for Automatic Transcription of Simple Polyphonic Music*. (Technical Report N. 385), Cambridge, MA: MIT Media Lab, Perceptual Computing Section.

Meddis, R., and Hewitt, M.J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification. *The Journal of the Acoustical Society of America*, 89(6), 2866—2894.

Moelants, D., and Rampazzo, C. (1997). A computer system for the automatic detection of perceptual onsets in a musical signal. In A.Camurri (Ed.), *KANSEI — The Technology of Emotion*, (pp. 140—146). Genova: AIMI/DIST.

Moorer, J.A. (1975). On the transcription of musical sound by computer. *Computer Music Journal*, 1(4), 32—38.

Schloss, W.A. (1985). *On the Automatic Transcription of Percussive Music. From Acoustical Signal to High Level Analysis*. (Report STAN-M-27). Stanford: Stanford University, Dept. of Music.

Solbach, L., Wöhrmann, R., and Kliewer, J. (1998). The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis. In D.F. Rosenthal and H.G. Okuno (Eds.), *Computational Auditory Scene Analysis* (ch. 18). Mahwah, N.J.: L. Erlbaum Assoc.

Sundberg, J., and Tjernlund, P. (1970). A computer program

for a notation of performed music. In *Report STL-QPSR, 2-3/1970* (pp. 46—49). Stockholm: Royal Institute of Technology.

Tanguiane, A. (1993). *Artificial Perception and Music Recognition*. Berlin/Heidelberg: Springer-Verlag.

Van Immerseel, L., and Martens, J.P. (1992). Pitch and voiced/unvoiced determination with an auditory model. *Journal of the Acoustical Society of America,* 91(6), 3511—3526.

Zenner, H.P. (1994). *Hören — Physiologie, Biochemie, Zell- und Neurobiologie*. Stuttgart: Georg Thieme Verlag.

APPENDIX A: The Learning Paradigm of the SOM

The learning process is performed by means of a cyclic set of steps. Each input sample is presented to the network and the distance to each neuron of the net is calculated using a certain metric. The neuron with the minimum distance is selected and the values of all the neurons with a distance less than a radius value are updated (adaptation process). The value of the radius, that has a large initial predefined value, is decreased to 1 according to a function that depends on the number of input samples yet to be processed. The training set is presented a fixed number of times during the learning process. At each time, the set is randomized. The initial values of the neurons of the grid are chosen by taking the mean value and the variance of the samples of the training set. In this study, the input samples and the neurons are 56-dimensional vectors (=images). The initial value of the radius is 50, the number of epochs is 10, and the learning rate is 0.1. The distance metric is the Euclidean distance.

The function that governs the value of the radius is:

$$vfr(radius = 1) = \text{Epochs} \cdot \frac{\text{Tv}}{3}$$

$$vfr(radius) = \text{round}(\text{Epochs} \cdot \frac{2 \cdot \text{Tv}}{3} \cdot \frac{\log(1 + \frac{1}{\text{radius}-1})}{\log(\text{m\_r})})$$

$$radius = 2 \ldots m\_r$$

(here, Tv is the total number of the vectors of the training set):

The best-matcher neuron, with coordinates (bx, by), is found with the following relationship:

$$(b_x, b_y) = \arg \min_{(A,B) \in Gb} ||\text{synaps}(A, B) - \text{actual\_input}||$$

Where **actual_input** and **synaps**(A,B) represent vectors and **G**b is the grid considered as a torus surface.

The neurons within the radius distance from the best-matcher one are defined as:

$$\text{Iv}(b_x, b_y) = \{(X, Y) \in \text{G}_\text{b} : \text{dist}(\ (X, Y), (b_x, b_y)\ ) \leq actual\ \_radius\}$$

And the adaptation is performed according to this law, where $\lambda = 0.1$ is the learning rate:

$$\text{synaps}(X, Y) = (1 - \lambda) \cdot \text{old\_synaps}(X, Y) + \lambda \cdot \text{actual\_input} \quad (\forall (X, Y) \in \text{Iv}(b_x, b_y))$$

APPENDIX B: Synoptic table with test results

	1	2	3	4	5	6	7	8	9
Bach, Prelude 8 in e flat (WTK T1)*	1-4	12	24	28	78	22	18	14	82
Mozart, Piano Sonata KV282 in E flat	1-8	38	103	103	73	27	16	9	86
Mozart, Piano Sonata KV331 in A (version 1)*	1-4	11	19	19	78	22	15	26	84
Mozart, Piano Sonata KV331 in A (version 2)*	1-4	11	19	19	80	20	7	37	68
Beethoven, Piano Sonata Op2, Nr1 in f (vers. 1)	1-8	9	53	47	62	38	17	21	77
Beethoven, Piano Sonata Op2, Nr1 in f (vers. 2)	1-8	9	53	40	48	52	17	13	92
Chopin, Prelude Op28, Nr4 in e	1-10	52	76	72	54	46	11	7	68
Schubert, Moments Musicaux Op94, Nr1 in C	1-8	13	41	38	74	26	24	25	84
Schumann, Op15, Träumerei in F (version1)*	1-4	17	28	29	67	33	6	48	81
Schumann, Op15, Träumerei in F (version2)*	1-4	17	28	25	65	30	12	40	76
Schumann, Op9, Valse Noble in B flat	1-8	9	38	32	66	38	21	6	75
Brahms, Op118, Nr2, Intermezzo in A	1-8	23	51	49	75	25	7	25	90

*Detailed analysis of these pieces can be found in the text.

1. measures in the score
2. duration in seconds
3. amount of expected onsets following the score
4. amount of onsets obtained by the onset detection process
5. percentage of correct notes found by the automatic description process
6. percentage of notes that were not found by the automatic description process
7. percentage of mistakes made in the description (added chord notes were not regarded as mistakes)
8. percentage of events with correct description of the chord notes
9. percentage of correct estimations of the fundamental

Francesco Carreras
Multimedia Systems Department
CNUCE, via S.Maria 36
56126 Pisa, Italy
Tel.: +39-050-593224

Dr. Francesco Carreras received a degree in Mathematics from the University of Florence, Italy and specialised in Computer Science. Since 1972 he is researcher at the Institute CNUCE, Pisa, of the Italian National Research Council. Until 1980 he was responsible for the operating systems of the computing Centre of CNR and was actively involved in the European organisations SEAS and GUIDE for scientific computing. From 1980 he collaborated with the computer music group in Pisa, headed by P.Grossi. He worked in the Parallel Computing group of CNUCE up to 1990. He was involved in several computational musicology projects and, since the early 1990's, he co-operates with Marc Leman of IPEM on several projects.

Marc Leman
IPEM-RUG, Department of Musicology
Blandijnberg 2
9000 Ghent, Belgium
E-mail: Marc.Leman@rug.ac.be
http://www.ipem.rug.ac.be/staff/marc

Dr. Marc Leman is research leader at the Fund for Scientific Research and Professor at the University of Ghent. He is director of the Institute for Psychoacoustics and Electronic Music, head of the Research Society for the Foundations of Music Research (sponsered by FWO), and editor-in-chief of the Journal of New Music Research (published by Swets & Zeitlinger, the Netherlands). His research interest is focused on the epistemological and methodological foundations of cognitive and systematic musicology.

Micheline Lesaffre
IPEM-RUG, Department Musicology
Blandijnberg 2
9000 Gent, Belgium
Tel.: +32-9-2644121
Fax: +32-9-2644143
E-mail: micheline.lesaffre@rug.ac.be
http://www.ipem.rug.ac.be/staff/micheline

After studying musicology at the University of Ghent Micheline Lesaffre became consecutive scientific collaborator at the Museum of Contemporary Art of Ghent and curator at a private gallery in the same city. Besides guest lectures she has also tought organized courses on contemporary art, electronic music and acoustics. Since 1999 she works as a scientific collaborator in musicology at IPEM, the research center of the Department Musicology at the University of Ghent and is a doctoral candidate. Her main concentration over the last period at IPEM has been on musical content description. Presently she is working on the analysis of modules developed at IPEM for automatic content extraction of musical features.