FIRST STEPS IN RELAXED REAL-TIME TYPO-MORPHOLOGICAL AUDIO ANALYSIS/SYNTHESIS

Norbert Schnell
IRCAM, CNRS - STMS
Norbert.Schnell@ircam.fr

Marco Antonio Suárez Cifuentes
IRCAM
Marco.Suarez@ircam.fr

Jean-Philippe Lambert
IRCAM
Jean-Philippe.Lambert@ircam.fr

ABSTRACT

This paper describes a real-time audio analysis/resynthesis system that we developed for a music piece for ensemble and electronics. The system combines real-time audio analysis and concatenative synthesis based on the segmentation of sound streams into constituting segments and the description of segments by an efficient set of descriptors adapted to the given musical context. The system has been implemented in Max/MSP using the *FTM & Co* and *MuBu* libraries and successfully employed in the production and performance of the piece. As more and more research in the domain of music information retrieval, we use the term of *typo-morhpology* to designate the description of sounds by morphologic criteria including the temporal evolution of sound features that also can provide pertinent means for the classification of sounds. Although, the article mainly insists on the technical aspects of the work, it occasionally contextualizes the different technical choices regarding particular musical aspects.

1. INTRODUCTION

The developments described in this article have been conducted in the framework of the production of the piece « Caméléon Kaléidoscope » for an ensemble of 15 instruments and electronics by Marco Antonio Suárez Cifuentes. The piece, commissioned by the *Ensemble Orchestral Contemporain*, IRCAM, and the GRAME, has been premiered at the *Biennale Musiques en Scène* in Lyon in march 2010.

The basic idea of this work was to create a real-time system that re-orchestrates and responds to solo instruments and instrument groups using sound materials that are either pre-recorded or recorded on the fly from the same ensemble. The musical writing of the piece features densely articulated musical structures and makes intensively use of contemporary playing techniques.

The real-time analysis sub-system that we developed in this context segments audio streams into elementary events and generates a description for each segment that represents its perceived duration, its energetic evolution, and its pitch content. The same technique and descriptors are used in the real-time analysis of the playing ensemble as for the description and retrieval of pre-recorded and pre-analysed materials [1] . This choice has been made to easily allow for an extension of the system that would extend in real-time the sound data base used by the resynthesis.

We have chosen the description to efficiently represent the most essential features of the given musical material. Although it cannot be considered as a symbolic representation it permits the composer to manipulate and transform the analyzed musical phrases as well as to generate new musical phrases from the recorded material using sequence patterns and generative algorithms.

In the setup of the piece, the analysis/synthesis system described in this article is embedded into an environment of further real-time audio processing applied to the 15 instruments of the ensemble such as spatialization, transposition, and distortion. The system analyzes a single input stream that can be switched to different instruments and instrument groups. The synthesis sub-system of the system concatenates automatically generated musical phrases from the pre-recorded data base in response to the analysis of the input stream. Up to 16 synthesis voices are used in parallel.

1.1 Typo-Morphology

The principals of this analysis picks up on current trends in music information retrieval research seeking to describe sounds by the temporal development of their features [1, 2, 3, 4, 5]. This research often refers to Pierre Schaeffer's *typo-morphology* [6] that provides a framework for the description of sounds independent of their origin. In the context of automatic description of sounds of all origins, and especially in the context of music, Schaeffer's work is an excellent source of inspiration in the elaboration of abstract sound representations that capture the essential characteristics of sounds from a particular "point of view". Many authors have proposed extensions of Schaeffer's description system [7] and adapted it to a particular context of application [8].

1.2 Relaxed Real-Time Analysis/Resynthesis

The system that we developed has all properties of a real-time audio analysis/resynthesis system. It analyzes an audio input stream and can generate an output stream as a real-time transformation of the input stream. Although,

[1] The pre-recorded materials used in the piece are solo and group recordings of instrument parts extracted from an early version of the score performed by the players of the ensemble that also performed the premiere of the piece.

the segment based approach of our system makes that the information about the incoming sound stream is available only at the end of each segment, which introduces a considerable input/output latency for the system's responds.

The approach to real-time audio processing we promote with this work inscribes itself into a current trend of real-time audio processing and interactive systems in music that relaxes the latency constraints to the benefit of introducing novel sound representations into this context. The challenge of this work is to provide musically pertinent descriptions of sounds that can be calculated in real-time and to provide the means to manipulate these representations as a part of a musical work.

In this sense, we'd like to refer to this approach as *relaxed real-time* audio analysis/synthesis [2] .

This work can also be seen as an extension of existing concatenative synthesis [9] and audio musaicing [10, 11, 12] systems introducing an efficient set of descriptors representing the evolution of sound features within a sound segment.

The implementation of the system is entirely based on Max/MSP using the *FTM & Co* [13] libraries *Gabor* [14] and *MnM* [15] as well as a set of modules recently developed in the framework of the *MuBu* project [16].

2. SEGMENTATION AND DESCRIPTION

Sound descriptions create a particular "point of view" on sound materials that has to be optimized in order to fit a particular application. Similar to other domains of application, the work in the context of music composition reveals particular aspects of sound descriptions in terms of their correspondence to actually perceived sound features and, beyond that, their correspondence to a particular vocabulary of musical writing. While, for the retrieval of sounds using similarity – directly or after a transformation/manipoulation of the sound description – this correspondence can stay implicit, the more explicitly the description reflects musically relevant characteristics corresponding to the vocabulary of musical writing, the more it can be integrated into a compositional process in articulation with a written musical score.

An important challenge of this work was to optimize the representation of the audio streams as a sequence of described segments according to multiple aspects:

- Pertinence regarding the musical material (i.e. instrument sounds, style of musical writing and performance)
- Potential in terms of its manipulation as part of the compositional process
- Efficient real-time calculation of the analysis and availability of optimized analysis operators in Max/MSP

[2] Technically, the overall Max/MSP system stays a synchronous real-time audio processing system with an audio input/output latency below 50 milliseconds that also applies further (low latency) real-time transformations to the incoming sounds. Nevertheless, this approach to real-time processing also generates new requirements for the processing environment concerning the possibility to integrate different streams of processing on different levels of scheduling and possibly having different representations of time.

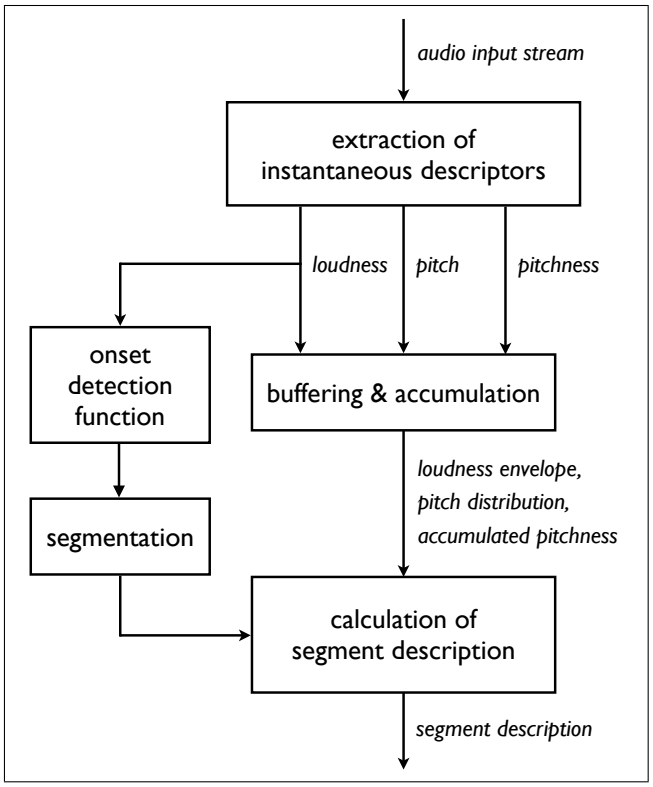

Figure 1. Schematic overview of the analysis sub-system

Figure 1 shows a schematic overview of the analysis sub-system and its different stages. As in similar systems, the analysis is organized in multiple stages:

- Extraction of instantaneous (low-level) descriptors
- Segmentation
- Calculation of the segment description

The following subsection will describe and discuss the techniques that have been employed.

2.1 Instantaneous Audio Descriptor Extraction

As in many other music information retrieval systems, the first stage of our analysis sub-system extracts instantaneous audio descriptors from the incoming audio stream. The following low-level descriptors are calculated on frames of 2048 samples with a hop size of 256 samples on the 44.1 kHz input audio stream generating regular streams of audio descriptors with a period of 5.8 milliseconds:

- Loudness extracted from the power spectrum
- Monophonic pitch
- Pitchness

The loudness is is extracted from the power spectrum using dBA weighting and the pitch and pitchness are calculated using the *Yin* algorithm [17].

The following stages of the sub-system perform segmentation and integrate the loudness and pitch information into a compact description of the segments.

2.2 Onset Detection and Segmentation

The automatic segmentation implemented in the system in based on an onset detection function calculated as the dif-

ference of loudness between the current frame and the median of the loudness over multiple previous frames (typically 3 to 5). A positive threshold is applied to this function to determine the instant of segmentation. The parameters of the algorithm are the length of the median filter and the segmentation threshold.

A variant of this algorithm had been developed on singing voice in a former unpublished research project [18]. The original algorithm had been inspired and successfully validated against other published methods [19, 20] for singing and spoken voice. Although we tested variants of the segmentation algorithm using a spectral description (i.e. MEL band and MFCC coefficients), the performance on the given sound materials was not significantly improved compared to the loudness based approach so that it would have justified the extraction of further spectral descriptions in the final system [3] .

On the given sound materials, the segmentation performs satisfactory although it evidently fails in two cases that clearly would be represented as distinct musical events in a score: extremely soft onsets (i.e. sound segments very successively appearing from silence) and smooth transitions between distinguishable events. While the latter is not an issue for the musical approach for which the system has been designed, the former remains an unsolved challenge to be further investigated on.

It was was surprisingly easy to find a parameterization of the onset detection algorithm for all used sound materials that well distinguishes local singularities of the input signal due to sound texture (i.e. roughness) from the onsets of musical events.

Theoretically, the onset detection function could also be used to determine the end of a segment by applying a negative threshold. Although, this technique has the tendency to cut off resonances and reverberations beyond the main body of particular sound events (e.g. partially attenuated chords or plates). In addition, it does not give an appropriate estimation of the actual perceived duration of the sound event. Since an overlap add technique is used for the synthesis allowing for the synthesis of almost arbitrarily overlapping segments it was not necessary to determine the end of a segment before the beginning of a the next giving the possibility to preserve resonances and reverberations wherever possible.

Consequently, a segment is defined by two successive onsets. Apart from its total duration, a set of eight descriptors representing its loudness envelope and pitch content is associated to each segment.

2.3 Loudness Envelope Description

The most essential descriptors of a segment in the given context concern the perceived energy. During the analysis, the characteristics of evolution of loudness between two onsets is recorded into a vector in order to extract the following descriptors:

- Maximum loudness
- Effective duration
- Envelope skewness

The maximum loudness represents well the perceived energy of the segment. The module used to calculate the other two descriptors actually calculates the first three standardized moments of the loudness envelope. The effective duration is derived from the spread (actually from the standard deviation) of the envelope by multiplying it with a constant factor and clipping it to the segment duration as defined by the inter-onset time. The multiplication factor that has been found by empirical experimentation (see 4) so that the descriptor represents very well the perceived duration of a segment in comparison to segments with equal or similar descriptor values and that it can be used to concatenate a sequence of sound events eliminating or equalizing the gaps between the end of one event and the beginning of the next.

The envelope skewness turns out to be an efficient and robust descriptor to distinguish whether and to which amount the perceived energy of sound event represented by a given segment raises or falls. The examples in section 4 illustrate well this descriptor.

For convenience in the processing of the descriptors, loudness is represented in the implementation by positive numbers corresponding to the actually measured loudness in dBA with an offset of 72 and clipped between 0 and 72 reducing the dynamics range used in all calculations to 72 dB.

2.4 Pitch Content Description

A second set of descriptors that describe a segment concerns the pitch. Several options have been considered for the extraction of the pitch content of a segment and the evolution of pitch within a segment. Given the sound material, the extraction of a pitch contour has been excluded. The majority of segments correspond to contemporary playing techniques without a clear pitch or multiple pitches (i.e. multiphonics or piano chords). Relatively few segments that actually have a monophonic pitch contain glissandi that would be worth to be described by a contour. The description we found describing best the pitch content of a segment is a distribution table accumulating the output of a pitch tracker over a segment.

As mentioned above, in the current version of the system we use a monophonic fundamental frequency estimation module based on the *Yin* algorithm, that also outputs a harmonicity coefficient and the sound energy (calculated as the first coefficient of the auto-correlation). For strictly monophonic harmonic sounds the module outputs precise estimation of the pitch and a quality estimation is close to 1. For slightly inharmonic or noisy but pitched sounds (e.g. due to damped resonances) the harmonicity coefficient decreases and for unpitched sounds it is close to 0. If a sound contains multiple pitches, the module tends to jump from one pitch to another still representing rather well the perceived pitches and strong resonances present in the segment. The product of the harmonicity coefficient and the energy of a frame corresponds to the relative salience of

[3] Since in our research project following up on the work described in this article we seek to include the description of timbre that anyway requires the extraction of further spectral representations, we are currently reconsidering this question.

the measured pitch. These salience values are accumulated in the pitch distribution table indexed by the estimated pitch quantized to a quarter-note scale. At an detected onset the table is evaluated for the last segment and cleared to accumulated the pitch distribution of the next segment. The pitch with the highest weight in this distribution and its quarter-tone pitch class as well as the centroid and the standard deviation calculated from the table are representing the pitch content among the set of descriptors of a segment.

In addition, the system calculates the mean value of the harmonicity coefficients for all frames with a loudness above a certain threshold (typically -72 dB) over the segment.

In summary, the following five descriptors have been chosen to represent the pitch content of a segment:

- Most salient pitch in quarter tones represented in floating-point MIDI pitches
- Pitch class of the most salient pitch
- Centroid of the pitch distribution table
- Standard deviation of the pitch distribution table
- Pitchness

These five descriptors can be seen as a compromise between descriptors that actually mean something for the user of the system (i.e. the composer) and descriptors that represent implicitly the most salient features of a segment.

3. SOUND RETRIEVAL AND RESYNTHESIS

The synthesis sub-system relies essentially on a recently developed set of modules that are designed together with an optimized data container for Max/MSP called *MuBu* [16]. In the Max/MSP implementation of the system, the data base of pre-recorded sound materials is stored in the *MuBu* data container module. The data is aligned to the audio data and associated to the onset time of the respective segment.

The interaction with the data base of pre-recorded sounds relies on a k-nearest neighbor method. The module used in the system implements a KD-tree allowing for an efficient retrieval of the segments of the data base of pre-recorded sounds that comes closed to a given description. For the retrieval, each descriptor can be given a weight to define its importance or even to exclude it from the query. The description for querying sounds can be directly given by the analysis of an audio input, generated arbitrarily or by transformation of the output of the analysis.

A basic analysis/resynthesis system is created when using the descriptors generated by the analysis of an input stream in real-time directly for the query of a the sound with the closest description in the data base and playing immediately the retrieved sound segment. The result of this setup is an output sound concatenated from sounds present in the data base that reflects at the same time the material in the data base as well as the "point of view" on the sound material – and sound in general – that is incarnated by the set of descriptors [4] . In the design phase of the system, we have intensively made use of this simple setup to permanently evaluate the performance of the system regaring the pertinence of the chosen description.

The most basic transformation of descriptors that we are experimenting with is the scaling of descriptors values. Scaling allows, for example, for adapting the range of the descriptors produced by the analysis of one instrument or playing style to another. We have calculated for each instrument in the data base the mean and standard deviation as well as the minimum and maximum values of each descriptor over all segments. Dependent on the descriptor and the desired effect one would use either the extreme values or the statistical distribution as a reference for scaling.

4. EVALUATION AND DEMONSTRATION

The system has been permanently evaluated in listening experiments during its design. In addition to these experiments that compared the behavior of the system for different sound materials, we have developed a small environment around the analysis sub-system to visualize the analysis data in real-time and offline in Max/MSP. This application records the loudness envelope and the pitch distribution tables as well as the onset times of the segments in addition to the descriptor values. The recorded data and the waveform of the corresponding sound segment can be visualized from multiple points of view using a dedicated graphical module of FTM & Co and played back with different options [5] .

The fact that programming, real-time and offline analysis/resynthesis, and visualization are possible in the same environment significantly facilitated the design and implementation of the system. Although this is common for many applications using Max/MSP (or similar environments) as rapid prototyping and execution environment, the availability of a large number of efficient analysis and statistical modeling operators and visualization tools in the FTM & Co libraries adds to this an additional dimension for the design of analysis/resynthesis systems.

The screen-shots presented in the figures 2 to 6 show segments analyzed by the system that are representative for the sound materials that we have worked with and that permits to briefly discuss the representation of the sound segments by the chosen set of descriptors. We selected different instruments and playing modes.

Each figure shows the loudness curve and pitch distribution table superposed to the waveform of the corresponding segment as well as the nine descriptor values calculated for the segment. While the standard deviation of the pitch distribution table has been calculated on the original values in a scale proportional to the power of the signal, in the figures below the values are visualized in logarithmic scale. The effective duration is additionally marked with a small vertical line crossing the zero-level of the waveform display.

Even if the segments may appear isolated they are in fact all real-time segmentations of the original recordings.

[4] Since the vector of descriptors is output by the analysis stage at the end of each segment the timing (i.e. rhythm) of the synthesized sound does not correspond to the input sound unless the onset times are corrected using an estimated maximal segment duration.

[5] For the purpose of this publication the visualization has been simplified and superposed in a single window.

The musical writing of the scores used for the recordings is rather dense, but privileges very clearly articulated relatively short events over longer sounds with a continuous evolution.

Figure 2 shows a pitched marimba note. The pitch distribution is reduced to a single peak at the pitch of the note (69) that also corresponds to the low standard deviation value (0.26). The envelope skewness (0.54) indicates a decaying loudness. The relatively low pitchness value corresponds to the inharmonicity of the marimba sound.

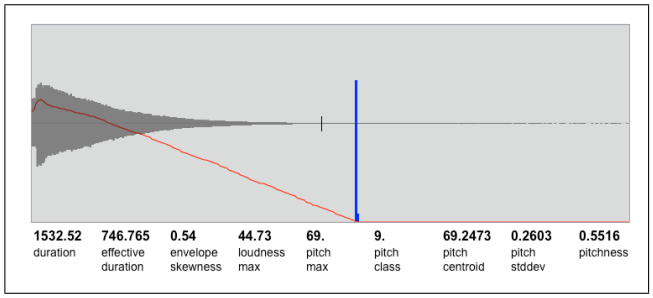

Figure 2. Waveform, and description data of a segment of a marimba note.

The pitch distribution of the segment of a flute note played with flatterzunge shown in figure 3 represents well the two salient pitches that are audible when listening to the sound. The descriptors calculated from the distribution still represent the pitch of the note (75.5) derived from the maximum of the distribution, but reduce the two pitches to an augmentation of the standard deviation (2.94) and a centroid below the maximum pitch. The low value of the loudness envelope skewness (0.11) indicates a rather flat envelope.

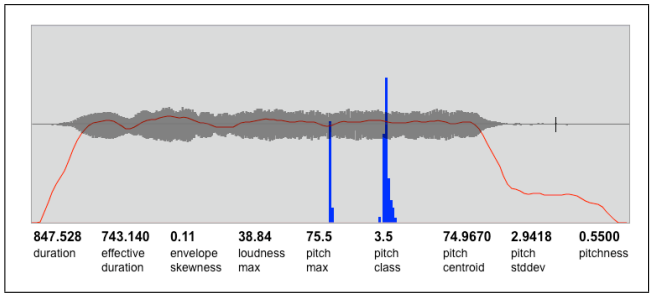

Figure 3. Waveform, and description data of a segment of a flute note with flatterzunge".

The segment represented in figure 4 corresponds to a clarinet multiphonic emerging from silence with a strong crescendo. The note onset of this example is strong enough to be detected by the system. Apart from the strongest peak that in fact corresponds to the strongest perceived pitch, the pitch distribution in this example only very vaguely corresponds to the pitch content of the multiphonic. Nevertheless, some overall characters of the pitch content are represented by the high standard deviation (5.8) combined with a high pitchness value (0.82). The crescendo is well represented by the very low negative skewness value (-1.17). The segment is cutoff by the onset of a staccato note following the crescendo.

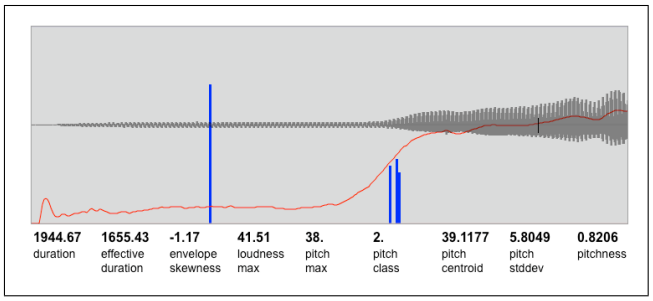

Figure 4. Waveform, and description data of a segment of a clarinet multiphonic.

The pitch distribution of the trombone glissando in figure 5 represents well the pitch range of the glissando. The additional peaks on the left and the right are octave errors of the pitch estimation that do not significantly change the standard deviation of the pitch distribution calculated on the linear scale values of the table (2.04) corresponding to 2 semitones. The slight crescendo over the segment is expressed in the negative skewness value (-0.22).

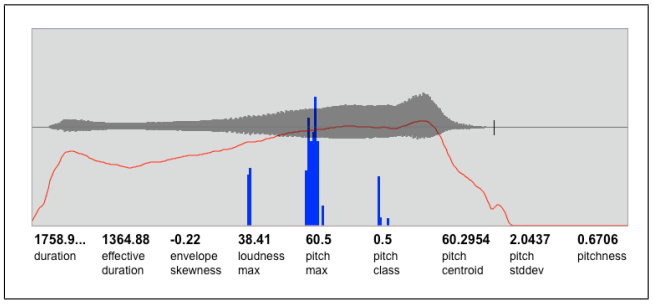

Figure 5. Waveform, and description data of a segment of a trombone glissando.

The last example in figure 6 visualizes the segmentation and analysis of a short bass note played “ecrasé”. Even though the effective duration is correctly estimated its display starting from the beginning the segment is not appropriate in this case. The low pitchness value (0.26) witnesses of the noisy character of the sound segment that corresponds to a note in the middle of a staccato sequence. Nevertheless, the most salient perceived pitch of the segment is represented by the indicted maximum of the distribution (35.5).

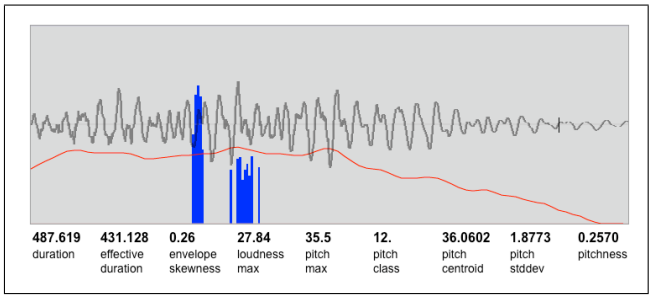

Figure 6. Waveform, and description data of a segment of double bass note played “ecrasé”.

5. CONCLUSIONS AND CURRENT RESEARCH

With the work described in this article, we have achieved a first step in the development of a real-time analysis/resynthesis using a segment based description.

Staying rather simple, the description developed for the context of a contemporary music piece has been proven its efficiency in the context of the given musical production and opened for us an interesting field of further investigation.

Although the system does not perform an explicit classification, it permits to access to the data base by typomophological criteria.

We are currently working on several improvements and extensions of the system. The most important improvements mainly concern the detection of soft note onsets and smooth note transitions as well as the development of a compact description of timbral aspects.

Further experiments and developments concern the transformation of the segment description and the resulting possibilities in composition. For example, the descriptor values can be normalized and scaled or inverted in order to create corresponding variations in the retrieval of sound segments and resynthesis.

While we currently do not apply any analysis and modeling of the sequence of segments, the given representation has an interesting potential to be used in conjunction with techniques such as *Bayesian networks* and *Factor Oracle*. Sequence modeling may require a more explicit classification of segments that can be easily derived by clustering in the descriptor space.

Even if a segment based approach was an obvious choice regarding the very strongly articulated character of the piece we are currently considering alternative techniques that do not require a segmentation at the analysis stage and still allow for the manipulation and retrieval of musically relevant sound segments represented by a temporally integrated description corresponding to the temporal evolution of sound features.

6. REFERENCES

[1] J. Ricard and H. Perfecto, "Using Morphological Description for Generic Sound Retrieval," in *International Conference on Music Information Retrieval (ISMIR)*, 2003.

[2] J. Ricard and H. Perfecto, "Morphological Sound Description Computational Model and Usability Evaluation," in *Proceedings of the AES Convention*, 2004.

[3] G. Peeters and E. Deruty, "Automatic Morphological Description of Sounds," in *Proceedings of Acoustics 08*, 2008.

[4] G. Peeters and E. Deruty, "Sound Indexing Using Morphological Description," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[5] J. Bloit, "Interaction musicale et geste sonore : modélisation temporelle de descripteurs audio," PhD Thesis, 2010.

[6] P. Schaeffer. Paris, France: Seuil, 1966.

[7] M. Chion. Paris, France: INA/GRM, 1966.

[8] L. Thoresen and A. Hedman, "Spectromorphological Analysis of Sound Objects: An Adaptation of Pierre Schaeffer's Typomorphology," *Organised Sound*, vol. 12, pp. 129–141, 2007.

[9] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, "Real-Time Corpus-Based Concatenative Synthesis with CataRT," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2006.

[10] T. Jehan, "Creating Music by Listening," PhD Thesis, 2005.

[11] M. Casey, *Soundspotting: A New Kind of Process?* Oxford University Press, 2009.

[12] P. A. Tremblay and D. Schwarz, "Surfing the Waves: Live Audio Mosaicing of an Electric Bass Performance as a Corpus Browsing Interface," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, June 2010.

[13] N. Schnell et al., "FTM — Complex Data Structures for Max," in *Proceedings of the International Computer Music Conference (ICMC)*, Septembre 2005.

[14] N. Schnell et al., "Gabor, Multi-Representation Real-Time Analysis/Synthesis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Septembre 2005.

[15] F. Bevilacqua, R. Muller, and N. Schnell, "MnM: A Max/MSP Mapping Toolbox," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, Mai 2005.

[16] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi, "MuBu & Friends - Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP," in *Proceedings of the International Computer Music Conference (ICMC)*, August 2009.

[17] A. de Cheveigné and H. Kawahara, "YIN, A Fundamental Frequency Estimator for Speech and Music," *JASA*, vol. 111, pp. 1917–1930, 2002.

[18] J.-P. Lambert and N. Schnell, "Internal Report of the ANR Project *VoxStruments*," tech. rep., IRCAM, Mai 2009.

[19] P. Brossier, J. P. Bello, and M. D. Plumbley, "Real-time Temporal Segmentation of Note Objects in Music Signals," in *IEEE Transactions on Speech and Audio Processing*, 2004.

[20] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, D. M., and M. B. Sandler, "A Tutorial on Onset Detection in Music Signals," in *IEEE Transactions on Speech and Audio Processing*, 2005.