Contemporary Music Review, 2013
Vol. 32, No. 1, 49–59, http://dx.doi.org/10.1080/07494467.2013.774513

Routledge
Taylor & Francis Group

Composing Morphology: Concatenative Synthesis as an Intuitive Medium for Prescribing Sound in Time

Benjamin Hackbarth, Norbert Schnell, Philippe Esling and Diemo Schwarz

This article presents AudioGuide, an innovative application for sound synthesis which aims to heighten compositional control of morphology in electronic music. We begin with a discussion of the challenges of managing detail when composing with computers, emphasizing the need for more tools which help the composer address the intricacies of sonic evolution. AudioGuide's approach—using a soundfile as a method for specifying morphological shape—provides a simple yet exacting medium for representing temporal ideas. Using the spectral structure of a soundfile, AudioGuide organizes a user-defined collection of pre-recorded sounds to create a similar morphological contour. Our matching strategy accounts for the spectral content, temporal evolution, and superimposition of sonic elements. We provide two examples which illuminate the capabilities of the algorithm from within a musical context.

Keywords: Concatenative Synthesis; Audio Features; Timbre; Gestural Transcription

Prelude

When writing music with computers, an ever-present challenge of using software for sound synthesis is navigating the vast array of technical possibilities and choices in order to realize concrete ideas. In many cases, composers are drawn to computer music by a desire to exercise a heightened degree of control over sound and to break free from a variety of restrictions imposed by instrument-based compositional paradigms. This freedom is strikingly evident in the plethora of inventive and exacting methods for crafting sound, often permitting the composer to create a seemingly infinite variety of sonic qualities and timbres. However, a majority of software does not aid the composer in organizing combinations of sounds intelligently or in structuring

sonic relationships across time. Rather, decisions regarding how sound is composited (either as vertically aligned mixtures or horizontally linked sequences) usually occurs on a separate programmatic plane. Regardless of whether temporal relationships are created by hand, through generative algorithms or with real-time signals, there is often a lack of interrelation, interaction, and integration between the spectral quality of sonic elements and their organization in time.

The challenges posed by the lack of strategies for mediating sound and temporality have been compounded by a gradual increase in the complexity of computer music algorithms and programs. In the early days of electronic music, there were certain hard and fast limits on user-supplied decisions, stemming from the relative simplicity of algorithmic inputs and limitations in layering sound imposed by magnetic tape. During the recent decades, tools have become decidedly digital and, as computational power, sophistication, and complexity have increased, the resolution with which composers can parameterize and superimpose sound has risen in tandem. While there are many benefits to an increased precision of control, there are also risks. When realizing a musical passage, a task not unlike assembling a sonic jigsaw puzzle, one can become overly immersed in shaping individual pieces without sufficient awareness of how they fit together to create a cumulative image. Consequently, one can lose sight of compositional intuition and intent in a landscape densely populated by choices and details. This type of unmediated engagement with complexity creates two primary aesthetic hazards:

(1) Neglecting the integrity of sonic relationships due to the complexity of crafting individual elements.
(2) The onset of 'creative decision fatigue', resulting in a slowing of compositional metabolism that threatens to undermine inspiration, exploration, and flexibility.

These hazards can lead to a scenario where technological management stipulates compositional engagement and not vice versa. Under these circumstances, rendering creative ideas with enhanced precision, an alluring promise of computer control, is difficult to realize due to a lack of tools to help composers manage the complexity of sonic superimposition intuitively and prevent technical minutia from impeding creative impulse. We propose that problems arise not necessarily due to the sheer volume of decisions and possibilities, but due to a lack of strategic methods to hierarchize decision-making and to interconnect programmatically different strata of sonic experience.

The Orchestrator's Keyboard

When searching for creative strategies to manage the complex interdependency of sound and time, a useful analogue is composing music for a large group of acoustic instruments. Like composing with computers, the act of orchestration can be a cumbersome and detail-rich affair. Many of the aforementioned aesthetic hazards—

distraction by the density of choice and the disruption of creative continuity—present similar challenges to the orchestral composer.

Setting aside aesthetic motivations, examining orchestration is particularly useful since composers have, over time, developed numerous strategies to manage the complexity of orchestral resources. Of particular interest is what could be termed the 'orchestrator's keyboard'—an intermediary structured interface (such as a piano) used to assist in the composition of music for a larger set of forces. Under this scenario, the composer is engaging with a meta-instrument in the sense that each key of the keyboard is a sonic substitute for a collection of orchestral sounds which share the same pitch. The power of this approach lies in the fact that the composer is able to shift easily between focusing solely on the creation of relational constructs (melodies, chords, and harmonic progressions) and contemplating which instruments will articulate particular components. By composing orchestral music with the aid of a keyboard, temporal relationships can be decoupled from the specific sonic qualities that will ultimately articulate the musical surface.

An argument in favour of using a keyboard as a tool in this regard is that it can serve to buffer the composer, in the moment of creation, from engaging with the full spectrum of necessary decisions. This buffer enables certain types of compositional approaches and explorations that would arguably not survive a scenario in which the totality of choice is confronted in a single moment. Thus, one is able more easily to think globally rather than locally, cumulatively rather than individually. Through the keyboard the composer is able to have, at the tips of his or her fingers, a simplified interface for experimentation and exploration; a medium where one can audition, react to, and refine, unfettered by the full dimensionality of compositional decision-making. To return to the metaphor of the sonic jigsaw puzzle, it enables the composer to concentrate more effectively on the assembled image rather than only on the shapes of individual pieces. Writing orchestral music with the aid of a keyboard is successful in part due to certain experiential realities of layering sound. The dense superimposition of instrumental actions, commonplace in most orchestral music, often leads to the masking of the prosaic details of individual performers. As such, the ability to shift focus from the individual to the cumulative gains additional strategic import when writing for large ensemble.

While the analogy to orchestration reveals a useful construct for managing detail-rich resources, a physical keyboard is likely not sufficiently robust to cope with the magnitude of decisions that the computer music composer confronts. While acoustic composition utilizes a system of symbolic notation which is interpreted by performers, electronic composers must manage sound production more intimately. A comparable interface must not only aid in the prescription of pitch, dynamics, and rhythms, but also timbre, morphology, and the spectral connectivity of adjacent sounds. In search of a sufficiently robust medium, our goal was not to create a tangible performative interface open to the tactile responsiveness and improvisatory caprice of a keyboard. Rather, our aim was to create software which permits sonic resource and temporal articulation to be prescribed separately yet to be realized with a degree of integration

and interdependence. By permitting the composer to address sound and time in a stratified fashion, the scope and resolution of technical engagement can be more responsive to high-level compositional intent.

AudioGuide

Aesthetic Considerations

AudioGuide[1] was conceived as an attempt to create an interface for arranging sonic resources according to a prescribed morphology. Like the orchestrator's keyboard, the program was designed to function as a medium which assists in the creation of cumulative entities from a large collection of sonic elements. The software was developed with the support of IRCAM and is a collaborative effort between Benjamin Hackbarth (composer), Norbert Schnell and Diemo Schwarz (*IMTR Team*), and Philippe Esling (*Musical Representations Team*). Development began when Hackbarth was composer in residence for musical research at IRCAM during 2010.

The program is structured around two main aesthetic interests. First is a predilection for creating electronic music from sampled sounds: more specifically, through using recordings that capture a nuanced array of sonic variation created by multiple iterations of virtually identical instrumental actions. This collection of recorded sound is hereafter referred to as a *corpus*. Taken as a whole, a corpus contains an out-of-time sonic repertoire that can serve as a reservoir for the creation of electronic sound. Second is an interest in the ability to layer the sounds of a corpus such that vertically and horizontally overlapping elements create time-varying characteristics which are evocative of a cumulative morphology. Such a morphology is rendered as experience when the phenomenological identity of individual sounds (e.g. notes) is outstripped or superseded by the totality of sonic information (streams, chords, and progressions).

Morphological Control

One of the difficulties in strategizing an intuitive form of control of sonic morphology is that qualities such as timbre are not readily disposed to manipulation with physical interfaces like a keyboard, which emphasize the importance of a single dimension. Research has shown that our perception of timbre is best described in higher dimensional spaces (Grey, 1977). Dimensional complexity problematizes compositional control—one must devise a medium where morphology can be intuitively prescribed yet contain the level of detail required to represent the complexity of sound over time.

In the case of AudioGuide, it was decided that an intuitive and exacting medium for prescribing morphology would be the use of sound itself. The program uses a user-specified soundfile, called a *target*, as a spectral template whose time-varying sonic qualities delineate a morphological structure. To capture this structure, AudioGuide analyses the target with a variety of formulas which yield time-varying measurements

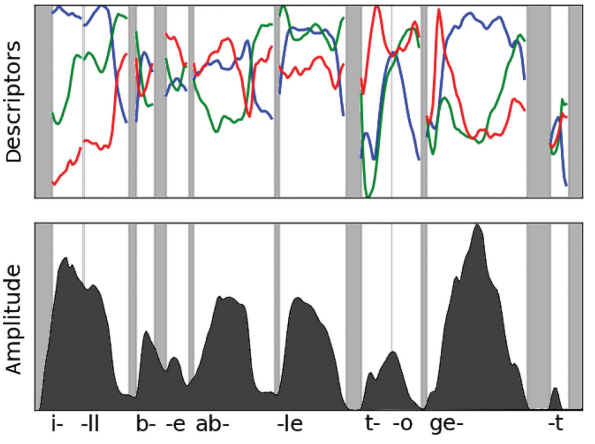

Figure 1 An Analysis of the First Three Mel-Frequency Cepstral Coefficients for Spoken Text 'I'll be able to Get.'

that describe sonic characteristics. AudioGuide provides 20 such measurements, termed audio descriptors, which approximate qualities such as pitch, loudness, and higher dimensional attributes that describe different aspects of timbre. For instance, Figure 1 shows the amplitude of a four second soundfile of speech along with the first three mel-frequency cepstral coefficients, a set of descriptors which describe timbral qualities. The white and grey regions show an algorithmic segmentation of the sound into acoustically viable chunks—the grey regions indicate that the sound has fallen below the threshold of audibility.

The complexity manifest in a multi-descriptor analysis reveals a high degree of discriminability among sonic slices—each moment is represented by different values and different relationships between values. Shown in Figure 2, each segment of speech is plotted as a coordinate in three-dimensional space according to averaged values for each of the three descriptors. The resulting geometry of variability, hereafter referred to as a *sound-space*, permits a computational measurement of similarity between different segments—neighbouring points are more similar while distant points are less similar. While a morphological trajectory (such as that shown in Figure 2) may be difficult to create manually or generatively, it is automatically obtained through an analysis of a recording. Therefore, using sound as a medium to prescribe temporal ideas is both flexible and intuitive since a high-level morphological structure can be procured from any type of sonic material (i.e. synthetic, acoustic, performed, improvised, etc.).

A Concatenative Approach

AudioGuide utilizes concatenative synthesis (Schwarz, 2006), similar to that proposed by Schwarz (2007), as a way of arranging sounds from a corpus according to a target soundfile's morphology. Programmatically, the user creates a concatenation

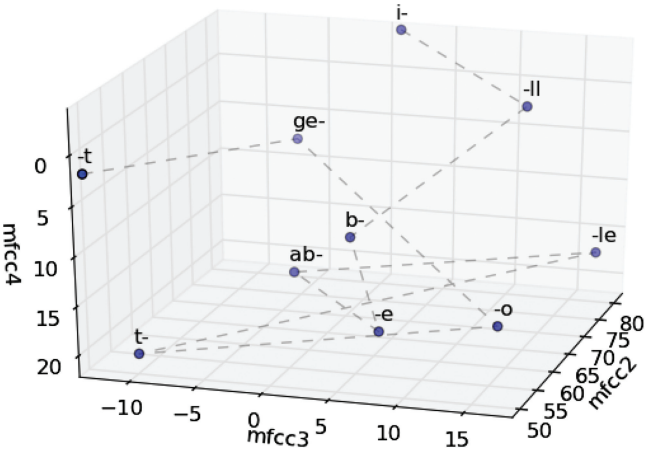

Figure 2 A Three-Dimensional Sound-Space Showing Each Segment from Figure 1 as a Single Coordinate.

by selecting a target sound, a set of corpus sounds, and a set of *N* audio descriptors which define an *N*-dimensional geometrical space for measuring similarity between corpus and target sound segments. Based on the work of Schwarz and Schnell (2010), the algorithm temporally arranges corpus sounds according to which combination of segments best matches the target's time-varying descriptors.

One of the primary challenges of using concatenative synthesis as a strategy for controlling morphology is that matching descriptor values is largely an automated process. In contrast to being confronted with an overabundance of choices and parameters, many sonic details are created algorithmically, imposing a readymade aura which can prove aesthetically troublesome. In order to give the user a higher degree of control over selection, AudioGuide provides different methods for defining and manipulating computational similarity to both aesthetic and pragmatic ends. Consider Figure 3, which shows a set of target segments and a set of corpus segments plotted in a three-dimensional space according to timbral descriptors. Using this raw descriptor data to match target and corpus sound segments maximizes imitative precision: for each target coordinate, the closest corpus coordinate in geometrical space would be the 'best' match.

A drawback of this approach comes when the variability and distribution of the corpus do not match the target's sound-space. As a result, only a small region of the corpus is selected during concatenation. In addition, the fidelity of geometric similarity breaks down when a portion of the target's sound-space is unoccupied by corpus segments.[2] In essence, the morphological variability of the target becomes less experientially significant when the target and corpus sound-spaces do not occupy the same geometrical ranges. Building on the work of Schnell, Suárez Cifuentes, and Lambert (2010), AudioGuide provides several methods for maximizing morphological mapping through warping the corpus sound-space such that it more closely

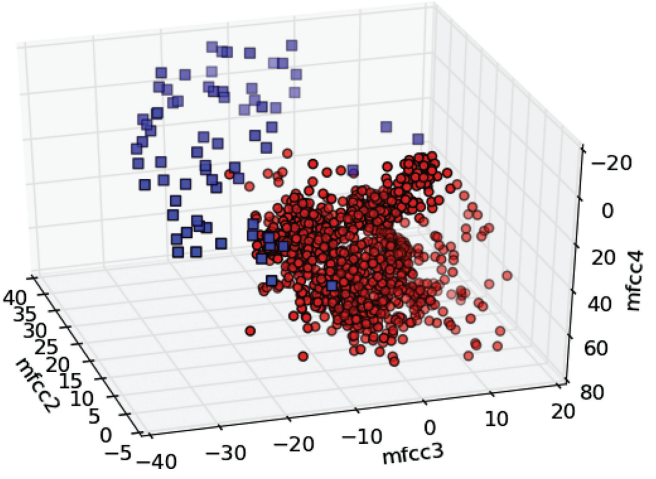

Figure 3 A Target Sound-Space (Blue Squares) and a Corpus Sound-Space (Red Circles) Graphed According to the First Three Mel-Frequency Cepstral Coefficients.

matches the target's dimensionality and distribution. Figure 4 shows the same target and corpus as Figure 3, but employs a hierarchical clustering normalization algorithm to modify the corpus sound-space so that it better correlates with the target's dimensionality. This manipulation of descriptor data changes the outcome of the concatenative process. Rather than *imitating* the target morphology—i.e. selecting corpus segments which best match raw descriptor values—the concatenative algorithm matches morphological variability, a methodology which we term *gestural transcription*.

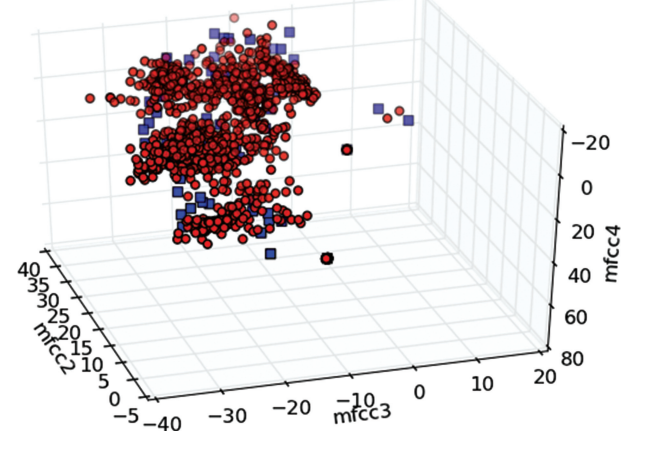

Figure 4 The Corpus Sound-Space is Warped to Better Match the Variability and Distribution of Target Segments.

The transcriptional approach moves away from real-world descriptor values to treat the data of the target and corpus as conceptual entities. When instantiated experientially, concepts are often treated with a degree of variability. For instance, consider the ways in which the experiential properties of the concept 'red' are modified depending on context: red hair, red wine, and redwood. Like colour, our semantic understanding of instruments can often be framed in a conceptual manner: a *high* note on a trombone and a *high* note on a piano are not equivalent in terms of real-world pitch, but relatable only if each phenomenon is normalized within each instrument's prototypical range of sonic variability. Assuming that the sound-spaces formed by descriptor variability can be treated as conceptual boundaries, this method for warping corpus descriptors can be thought of as reshaping the corpus such that it fits into a conceptual mould defined by the target. At the expense of attempting to match exact pitches, dynamics, and timbres, this warping strategy ensures that the expressive nuance of the target is more completely encoded in the sonic world of the corpus.

A final point of critical importance is AudioGuide's ability to account for layered sounds, both as horizontally overlapping segments and vertically stratified complexes. AudioGuide utilizes formulas proposed by Damien Tardieu which predict the descriptors of an audio mixture based on the descriptors of individual segments (Tardieu, 2008). During a concatenation, AudioGuide is able to account for the summation of previously selected segments when evaluating the similarity of subsequent selections. This permits each sonic chunk of the target to be realized as a composite of corpus sounds. Depending on the energy profile of the target and the parameterization of the algorithm, this can result in either simultaneous events (similar to chords) or overlapping selections where shorter corpus segments are chained together to fit the target's contour.[3]

Examples

Two examples of AudioGuide's use in a musical context demonstrate both the algorithm for sound-space transcription and the algorithm for simultaneous selection. Each passage comes from *Am I a Particle or a Wave?*[4] for two percussionists and 'imaginary pianist' by Benjamin Hackbarth. The piano part is a computer rendering of densely layered piano samples, recorded as single notes, which are arranged in time according to the morphology of target soundfiles. Electronic passages are notated on an 18 line staff which documents pitch, amplitude according to notehead size and playing technique according to notehead shape. The passage in Figure 5 was created with AudioGuide using a pre-recorded performance of the percussion parts as the target and an extensive database of piano sounds as the corpus. From an experiential point of view, these two sonic collections have little in common by way of timbre. As a result, the geometric relationship of sound-space distributions is not unlike that shown in Figure 3. In the case of this example, the transcriptional approach is used to warp the piano's sonic dimensionality to match the dimensions and distribution of the percussive gesture. The resulting sequence of piano chords, each generated from a single

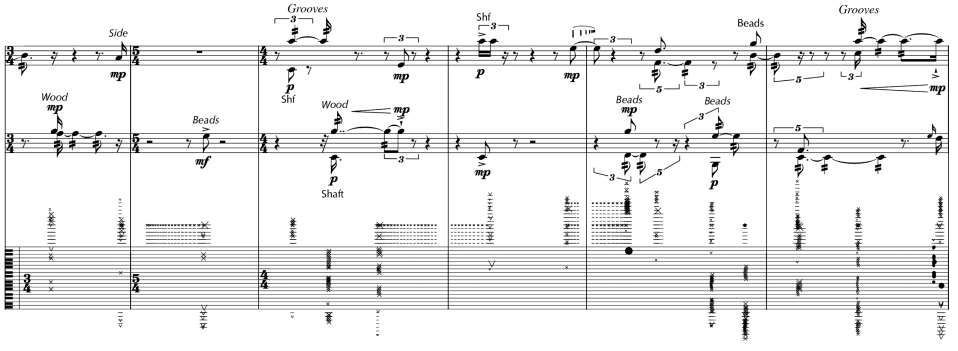

Figure 5 *Am I a Particle or a Wave*?, mm. 146–151. Audio: http://crca.ucsd.edu/~ben/cmr/5.mp3

percussive event, do not imitate the percussionist's spectra per se, but the spectral variability of the percussionist's gesture rescaled to maximally exploit the piano's expressive capacity.

In this passage, the desired result was not to create prototypical piano music, but to create timbral objects forged from the superimposition of piano sounds. AudioGuide was calibrated to select between 10 and 80 piano notes for each percussive event depending upon its spectral energies. The synthesis of timbral objects from highly pitched sound sources requires not only a high level of stratification, but also a refined interaction between overlapping sounds. While each sound of the corpus has its own timbre, corpus samples must be selected such that the composite of chosen sounds yields an experientially fused outcome. Because audio descriptors afford a high degree of discrimination between sonic categories, the relationship between each piano note's colour, technique, and register was picked such that the summation of notes projects a cumulative timbre (Figure 6).

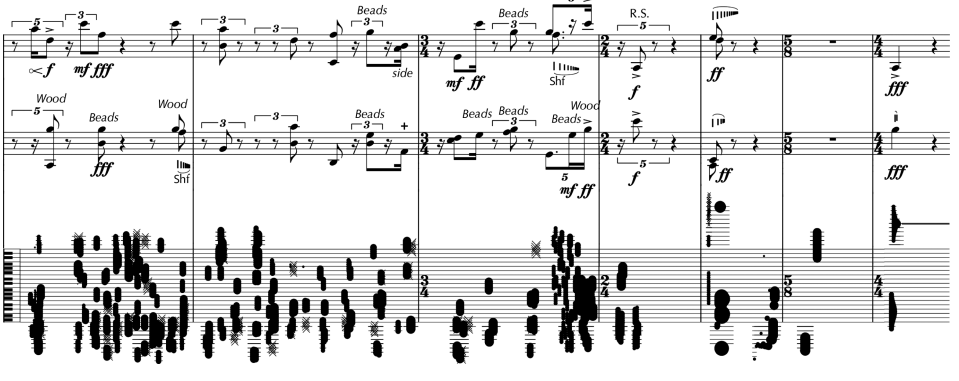

Figure 6 *Am I a Particle or a Wave*?, mm. 108–114. Audio: http://crca.ucsd.edu/~ben/cmr/6.mp3

The second example exhibits a different type of morphological divergence between the target and the corpus. A vocal recording was used as the target and, as a consequence, segments from the piano corpus are layered such that a rapid succession of onsets gives the illusion of a piano capable of speech-like dynamism. Like the previous example, the cluster-based normalization strategy is deployed so that the piano's sound-space is warped to fit the morphological variability of the recorded voice. AudioGuide's ability to account for the superimposition of corpus segments instills the result with a global shape which outstrips the phenomenology of any individual piano note found in the corpus. The resulting relationship of sonic complexes exhibits an integration of sound and gesture—the morphological shape of the vocal recording is inextricably linked to characteristics of the piano's sonic world.

Conclusion

These examples demonstrate AudioGuide's potential as a tool for generating electroacoustic gesture by arranging an out of time sonic repertoire in service of a global morphology. The inherent complexity of morphological control is addressed through the use of soundfiles: morphological shapes can be created intuitively, revised, and transformed. The sounds of the corpus can be readily organized, filtered in response to compositional ideas, and deployed with various kinds of restrictions in order to give the composer differing degrees of control over the performative resource. Results can often exceed a level of complexity that would be possible if individual sounds were auditioned, selected, and arranged by hand. By using a medium to manage sound selection, the composer can create and revise temporal ideas uninhibited by the totality of sonic possibilities. Similar to the orchestrator's keyboard, the ability to prescribe sound and time in a manner which is conceptually detached yet sonically integrated permits the composer to sculpt gesture intuitively in a detail-rich environment.

Acknowledgements

We would like to thank Arshia Cont for his assistance in the preparation of this article.

Notes

[1] Information and examples may be found at http://crca.ucsd.edu/~ben/audioGuide/

[2] In the case of Figure 3, the morphological information present in the target's upper-left quadrant will likely not be well-represented in a resulting concatenation.

[3] This approach builds on the design of other programs for concatenative synthesis. In particular, it can be seen as adding a method for simultaneous selection to an application similar to CataRT; alternatively, it can be seen as adding a time model to a program like Orchidée.

[4] http://crca.ucsd.edu/~ben/aipow/

References

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, *61*(5), 1270–1277.

Schnell, N., Suárez Cifuentes, M. A., & Lambert, J.-P. (2010). *First steps in relaxed real-time typo-morphological audio analysis/synthesis*. International conference on sound and music computing, Barcelona, Spain.

Schwarz, D. (2006). Concatenative sound synthesis: The early years. *Journal of New Music Research*, *35*(1), 3–22.

Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, *24*(2), 92–104.

Schwarz, D., & Schnell, N. (2010, June). *A modular sound descriptor analysis framework for relaxed-real-time applications*. Proceedings of the International Computer Music Conference, New York, USA.

Tardieu, D. (2008). *Modèles d'instruments pour l'aide à l'orchestration* (Unpublished doctoral dissertation). Université Pierre & Marie Curie, Paris 6.