MAKING AN ORCHESTRA SPEAK

Gilbert Nouno, Arshia Cont, and Grégoire Carpentier
Ircam-Centre Pompidou
{nouno,cont,carpentier}@ircam.fr

Jonathan Harvey
Composer
jharvey@solutions-inc.co.uk

ABSTRACT

This paper reports various aspects of the computer music realization of "Speakings" for live electronic and large orchestra by composer Jonathan Harvey, with the artistic aim of making an orchestra *speak* through computer music processes. The underlying project asks for various computer music techniques: whether through computer aided compositions as an aid for composer's writing of instrumental scores, or real-time computer music techniques for electronic music realization and performance issues on the stage with the presence of an orchestra. Besides the realization techniques, the problem itself brings in challenges for existing computer music techniques that required the authors to pursue further research and studies in various fields. The current paper thus documents this collaborative process and introduce technical aspects of the proposed methods in each area with an emphasis on the artistic aim of the project.

1 INTRODUCTION

This paper documents a collaborative work surrounding the production and realization of "Speakings" for large orchestra and live electronics, between the composer Jonathan Harvey and several researchers and computer music designers at IRCAM [1] . The central idea behind the project is the composer's aim to bring in speech and music structures together through live electronics and orchestral writings and with the aid of computer music formalizations and realizations. We therefore devote this introduction to the clarification of artistic goals of the project.

Speech and music are very close and yet also distant. The musical quality of speech has long been known to composers and artists. Today, sound analysis tools show us that speech signals not only contain melodic information but also harmonic and inner rhythmic and dynamic qualities pertained to complex musical structures. Recent research has uncovered common trajectories between the evolutionary roots of music and speech [9]. The central idea of this project is to bring together orchestral music and human speech but not merely through realistic speech synthesis and semantic contents of the speech, but to emphasize non-verbal aspects of speech structures in music composition. It is as if the orchestra is learning to speak, like a baby with its mother, or like first man, or like listening to a highly expressive language we don't understand. The rhythms and emotional content of speech are not only formed by semantics, but also (or probably more) formed by specific spectral dynamics of speech signals despite the semantic context. Therefore, making an orchestra speak in this sense is not to reach the semantic values of speech through computer music processes, but to emphasize the non-verbal structures in speech and realize them through instrumental writing as well as live electronic processes. Starting from baby screaming, cooing and babbling, an evolution of speech consciousness through frenzied chatter to mantric serenity becomes the basic metaphor of the half-hour work's trajectory.

With this respect, speech structures are introduced into musical patterns through two distinct processes: (1) With the use of computer-aided composition techniques to enhance instrumental writing. Through this process, the orchestral instruments - soli and ensembles - would imitate speech patterns, full of glissandi, fast, and a mixture of percussive consonants and sliding vowels. Computer music techniques that allow such a passage are automatic transcription of speech signals to symbolic music notation, as well as a novel automatic orchestration technique introduced later in the paper. And (2) using of real-time analysis/resynthesis techniques to reshape the orchestral audio signals to speech structures, through which the orchestral discourse, itself inflected by speech structures, is electro-acoustically shaped by the envelopes of speech taken from largely random recordings. The vowel and consonant spectra-shapes flicker in the rapid rhythms and colors of speech across the orchestral textures.

This paper is organized as follows: In the incoming section, we discuss previous works and their relation to the presented paper. Section 3 details the first phase of this work or *computer-aided composition* techniques as means to provide musical material for instrumental writing. Section 4 details real-time audio processing techniques employed for the computer music realizations and their relations to the instrumental writing both at the score level and during performance. Section 5.1 discusses performance and synchro-

[1] http://www.ircam.fr/

SMC 2009, July 23-25, Porto, Portugal

nization issues between live electronics and the orchestra, and we conclude the paper by remarks and discussions on further developments of techniques introduced in the paper.

2 PREVIOUS WORKS

The idea of bringing music and speech together in orchestral composition is not new and has a long history that goes beyond the scope of this paper. For example, the russian composer Modest Mussorgsky's orchestral writing was highly influenced by the relations between speech and music. He went further to claim that the aim of musical art as the reproduction in musical sounds not only of modes of feeling but mainly of the reproduction of modes of human speech [7]. Another more recent example in computer music, is Clarence Barlow's *Synthrumentation* by spectral analysis of speech and their resynthesis to acoustic instruments [2]. Another similar but more recent work is Claudy Malherbe's piece *Locus* for real and virtual voice (1997). Malherbe's work make use of voice analysis techniques to deduce symbolic music materials used during composition and through a formal development compromising voice, speech and noisy structures (see [8] for details and documentations). Both Malherbe and Barlow's attempts to deduce speech structures in music could be categorized within the realm of *Computer-Aided Composition*, where music materials emerge out of composers' formalizations and offline treatment of music materials and/or eventually sound synthesis. The work presented here partially adopts both approaches in [2] and [8], see section 3.1, but takes one step further by enhancing hidden speech structures through real-time analysis of orchestral sounds and their timbre-stamped synthesis using known speech structures, detailed in section 4. This addition is not only due to artistic aims of the project, but for limitations of formal analysis techniques where non-verbal structures and inner rhythmic content of speech are often lost.

3 COMPUTER-AIDED COMPOSITION TECHNIQUES

This section details the first phase of our realization, to provide preliminary musical materials taken out of analysis of voice and speech structures, and an aid to orchestral writing. Such activities are generally referred to as *Computer-aided Composition (CAC)*. The output of this phase are raw symbolic score materials that help the composer realize the orchestral score. Speech samples used for this phase are partially random and chosen recordings of radio interviews, natural baby babbles and sounds, and poetry readings chosen by the composer. We discuss this phase within two steps: In the first, we simply transcribe melodic and harmonic structures of speech and voice (if any) through sound analysis and symbolic music score. Afterwards, we incorporate non-harmonic speech structure such as formants and timbral dynamics to provide an aid for orchestration.

3.1 Melodic and Harmonic Voice Transcription

The simplest way to extract musical information out of voice (or any audio) signal is to transcribe the melodic and harmonic structures into symbolic scores. For voice and speech, this information does not illustrate most interesting internal structures (such as formants) but is nevertheless important as a first insight. Extracting the melodic part of any speech signal amounts to running a simple pitch detector on the audio available in most computer music systems. However, a better way to extract melodic pitch information on speech, and in order to be coherent with the inner-structure, is to extract melodic contours on the level of syllabic segmentations. To this end, we chose the commercially available *Melodyne* editor [2] that automatically performs syllabic segmentations and allows further refinement of results through its intelligent graphical user interface. The results of this melodic transcription are then saved as MIDI files which will be mixed later with the harmonic transcriptions.

By harmonic transcription of speech and voice signals, we aim at transcribing a partial tracking analysis of the audio spectrum into polyphonic music scores. For this aim, we pass the audio recording to a transient detector and partial tracking module based on [11], and then translate "loud" enough partials into symbolic notation followed by rhythmic quantization. The whole procedure is done in one shot and in the *OpenMusic* programming environment, and using its default libraries [1]. Figure 1 shows a snapshot of the patcher used for this procedure, starting at the top (the audio) to the bottom (symbolic transcriptions).

Combining both melodic and harmonic transcription results would result into a combined score that reveals the harmonic structures of a speech signal through symbolic music notation. Figure 2 shows a sample score result of this process. Note again that this process reveals only elementary harmonic structures of the signal through pitched notes, and does not reveal any interesting timbral or formant structures. This latter is the goal of the next section.

3.2 Automatic Orchestration

Among all techniques of musical composition, orchestration has never gone further than an empirical activity. Practicing and teaching orchestration – the art of blending instrument timbres together – involve hard-to-formalize knowledge and experience that computer music and composition systems have for years stayed away from. Although several recent attempts to design computer-aided orchestration tools should be mentioned (see [3] for a review), those systems only offer little ability to finely capture musical timbre. Moreover, they

[2] http://www.celemony.com/

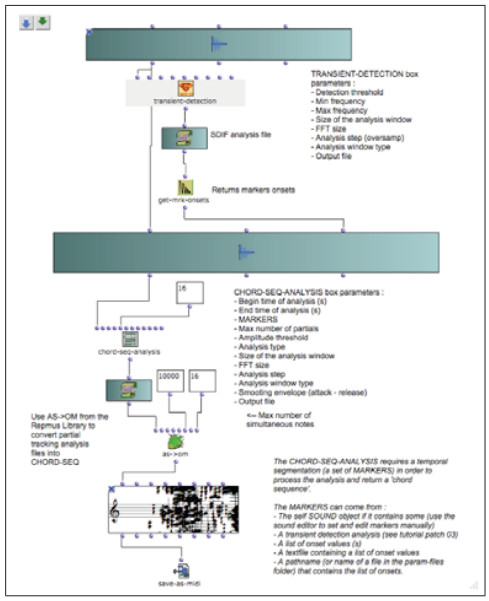

Figure 1. OpenMusic Snapshot of partial analysis tracking for harmonic transcription

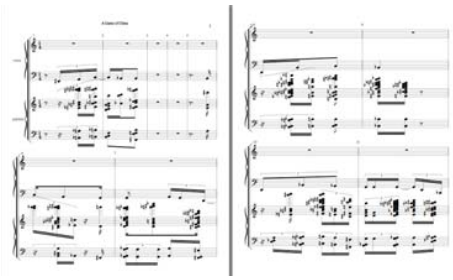

Figure 2. Resulting score sample

are often limited to small-size problems due to the cobinatorial complexity of orchestration (the set of playable sound mixtures in a large orchestra is virtually infinite).
The piece of music presented in this paper is the first to benefit from the most recent advances in automatic orchestration research [3]. With the aim of composing instrument textures that imitate the timbre of sung vowels we used the `Orchidée` [4] orchestration tool. `Orchidée` is a MATLAB-based server application that communicates with traditional computer-aided composition environments through OSC [13] messages (see figure 3). `Orchidée` embeds both a representation of instrument capabilities – obtained from prior analysis and indexation of large instrument sound sample databases – and a set of efficient orchestration algorithms. Given an input target sound, `Orchidée` outputs a musical score for imitating this sound with a mixture of traditional instruments.

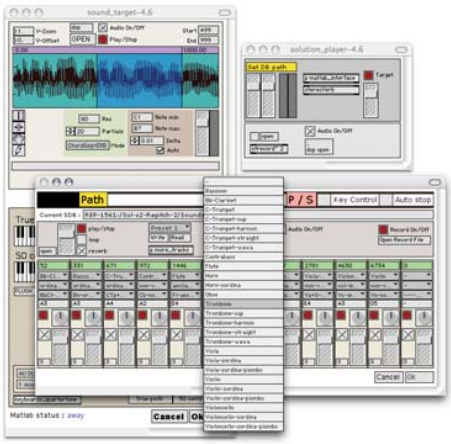

Figure 3. Controlling `Orchidée` with Max/MSP

Compared to its predecessors `Orchidée` offers many innovative features. First, the instrumental knowledge in `Orchidée` is represented by a sound description database in which each item is a an instrument sound sample associated with *musical attributes* (musical variables such as instruments, dynamics etc.) and *perceptual features* (such as brightness, roughness etc.). Second, `Orchidée` embeds a *timbre model* that efficiently estimates the joint features of any instrument sound mixture. The resulting features may then be compared to the target's features and a similarity estimate along each perceptual dimension may be computed. Last, `Orchidée` explicitly addresses combinatorial issues and tackles the orchestration problem in its inner complexity. The system comes with a time-efficient evolutionary orchestration algorithm allowing the exploration of non-intuitive sound mixtures and the fast discovery of nearly-optimal orchestration proposals. By iteratively capturing and refining users' implicit preferences, the algorithm may quickly identifies the most relevant configurations for a given orchestration situation. For more details see [4].
For the realization of the piece discussed here, `Orchidée` was used to write orchestral background textures that imitate sung vowels. The starting point was a simple three notes mantra sung and recorded by the composer. To each note corresponded a given vowel: *Oh/Ah/Hum* (see Fig. 4). The goal was to imitate the sound of the sung mantra with

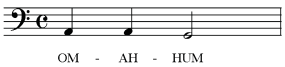

Figure 4. Mantra used as an input for `Orchidée`

a ensemble of 13 musicians. The composer wanted the orchestra to sing the mantra 22 times, and wished the resulting

timbre to evolve along the ostinato in the following manner: The sound was to become louder and louder, brighter, and closer over time to the target vowel. The orchestration was to use progressive pitches with harmonic richness. Feature optimization and constraints specification and handling techniques provided by `Orchidée` were jointly used to generate a continuously evolving orchestration. Figure 5 shows an excerpt of the overall result.

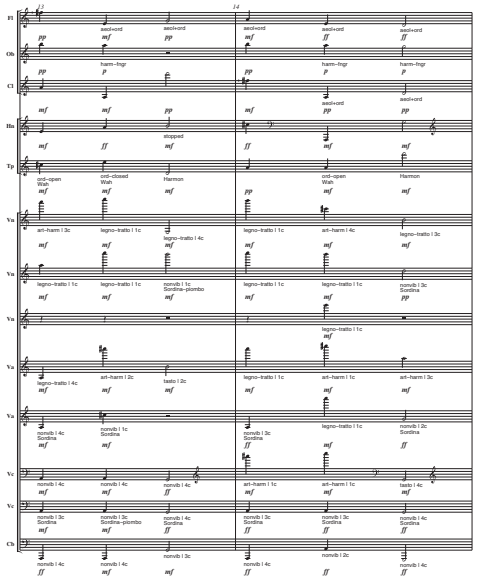

Figure 5. Excerpt of mantra orchestration by `Orchidée`

4 REAL-TIME ANALYSIS/SYNTHESIS TECHNIQUES

In the previous section, we showed how some inherent speech structures such as harmonic and formant structures could be translated to raw musical material and used during orchestral writing. Despite the significance of the information provided in this phase of work, there seem to be a lot of interesting structural information that are seemingly lost in this process, or would be lost during a realistic orchestral performance. To overcome this, we propose integrating the spectral dynamics of speech structures directly onto the orchestral spectrum through real-time analysis and resynthesis, and without passing through any formalization scheme as was the case in section 3. This way, we hope to inherit directly the inner-rhythmical and dynamic formant structures of speech onto the transformed orchestral sound.
To this aim, we couple the real-time orchestral audio, as the audio result of the transcription and orchestration process of section 3 out of given speech samples, with the speech sample itself through an analysis/synthesis scheme. The analysis/resynthesis used for this proposal is based on formant envelope computation of the speech signal using Linear Predictive Coding (LPC) and enforcing it onto the orchestral's natural spectrum envelopes arriving in real-time. The real-time implementation of this process is done using the *Gabor* libraries [12] in the *MaxMSP* programming environment [3]. Figure 6 shows a snapshot of the realtime process with visualizations of deduced formant envelopes, orchestral input and formantized orchestra on one analysis frame.

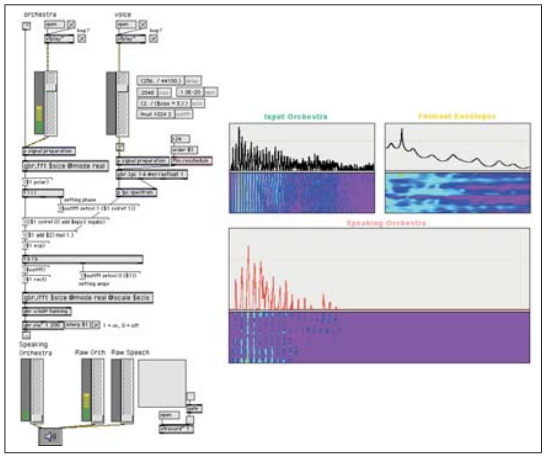

Figure 6. Realtime Formantization MaxMSP patcher

Using this process, the inner-rhythmical structures of speech are stamped into the orchestra and could be diffused as live electronics through a spatialization process. This process, hence introduces a second phase of composition involving the coupling of pre-recorded speech with sections in the orchestra and parametrizing the live electronics for diffusion.

5 PERFORMANCE ISSUES

5.1 Live electronic Synchronization

In an orchestral setting, human musicians' temporal synchronization is assured during the live performance through constant and active coordination between themselves, the music score and a conductor. Adding live or fixed electronics into the equation should not undermine the importance of this synchronization process, which is one of the main responsible factors for musical expressivity.
Traditionally, synchronization between electronics and instrumental components of a mixed piece of music has been done either by human performers through adhoc cueing, by

[3] http://www.cycling74.com/

score following paradigms, or by a combination of both. Either way, this type of synchronization usually assures the starting point (or correct triggering) of processes in time but not necessarily their temporal life-span. The following simple example can illustrate this important problem: Imagine an instrumental score accompanied by a fixed electronic (audio file) over three measures. Assuming that the score has a time-signature of $4/4$ with a tempo of $60BPM$, the corresponding audio should ideally have an initial length of 12 seconds. During live performance, the human player might for many reasons vary the initial tempo of 60 from the very beginning to the end of the third measure. In such situations, although the onset trigger could be easily made accurate, synchronization of the overall electronics could not be assured unless the electronic process detects and undertakes the same temporal dynamics of the human performance over the electronic score, as if two humans were interacting.

Given the temporal nature of the analysis/resynthesis technique described in section 4 we are in the schema of the example above: The speech samples behind the live processing should not only be triggered on-the-fly, but also temporally aligned to the orchestra during their life-span. To achieve this, we use `Antescofo`, a score follower that aligns score positions and also decodes an anticipatory tempo of the performance [5]. Using this information each sound file is then played back using `SuperVP` [4] advanced phase vocoder technology [11] to preserve their quality upon temporal adaptation. To achieve this, a keyboard player in the orchestra plays its own score along with the orchestra and synchronous to the tempo given by the conductor. The `Antescofo` score of the keyboard part contains not only the instrumental score, but also electronic commands written in relative beat-time, translated to clock-time at each tempo change. This way, getting back to the simple example described above, we can be certain that upon continuous tempo change of the instrumental section, the audio playback is assured to change time span during live performance and up to an acceptable precision. This procedure is applied to all live treatments and sound diffusions.

5.2 Live Electronics Interpretation

The coexistence of orchestral and computer music parts emerge from specific artistic purposes. Composer's initial idea in employing live electronics was to enhance the sonic space of the piece with a network of relations between electronic and instrumental sounds. Such internal or intimate sound relations emerge not only through compositions but also through interpretation of the electronics during performance. With this respect, live electronics is considered as an additional *instrument* in the piece whose performed gestures during any executation contribute to the wholeness of the sonic space created within the piece. We have used the `Lemur` [5] multi-touch screen interface to easily map hand gestures onto the sound processes, and to control the amounts of sound processing like a musician would control the amount of a muted sound listening to what she is producing in interactions of his movements with what she hears and expects. This feedback behavior goes farther than a simple mixing of the audio signals as it enables a direct interpretation of the musical choices. Figure 7 shows the main control screen designed for live performance as about the size of two hands.

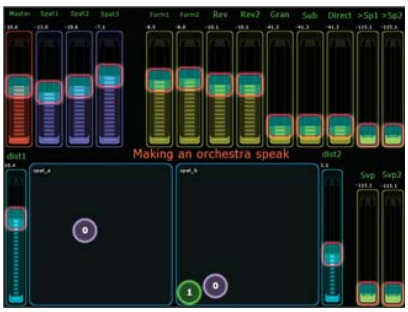

Figure 7. The *Lemur* multi-touch device interface

Human intervention of this kind for interpretation of live electronics is in no way in contradiction to the automaticity of cue-lists or the use of score following paradigms, but rather complimentary. The passage from any automatic process to "music" is beyond the process itself and depends on our hearing abilities and interaction with the sound itself as musicians. This is inherently the purpose of the simple interactive framework using the `Lemur` interface setup of figure 7. The act of making live electronic music design is then to find the balance between automatism and interference to make interactions more musical.

5.3 Enhancing Rhythm with Space

The interactive surface display of the `Lemur` in figure 7 offers extended possibilities rather than mere mixing of effects, to control spatial movements either automatically or manually through a virtual 2D space. Sound spatialization is achieved using dedicated intensity panning of `Spat` [6] real-time modules [6]. Using spatialization, we bring the sound into the audience space and put the emphasis on some theatrical characters of the sound, thus reinforcing and converging toward the voice quality of the sound processing. It is also a mean to musically enhance the rhythmic perception through fast movements, either within a rhythmic counterpoint with the orchestra or in phase with what it is

[4] http://forumnet.ircam.fr/708.html?&L=1

[5] http://www.jazzmutant.com

[6] http://forumnet.ircam.fr/692.html

being performed. It is also an extension of the score writing techniques as space is considered as a compositional parameter [10]. An advantage of employing `Spat` modules for this 8-channel work is its ability to *perceptually control* the spatial quality rather than mere positioning of sources in space and in realtime. `Spat` movements are algorithmically composed using modules offered by the `ICST` [7] interface tools originally developed for ambisonics spatialisation techniques, here adapted to the `Spat` modules. Figure 8 shows the diagram of the real-time performance setup implemented in the MaxMSP programming environment based on the considerations discussed in this section.

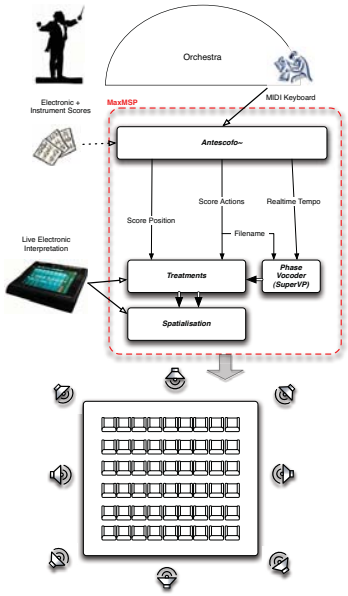

Figure 8. Real-time Performance diagram

6 CONCLUSION

In this paper, we documented a collaborative work between a composer and researchers to create an orchestral piece with live electronics with the aim of enforcing speech structures over an orchestra. "Speakings" was first performed by the BBC Scottish Orchestra and IRCAM in the BBC Proms festival on August 2008 and has since had several more performances. An audio recording is under publication.

7 REFERENCES

[1] Gérard Assayag, Camilo Rueda, Mikael Laurson, Carlos Agon, and O. Delerue. Computer Assisted Composition at Ircam: From PatchWork to OpenMusic. *Computer Music Journal*, 23(3), 1999.

[2] Clarence Barlow. On the spectral analysis of speech for subsequent resynthesis by acoustic instruments. *Forum phoneticum*, 66:183–190, 1998.

[3] Grégoire Carpentier. *Approche computationnelle de l'orchestration musicale – Optimisation multicritère sous contraintes de combinaisons instrumentales dans de grandes banques de sons*. PhD thesis, UPMC Paris 6, Paris, 2008.

[4] Grégoire Carpentier and Jean Bresson. Interacting with Symbolic, Sound and Feature Spaces in Orchidee, a Computer-Aided Orchestration Environment (accepted for publication). *Computer Music Journal*, 2009.

[5] Arshia Cont. Antescofo: Anticipatory synchronization and control of interactive parameters in computer music. In *International Computer Music Conference*, North Irland, Belfast, Août 2008.

[6] Jean-Marc Jot and Olivier Warusfel. A real-time spatial sound processor for music and virtual reality applications. In *ICMC: International Computer Music Conference*, pages 294–295, Banff, Canada, Septembre 1995.

[7] Leslie Kearney. *Linguistic and musical structure in Musorgsky's vocal music*. PhD thesis, Yale University, 1992.

[8] Claudy Malherbe. *The OM Composer's Book*, volume 2, chapter Locus: *rien n'aura eu lieu que le lieu*. Editions Delatour France, 2008.

[9] Steve Mithen. *The Singing Neanderthals: The Origins of Music, Language, Mind and Body*. Weidenfeld & Nicolson, 2005.

[10] Gilbert Nouno and Carlos Agon. Contrôle de la spatialisation comme paramètre musical. In *Actes des Journées d'Informatique Musicale*, pages 115–119, Marseille, France., 2002.

[11] Axel Roebel. Adaptive additive modeling with continuous parameter trajectories. *IEEE Transactions on Speech and Audio Processing*, 14-4:1440–1453, 2006.

[12] Norbert Schnell and Diemo Schwarz. Gabor, multi-representation real-time analysis/synthesis. In *COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 122–126, Madrid, Spain, Septembre 2005.

[13] Matthew Wright, Adrian Freed, and Ali Momeni. OpenSound Control: State of the Art 2003. In *Proceedings of the 2003 Conference on New Interfaces for Musical Expression (NIME-03)*, Montreal, Canada, 2003.

[7] http://www.icst.net