Computer-Assisted Orchestration, Format Theory, and the Construction of Timbre in New Media Culture

February 4, 2021, IRCAM Forum Workshops @ Montreal

---

1. Introduction

[SLIDE 1] Hi, everyone! I'm Landon Morrison and I'm happy to be presenting remotely here at the IRCAM Forum workshops! I want to thank you in advance for listening and I look forward to our discussion at the live round table.

[SLIDE 2] In his 2008 work *Speakings* for orchestra with live electronics, the late British composer Jonathan Harvey used newly developed software at IRCAM to model his instrumental writing on vocal gestures, realizing his "artistic aim of making an orchestra speak through computer music processes."[1] The work dramatizes this goal by means of an audible program, first heard in the opening movement when the oboe and strings join forces with what the score describes as "baby screaming, cooing, and babbling." [X]

The extramusical plot continues in the second movement, where we hear a trombone simulating "adult chatter." [X]

And finally, this progression of speech genres arrives on a tutti orchestration of the Tibetan Buddhist mantra "OM – AH – HUM," [X] completing a narrative arc, in which, according to Harvey, "it is as if the orchestra is learning to speak."

---

2. Formatting Timbre in Assisted Orchestration Software

[SLIDE 3] In this paper, I'd like turn this narrative on its head by showing how, before it was possible to *make an orchestra speak*, it was necessary to *make software listen*. Drawing on archival materials housed at IRCAM and the Paul Sacher Foundation, I will track the Harvey's creative process as he transitions from a hybrid software configuration to the newly-developed Orchidée program. [X] The old setup entailed use of a commercially available application, Melodyne, to transcribe the fundamental pitch of speech patterns, and a custom partial-tracking application to analyze higher overtones within the spectrum. [X] By comparison, Orchidée coupled spectrum analysis with the logic of music information retrieval (MIR) systems, allowing low-level acoustic features to be extracted from an audio signal, translated into high-level semantic descriptors, and cross-referenced against a database of pre-analyzed instrumental samples. Crucially, the indexical operations translating between signals and semantics in this process relied on definitions of timbre encoded in standard file formats, like MPEG-7. So, to learn more about what software like Orchidée is "hearing," we need to think about how the idea of timbre is constructed in the program's surrounding information infrastructure.

[1] Nouno, Harvey, et al. 2009.

[X] Borrowing from sound and media studies, I approach timbre classification from the perspective of *format theory* which, as proposed by Jonathan Sterne, dwells on "smaller registers like software, operating standards, and codes, as well as larger registers like infrastructures, international corporate consortia, and whole technical systems."[2] Extended to assisted orchestration software, format theory provides a valuable tool for understanding how the categories used to classify timbre are themselves contingent on the metrics of a multi-dimensional "timbre space," as well as on the wider network of research institutions, private industry, and governmental organizations involved in negotiating how sound is represented in new media. In the final section of my talk, I will consider in greater detail how the encapsulation of psychoacoustic models in formatting standards acted as an essential pre-requisite to the functioning of software like Orchidée. But to begin, I'd like to demonstrate how Harvey's two modes of instrumental voice synthesis worked in practice by revisiting two of the passages from *Speakings* that we heard at the top.

---

3. Analytical Examples from Harvey's *Speakings*

[SLIDE 4] + [SLIDE 5] As an example of the pre-Orchidée setup, the passage here is one that Harvey created using Melodyne and partial-tracking software, [X] implementing a two-part process evident in the score by a division of the orchestral texture into a solo trombone line (for the voice) and tremolo strings (for the higher partials). The underlying model is adult "chatter," and as you will hear in this slightly longer cut, there is a kind of call-and-response between the orchestra and its playback in the live electronics, where instrumental sounds are filtered through an additional speech envelope before being spatially diffused into the auditorium through loudspeakers. [PLAY + 2x].

Like Charlie Brown's teacher, the orchestra remains indecipherable, but is nevertheless animated by a speech-like impulse—a kind of subterranean logic that drives the morphological flow of the passage. Though not speaking from experience, I imagine this phantasmagoric effect is further amplified in live concerts, where audiences can plainly see the instruments responsible for making sound, while still sensing an invisible presence behind the orchestra. Adopting Brian Kane's formulation of *acousmatic*, or unseen, sounds as those that feature a "spacing of source, cause, and effect,"[3] this music might be understood as invoking a kind of *virtual acousmaticity*, where two sets of sources, causes, and effects exist simultaneously: one for the instrumental performance and one for the underlying sound model.

[SLIDE 6] The recordings Harvey used as source material in this piece vary widely, with some, such as a recitation of T. S. Eliot's "The Wasteland," imbuing the music with heady overtones, while others are decidedly less serious. The excerpt we just heard is based on an interview with Matt Groening, creator of *The Simpsons*. [X] Shown here, Harvey's transcription features the fundamental line of Groening's voice with penciled-in text on the upper staff of each

---

[2] Sterne outlines "Format Theory" in Chapter 1 of his book, *MP3: The Meaning of a Format* (Durham: Duke University Press, 2012), 1–31.
[3] Brian Kane, *Sound Unseen: Acousmatic Sound in Theory and Practice* (New York: Oxford University Press, 2014), 114.

system, while higher partials are clustered on the grand staff below. Try to follow along as we listen to a short clip: [PLAY + 1x]. Impressively detailed at first glance, it's worth noticing how this transcription flattens the voice's dynamic profile and quantizes its frequency and rhythmic content into twelve-tone equal temperament and a 2/4 metric grid, respectively.

[SLIDE 7] To hear the connection between Groening's voice and the trombone solo, I'd like to re-play two abbreviated audio clips back-to-back. [X] I can also zoom in on the score to give you a better view. Here is Groening [X]. And now the orchestra [X]. There is a clear ressemblance between the voice and ensemble, despite a lack of technological specification for suitable combinations of instruments, articulations, and dynamics. Rather, in the approach taken here, timbral qualities were represented as a combination of the voice's fundamental frequency paired with upper partials, leaving largely untouched the kinds of perceptual attributes that orient multi-dimensional timbre-space models, and which are incorporated in programs like Orchidée.

[SLIDE 8] + [SLIDE 9] My second example is one that Harvey created using Orchidée, which at the time, pegged sound translation to three features: [X] the main resolved partials (or MRPs), spectral centroid (i.e., mean frequency), and spectral spread (i.e., standard deviation), as shown in this diagram by the Orchidée team.[4] Moreover, before measurements were taken for these features, sounds passed through a series of algorithms designed to detect spectral peaks, model the transfer function of the inner ear, and account for perceptual loudness. Thus, the program modeled not only the signal, but the perception of that signal, in effect, "listening" for the composer through what Bruno Latour would describe as processes of delegation.

[X] Packaged in the program's user interface were a number of workflow possibilities, many of which the research team depicted in an algorithmic map of orchestration scenarios.[5] [SLIDE 10] The simplest path proceeded directly from a target sound, through the feature extraction process, into a search for similar samples within a database, and finally to a set of possible solutions, which the composer could navigate and interact with through the software interface. [X] Alternate routes allowed the composer to verbally define a desired timbre using recognized audio descriptors, [X] to define a set of instrumental constraints on the search process, [X] or to specify listener preferences that order final solutions according to one's own musical priorities.

[SLIDE 11] Ultimately, a target sound entered into the program is converted to a notated score; here are Orchidée solutions that Harvey used to orchestrate the aforementioned Buddhist mantra, "OM-AH-HUM," which he chanted himself in this source recording [PLAY]. As you can see, unlike the technologies used in our previous example, Orchidée indicated specific instruments, along with articulations, dynamics, and special playing techniques. This more detailed solution was obtained by taking the unique spectral properties of each instrument into account for a calculation that would be impractical without the aid of a computer. The tradeoff, of course, was that early iterations of Orchidée were only able to process steady-state sounds; so

[4] Carpentier et al., "Predicting Timbre Features of Instrument Sound Combinations: Application to Automatic Orchestration," *Journal of New Music Research* 39, no. 1 (2010): 47–61; Carpentier and Bresson, "Interacting with Symbol, Sound, and Feature Spaces in Orchidée, a Computer-Aided Orchestration Environment," *Computer Music Journal* 34, no. 1 (2010): 10–27;
[5] Ibid.

in these scores, each measure represents a different orchestral solution for a static spectrum. Working around these limitations, Harvey applied a series of different constraints during the search process, varying the list of instruments, dynamics, techniques, and the range of partials to be included in the results.[6] He then filtered this process to create the impression of an over-arching timbral transformation, repeating the phoneme sequence many times and tweaking each iteration to make the timbre grow louder, brighter, and closer to the target sounds. Through this imposition of constraints, he directed a timbre-driven motion that shapes the orchestration of a lengthy passage at the end of the second movement.

[SLIDE 12] I will play an excerpt from this passage, so you can hear the spectral evolution of the repetitive "OM-AH-HUM" chant grounded in the bass, while seeing the idiosyncratic markings Harvey includes for in each instrument, which change erratically from measure-to-measure. [PLAY] Whether or not this passage succeeds in a clear recitation of the mantra, it is still remarkable for its use of a new technology to produce the orchestra's flickering timbral allusions to language.

---

4. Tracing the ACTOR Network

[SLIDE 13] Since the premiere of Harvey's *Speakings* over a decade ago, Orchidée has evolved into the more advanced Orchidea, boasting dynamic sound analysis and machine-learning algorithms.[7] [X] No longer the purview of a small collaborative team at IRCAM, the program has been assimilated into the broad framework of the "ACTOR Project," for the Analysis, Creation, and Teaching of Orchestration. As many in the audience will know, this project brings together experts from a number of academic, private, and government institutions to advance new methods and tools for the study of timbre in orchestration. Given its expansive reach, it is convenient to think of the ACTOR partnership in Latourian terms as an "ACTOR Network," within which Orchidea and its supporting file formats play a mediating role as dual scientific and musical instruments, shaping the way composers and researchers alike interact with sound. [X] Rewinding the history that feeds into this network, we're taken back to the idea of "timbre space" that emerged in the mid-seventies along with multi-dimensional scaling techniques (MDS) and developed into familiar formats like MPEG-7 over the course of the next few decades.

[SLIDE 14] Shown here, an MDS graph from a 1977 study by John Grey rates a group of instruments rated according to scales of dissimilarity in three dimensions—brightness, spectral flux, and attack—with scores based of the listening tests of "musically sophisticated" subjects.[8] These highly controlled tests have since been criticized by researchers like Reinier Plomp for focusing too narrowly on the perception of isolated synthetic tones in "clean" laboratory spaces cut-off from the "dirty" conditions of everyday listening,[9] but they remain a cornerstone of timbre-space studies.

---

[6] Harvey et al. 2009.
[7] Esling et al. 2010.
[8] Grey 1976, 1272.
[9] Plomp 2002.

[SLIDE 15] By the late-nineties, the timbre-space model was being incorporated into new classification standards and employed in a more predictive capacity. A decisive event in this shift was the formation of a European project dubbed CUIDADO (for Content-based Unified Interfaces and Descriptors for Audio Databases Online), which was initiated to produce the framework for MPEG-7, a.k.a. the *Multimedia Content Description Interface*. Approved by the ISO in 2002, MPEG-7 standardized a set of descriptive categories for representing sound as metadata using XML markup language.[10] Shown here, it subdivided low-level descriptors based on temporal, energy, spectral, harmonic, and perceptual features, setting up a framework that informed later applications like the Timbre Toolbox for MatLab.

[SLIDE 16] As seen in this color-coded diagram, the Timbre Toolbox retains over forty audio descriptors, [X] including both global and time-varying aspects of sound; [SLIDE 17] but of these features, it is only possible to distinguish ten independent classes.[11] This kind of variance hints at a problem of correlation between statistical models and timbre perception, and it points to the potential for larger controversies between disciplines with opposing objectives. For instance, as has been noted by many scholars in both fields, MIR-based research and cognitive psychology tend to diverge on which parts of the audio signal should be used as physical correlates of timbre perception, as well as to what extent one can relate these to higher-level constructs, such as genre, mood, instrumentation, etc.[12] As a result, MIR researchers may analyze hundreds of audio features to infer semantic descriptions of the music based on established statistical correlations, while cognitive psychologists consider only a handful of acoustic correlates relevant because the field is concerned with a whole different set of evaluative criteria, such as the perceptual and physiological accuracy of the model. Such discrepancies bring to light the often-hidden processes of negotiation that underwrite formatting standards, opening up space to interrogate why one particular set of audio features, instead of another, should become constitutive of timbre.

---

5. Concluding Thoughts

[SLIDE 18] Timbral taxonomies are important because they crystallize a set of assumptions about sound, filtering signals through psychoacoustic data and infusing them with cultural understandings of audition. At the level of formats, timbre gets of reduced to a set of perceptual categories that performs a consensus of scientific study and industry regulation, which I have characterized here in terms of an ACTOR network. The controversies that arise within this network raise important questions about how timbre formats elide diverse sonic practices within

[10] ISO/IEC 15938.
[11] Peeters et al. 2011.
[12] See Jean-Julien Aucouturier and Emmanuel Bigand, "Mel Cepstrum and Ann Ova: The Difficult Dialog Between MIR and Music Cognition," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 397–402 (Porto, Portugal: FEUP Ediç oes, 2012). See also Kai Seidenburg, Ichiro Fujinaga, and Stephen McAdams, "A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology," *Journal of New Music Research* 45, no. 1 (2016): 27–41.

the fuzzy bounds of a contested nomenclature. For when timbre classifications are locked into a standard, they foreclose on future opportunities for difference and variety by naturalizing a set of listening techniques, as though they were universal and not culturally specific. Thus, we must remember that what seems to pass as a general theory of timbre in formats like MPEG-7 is, in fact, a theory of orchestral timbre that is shot through with a history of instrumental playing (and listening and recording) techniques, along with a way of talking about these things—all of which emerged from a common substrate of Western musical values. By reading the timbre classifications and standards inscribed in file formats against the absence of alternatives, we become aware of the historical specificity built into such representations, and we gain critical perspective on how programs like Orchidée mediate between a composer's creative process and their access to an invisible information infrastructure.

Thank you for listening. I am happy to respond to any questions or comments you might have, so please feel free to send me an email, and I hope to see you at the live round table.

[SLIDE 19 + 20]