Aaron Einbond,* Thibaut Carpentier,† Diemo Schwarz,† and Jean Bresson**

*Department of Performing Arts
City, University of London
Northampton Square, London EC1V 0HB, UK
†Institut de Recherche et de Coordination Acoustique/Musique (IRCAM)
STMS Lab, CNRS, Sorbonne Université, Ministère de la Culture
1, place Igor Stravinsky, 75004 Paris, France
**Ableton
Schönhauser Allee 6–7, 10119 Berlin, Germany
Aaron.Einbond@city.ac.uk,
{thibaut.carpentier, diemo.schwarz, jean.bresson}@ircam.fr

Embodying Spatial Sound Synthesis with AI in Two Compositions for Instruments and 3-D Electronics

Abstract: The situated spatial presence of musical instruments has been well studied in the fields of acoustics and music perception research, but so far it has not been the focus of human–AI interaction. We respond critically to this trend by seeking to reembody interactive electronics using data derived from natural acoustic phenomena. Two musical works, composed for human soloist and computer-generated live electronics, are intended to situate the listener in an immersive sonic environment in which real and virtual sources blend seamlessly. To do so, we experimented with two contrasting reproduction setups: a surrounding Ambisonic loudspeaker dome and a compact spherical loudspeaker array for radiation synthesis. A large database of measured radiation patterns of orchestral instruments served as a training set for machine learning models to control spatially rich 3-D patterns for electronic sounds. These are exploited during performance in response to live sounds captured with a spherical microphone array and used to train computer models of improvisation and to trigger corpus-based spatial synthesis. We show how AI techniques are useful to utilize complex, multidimensional, spatial data in the context of computer-assisted composition and human–computer interactive improvisation.

How can one situate the listener inside a virtual musical instrument, and how can the interaction of human and AI help to realize this goal? Answering these questions requires consideration of the complex ways acoustic musical instruments interact with the space in which they are situated. Previous research has elucidated the 3-D radiation patterns of many instruments and pointed to their dependence on instrument construction, orientation, and performance (Hohl and Zotter 2010; Shabtai et al. 2017). Attempts have been made to reproduce these patterns synthetically, but primarily in the context of research rather than artistic creation (Zotter 2009; Noisternig et al. 2011). Conversely, artists have deployed sound spectra spatially through techniques they have termed "timbre spatialization" (Normandeau 2009) and "texture composition" (Hagan 2017), or used embodied gesture to control spatial synthesis (Goeschke 2022), but without reference to measured instrumental radiation patterns.

Computer Music Journal, 46:4, pp. 43–61, Winter 2022
doi:10.1162/COMJ_a_00664

Artificial intelligence could be a promising tool to apply to this problem. Yet, despite the importance of the spatial presence of musical instruments for situated perception (Schmeder 2009), spatial sound has, as yet, not been the focus of research in AI. Most AI musical applications do not consider the spatial presence of performers, instruments, and listening subjects, ignoring how sound is recorded and reproduced: whether with headphones or loudspeakers, or whether in monophonic, stereophonic, or multichannel formats. We respond critically to this omission, situating the performer and listener at the center of human–AI interaction by attempting to reembody the spatial presence of musical instruments using AI. This approach requires rich spatial data that we derive from natural acoustic phenomena by exploiting measured radiation patterns of orchestral instruments as models for the diffusion of synthesized sounds. We investigate how these instrumental radiation patterns can be used as a training dataset for a machine learning (ML) model that then responds live to the human performer, synthesizing each sound with its own detailed 3-D radiation pattern. We build on previous work (Einbond et al. 2021), in which we applied this approach to diffusion with an Ambisonic loudspeaker dome, now extending it to a compact spherical loudspeaker array (SLA).

A further addition to our previous work is to extend the sequence of spatial forms generatively over time. Again, this can be compared to an acoustic instrument, whose complex spatial presence is not a steady state but changes dynamically as the performer produces different sounds, each with a unique radiation pattern (Meyer 2009). We examine how to model this temporal spatial dimension using human–AI interaction, enabling the computer to analyze and continue spatial gestures. Building on previous research by our team and others in computer improvisation (Einbond et al. 2016), we train an ML model on the sequence of spatial sounds during the live performance. This model, both trained and performed in real time, allows the computer to "improvise" a continuation based on the context of the preceding sounds. Taken together with the synthesis of spatial patterns, these interacting ML agents allow the computer to produce an independent spatial improvisation in a real-time response to the live performer.

Musical Motivation

Our approach to human–AI interaction was motivated by two musical compositions by the first author that draw upon natural acoustic phenomena as source material for spatial sound. One source is a 3-D microphone array, the em32 Eigenmike by mh acoustics, a 32-channel array (https://mhacoustics.com/products) used to capture the live instrumental performance and diffuse it spatially. The other is generative spatial sound synthesis produced through ML of an existing large database of radiation measurements for acoustic instruments (Shabtai et al. 2017; Weinzierl et al. 2017). These two sources of spatial sound are intentionally overlapped and fused so the listener cannot easily distinguish or segregate the sources. The aesthetic goal is to create a setting for curious and detailed listening, where one may not discern the "sleight of hand" between the live performer and computer, as suggested by the title of one of the works, "Prestidigitation."

Piano

"Cosmologies" for piano and 3-D electronics began with a proposal, made by Einbond, to situate the listener inside a larger-than-life virtual grand piano in order to experience its secret inner life. To achieve this, the sound field inside the body of the piano needed to be captured from a virtual listening position. This led to the decision to position the em32 above the center of the instrument's frame with the lid removed, a listening position that would be difficult to achieve acoustically. The pianist acts on the instrument with found materials of contrasting textures and densities: aluminum foil, metal knitting needles, a vegetable scrub brush, a guitar plectrum, a rubber inner tube, and a Superball (see Figure 1). The same materials and performance techniques, realized by the composer, were recorded to produce the score and source samples for the electronics. As the performance unfolds, these preparations are gradually placed inside the piano and excited cumulatively. Like a puppeteer or Foley artist, the pianist manipulates the preparations while exploring the space within the piano and around the microphone. This microchoreography is magnified and projected to the listener with a surrounding Ambisonic dome. In the premiere performance by pianist Alvise Sinivia, produced by the Institut de Recherche et de Coordination Acoustique/Musique (IRCAM) in the Grande Salle of Centre Georges Pompidou, Paris, a configuration with 27.2 channels was used. A video and binaural recording of this performance can be viewed at https://youtu.be/jKIWLwPrun4.

Percussion

"Prestidigitation" for percussion and 3-D electronics was motivated by a related idea, placing the listener

Figure 1. Performance of "Cosmologies" by pianist Alvise Sinivia showing the positions of the Eigenmike, objects, and preparations. (Photo by Quentin Chevrier.)

virtually in the middle of a sculptural percussion setup to hear sonic details normally only audible to the percussionist. Working in close collaboration, Einbond and the performer, Maxime Echardour, constructed a purpose-built frame and suspended small percussion instruments surrounding the em32 microphone. Together they selected found and handmade instruments of different materials, including chimes and shakers made of seashells, nutshells, metal, bamboo, and wood. Similar to "Cosmologies," over the course of the performance, the percussionist gradually "builds" the instrument by adding each element to the frame. A large frame drum is finally positioned beneath the microphone to complete the setup (see Figure 2). Unlike the larger-than-life Ambisonic dome of "Cosmologies," however, a contrasting diffusion system was chosen to fit the more intimate scale of the percussion setup: the IKO icosahedral loudspeaker, a compact SLA (cf. Zotter et al. 2017). This permits the projection of spatial sounds and gestures captured by the em32 to a diffusion system of physical size and radiation characteristics similar to the percussion instruments themselves. Like an acoustic instrument, the IKO is situated in the middle of the space, allowing for complex interaction between the bodies of the performer, instruments, loudspeaker, listeners, and acoustical environment. Unlike "Cosmologies," for the first performance of "Prestidigitation" (in IRCAM Studio 5), Einbond and Echardour decided to free the listeners to navigate the space around and between the IKO and live percussionist, changing the relationships between their bodies and the environment, and experiencing different perspectives of the real instruments and their projected "doubles." A video and binaural recording of "Prestidigitation" can be viewed at https://doi.org/10.1162/comj_a_00664.

Figure 2. Percussion setup for "Prestidigitation" showing the Eigenmike in the center and the IKO in the background.

Technical Realization

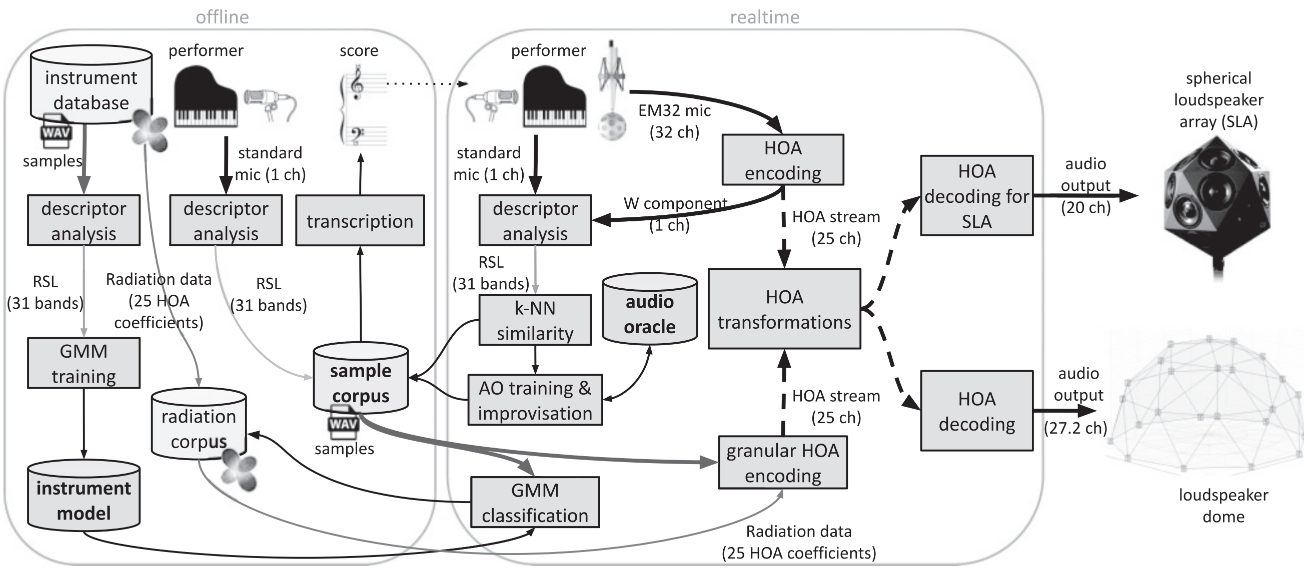
The approach to human–AI interaction motivated by these compositions is implemented through the dialogue of the human performer with three machine listening or machine learning agents that can be used together or separately. One agent reacts to the live performance with electronic sounds whose 3-D radiation patterns are learned and reproduced from acoustic instruments. Another learns from the sequence of the performer's timbral and spatial gestures and extends them through computer improvisation. The third, upon which the others both rely, is an underlying system of machine listening utilizing audio features to analyze timbral descriptors of each sound and connect them to the sound's spatial characteristics. Although the machine learning algorithms used are not themselves new, we argue that their combined creative application leads to novel possibilities for interactive and generative synthesis of spatial sound. At the same time, the choice of relatively simple algorithms—as compared to algorithms that are more computationally intensive—presents advantages for interactivity in terms of relatively small training sets and fast, or real-time, training. We implement these models using software tools connecting, for the first time, the computer programs Max, Python, and OM# (the last of these is described more fully in in the Computer-Assisted Composition section), with the associated packages Spat (Carpentier 2018) and Mubu (Schnell et al. 2009), as shown in Figure 3.

Machine Listening and Corpus-Based Sound Synthesis

Machine-listening algorithms, together with efficient pattern recognition, can serve as an effective technique for the measurement and classification of audio similarity. Machine listening is not the same as human listening, however, and it requires subjective human input. In particular, what is measured by "similarity" is not a neutral decision, but instead "aesthetically situated" (see Gioti et al. 2022, the companion article in this issue of *Computer Music Journal*). An inspiration for our approach is the concept of timbre space (Grey 1977; Wessel 1979; McAdams 1999), which proposes a multidimensional space of audio features as a perceptual model for timbre, or sound color, in which relative distances between sounds in each spatial dimension are comparable to dissimilarity judgments about relative timbre by listening subjects. Influenced by these advances from early AI research, we can achieve a strategy of sound synthesis based on the selection of sounds from a large collection of recorded audio, the "corpus," in which timbral features are foregrounded.

The resulting method, corpus-based concatenative synthesis (CBCS; cf. Schwarz 2006, 2007), is based on descriptor analysis of audio-content for

Figure 3. Signal and data flow between software and hardware components for both compositions: thin lines indicate data, marked with higher-order Ambisonic (HOA) icons for radiation data and relative specific loudness (RSL) for descriptor data; thick lines indicate audio, marked with microphones for live audio and wav icons for samples; thick dotted lines indicate HOA streams; and boldface text indicates ML models.



any number of preexisting or live-recorded sounds, and synthesis is guided by selection and playback of sound segments from the database, matching user-chosen sound characteristics. It has been used in various contexts of music composition (Einbond et al. 2009, 2016, 2021), live performance (Schwarz 2012), sound design, and installation (Savary et al. 2012). It allows exploration of a corpus of sounds interactively by composing paths in a multidimensional model of timbre space, and thus the creation of timbral evolutions while maintaining the richness and detail of the original sounds. The CBCS method can be seen as a content-based extension of granular synthesis, providing direct access to specific sound characteristics with perceptual control of the timbres of the played grains. At the same time, it recognizes limitations of machine listening as an objective model, offering instead flexibility for the human user to choose timbral characteristics based on subjective listening as an expressive artistic resource.

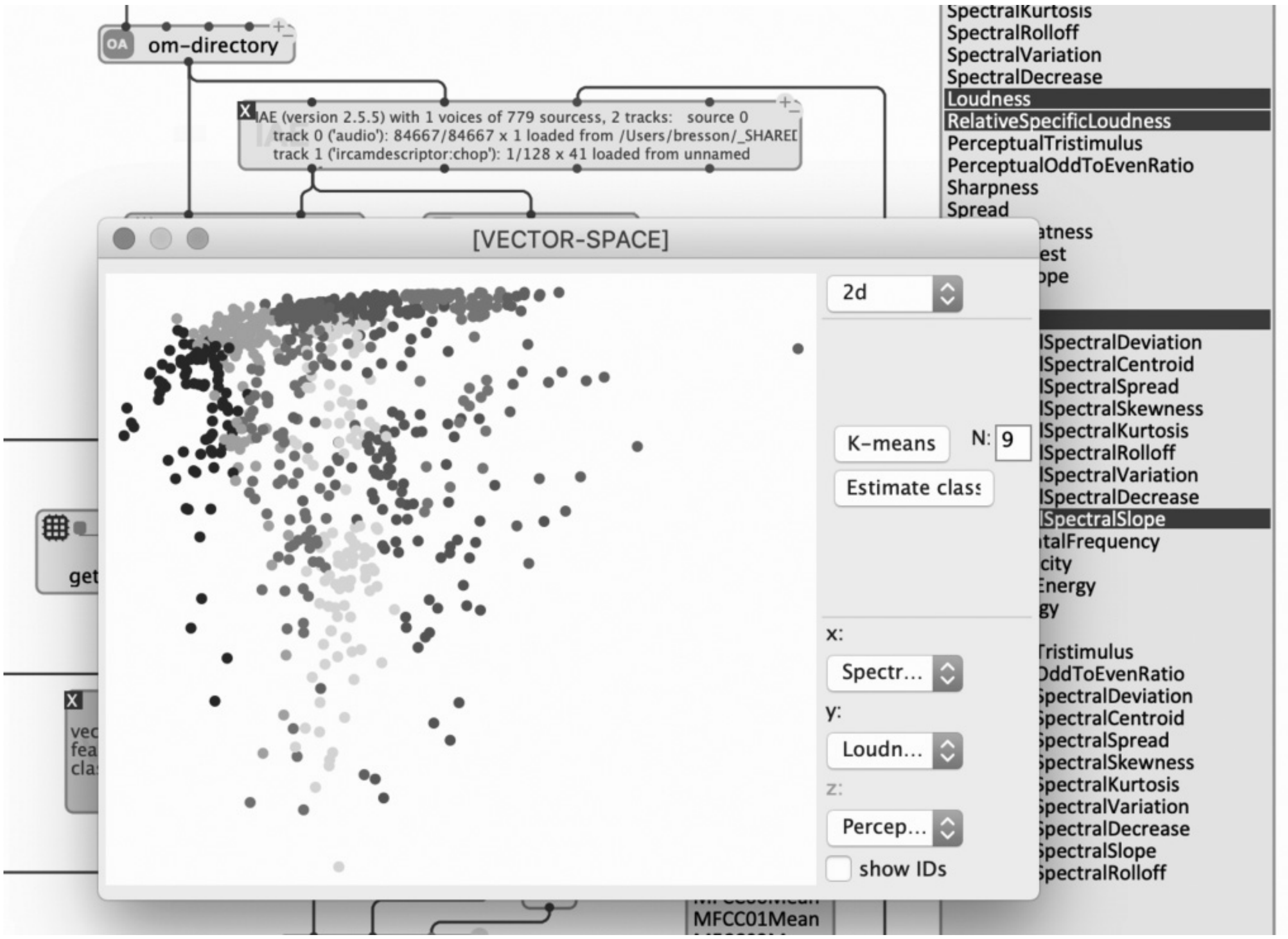
In the two compositions described here, CBCS is controlled by descriptor analysis of live audio from the performances of the instrumentalists, used to search for prerecorded sounds by a pattern recognition algorithm based on a multidimensional search tree: a k nearest neighbor (k-NN) query. In this way, the diffused sound samples follow the human player according to chosen timbral descriptors. This similarity judgment represents a simple form of human–AI interaction, in which the computer responds to the live performer with a sound-object recognition task. It can be referred to as live *audio mosaicking*, where many short samples are concatenated to reproduce the timbral features of a longer live performance.

Computer-Assisted Composition

We used the computer-assisted composition (CAC) software OM# to carry out preliminary experiments and prototype the AI models used in real time in the performance system.

OM# is a visual programming environment derived from OpenMusic (Bresson et al. 2017), offering close interactions with the IAE concatenative synthesis audio engine (named for the Sound Music Movement Interaction research group at IRCAM; cf. Schnell et al. 2012), the Spat framework, and the OM-AI library for ML applications (Vinjar and Bresson 2019), with associated data processing and visualization tools. It allowed us to streamline the development processes for the audio mosaicking

Figure 4. Clustering and visualization of audio feature vectors in OM# using the OM-AI tools. In the Vector-Space editor, each dot represents an audio segment (internally encoded as a vector of audio features). Identified clusters appear with different colors or grayscale values.

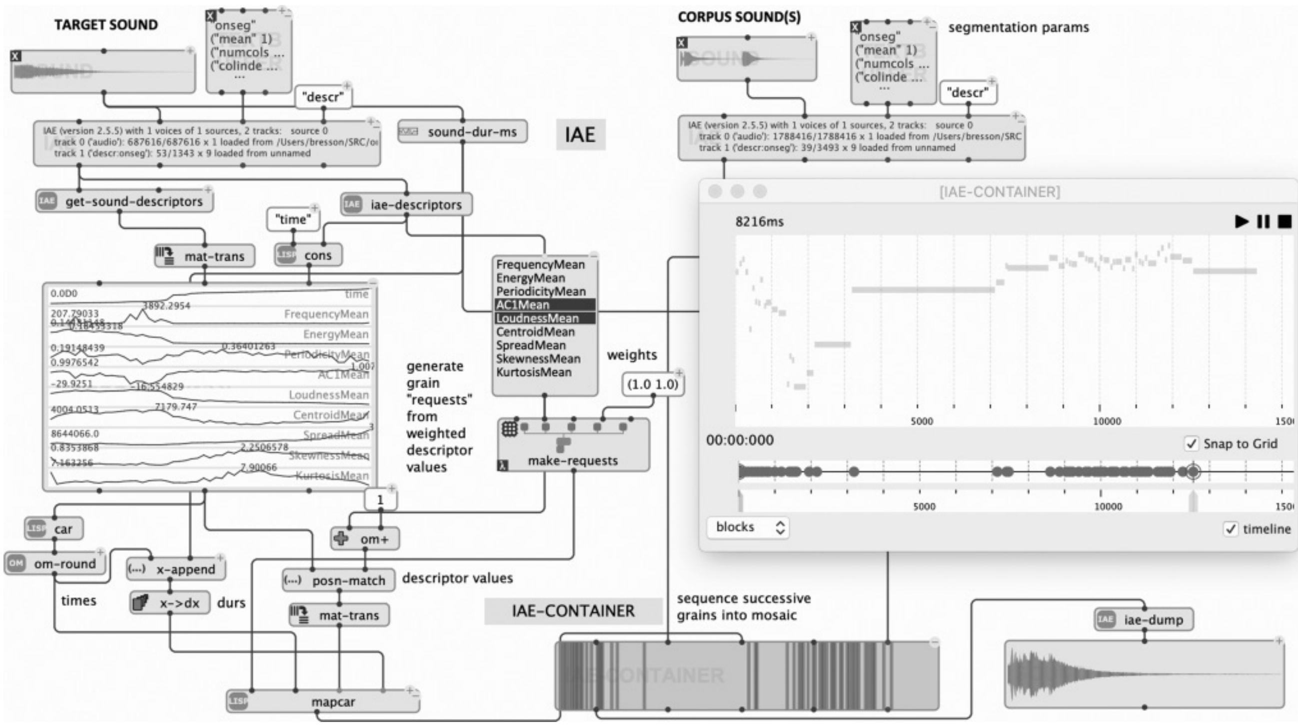


and spatialization techniques we employed, and to select the most effective descriptors for the machine learning models.

Selection of Audio Descriptors

In the machine-listening process, similarity is evaluated as proximity of audio descriptor values; and, as we argue above, our choice of descriptors has a decisive effect on the musical outcome. The IAE tools first enabled the extraction of audio features from the samples of a database of acoustic orchestral instruments produced by the TU Berlin (TUB; cf. Shabtai et al. 2017; Weinzierl et al. 2017) using the audio feature set from the IrcamDescriptors library (Peeters 2004). The extracted data was processed in OM# using OM-AI, which performs a basic *k*-means clustering algorithms on any dataset encoded as feature vectors and provides tools for visualizing the results (see Figure 4). By selecting different feature combinations on the 2- or 3-D axes of the graphical representation, we could compare desired features subjectively and test the classification against ground truth samples from the training dataset. The best results were obtained with mel-frequency cepstrum coefficients (MFCCs) and relative specific loudness (RSL, a multiband loudness curve comprising the loudness of a specified number of frequency bands, each normalized by the total loudness, cf. Peeters 2004). Intuitively, RSL can be

Figure 5. Audio mosaicking with OM# and IAE: audio segmentation and descriptor analysis of a target (top left); generation of descriptor queries (lower left); selection of the most similar grains from a corpus (top right); and visualization and synthesis of a new sequence from the corpus (lower left).



thought of as an equalization curve for the sound, capturing rich timbral information. It was chosen in favor of MFCCs for compatibility with the machine learning model for spatialization discussed below, in the Machine Learning of Radiation Patterns section.

Some aspects of the spatial rendering process were also prototyped using the Spat modules integrated into OM# (Garcia et al. 2016). To explore the possibility of spatial filtration, discussed further below in the Spatial Filtering section, synthesized sounds were filtered into different numbers of frequency bands, and each band was spatialized separately by application of an Ambisonic-encoded radiation pattern from the TUB database. This prototyping phase was performed iteratively for various inputs and parameter configurations. In the CAC environment it was also possible to simulate the multichannel diffusion setup with a binaural preview, enabling more-informed tuning of the spatial synthesis.

Audio Mosaicking

The scores of both compositions composed by Einbond were based on audio mosaics, similar to the live CBCS process, in which a "target" sound is imitated by concatenating small sound segments selected from the corpus (Einbond et al. 2009). Longer samples or improvisations from the acoustic instruments, as well as field recordings in "Cosmologies," were used as targets, and shorter instrumental samples constituted the corpus database: prepared piano performed by the composer for "Cosmologies" and percussion performed by Echardour for "Prestidigitation." The CAC prototypes produced with OM#-IAE and MuBu were used to evaluate different versions of the resulting mosaics (see Figure 5). IAE enables various options for automatic or parameterized segmentation of target sounds, as well as options for the generation of grains matching selected features of subsequent segments by similarity search in a corpus sound database. Processed offline and wrapped into the visual programming framework, such features offer unprecedented possibilities for the control and visualization of audio-mosaic structures, as well as the production of sound-file mosaics by concatenation of short grains. Thanks to the identical algorithms used by the IAE API and MuBu for Max, the two platforms can produce mutually informative results.

The generated audio mosaics were eventually converted to symbolic values, associating each concatenated sample with a timestamp and information about the performance techniques used to produce it. This information was then exported to music notation software for subjective edition and composition of the instrumental performance score, which is reinterpreted in live performance (see Figure 6).

Spatial Sound Synthesis

One of the novel features of our research is the use of data from natural acoustic phenomena as a model for ML and spatial sound synthesis. Significantly, the training dataset we used did not include piano and percussion instruments, so our approach is not expected to mimic the physical instruments on stage, but rather as a source of rich spatial information to synthesize novel interactive gestures.

Instrumental Radiation Data

Spatial information is derived from the TUB database of measured radiation patterns of 41 orchestral instruments, both modern and historical. The data include recordings of each note of each instrument or voice performed at two dynamic levels, recorded in anechoic conditions with a surrounding sphere of 32 microphones. From these measurements, the researchers estimated the spherical Fourier coefficients, that is, a compact representation of the radiation patterns encoded in the spherical-harmonic domain, and suitable for applications with higher-order Ambisonics (HOA; cf. Zotter and Frank 2019) up to fourth order. Radiation patterns are available for each of the first ten partials of each performed note, as well as for 31 third-octave frequency bands, obtained by averaging radiation data for all partials that fall within the corresponding narrow band (Shabtai et al. 2017). We used these third-octave band data for compatibility with the noise-rich piano and percussion sound material used in the compositions. Although we studied all 41 instruments and voice for testing and prototyping, only 21 modern instruments and voice were used for final ML training and realization of performance patches. These instruments were chosen for their larger pitch ranges, permitting a wider variety of radiation patterns available from each instrument. For a visualization of one of these patterns see Figure 7b.

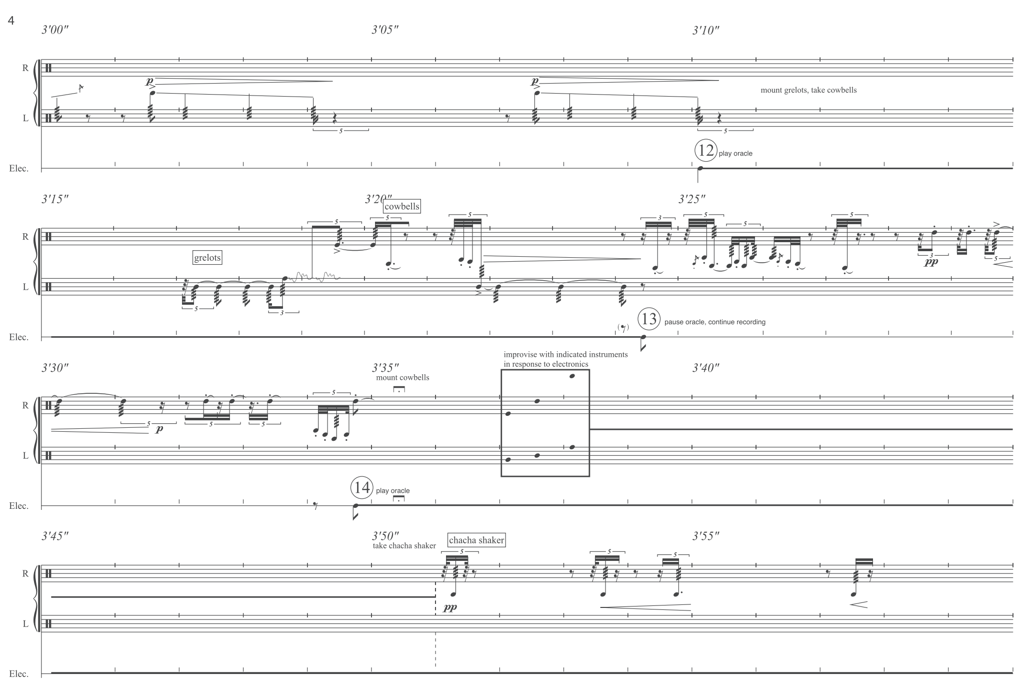
Machine Learning of Radiation Patterns

The TUB database was used to train an ML model to respond to a monophonic sound from live input with a predicted radiation pattern as follows: Monophonic files extracted from each instrumental sample in the TUB database were segmented into 100-msec units, and RSL descriptors (Peeters 2004) were calculated, with band limits corresponding to the 31 third-octave bands. As RSL is normalized by the total loudness of the unit, it provides a robust descriptor, independent of the varying dynamic range of the instruments in the corpus. We used the resulting database of 1,788 samples and their descriptor data to train a Gaussian mixture model (GMM, cf. Françoise et al. 2014) in a supervised classification task to predict the source instrument of an unknown monophonic input. The GMM model parameters were adjusted for a tightly fitting classification, with a relatively large number of Gaussian distributions (ten). Training was carried out offline and identical parameter settings for RSL were used to analyze the live audio input, facilitated by the software architecture of the MuBu for Max package.

To test the accuracy of the classification, we applied it to samples from the training set: in a representative test set of 72 samples (all concert A pitches for each instrument or voice), 66 samples (92 percent) were correctly matched for at least some of their 100-msec segments. Classification accuracy was not a primary artistic concern in the two compositions, however. Since there are no piano or percussion samples in the TUB database,

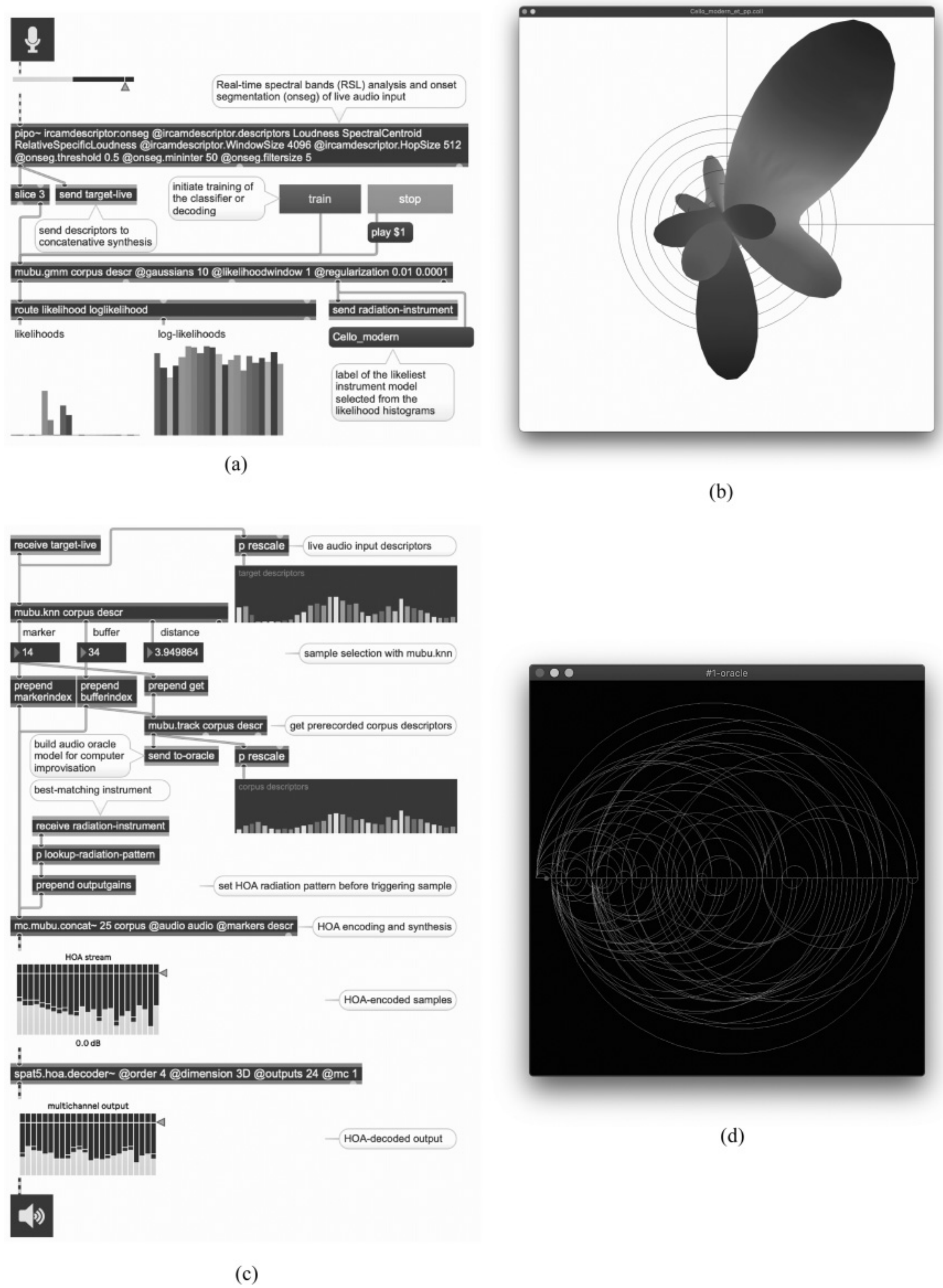
Figure 6. Excerpts from the performance scores of "Cosmologies," showing the audio mosaic of a field recording transcribed for prepared piano (a) and "Prestidigitation," with a transcription of recorded percussion improvisation (3′ 15″–3′ 35″), used to train computer improvisation, followed by live human improvisation (3′ 35″–3′ 50″) (b). (Copyright © Edition Gravis, Germany. Reprinted with kind permission.)

(a)



(b)

Figure 7. Images based on the concert patches showing descriptor analysis of the live audio input and instrument classification by `mubu.gmm` *(a), visualization of the resulting 3-D radiation pattern by* `spat5.hoa.plot` *(b), synthesis of the sample, encoded using higher-order Ambisonics (HOA), by* `mubu.concat~` *and HOA decoding for the loudspeaker output (c), and addition of this state to the audio oracle model for subsequent computer improvisation (d).*



no ground truth is available for comparison, and the radiation patterns applied to these samples were not intended to reproduce a real instrument. Subjectively, the classification nevertheless leads to plausible associations between sample timbres and radiation patterns. For example, a Superball drawn along the piano's bass strings was mapped to tuba and double bass, and high seashell chimes were mapped to oboe and violin.

In performance, once an instrument is identified based on the GMM classification, that instrument's radiation pattern is chosen with the best matching dynamic level (***pp*** or ***ff***) and the frequency band in which the input sound has the greatest loudness. Alternatively, to produce more variation in the output radiation pattern, we use the average of the n closest third-octave frequency bands weighted by loudness. Generally, for small values of n ($n \leq 3$), these bands are adjacent or nearby in frequency and therefore their radiation patterns are relatively similar. Consequently, the weighted average is a relatively small perturbation of the radiation pattern of the loudest band but nevertheless presents the musical advantage that output radiation patterns for different sounds are never identical. A further alternative is to filter the sound and spatialize different filtered bands independently, as discussed further below in the IKO section.

Ambisonic Encoding

As reported previously (Einbond et al. 2021), the MuBu for Max package is ideally structured to implement CBCS with HOA spatialization. The live audio input from the instrumental performer is analyzed using `pipo.ircamdescriptors~` and segmented using the onseg algorithm for onset detection. This module for live analysis is set to identical parameters used to analyze the prerecorded corpus with `mubu.process~` to facilitate matching between live and prerecorded sounds. The descriptors include RSL (as described in the Selection of Audio Descriptors section), loudness (used for sample segmentation), and spectral centroid (chosen subjectively for some audio mosaicking tasks instead of RSL). The descriptor values for each segment of the live input are sent to `mubu.knn` to choose the closest matching sample segment, to the Audio Oracle (AO; as described in the Factor and Audio Oracle section) to add the next state for computer improvisation, and then to `mubu.gmm` to generate the radiation pattern. The chosen monophonic sample is then encoded into an HOA stream with `mubu.concat~` by applying a list of output gains that we derive from the TUB database, effectively delivering a 25-channel fourth-order HOA stream, or 16 channels for third-order streams. MuBu's overlap-add algorithm allows an arbitrary number of sample segments to be superposed, each with its own radiation pattern, theoretically enabling unlimited spatial polyphony. See Figure 7 for screenshots of the steps of this signal and data flow.

Computer Improvisation

A further dimension of human–AI interaction is added in "Prestidigitation" to complement the GMM: A computer improvisation algorithm produces CBCS sequences generatively. This allows the computer to learn from a sequence of input sounds and respond with a continuation of the sequence, joining sounds that previously appeared in a similar context. This adds an important temporal dimension to CBCS and spatialization with GMM: These two processes are "in the moment," responding to input from the live performer to select and synthesize an audio segment, then moving on to the next segment with no memory of the preceding segment. With the addition of a generative ML model, the computer can record the sequence of segments that have been previously selected and synthesized, analyze this record for patterns, and output a generative continuation of these patterns. Even when the input from the live performer is suspended—either because of an extended silence or by deliberately closing the microphone input—the computer improvisation agent can continue its sequence indefinitely. This offers significant creative possibilities for human–AI interaction, allowing the computer not only to react to the live performer but also to produce a greater impression of agency in its generative improvisation.

Factor and Audio Oracle

Our approach builds upon previous research in computer improvisation as implemented in the OMax family of software tools (Assayag et al. 2006). Like OMax, we use the factor oracle (FO) aglorithm (Allauzen et al. 1999; Assayag and Dubnov 2004) and its extension to the audio oracle (AO) as implemented in the PyOracle library for Python (Surges and Dubnov 2013). These models are particularly suited to musical applications, owing to their capacity for real-time computation. Unlike algorithms that require offline training, the ML model is calculated incrementally in real time with the addition of each successive state. This is an advantage of these relatively simple early ML algorithms over more-recent, computationally intensive alternatives.

The audio oracle extends FO with the concept of information rate (IR) "to measure the amounts of complexity and repetition in the signal over time, and . . . to find the ideal AO model" (Surges and Dubnov 2013). The IR threshold value is the Euclidean distance in multidimensional descriptor space used to cluster audio segments into states of the oracle. A relatively short duration of simulated audio input (in our case, about one minute), similar to the audio input expected in the live performance, is used to calculate the ideal threshold. Once this value has been set, subsequent phases of AO learning and improvisation are carried out in real time. These phases may overlap, a possibility used in "Prestidigitation" to blur the boundaries between learning and improvisation and dovetail exchanges between human and computer improvisers. (Figure 6b showed a relevant score excerpt and Figure 7d displayed a visual representation of the AO.)

We have integrated AO with the CBCS workflow in our tool CatOracle (available at https://forum.ircam.fr/projects/detail/catart-mubu), based on MuBu for Max, and first introduced in Einbond's compositions "Xylography" and "Graphology" for violoncello, ensemble, and electronics, composed 2015–2016 (Einbond et al. 2016). CatOracle combines the advantages of AO for real-time pattern recognition and generative continuation with the flexibility and customizability of descriptor calculation and synthesis in CBCS. One of its distinctive features is access to the full list of audio descriptors available in the IrcamDescriptors library. This means that, unlike some tools that base computer improvisation primarily on pitch and duration, CatOracle permits training on a much wider range of sonic materials. As "Prestidigitation" already uses live analysis and classification with RSL descriptors for CBCS and spatialization, we use the same descriptors for AO learning and improvisation, and we could refer to the results as "computer noise improvisation." To implement CatOracle in "Prestidigitation," we updated the original PyOracle library from Python version 2.7 to version 3.11 and implemented communication between Max and Python using the Open Sound Control (OSC) protocol.

Ambisonic Diffusion

Both compositions rely on a system of Ambisonic audio spatialization to diffuse the 3-D gestures produced through live performance and interactive ML and situate them in the performance space with the audience. This process differs between the Ambisonic dome in the performance of "Cosmologies" and the IKO SLA used to perform "Prestidigitation." In both cases, subjective listening to the sounding result during rehearsals and performances was necessary to shape the spatial gestures in situ.

3-D Amplification

In both works, the output of the interactive electronics is complemented by 3-D amplification of the live acoustic instruments, captured by the em32 microphone, as was illustrated in Figures 1 and 2. The em32 was positioned upside down, oriented downwards toward the piano frame or frame drum. Vertical reflection was not applied to the HOA stream, such that the most resonant parts of the instruments were mapped to the zenith of the diffusion systems: in "Cosmologies" that was the center of the piano frame; in "Prestidigitation," the frame drum head. In both cases this required the HOA stream to be mirrored along the left–right axis to preserve the apparent orientation of the instruments (Kronlachner and Zotter 2014).

Human–AI interaction is triggered by a monophonic signal captured from the live instrument. In "Prestidigitation," separate cardioid condenser microphones are positioned in the piano case. In "Prestidigitation," however, to conserve limited space in the percussion setup, the omnidirectional channel (W) from the em32 HOA–encoded stream was used, despite reduced timbral quality and additional latency due to the encoding filters. This was judged acceptable for triggering purposes, however. In both compositions, the input signal from the em32 microphone is processed through encoding filters (Moreau, Daniel, and Bertet 2006) to produce a fourth-order HOA stream. This layer is then mixed with the synthesized interactive electronics stream, also encoded into HOA, and the combined layers are decoded together.

Spatial Paradigms

Comparing the projects (and following terminology introduced in Zotter 2009; Noisternig, Zotter, and Katz 2011), we can contrast *exterior problems* with *interior problems*. The former encompasses situations in which a central acoustical source is surrounded by peripheral listening positions; the latter covers situations in which acoustical sources surround a central listening position.

The data from the TUB represents an exterior problem, in which acoustical sources were located inside a surrounding sphere of 32 microphones, and their radiation projecting outward was recorded and analyzed. This situation is best suited to reproduction with an SLA such as the IKO, which is also positioned in the center of an acoustical environment and projects sound outward, as in "Prestidigitation." In contrast, the sound captured by the em32 represents an interior problem, in which acoustical sources are positioned outside the spherical microphone and the recorded and processed sound is projected inward. This is best suited for reproduction with an Ambisonic dome, as in "Cosmologies." Yet, both works take advantage of both sources of spatial information, intentionally "collapsing" the contrasting acoustical scenarios to artistic ends, and requiring additional technical adjustments to negotiate their differences.

Ambisonic Dome

In "Cosmologies," the fourth-order HOA streams produced from both ML spatialization and encoding the input of the em32 microphone are combined and diffused, implementing energy-preserving decoding (Zotter and Frank 2019) directly to the 27.2-channel Ambisonic dome installed for the work's first performance. Listening tests in the concert hall suggested subjectively that sound was perceived disproportionately from the back of the concert hall, however, pointing to the necessity to "warp" the HOA stream slightly toward the front of the Ambisonic dome (Kronlachner and Zotter 2014). This was due to the unusually steep layout of the concert meant that the Ambisonic dome had to be installed at an angle over the audience, as well as to our desire to point listeners' attention toward the live pianist positioned onstage in front of the dome.

Cosmologies I, II, III

In addition to "Cosmologies" Einbond used similar musical materials to produce two other modular "movements" that may be performed together or separately. "Cosmologies II" is an interactive sound installation that is intended to be performed before the other movements as the audience enters the concert hall, as realized before the first performances of "Cosmologies" and "Cosmologies III." The gains of the cardioid condenser microphones positioned in the instrument, or around the concert hall, are turned up to capture the ambient sounds of the audience and trigger short grains from the prepared piano corpus that are spatialized with the GMM model. The audience experiences the 3-D electronics while being free to move within the space where the piano is silent. This is in contrast to the live performance, during which the pianist is in motion and the audience is stationary. The audience joins in human–AI interaction both by triggering spatial synthesis directly with their incidental sounds and by reacting to these sounds by changing their perspective in the listening environment, similar to "Prestidigitation" (as discussed below).

In "Cosmologies III" for fixed 3-D electronics, created during a later residency in the Kubus at the Center for Art and Media Karlsruhe (ZKM), CBCS and GMM techniques were combined with a wider collection of em32 samples of prepared piano and field recordings. This work may be performed on its own or directly following "Cosmologies" as the live performer leaves the stage. Although the human–AI interaction in the fixed work is no longer live, it remains as a trace of the work's creation process, refracting the human performer's presence behind the spatial audio recordings. A binaural version was also prepared using `spat5.virtualspeakers~` based on virtualization of the 43.4-channel loudspeaker dome of the Kubus, allowing the listener to experience the 3-D electronics over headphones as intimate virtual chamber music. (A video is available at https://youtu.be/sooNxK6oQ4c?t=14300.)

IKO

The production of "Prestidigitation" required different processing steps for diffusion of the electronic output with the IKO. As the 20-channel IKO can only reproduce HOA sounds up to third order, all electronic layers were reduced to that maximum order, which also helped reduce the CPU load. A further significant CPU reduction was enabled by shortening the FIR filters to 1,024 taps instead of the version, provided by the IKO manufacturer, using 4,096 taps. We also enabled Max parallel processing to render the spatial grains concurrently. Furthermore, we experimented with filtering the CBCS output so different frequency bands could be spatialized separately, as well as focusing beams from the em32 signal to enhance directionality.

Spatial Filtering

In the simplest case of ML of spatial synthesis, a single radiation pattern is applied to each sample segment from the prerecorded corpus. Pursuing the analogy further to acoustic instruments, however, different spectral bands of each sample segment could be spatialized separately, just as different partials of each instrument in the TUB database have different frequency-dependent radiation patterns. Using OM# we carried out tests to filter each sample into different numbers of spectral bands. Although the TUB radiation data is available in 31 third-octave frequency bands, using such fine resolution did not produce musically convincing results, as the perceived change in the radiation pattern from sample to sample was limited. An effective compromise, however, was to filter each sample into three frequency bands using `spat5.complementarybank~` (Favrot and Faller 2010) and to spatialize each band separately according to the radiation pattern of the subband with the highest RSL value. The results were audibly different from diffusing the entire sample with one radiation pattern: Subjectively, spatial motion within three bands could be described as more complex and multifaceted, whereas spatialization with a single radiation pattern sounded more directional and abrupt. For initial performances of "Prestidigitation," however, spatial filtration was not implemented owing to the higher CPU load.

Beamforming

Although the interior problem of diffusing the HOA-encoded signal from the em32 microphone directly to the Ambisonic dome translated effectively with little additional treatment, we found that this was not the case with the IKO, for which changes of radiation pattern are not as perceptible by a listener at a fixed position in the performance space. One solution is to invite the listener to move with respect to the IKO, as discussed below in the Interactive Listening section. Another is to apply "beamforming" to the HOA stream diffused by the IKO, emphasizing the point in the 3-D field with the greatest intensity. This method was inspired by composer Natasha Barrett, who uses a similar technique to extract moving sources from fixed recordings of the em32 sound field (personal communication). We apply this idea in real time and mix the focused beam back in with the original em32 HOA stream, taking advantage of both the directional beam and the rich detail of the em32 field.

The focused beam is produced by steering a virtual microphone in a particular direction on an HOA-encoded signal (Rafaely 2019), with a hypercardioid microphone pattern chosen for maximum directivity. The object `spat5.hoa.intensity~` is used to estimate the direction of arrival (DOA), based on the instantaneous sound intensity of the em32 HOA stream (Merimaa and Pulkki 2005), and further provides a measure of the "diffuseness" of the sound field. For highly directive sound fields (i.e., those with low diffuseness) with one predominant source, the DOA estimation is relatively accurate and stable. The hypercardioid beam pattern is steered in the estimated DOA using `spat5.hoa.beam~` and the extracted monophonic signal is then reencoded in HOA and mixed with the HOA stream with a subjectively adjusted gain.

Directivity Database Browser

Our work with instrument radiation patterns led to the development of novel tools in the Spat5 package, including a new approach to computing the correlation between patterns (Carpentier and Einbond 2023). This was motivated by the desire to visualize and to navigate among radiation patterns for precompositional testing, with potential future applications to the ML calculations. The newly developed object `spat5.hoa.correlate` allows estimation of the correlation between two arbitrary radiation patterns and of the angle of rotation necessary to minimize this correlation. After calculating the pairwise correlations of a selection of radiation patterns, these patterns can be visualized in a low-dimensional space using the technique of multidimensional scaling (MDS). This technique is used to translate information about the pairwise distances (or dissimilarities) among a set of objects into an abstract Cartesian space, permitting visualization of the objects' level of similarity. Figure 8 shows an MDS plot of the 41 modern and historical instruments and voice of the TUB database for a single played pitch, B3. The visualization shows clustering of instruments by the expected radiation characteristic as proposed by Shabtai et al. (2017): instruments with one expected radiation point (such as brass), with several expected radiation points (such as woodwinds), and with full-body radiation (such as strings). Although visual navigation of the directivity database has so far been used only for precompositional exploration of subjective connections between radiation patterns, in the future it could be applied to ML algorithms themselves.

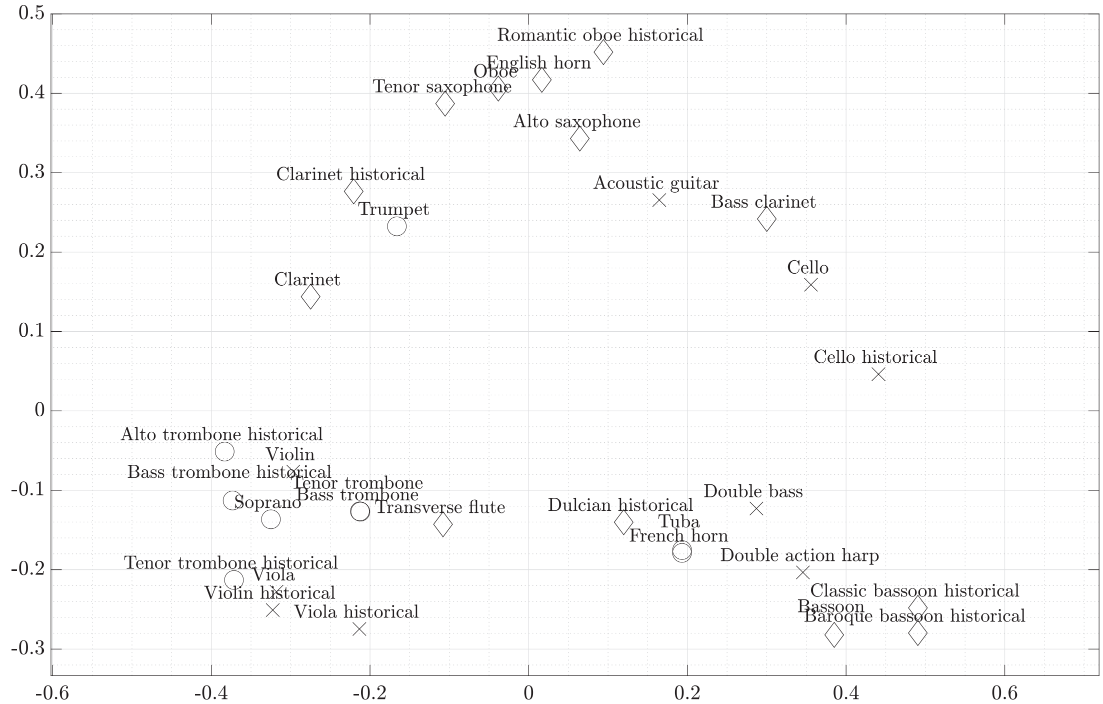
Discussion: Human–AI Interaction

As a central artistic goal in both compositions, the computer "learns" from the live performers' timbral and spatial gestures and reacts to complement the performers in a way that at times fuses or confuses them and at times underscores their difference. The performers, in turn, respond to the computer: both by shaping their expressive realizations of a notated score, which nevertheless leaves many details open to interpretation (for example, see Figure 6), and by elaborating the score with extended passages of guided improvisation. These passages inform the computer's further responses, shaping human–AI interaction in both directions. Significantly, this approach questions the assumption of a strict division between notation and improvisation, instead positioning the practices in a continuum that welcomes an interactive role by both performer and computer. The audience joins the multimodal interaction of human and AI as active listeners, who must intentionally engage by attending to the shifting 3-D relationships between live performer, computer, and the space in which they are all situated.

Live Audio Mosaicking

As introduced above in the Machine Listening and Corpus-Based Sound Synthesis section, audio mosaics with the k-NN algorithm rely upon a simple form of machine listening. When the target for the mosaic is the input from a live performer, this introduces a degree of human–AI interaction: The performer triggers a response from the computer and

Figure 8. Directivity database browser produced by mutidimensional scaling of dissimilarities from the TU Berlin database for played pitch B3; shapes indicate expected radiation characteristics: circle (◯) for brass-like, diamond (◇) for woodwind-like, and cross (×) for string-like.



listens to shape the performance further. Although the degree of interaction is basic, the results can still be musically stimulating, especially when the target and corpus samples are relatively long and more varied. For example, one passage of "Cosmologies" proceeds from a guided improvisation by the pianist with a Superball along the low piano strings and metal frame. The computer responds by choosing among relatively long samples (up to two seconds) that it estimates to be similar. Although the algorithm itself is not generative, small variations in the timbre and timing of the pianist's performance trigger unpredictable computer responses, which in turn influence the successive guided improvisation by the pianist, evolving in an expressive dialogue.

Human and Computer Improvisation

Although both scores include elements of guided improvisation, "Prestidigitation" takes this further by introducing extended computer improvisation through the AO algorithm. In response, the score for "Prestidigitation" incorporates an increased amount of freedom for the live performer to respond to the computer. These passages for human improvisation specify only their approximate lengths and the percussion instruments to be used, serving to smooth the transitions to and from fully notated passages (see Figure 6b). The notated passages themselves are based on a recorded improvisation by Echardour that Einbond transcribed subjectively through offline audio mosaicking. Echardour reinterprets these notated passages live, the AO learns from them and responds with computer improvisation, Echardour responds with live improvisation, which then affects the computer improvisation in a layered loop of performative feedback. Anecdotally, listeners to the first performances of "Prestidigitation" could not distinguish between notated and improvised passages, paralleling the work's goals of perceptual fusion in timbral and spatial dimensions as well. Although computer improvisation systems, such as Voyager by George Lewis (1999) and OMax, have long investigated collaboration between human and nonhuman improvisers, the unique permeability between notation and improvisation in "Prestidigitation" suggests novel compositional applications of ML.

AI–AI Interaction

The combination of GMM and AO in "Prestidigitation" presents powerful possibilities for continuation of spatial gestures through AI–AI interaction. In a sequence of sound segments synthesized by CBCS, each is associated with a spatial form derived from the TUB database. When this sequence is recorded by the AO, the sequence of spatial forms is also recorded. Subsequent computer improvisations by the AO reproduce patterns from this sequence and connect states that occur in similar contexts, (i.e., with a similar subsequence preceding or following). As a result, the AO's capability for pattern continuation is now applied to spatial patterns as well. The two communicating ML agents represent a form of AI–AI interaction that ultimately serves to enrich human–AI interaction: As the performer realizes a sequence of timbral-spatial gestures, the computer responds first by mapping the segments to spatial synthesis, then by improvising an extension of the sequence, resulting in a generative continuation of the spatial gestures.

Interactive Listening

Embodied listening is central to both compositions, which "Prestidigitation" explores further by situating computer sound synthesis in the midst of the performance space with the IKO. This means that the generative ML trained on instrumental radiation patterns, as well as the live performer, interact with the performance space in a different way than in the Ambisonic dome of "Cosmologies." This led Echardour and Einbond to invite the listeners to join in the interaction by standing and moving in the space around and between the IKO and live percussionist. Listeners explore the spatial presence and changing relationships of their bodies, the space, the percussion instruments, the performer, and the ML-generated spatial sound synthesis. This brings the listeners as active participants into the circle of human–AI interaction. Although we have received only limited responses from listeners so far, one personal account pointed to the performance's human scale—the similar size and height of the performer's body, the percussion setup, the IKO, and the listener's body—as a memorable feature of the experience (personal communication with Rémy Jannin).

Conclusions

We began with the artistic motivation to reembody electronic sound synthesis by drawing on the spatial presence of acoustic instruments and their performers, and we have demonstrated that human–AI interaction can play a decisive role in this process. Taken together, the two musical projects described here are, to our knowledge, the first to combine approaches to ML, CBCS, spatialization with HOA, and CAC. The human–AI interactions that pilot spatial sound synthesis and computer improvisation present a promising paradigm for dynamic and interactive control of electronic sound in an immersive performance.

Interaction is enhanced by mutual presence in a shared space, and each composition explores this possibility differently. In "Cosmologies," the space of the piano interior activated by the performer interacts both timbrally and spatially with the ML-informed process of spatial sound synthesis and the ambisonic dome surrounding the audience. In the interactive sound installation "Cosmologies II," the audience members take on the role of performers by triggering spatial synthesis with their own sounds and movements. In "Prestidigitation," interaction is taken further by the layers of generative improvisation by the computer and human performer, by the situated presence of the IKO in the shared space with the performer and listeners, and by the motion of the listeners who change their relationships to the other situated bodies in the room, in effect composing their own spatial listening experiences. In this light, we could argue that another implicit interaction is with the room, which mediates each of the other interactions. Although not the focus of this study, future work could examine the effects of situated presence on the performers and how it could enhance their engagement in human–AI interaction, as it does with listeners' engagement.

A further goal would be to use the directivity database visualization with MDS (see Figure 8) directly to train an ML model for spatial pattern regression. This would permit the ML algorithm to build models based directly on spatial data, enhancing the human–AI interaction between the timbrally rich live performance, training and improvisation of spatial gestures, and situated 3-D listening.

Acknowledgments

This work was supported by ERC advanced grant "MusAI—Music and Artificial Intelligence: Building Critical Interdisciplinary Studies" (European Research Council grant agreement no. 101019164, 2021–26) and by the Vertigo project as part of the Science, Technology, and the Arts (STARTS) program of the European Commission, based on technological elements from the project OM7/om-spat.

References

Allauzen, C., M. Crochemore, and M. Raffinot. 1999. "Factor Oracle: A New Structure for Pattern Matching." In *Proceedings of the Annual Conference on Current Trends in Theory and Practice of Informatics*, pp. 291–306.

Assayag, G., and S. Dubnov. 2004. "Using Factor Oracles for Machine Improvisation." *Soft Computing* 8(9):604–610.

Assayag, G., et al. 2006. "OMax Brothers: A Dynamic Topology of Agents for Improvization Learning." In *Proceedings of the ACM Workshop on Audio and Music Computing Multimedia*, pp. 125–132.

Bresson, J., et al. 2017. "Next-Generation Computer-Aided Composition Environment: A New Implementation of OpenMusic." In *Proceedings of the International Computer Music Conference*, pp. 253–258.

Carpentier, T. 2018. "A New Implementation of Spat in Max." In *Proceedings of the Sound and Music Computing Conference*, pp. 184–191.

Carpentier, T., and A. Einbond. 2023. "Spherical Correlation as a Similarity Measure for 3D Radiation Patterns of Musical Instruments." In *Acta Acustica 7*, 40. 10.1051/aacus/2023033

Einbond, A., D. Schwarz, and J. Bresson. 2009. "Corpus-Based Transcription as an Approach to the Compositional Control of Timbre." In *Proceedings of the International Computer Music Conference*, pp. 223–226.

Einbond, A., et al. 2016. "Introducing CatOracle: Corpus-based Concatenative Improvisation with the Audio Oracle Algorithm." In *Proceedings of the International Computer Music Conference*, pp. 141–147.

Einbond, A., et al. 2021. "Instrumental Radiation Patterns as Models for Corpus-based Spatial Sound Synthesis: 'Cosmologies' for Piano and 3D Electronics." In *Proceedings of the International Computer Music Conference*, pp. 148–153.

Favrot, A., and C. Faller. 2010. "Complementary N-Band IIR Filterbank Based on 2-Band Complementary Filters." In *Proceedings of the International Workshop on Acoustic Echo and Noise Control*. Available online at www.iwaenc.org/proceedings/2010/HTML/Uploads/975.pdf. Accessed November 2023.

Françoise, J., et al. 2014. "Probabilistic Models for Designing Motion and Sound Relationships." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 287–292.

Garcia, J., T. Carpentier, and J. Bresson. 2016. "Interactive-Compositional Authoring of Sound Spatialization." *Journal of New Music Research* 46(1):74–86. 10.1080/09298215.2016.1230632

Gioti, A.-M., A. Einbond, and G. Born. 2022. "Composing the Assemblage: Probing Aesthetic and Technical Dimensions of Artistic Creation with Machine Learning." *Computer Music Journal* 46(4):1–19.

Goeschke, F. 2022. "The iOSCahedron: Developing a Hybrid Spatialization Instrument." In *Proceedings of the International Audio Mostly Conference*, pp. 151–154.

Grey, J. M. 1977. "Multidimensional Perceptual Scaling of Musical Timbres." *The Journal of the Acoustical Society of America* 61(5):1270–1277. 10.1121/1.381428.

Hagan, K. L. 2017. "Textural Composition: Aesthetics, Techniques, and Spatialization for High-Density Loud-Speaker Arrays." *Computer Music Journal* 41(1):34–45. 10.1162/COMJ_a_00395

Hohl, F., and F. Zotter. 2010. "Similarity of Musical Instrument Radiation-Patterns in Pitch and Partial." In *Tagungsband der deutschen Arbeitsgemeinschaft für Akustik "Fortschritte der Akustik"*.

Kronlachner, M., and F. Zotter. 2014. "Spatial Transformations for the Enhancement of Ambisonic Recordings." In *Proceedings of the International Conference on Spatial Audio*, 5 pages.

Lewis, G. E. 1999. "Interacting with Latter-Day Musical Automata." *Contemporary Music Review* 18(3):99–112. 10.1080/07494469900640381

McAdams, S. 1999. "Perspectives on the Contribution of Timbre to Musical Structure." *Computer Music Journal* 23(3):85–102. 10.1162/014892699559797

Merimaa, J., and V. Pulkki. 2005. "Spatial Impulse Response Rendering I: Analysis and Synthesis." *Journal of the Audio Engineering Society* 53(12):1115–1127.

Meyer, J. 2009. *Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instruments Makers*. Springer, 5th edition.

Moreau, S., J. Daniel, and S. Bertet. 2006. "3D Sound Field Recording with Higher Order Ambisonics: Objective Measurements and Validation of Spherical Microphone." In *Proceedings of the 120th Convention of the Audio Engineering Society*, 24 pages.

Noisternig, M., F. Zotter, and B. F. G. Katz. 2011. "Reconstructing Sound Source Directivity in Virtual Acoustic Environments." In Y. Suzuki, D. Brungart, and H. Kato, eds. *Principles and Applications of Spatial Hearing*. World Scientific, pp. 357–373.

Normandeau, R. 2009. "Timbre Spatialisation: The Medium Is the Space." *Organised Sound* 14(3):277–285. 10.1017/S1355771809990094.

Peeters, G. 2004. "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the Cuidado Project." CUIDADO Project Report No. 54. Paris, IRCAM.

Rafaely, B. 2019. *Fundamentals of Spherical Array Processing*. Berlin: Springer, 2nd ed.

Savary, M., D. Schwarz, and D. Pellerin. 2012. "DIRTI: Dirty Tangible Interfaces." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 347–350.

Schmeder, A. 2009. "An Exploration of Design Parameters for Human-Interactive Systems with Compact Spherical Loudspeaker Arrays." In *Proceedings of the Ambisonics Symposium*. Available online at ambisonics-symposium.org/symposium2009/proceedings/ambisym09-schmeder-csphlsinteraction.pdf. Accessed November 2023.

Schnell, N., et al. 2009. "MuBu and Friends: Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP." In *Proceedings of the International Computer Music Conference*, pp. 423–426.

Schnell, N., et al. 2012. "IAE and IAEOU." In R. Cahen, ed. *Topophonie Research Project: Audiographic Cluster Navigation (2009–2012)*. Paris: École Nationale Supérieure de Création Industrielle, pp. 50–51.

Schwarz, D. 2006. "Concatenative Sound Synthesis: The Early Years." *Journal of New Music Research* 35(1): 3–22. 10.1080/09298210600696857

Schwarz, D. 2007. "Corpus-Based Concatenative Synthesis." *IEEE Signal Processing Magazine* 24(2):92–104. 10.1109/MSP.2007.323274

Schwarz, D. 2012. "The Sound Space as Musical Instrument: Playing Corpus-Based Concatenative Synthesis." In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 10.5281/zenodo.1180593.

Shabtai, N. R., et al. 2017. "Generation and Analysis of an Acoustic Radiation Pattern Database for Forty-One Musical Instruments." *Journal of the Acoustical Society of America* 141(2):1246–1256. 10.1121/1.4976071

Surges, G., and S. Dubnov. 2013. "Feature Selection and Composition Using PyOracle." In *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference*, pp. 114–121.

Vinjar, A., and J. Bresson. 2019. "OM-AI: A Toolkit to Support AI-Based Computer-Assisted Composition Workflows in OpenMusic." In *Proceedings of the Sound and Music Computing Conference*, pp. 84–85. 10.5281/zenodo.3249264.

Weinzierl, S., et al. 2017. "A Database of Anechoic Microphone Array Measurements of Musical Instruments: Recordings, Directivities, and Audio Features." Technical report. TU Berlin. 10.14279/depositonce-5861.2.

Wessel, D. 1979. "Timbre Space as a Musical Control Structure." *Computer Music Journal* 3(2):45–52. 10.2307/3680283

Zotter, F. 2009. "Analysis and Synthesis of Sound-Radiation with Spherical Arrays." PhD dissertation, IEM Institute of Electronic Music and Acoustics, Graz, Austria.

Zotter, F., and M. Frank. 2019. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Berlin: Springer.

Zotter, F., et al. 2017. "A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker." *Computer Music Journal* 41(3):50–68. 10.1162/comj_a_00429