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

Aaron Einbond,[1] Thibaut Carpentier,[2] Diemo Schwarz,[2] and Jean Bresson[2,3]

[1] Department of Performing Arts, City, University of London
Northampton Square, London EC1V 0HB, UK
Aaron.Einbond@city.ac.uk

[2] STMS Lab – IRCAM, CNRS, Sorbonne Université, Ministère de la Culture
1, place Igor Stravinsky, 75004 Paris, France
{thibaut.carpentier, diemo.schwarz, jean.bresson}@ircam.fr

[3] Ableton AG
Schönhauser Allee 6-7, 10119 Berlin, Germany

Abstract

The situated spatial presence of musical instruments has been well studied in the fields of acoustics and music perception research, but so far has not been the focus of Human-AI interaction. We respond critically to this trend by seeking to "re-embody" 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 where 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 3D 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 improvisation models and trigger corpus-based spatial synthesis. We show how AI techniques are useful to leverage complex, multidimensional, spatial data in the context of computer-assisted composition and human-computer interactive improvisation.

«BEGIN ARTICLE»

How can one situate the listener inside a virtual musical instrument, and how can the interaction of human and artificial intelligence (AI) help to realize this goal? To answer 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 three-dimensional (3D) 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.

Artificial intelligence (AI) 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 so far 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 and in mono, stereo, or multichannel. We

respond critically to this omission, situating the performer and listener at the center of human-AI interaction by attempting to "re-embody" 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 data set for a machine learning (ML) model that then responds live to the human performer, synthesizing each sound with its own detailed 3D radiation pattern. We build on our previous work in which we applied this approach to diffusion with an ambisonic loudspeaker dome (Einbond et al. 2021), 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 realtime, 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 realtime response to the live performer.

Musical Motivation

Our approach to human-AI interaction was motivated by two musical compositions by Aaron Einbond that draw upon natural acoustic phenomena as source material for

spatial sound. One source is a 3D microphone array, the mh acoustics Eigenmike 32-channel microphone array (EM32: 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 3D electronics began with Einbond's proposal to situate the listener inside a larger-than-life virtual grand piano to experience its secret inner life. In order 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 mallet (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 micro-choreography is magnified and projected to the listener with a surrounding Ambisonic dome (27.2 channels for the premiere performance by pianist Alvise Sinivia, produced by IRCAM in

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

the Grande Salle of Centre Georges Pompidou, Paris.) A video and binaural recording of the first performance can be viewed at the following link: https://youtu.be/jKlWLwPrun4.

Percussion

Prestidigitation for percussion and 3D electronics was motivated by a related idea to place the listener virtually in the midst of a sculptural percussion setup to hear sonic details normally only audible to the percussionist. Working in close collaboration, Einbond and performer Maxime Echardour constructed a purpose-built frame and suspended small percussion instruments surrounding the EM32 microphone. They together 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" his 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). However, unlike the larger-than-life Ambisonic dome of *Cosmologies*, a contrasting diffusion system was chosen to fit the more intimate scale of the percussion setup: a compact spherical loudspeaker

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

array (SLA), the IKO icosahedral loudspeaker (Zotter et al. 2017). This permits the projection of spatial sounds and gestures captured by the EM32 to a diffusion system of similar physical size and radiation characteristics to the percussion instruments themselves. Like an acoustic instrument, the IKO is situated in the midst 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 on the real instruments and their projected "doubles." A video and binaural recording of *Prestidigitation* can be viewed at the following link:
«TYPE: Please insert URL at MIT Press of *Prestidigitation* in the vol. 46 Sound Anthology».

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 and learning agents that can be employed together or separately. One reacts to the live performance with electronic sounds whose 3D radiation patterns are learned and reproduced from acoustic instruments. Another learns from the sequence of the performer's timbral and spatial gestures and 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 sounds' spatial characteristics. While the ML algorithms used are not themselves new, we argue that their combined creative application leads to novel possibilities for interactive and generative spatial sound synthesis. At the same time, the choice of relatively simple algorithms, as compared to more recent computationally intensive ones, presents advantages for interactivity in terms of relatively small training sets and fast—or realtime—training. We implement these models using software tools connecting for the first time computer programs Max, Python, and OM# with 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 audio similarity measurement and classification. However, machine listening is not the same as human listening, and requires subjective human input: in particular, what is measured by "similarity" is not a neutral decision, but instead "aesthetically situated" (see companion article by Gioti, Einbond, and Born in this volume). 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

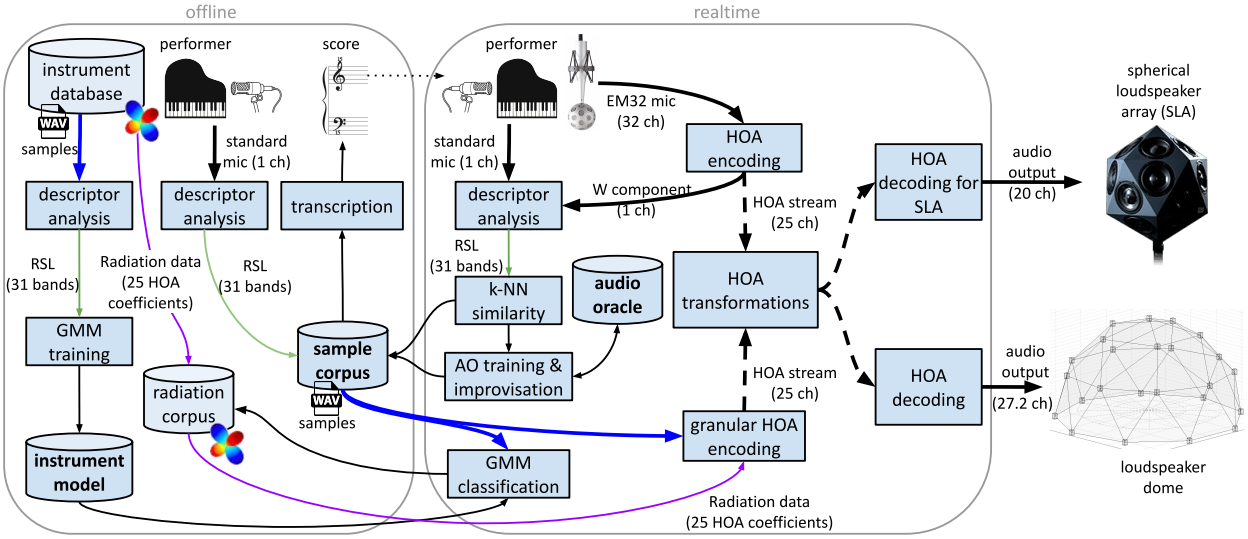


Figure 3. Signal and data flow between software and hardware components for both compositions: thin lines indicate data, marked with HOA icons for radiation data and RSL for descriptor data; thick lines indicate audio, marked with microphones for live audio and wav icons for samples; thick dotted lines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

perceptual model for timbre, or sound color, in which relative distances between sounds in each spatial dimension are comparable to relative timbral dissimilarity judgements 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) (Schwarz 2006, 2007), is based on audio content descriptor analysis of any number of pre-existing 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 multi-dimensional timbre space model, and thus to create timbral evolutions while maintaining the richness and detail of the original sounds. CBCS 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, instead offering 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 multi-dimensional 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 judgement 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 computer-assisted composition (CAC) software OM# to carry out preliminary experiments and prototype the AI models used in realtime 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 (Schnell et al. 2012), the Spat framework, and the OMAI 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 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 Technische Universität (TU) Berlin (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 OMAI, 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 2D or 3D 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 multi-band loudness curve comprising the loudness of a specified number of frequency bands, each normalized by the total loudness) (Peeters 2004). RSL can be thought of intuitively as an EQ-curve for the sound, capturing rich timbral information. It was chosen in favor of MFCCs for compatibility with the ML spatialization model discussed below.

Some aspects of the spatial rendering process were also prototyped using the Spat modules integrated in OM# (Garcia et al. 2016). To explore the possibility of spatial filtration, discussed further below, synthesized sounds were filtered into different numbers of frequency bands and each band was spatialized separately by application of Ambisonic-encoded radiation pattern from the TU 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 spatial synthesis.

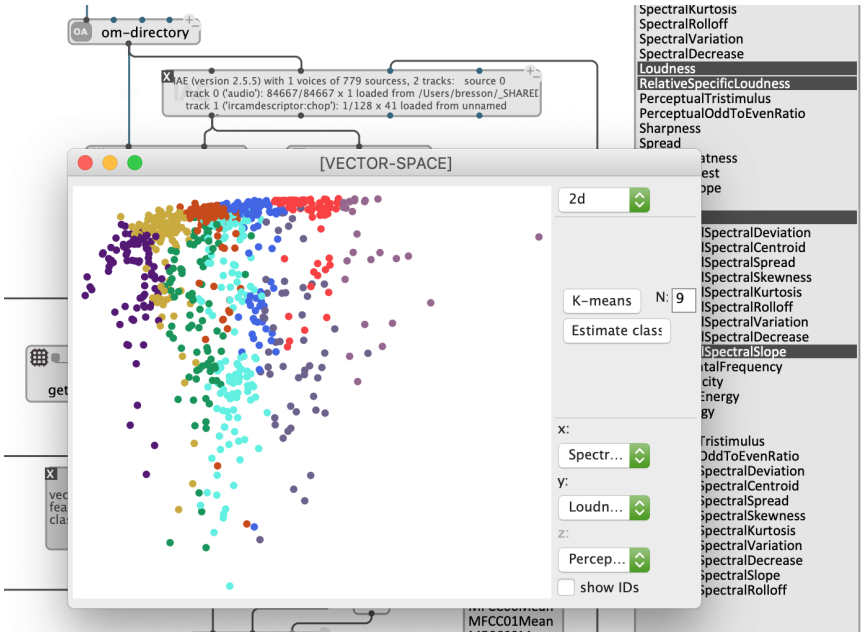


Figure 4. Clustering and visualization of audio feature vectors in OM# using the OMAI tools. In the vector-space *editor, each dot represents an audio segment (internally encoded as a vector of audio features). Identified clusters appear with different colors or grayscale values.*

Audio Mosaicking

Einbond composed the scores of both compositions 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, and 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.

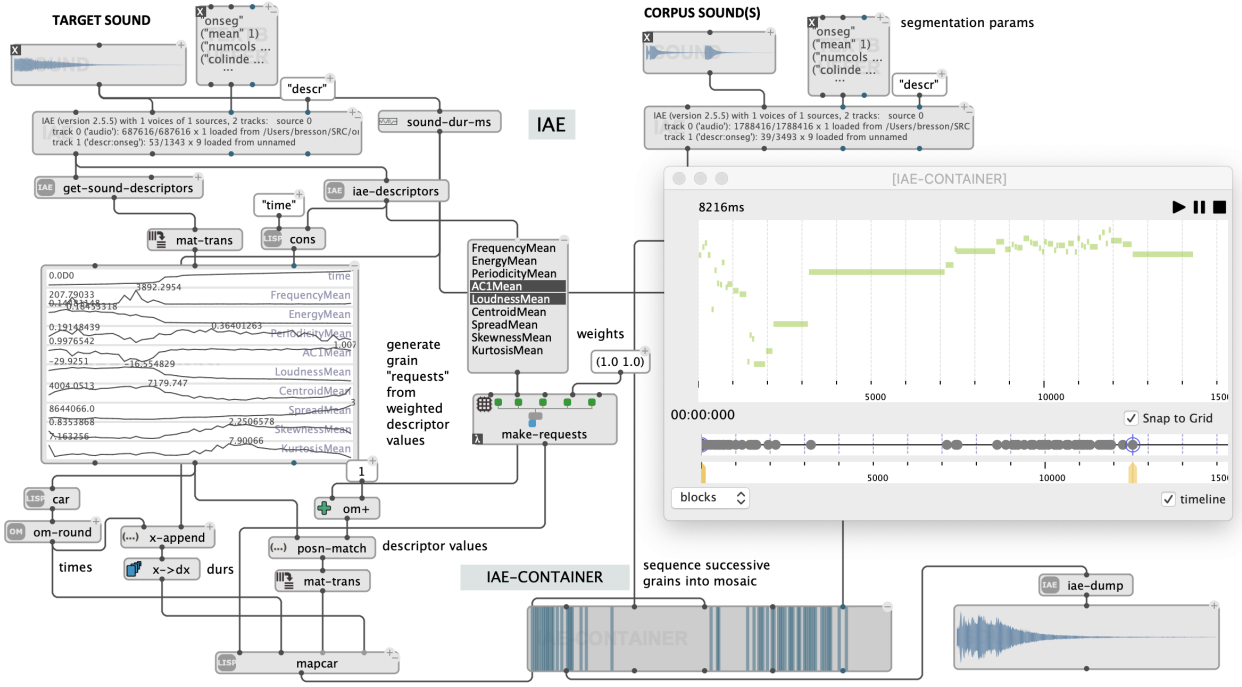
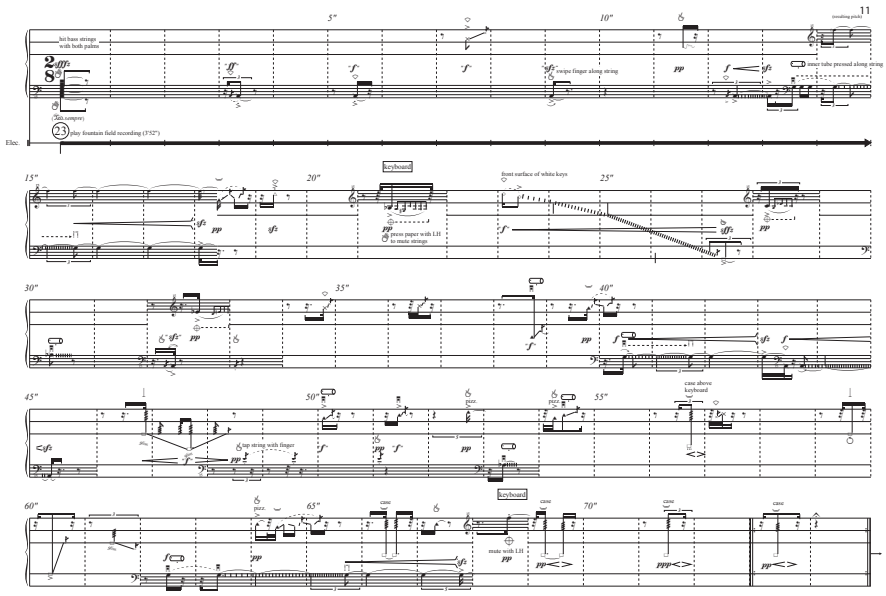


Figure 5. Audio mosaicking with OM# and IAE: at the top left, audio segmentation and descriptor analysis of a target; at the lower left, generation of descriptor queries; at the top right, selection of the most similar grains from a corpus; and at the lower right, visualization and synthesis of a new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

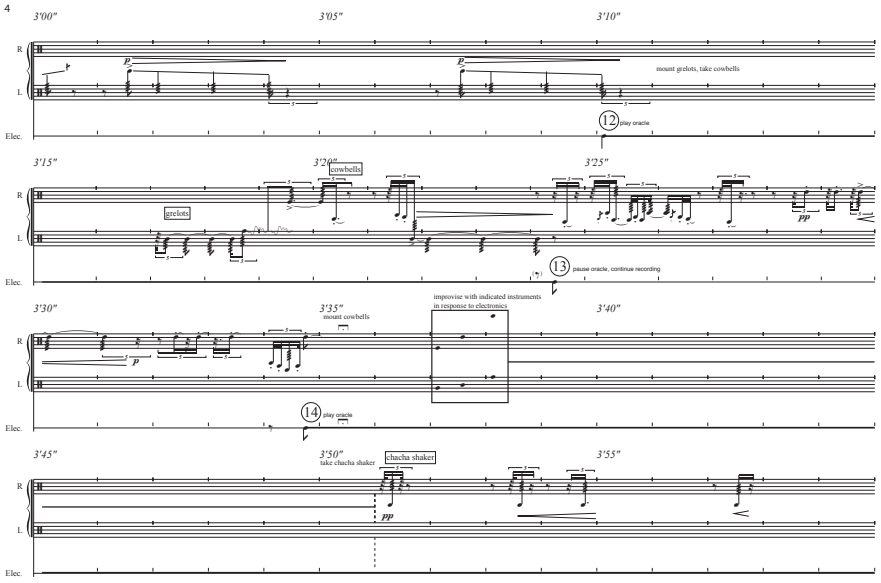
The generated audio mosaics were eventually converted to symbolic values associating each concatenated sample with a time stamp and information about the performance techniques used to produce it. This information was then exported to music notation software for subjective edition and composition of 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



(a) audio-mosaic of a field recording transcribed for prepared piano



(b) transcription of a recorded percussion improvisation (systems 2-3), used to train computer improvisation, and followed by live human improvisation (systems 3-4)

Figure 6. Excerpts from the performance scores of (a) Cosmologies *and (b)* Prestidigitation.

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 a database of measured radiation patterns of 41 orchestral instruments (modern and historical) and voice published and made publicly available by TU Berlin (Shabtai et al. 2017; Weinzierl et al. 2017). 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, i.e. a compact representation of the radiation patterns encoded in the spherical harmonic (SH) domain, and suitable for applications with Higher Order Ambisonics (HOA) (Zotter and Frank 2019) up to 4^{th} order. Radiation patterns are available for each of the first 10 partials of each performed note, as well as for 31 one-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 7c.

ML of Radiation Patterns

The TU database was used to train an ML model to respond to a mono sound from a live input with a predicted radiation pattern, as follows: mono files extracted from each instrumental sample in the TU database were segmented into 100 ms 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 1788 samples and their descriptor data to train a Gaussian Mixture Model (GMM) (Françoise et al. 2014) in a supervised classification task to predict the source instrument of an unknown mono input. The GMM model parameters were adjusted for a tightly fitting classification, with a relatively large number of gaussians (10). 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 (Schnell et al. 2009).

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%) were correctly matched for at least some of their 100 millisecond segments. However, classification accuracy was not a primary artistic concern in the two compositions: as there are no piano or percussion samples in the TU database, no ground truth is available for comparison, and the radiation patterns applied to these samples were not intended to reproduce a real instrument. Nevertheless, subjectively, the classification leads to plausible associations between sample timbres and radiation patterns: for example, a superball drawn along the piano's bass strings is mapped to tuba and double bass, or high seashell chimes 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, in order 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 context of the IKO.

Ambisonic Encoding

As we have 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* onset detection algorithm. This live analysis module 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 defined above, *loudness* which is used for sample segmentation, and *spectral centroid* which was 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 AO to add the next state for computer improvisation, and then to mubu.gmm to generate the radiation pattern. The chosen mono sample is then encoded into an HOA stream with mubu.concat~ by applying a list of *outputgains* that we derive from the TU database, effectively delivering a 25-channel 4$^{\text{th}}$-order HOA stream, or 16 channels for 3$^{\text{rd}}$-order. MuBu's overlap-add algorithm allows an arbitrary number of sample segments to be superposed, each with its own radiation pattern, enabling theoretically 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. It 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, and 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 due to 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

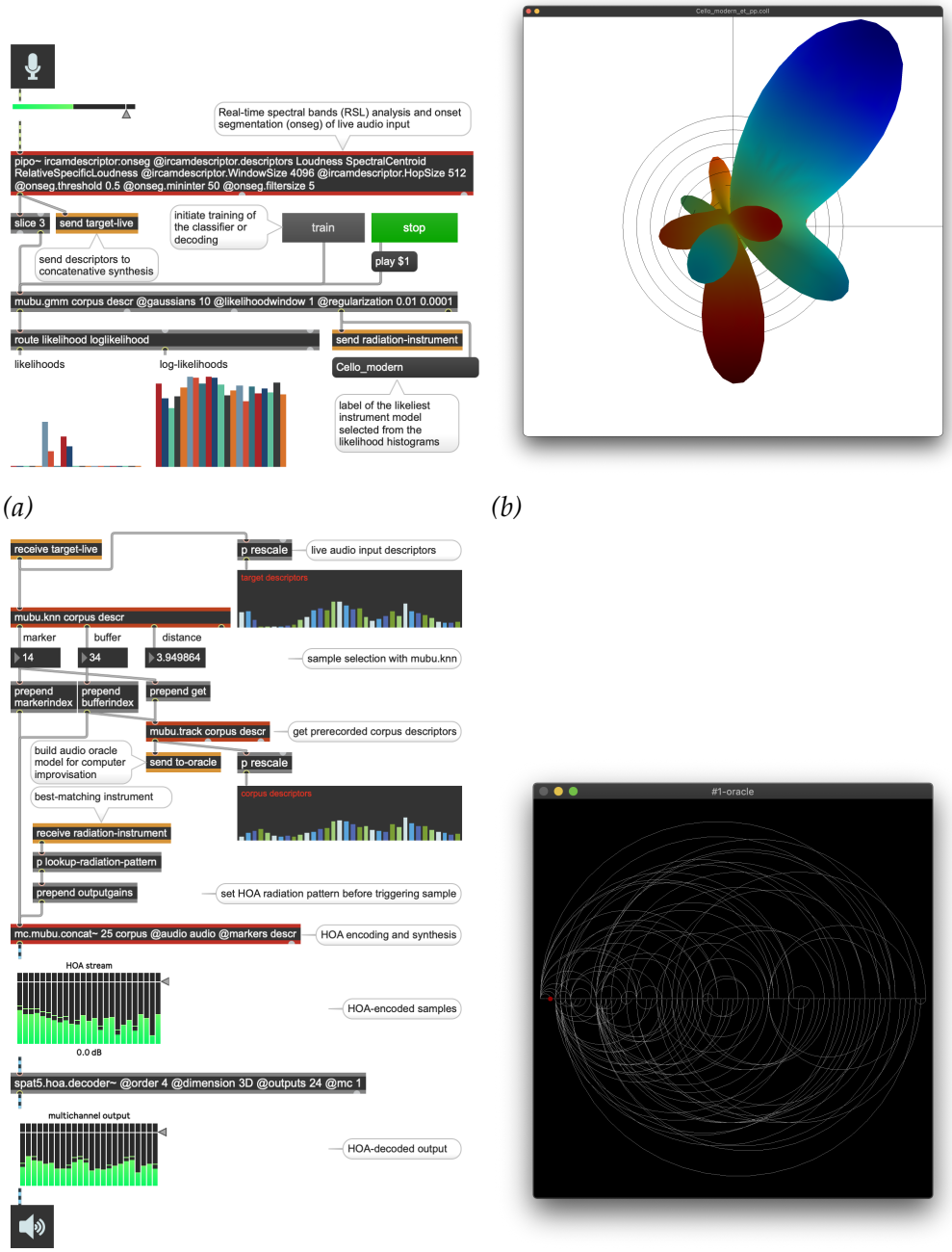


Figure 7. Images based on the concert patches showing (a) descriptor analysis of the live audio input and instrument classification by mubu.gmm*, (b) visualization of the resulting 3D radiation pattern by* spat5.hoa.plot*, (c) synthesis of the HOA-encoded sample by* mubu.concat~ *and HOA decoding for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

applications due to their capacity for realtime computation: unlike algorithms that require offline training, the ML model is calculated incrementally in realtime with the addition of each successive *state*. This is an advantage of these relatively simple early ML algorithms over more recent, computationally intensive, alternatives.

AO 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 ca. 1 min) similar to the audio input expected in the live performance is used to calculate the ideal threshold, and once this value has been set, subsequent phases of AO learning and improvisation are carried out in realtime. 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. (See Figures 6b for a relevant score excerpt and 7a for 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 in 2015-16 (Einbond et al. 2016). *CatOracle* combines the advantages of AO for realtime 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 2.7 to 3.11 and implemented communication between Max and Python via the Open Sound Control (OSC) protocol (Wright 2005)

Ambisonic Diffusion

Both compositions rely on a system of Ambisonic audio spatialization to diffuse the 3D 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*.

3D Amplification

In both works, the output of the interactive electronics is complemented by 3D amplification of the live acoustic instruments, captured by the EM32 as illustrated in Figures 1-2. 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* the center of the piano frame, and 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 *Cosmologies*, separate cardioid condenser microphones are positioned in the piano case. However, in *Prestidigitation*, 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. However, this was judged acceptable for triggering purposes. In both compositions, the input signal from the EM32 microphone is processed through encoding filters (Moreau et al. 2006) to produce a 4th-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, we can contrast *interior* problems, or situations where acoustical sources surround a central listening position, with *exterior* problems, or situations where a central acoustical source is surrounded by peripheral listening positions (Zotter 2009; Noisternig, Zotter, and Katz 2011). The data from the TU represents an exterior problem, where 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, where 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 4th-order HOA streams output from both ML spatialization and from encoding the EM32 microphone input 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. However, listening tests in the concert hall suggested subjectively that sound was perceived disproportionately from the back of the concert hall, 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 hall, which meant that the Ambisonic dome had to be installed at an angle over the audience, as well as 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 3D electronics while free to move within the space where the piano is silent, in contrast to the live performance where 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 3D electronics, created during a later residency in the ZKM Kubus, 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 3D electronics over headphones as intimate virtual chamber music (listen at the following link: 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 3rd-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 1024-tap length instead of the 4096-tap version provided by the IKO manufacturer, as well as by enabling 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. However, further pursuing the analogy to acoustic instruments, different spectral bands of each sample segment could be spatialized separately, just as different partials of each instrument in the TU database have different frequency-dependent radiation patterns. Using OM# we carried out tests to filter each sample into different numbers of spectral bands. While the TU radiation data is

available in 31 third-octave frequency bands, using such fine resolution did not produce musically convincing results, as there was limited perceived change in the radiation pattern from sample to sample. However, an effective compromise was to filter each sample into three frequency bands using spat5.complementarybank~ (Favrot and Faller 2010) and spatialize each band separately according to the radiation pattern of the sub-band 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, while spatialization with a single radiation pattern sounded more directional and abrupt. However, for initial performances of *Prestidigitation*, spatial filtration was not implemented due to a higher CPU load.

Beamforming

While the interior problem of diffusing the HOA-encoded signal from the EM32 directly to the Ambisonic dome translates effectively with little additional treatment, we found that this was not the case for 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. Another is to apply "beamforming" to the HOA stream diffused by the IKO, emphasizing the point in the 3D 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 realtime, 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 hyper-cardioid 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. with low diffuseness) with one predominant source, the DOA estimation is relatively accurate and stable. The hyper-cardioid beam pattern is steered in the estimated DOA using spat5.hoa.beam~ and the extracted mono signal is then re-encoded 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 2022). This was motivated by the desire to visualize and navigate among radiation patterns for pre-compositional testing, with potential future applications to the ML calculations themselves. 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). MDS 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 TU database for a single played pitch, B3. The visualization shows clustering of instruments by 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. While visual navigation of the directivity database has so far

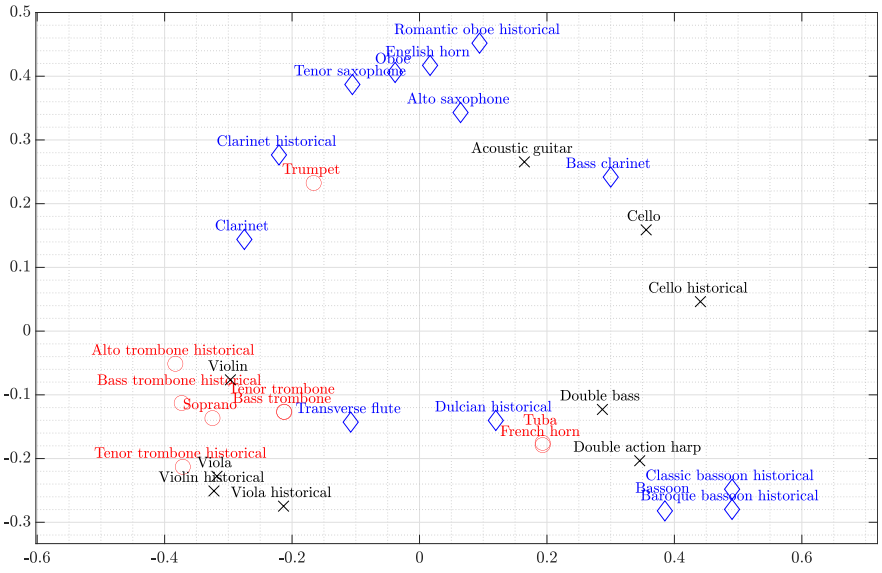


Figure 8. "Directivity database browser" produced by mutidimensional scaling of dissimilarities from the TU Berlin database for played pitch B3; shapes indicate expected radiation characteristics: circle for brass-like, diamond for woodwind-like, and x for string-like.

been used only for pre-compositional 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 underlines their difference. The performers, in turn, respond to the computer: both by shaping their expressive realizations of a notated score, that 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 3D relationships between live performer, computer, and the space in which they are all situated.

Live Audio Mosaicking

As introduced above, audio mosaics with the kNN 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 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 (up to 2 sec) samples that it estimates to be similar. While 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

While 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. While computer improvisation systems such as George Lewis's *Voyager* (Lewis 1999) and OMax have long investigated collaboration between human and non-human 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 spatial gesture continuation through AI-AI interaction. In a sequence of sound segments synthesized by CBCS, each is associated with a spatial form derived from the TU 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 sub-sequence 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. While we have received limited responses so far from listeners, 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 (Rémy Jannin, personal communication).

Conclusions

We began with the artistic motivation to "re-embody" 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 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 listeners'.

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 3D listening.

Acknowledgements

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 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 SOFSEM'99*, pp. 291–306.

Assayag, G., G. Bloch, M. Chemillier, B. M. Juin, A. Cont, and S. Dubnov. 2006. "OMax brothers: a dynamic topology of agents for improvization learning." In *Proc. of the 1st ACM workshop on Audio and music computing multimedia*, pp. 125 – 132.

Assayag, G., and S. Dubnov. 2004. "Using Factor Oracles for Machine Improvisation." In *Soft Computing*, vol. 8. Springer, pp. 1–7.

Bresson, J., D. Bouche, T. Carpentier, D. Schwarz, and J. Garcia. 2017. "Next-generation Computer-aided Composition Environment: A New Implementation of OpenMusic." In *Proc. of the International Computer Music Conference*.

Carpentier, T. 2018. "A new implementation of Spat in Max." In *Proc. of the 15th Sound and Music Computing Conference (SMC)*, pp. 184 – 191.

Carpentier, T., and A. Einbond. 2022. "Spherical correlation as a similarity measure for 3D radiation patterns of musical instruments." In *Proc. of the 16th Congrès Français d'Acoustique (CFA)*, pp. 1 – 6.

Einbond, A., J. Bresson, D. Schwarz, and T. Carpentier. 2021. "Instrumental Radiation Patterns as Models for Corpus-Based Spatial Sound Synthesis: *Cosmologies* for Piano and 3D Electronics." In *Proc. of the International Computer Music Conference*, pp. 148 – 153.

Einbond, A., D. Schwarz, R. Borghesi, and N. Schnell. 2016. "Introducing CatOracle: Corpus-based Concatenative Improvisation with the Audio Oracle Algorithm." In *Proc. of the International Computer Music Conference*, pp. 141 – 147.

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

Favrot, A., and C. Faller. 2010. "Complementary N-Band IIR Filterbank Based on 2-Band Complementary Filters." In *Proc. of the International Workshop on Acoustic Echo and Noise Control*.

Françoise, J., N. Schnell, R. Borghesi, and F. Bevilacqua. 2014. "Probabilistic Models for Designing Motion and Sound Relationships." In *Proc. 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* 46:1–13.

Goeschke, F. 2022. "The iOSCahedron: developing a hybrid Spatialization Instrument." In *AudioMostly*, 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.

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

Hohl, F., and F. Zotter. 2010. "Similarity of musical instrument radiation-patterns in pitch and partial." In *Proc. of the Fortschritte der Akustik (DAGA)*.

Kronlachner, M., and F. Zotter. 2014. "Spatial transformations for the enhancement of Ambisonic recordings." In *Proc. of the 2nd International Conference on Spatial Audio*.

Lewis, G. E. 1999. "Interacting with latter-day musical automata." *Contemporary Music Review* 18(3):99–112.

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

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 *Proc. of the 120th Convention of the Audio Engineering Society (AES)*.

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 Press, pp. 357 – 373.

Normandeau, R. 2009. "Timbre Spatialisation: The medium is the space." *Organised Sound* 14(3):277–285.

Peeters, G. 2004. "A large set of audio features for sound description (similarity and classification) in the Cuidado project." Technical report. IRCAM.

Rafaely, B. 2019. *Fundamentals of Spherical Array Processing – Second edition*. Springer.

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

Schmeder, A. 2009. "An exploration of design parameters for human-interactive systems with compact spherical loudspeaker arrays." In *Proc. of the 1st Ambisonics Symposium*.

Schnell, N., A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi. 2009. "MuBu & Friends – Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP." In *Proc. of the International Computer Music Conference*, pp. 423 – 426.

Schnell, N., D. Schwarz, R. Cahen, and V. Zappi. 2012. "IAE & IAEOU." In *Topophonie research project: Audiographic cluster navigation (2009-2012)*. ENSCI – Les Ateliers / Paris Design Lab, pp. 50–51.

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

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

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

Shabtai, N. R., G. Behler, M. Vorländer, and S. Weinzierl. 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.

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

Vinjar, A., and J. Bresson. 2019. "OM-AI: A Toolkit to Support AI-Based Computer-Assisted Composition Workflows in OpenMusic." In *Proc. of the Sound and Music Computing Conference*.

Weinzierl, S., M. Vorländer, G. Behler, F. Brinkmann, et al. 2017. "A Database of Anechoic Microphone Array Measurements of Musical Instruments - Recordings, Directivities, and Audio Features." Technical report. TU Berlin. URL http://dx.doi.org/10.14279/depositonce-5861.2.

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

Wright, M. 2005. "Open Sound Control: an enabling technology for musical networking." *Organised Sound* 10(3):193 – 200.

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*. Springer.

Zotter, F., M. Zaunschirm, M. Frank, and M. Kronlachner. 2017. "A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker." *Computer Music Journal* 41(3):50 – 68.