Touch Interaction for Corpus-based Audio-Visual Synthesis

Diemo Schwarz

Touch Interaction for Corpus-based Audio–Visual Synthesis

Diemo Schwarz
STMS Ircam–SU–CNRS
1 pl. Stravinsky
75004 Paris, France
schwarz@ircam.fr

ABSTRACT

Audio–visual corpus-based synthesis extends the principle of concatenative sound synthesis to the visual domain, where, in addition to the sound corpus (i.e. a collection of segments of recorded sound with a perceptual description of their sound character), the artist uses a corpus of still images with visual perceptual description (colour, texture, brightness), in order to create an audio–visual musical performance by navigating in real-time through these descriptor spaces, i.e. through the collection of sound grains in a space of perceptual audio descriptors, and at the same time through the visual descriptor space, i.e. selecting images from the visual corpus for rendering, and thus navigate in parallel through both corpora interactively with gestural control via touch sensing. The artistic–scientific question that is explored here is how to control at the same time the navigation through the audio and the image descriptor spaces with gestures, in other words, how to link the touch input to both descriptor spaces in order to create a multi-modal embodied audio–visual experience.

Author Keywords

audio–visual concatenative synthesis, cross-modal perception, 2D tactile interaction, audio and image descriptors

CCS Concepts

•**Applied computing** → **Media arts; Sound and music computing;** •**Human-centered computing** → *Gestural input;* •**Computing methodologies** → *Visual content-based indexing and retrieval;*

1. INTRODUCTION

This article presents a new method to create gesture-controlled audio–visual live performances, based on cross-modal perception and mobilising inter-modal analogies. It relates the joint realisation of the artistic idea and the technological development in this practice-led research project, supported by an art–science residency at the IMéRA institute of advanced research in 2022 called *CoCAVS*, for *Concatenative Corpus-based Audio–Visual Synthesis*. Its aim was to develop a gesture-controlled audio–visual performance instrument, that can be used in the context of musical and visual improvisation, and as a device for interactive installations. The system was developed in symbiosis with its artistic use, collecting, and experimenting with, audio and image corpora, which informed the advancement of the technological development, and during which a piece for audio–visual concatenative synthesis and violin was produced and performed in public at the Marseille Observatory's planetarium.[1]

The scientific aim of the CoCAVS project is to extend the principle of corpus-based sound synthesis to the visual domain, where, in addition to the sound corpus (i.e. a collection of segments of recorded sound (called *sound grains*) with a perceptual description of their sound character, such as pitch, brilliance, roughness, noisiness), the artist would use a corpus of still images with perceptual description (colour, texture, brightness, entropy, and other content-based image descriptors).

The artist then creates an audio–visual musical performance by navigating through one of these descriptor spaces, e.g. through the collection of sound grains in a space of perceptual audio descriptors, and at the same time through the other descriptor space, i.e. select images from the visual corpus for rendering, and thus navigate in parallel through both corpora interactively with gestural control using a touch pad. This evokes an aesthetic of acoustic and visual collage or cut-up. When navigation is local, it generates linked audio and image sequence from the two corpora with small changes over time. When the navigation jumps to different parts of the linked sound/image descriptor space, it will oppose contrasting sounds/images.

For example, a navigation from the left of the gestural control interface (typically a touch-pad like the *Sensel Morph*), to the right, might play a sequence of short sound grains chosen to start with dull and to end with brilliant sounding grains, while the chosen image sequence would start with dark greyscale and end with bright and saturated images. Likewise, navigating from the top to the bottom might start at stable pitched sounds and gradually go to noisy sounds, while the image selection goes from plain uniform content to increasingly finely textured image choices.

The fascinating artistic–scientific question that is explored here is how to control at the same time the navigation through the audio and the image descriptor spaces with gesture sensors—in other words, how to link the gesture input to both the image descriptors and the sound descriptors in order to create multi-modal audio–visual embodied per-

[1] *Performance for Audio–Visual Concatenative Synthesis & Violin*: `https://youtu.be/EFAN9fOofd0`

NIME'23, 31 May–2 June, 2023, Mexico City, Mexico.

formances and installations. Here, the questions of multisensory correspondences and synaesthesia to be explored in the mapping could lead towards new research directions in cross-modal perception.

The article will briefly present the concept of audio–visual corpus-based synthesis and will then confront the original scientific and artistic idea (or *ideal*) with the practical aspects of how to implement a technical system that allows to exploit and actually perform with audio–visual source material. It will notably address questions of making the joint navigation space maximally accessible for gestural control, for which we developed a new simple distribution method, and the question of pairing audio and visual elements (the A/V mapping).

This account sometimes takes a perspective inspired by autoethnography [2, 4], relating the subjective experience and knowledge gained during the elaboration of the idea, in the context of the author's 18 years of research and performance practice in audio CBCS with embodied gestural control, and the context of digital art.

1.1 Background and Related Work

The basis of this project is the scientific research on *Corpus-based concatenative sound synthesis* (CBCS) [29, 30], a method explored for the first time in a musical context in the author's PhD work [27] started in 1999. Here, sound synthesis 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 (units) from the database matching given sound characteristics. It is used in various contexts of music composition, live performance, sound design, installations, and allows to explore a corpus of sounds interactively or by composing paths in the timbre space, and thus to recreate novel timbral evolutions while keeping the richness and fine details of the original sound. CBCS can be seen as a content-based extension of granular synthesis, providing direct access to specific sound characteristics, allowing to create dense and lively textures with direct control of their timbre.

The open source CataRT software[2] [31, 32] created by the author, the ISMM team, and contributors, for the Max interactive media programming environment,[3] put this idea in the hands of electronic musicians, composers, and sound designers [3], and with the help of gestural controllers, allows expressive and intuitive interaction with rich sound corpora [30]. Audio CBCS draws on the well-known artistic techniques of collage, cut-up, and citation, augmenting them by digital means, automatic content analysis, and computational models [28].

Visual collage in the avant-garde art movements [17] draws on a long history of artistic exploration based on film editing techniques by *montage* [13, 20]. In digital visual arts, audio+video collage has been based on digital means in video samplers, such as VJAMM[4] [19] created by the VJ artist *Coldcut*, or the klipp av [7] system for synchronous and cross-mapped cut-up of live or composed audio and video streams. However, audio–visual collage has up to now only rarely been augmented by automatic content analysis or computational models. A first step in this direction of content-based processing of audio–visual data was taken by Collins [6] on video footage as source for the linked audio–visual corpus, analysed by 5 audio and 5 visual descriptors.

[2] `http://ismm.ircam.fr/catart`
[3] `http://cycling74.com/max`
[4] `https://www.vjamm.com`

Regarding human audio–visual perception, there is a large literature on cross-modal perception of images (or shapes) and sounds [40, 15, 1, 34, 11, 35, 23, 16]. Tsiros [36] summarises the complexities of audio–visual associations to gather insights for their efficient and intuitive design, stressing structural and functional isomorphisms (linked to Gestalt theory). Indeed one can argue that very early onwards in the human neural processing pipeline, both acoustic and visual information is essentially represented in the same way [36, 37].

However, in this project, we are not interested in multimodal integration *per se*, but rather in *intermodal analogies*, mobilising amodal, i.e. intersensory qualities. CoCAVS plays with human multi-modal perception and the expectation of cross-modal correspondence, but tests its limits by sometimes going against it, i.e. exploring all three levels of audio–visual correspondence in multimedia proposed by Cook [8]: *conformance* (one-to-one correlation), *complementation* (of the broad meaning but not contained events), and *contest* (opposition).

2. SYSTEM OVERVIEW

The CoCAVS project has the following components, shown in figure 1 which will be described in detail in the next sections:

1. an audio corpus, populated by sound files, segmented and analysed by a number of audio content descriptors,
2. an image corpus, populated by a collection of images, each image analysed by a number of image content descriptors,
3. an audio–visual mapping M, linking some of the audio descriptors with image descriptors, possibly with weightings and inversions,
4. a control input that provides a target position in the AV space (linked by mapping M) (this is typically coming from a gestural input on a 2D touch input device, allowing to create trajectories with dynamics through pressure sensing),
5. a selection algorithm that finds the audio or visual element closest to the position given by the control input above,
6. a granular rendering engine for the audio part that plays the selected audio segment with possible transformations of envelope, pitch, spectrum, and dynamics,
7. an image rendering engine that displays the selected image with a video fade in and fade out, blended with the possibly still visible previous images,

2.1 Audio Analysis

When sound files are imported into the audio corpus, they are put through an analysis stage which calculates the $n_a = 8$ audio descriptors listed in table 1: fundamental frequency (pitch), periodicity (corresponding to the continuum between tonal and noisy sounds), and first order autocorrelation coefficient (expressing spectral tilt), estimated by the time-based *yin* algorithm [10]; loudness, and spectral descriptors centroid (corresponding to the perceived brightness of a sound), spread, skewness, kurtosis (describing the

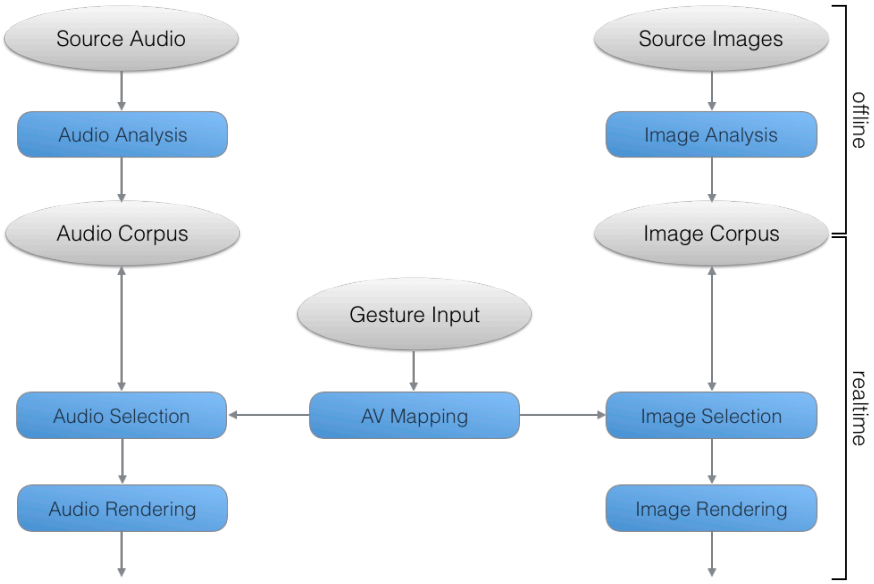

Figure 1: Data flow diagram of the CoCAVS system with its two parallel strands of processing, linked by the gesture input and AV mapping M. The upper analysis part runs offline, producing the audio and image corpora, while the lower part runs interactively in real time, using the corpora for real-time audio–visual rendering.

overall shape and distribution of the spectrum). The segmentation of the source audio files depends on the type of content. Material with clearly defined events and transients are segmented by an onset detection algorithm based on the loudness descriptor. More uniform material can be segmented by regular intervals.

The time-varying instantaneous descriptors calculated at a period of 6 ms have to be condensed to summarise each of the varying-length segments by a fixed number of values. This *temporal modeling* stage typically calculates the mean value of each descriptor over each segment, but the standard deviation, min, max, or slope can also be used to more precisely summarise the evolution of each descriptor over the duration of a segment.

Finally, for each segment, the temporal position and extent, source sound index, and summary descriptor values are stored in data tables, constituting, together with the source sounds themselves, the audio corpus.

2.2 Visual Descriptors

The image corpus is constituted of digital still images, and $n_v = 7$ image descriptors, listed in table 2, which characterise the perceptive qualities of each image with numerical values. These are calculated as the mean and standard deviation of each channel of the HSL colour representation over all pixels, expressing the average colour (hue), saturation, and luminosity of an image, and the degree of their disparity (variance). As the average hue does not necessarily correspond to any colour existing in the image, the *dominant colour* descriptor is calculated as the mode of the hue histogram weighted by luminance, i.e. the hue which occurs in the most pixels in the image, where bright pixels count more than dark pixels.

The last five image descriptors are derived from a *keypoint* or *salient feature* computer vision analysis of the image by CV.JIT, detecting "interesting" image details by edges and their intersections. The number of these keypoints found expresses the overall complexity of the image, and their average position and variance the spatial centre and extent of the region with the most salient details in the image.

2.3 Selection and Rendering

Both audio and image corpora are visualised separately as 2D scatterplots by choosing two descriptor dimensions as x and y axis, and two more for size and colour of the points, each point representing one *unit*, i.e. an audio segment or an image. Fig. 2 shows examples of audio and image corpora 2D visualisations. This visualisation also serves as the navigation interface for interactive concatenative audio or video synthesis, described in section 2.5, where an external controller such as a 2D touch pad allows the player to point to a position in the 2D visualisation. This position is taken as the selection *target*, i.e. while moving through the 2D space, the segment or image closest to the target position is rendered once.

Audio is rendered by playing the selected audio segment with a given fade in and out envelope to avoid clicks in the audio signal. When the movement through the audio corpus is fast, this can generate clouds of overlapping sound grains. The selected grains can be transformed prior to rendering by filtering or transposition, both with fixed or randomised values, and are typically played at a position in the stereo or surround sound panorama corresponding to their position in the 2D navigation space, possibly also with randomisation. This way, the continuum between preciseness and diffuseness in grain timbre and spatialisation can be explored artistically in expressive performance.

A grain can be triggered when a new point comes closest to the target position, akin to strumming strings, or be triggered at a regular period for repetitive sound or expressive accelerandi or decelerandi, allowing for a rich and varied creation of temporal structure.

The visual rendering uses the simple but efficient scheme of the framebuffer fade out effect to achieve mixing of any number of images, where the newly selected image is superposed on the current framebuffer content with a certain

Audio Descriptor Name	Explanation
Fundamental Frequency	Mean pitch over segment [Hz]
Periodicity	Mean pitch strength over segment
AC1	Mean first order autocorrelation coefficient over segment
Loudness	Mean A-weighted loudness over segment [dB]
Spectral Centroid	Mean brightness over segment [Hz]
Spectral Spread	Mean spectral spread over segment
Spectral Skewness	Mean spectral skewness over segment
Spectral Kurtosis	Mean spectral skirt width over segment

Table 1: List of CoCAVS audio descriptors.

Image Descriptor Name	Explanation
HueAvg, HueVar	Mean and standard deviation of pixel hue (colour value)
SaturationAvg, SaturationVar	Mean and standard deviation of pixel saturation
LuminanceAvg, LuminanceVar	Mean and standard deviation of pixel brightness
Color	Dominant colour
Complexity	Number of keypoints
XAvg, XVar	Mean and standard deviation of keypoint x positions
YAvg, YVar	Mean and standard deviation of keypoint y positions

Table 2: List of CoCAVS image descriptors.

luminance scaling (*attack strength*), either replacing it or mixing in the new image. The framebuffer pixels' RGB values are then multiplied by a *decay factor*, achieving a fade out in a given time. When triggered, each image is added to the framebuffer during a number of frames given by the *hold time* parameter, which, together with the attack strength, can be used to achieve a fade in of a new image. As for audio rendering, either images are triggered based on movement, when the target position comes closer to a new point, or at regular intervals the currently closest image is triggered.

2.4 Audio–Visual Mapping

The key aspect of artistic research of the CoCAVS project is to create and explore links and correspondences between the audio and visual perceptual modalities in the media streams originating from independent source corpora by joint navigation in linked descriptor spaces. The concrete mapping choice and implementation presents three practical problems due to the discreteness, multi-dimensionality, and unrelatedness of the audio and visual source corpora, discussed in the following.

2.4.1 Cross-modal Linking of Perceptual Dimensions

The audio–visual mapping needs to assign the audio descriptor dimensions to the visual descriptor dimensions. Rather than being strictly grounded on human cross-modal perception, these links can mobilise intermodal analogies [9] (for instance, audio brightness as expressed by spectral centroid is mapped to image brightness expressed by average luminance). They also are subject of artistic exploration and development of narrative, and can thus change over the duration of a piece. They can create surprise by possibly being the opposite of the "expected" link (e.g. audio brightness being mapped to inverse image brightness — such that a brilliant sound will be accompanied by a dark image, and vice versa), thus underlining the independence of each medium by mutually creating space for the other.

Technically, in the short time frame of the author's art–science residency, an explicit one-to-one mapping from a subset of the audio perceptual dimensions to visual perceptual dimensions was chosen. For the concert performance at the end of the residency,[1] the two descriptor spaces were linked in 2D for clarity, with the mappings and image corpora per section of the resulting piece shown in table 3. The mappings were chosen empirically to exploit as best as possible salient differences in the rather small image corpora. Indeed, the image corpora used in the residency were mostly collected on site, and counted from around 50 to 450 images, while the audio corpora typically range in the thousands of segments.

2.4.2 Cross-modal Pairing of Sound to Image Units

Once the perceptual dimensions have been linked between audio and image corpora, another question arises: How to link the selection and triggering of sound and image units from the two corpora? There are two different ways to realise this pairing: pre- or post-selection. *Pre-selection pairing* will use the player-controlled 2D position in the navigation space to derive separate audio and visual target positions. The closest audio and visual units will then be selected separately, which means that there isn't a one-to-one correspondence between audio and visual units.

Post-selection pairing uses the target position from navigation for audio selection, and the selected unit i's audio descriptors a_i are then mapped to the visual target position t_v via the cross-modal mapping above. The visual unit j closest to t_v is then selected. This leads to a unique injective mapping from the audio to the visual corpus: each audio unit corresponds to one specific visual unit (although the same visual unit can be associated to different audio units, and vice versa, and some visual units might not be mapped to any audio unit).

2.4.3 Density Equalisation

A third problem arises from the unequal typical distribution patterns of units in audio and image corpora, as can be seen in figure 2. There are actually two aspects to this problem: First, the touch input controller's navigation space is underused, and second, the cross-modal pairing is highly distorted and uneven, as can be seen in figure 4 (left).

Existing solutions use mass–spring models and triangulation to completely equalise the density of a point cloud [18,

Figure 2: Example of an audio corpus (left), visualised using fundamental frequency (x), periodicity (y), and an image corpus (right), visualised using average luminance (x), dominant colour (y). Colour and size of the points reuse the same descriptors as x and y to illustrate the distortion, compared to figure 3.

Audio Corpus	Audio Descriptors X/Y	Image Descriptor Mapping X/Y	Image Corpus
Prepared Piano	SpectralCentroid FundamentalFrequency	LuminanceAvg Color	Skies 1
Wind, Waves		Complexity SaturationAvg	Ruins
—	— —	LuminanceAvg SaturationAvg	Bioluminescence
Metallic Attacks		XAvg Complexity	Plants
Prepared Piano		Color SaturationAvg	Natural textures
Bowed Piano		Complexity Color	Flowers
Extended Violin	SpectralCentroid Periodicity	LuminanceAvg Color	Skies 2

Table 3: List of audio to visual descriptor mappings used in the performance for the 7 pairs of audio and image corpora. In the third part, the performer did not play sounds, but only controlled the image selection.

21]. This certainly fully optimises the use of the navigation space, but the distortion of the perceptual descriptor spaces is quite high, exacerbating as a result the distortion of the cross-modal pairing.

Therefore, a simple but sufficiently efficacious intermediate approach to density equalisation was chosen, which sorts, separately for each dimension, the descriptor values, and replaces them by their sort index (or *rank*). For instance, the smallest pitch value will be replaced by zero, and the highest one by the number of units minus one. Fig. 3 shows the corpora from figure 2 after density equalisation by sorting. One can see that this approach keeps clusters of similar units together and thus allows better visual navigation and cluster-to-cluster cross-modal pairing: The vertical lines in the top-left corner of figure 2 (left) are stable individual notes from the bowed piano corpus, which can be accessed deliberately to play expressive melodies. These notes stay together in slightly oblique clumps in figure 3 (left). This distribution also clearly improves the cross-modal pairing in figure 4 (right).

2.5 Interaction

The normalised touch input coordinates from the gesture sensing device, here a *Sensel Morph*[5] pressure-sensitive 2D touch pad to be played with one or more fingers, are mapped to the audio and/or visual navigation space by a direct translation to normalised coordinates in the chosen 2D projection of the audio and visual descriptor spaces.

[5] `https://sensel.com`

Multiple finger touches very naturally trigger multiple streams of audio grains. On the video side, there is currently no special handling for these multiple target positions (which would assign a separate layer for parallel rendering of multiple images to each touch input). Yet, the current implementation will cycle through the touch points at each rendering frame, which, together with the framebuffer fade-out technique, leads to mixed streams of images per finger when the attack strength and decay factors are at medium values.

The pressure measurement of each touch point is mapped to a per-grain gain factor for audio rendering, and to the attack strength for image rendering. This leads to the fundamental link of the performer's gesture energy (physical pressure) to overall loudness of the audio stream and overall brightness of the image stream, i.e. to acoustic and visual intensity. This mapping built into the CataRT musical instrument is a foundational element of the author's music performance work, and, augmented by the visual stream, allows to play intertwined audio–visual forms as dynamic musical gestures.

3. IMPLEMENTATION

The CoCAVS performance system is implemented in Max[6], with the MuBu[7] [26] and CataRT[8] [31, 32] extension libraries, and making use of Max's Jitter operators for han-

[6] `http://cycling74.com/max`
[7] `https://forum.ircam.fr/projects/detail/mubu`
[8] `http://ismm.ircam.fr/catart`

Figure 3: Example of the corpora in figure 2 after density equalisation. The colour and sizes are mapped to the same descriptors as x/y to show the low distortion this method entails.

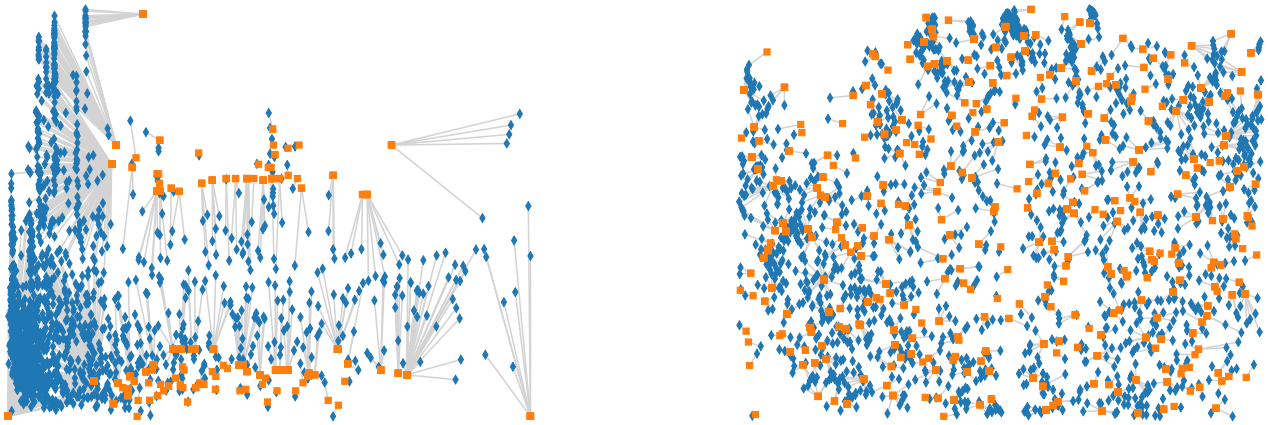

Figure 4: Example of AV pairing of the corpora in figure 2 without (left), and with density equalisation (right), showing audio units as blue lozenges and image units as orange squares.

dling, processing, and rendering image data (on the CPU and GPU).

It was straightforward to adapt CataRT and its underlying MuBu container for time-based media and data to handle image references and descriptors, instead of audio segments, keeping the same structuring (one MuBu buffer for each external file, one data frame for each unit). This can later be extended to handle multiple sub-images in one image file, or frames extracted from a video file. The only difference is that the image data itself is stored in a Jitter matrix set, and MuBu just stores the matrix index as a reference. CataRT's architecture for importing, analysing, selecting, and rendering audio units was easily adapted to handle image data. Especially the scalable kNN unit selection [33] is completely agnostic of the media type, since it finds nearest neighbours in a Euclidean descriptor space and outputs the ID of the closest unit(s), which is a segment ID for audio, and an image index for visual corpora.

The image rendering part of CoCAVS runs in the Ableton Live performance sequencer as a Max for Live (M4L) device. The device receives control data via the Open Sound Control (OSC) protocol from the audio rendering part running in a separate Max instance. This is to distribute the CPU load of graphical rendering and UI feedback for the audio part over multiple processors.

4. DISCUSSION AND FUTURE WORK

In preparing and rehearsing the performance with the CoCAVS system, one important observation was that the experience of the cross-modal AV mapping between descriptor spaces was almost always overshadowed by the dynamics mapping of touch pressure to audio volume and image intensity. Presumably, this is due to the strong congruence of this mapping (from gesture physical intensity to audio and visual intensity) and its familiar use in musical performance. This could in the future be balanced by a deliberate but not too demonstrative introduction of the AV mapping, with fixed intensity, at some early part of the performance.

Another experience is that the explicit, understandable, and perceptually valid descriptor dimensions of the two corpora are an important enabling feature for the design and artistic creation of cross-modal performances or compositions. This should in future work be compared to data-driven methods from machine learning, e.g. mapping the

anonymous dimensions of two latent spaces from two variational auto-encoder (VAE) deep neural networks, one for each modality, as has been explored recently for mapping drawn shapes to temporal sequencer patterns [39]. However, more literature research needs to be done here to find the most appropriate network architectures for linking disjoint discrete spaces, or even where the mapping could be conditioned on a few human-provided examples.

Regarding the technical development, the next steps could be to complement the image descriptors by, first, an image texturedness estimation [38], also usable in the audio domain [12], second, the detection of salient object(s) with their size, position, and shape, and third, technical image metadata recorded by the camera (year, time of day, focal length, shutter speed, light measurement, zoom factor).

Further, the corpus navigation view of the project, produced at the moment by projection, can be extended in a number of ways of increasing complexity: The techniques of dimensionality reduction such as Principal Component Analysis (PCA), Self-Organising Maps (SOM), t-Distributed Stochastic Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP) [22], already in use for corpus-based sound synthesis, could be applied analogously to the visual descriptor space, to allow to exploit the full set and structuring of descriptors for multi-modal expressive navigation in both the audio and visual descriptor spaces from a touch controller.

Regarding the cross-modal pairing problem in section 2.4.2, algorithms such as the Kuhn-Munkres ("Hungarian") algorithm for optimal assignment could produce a "better" pairing of audio to image units. However, to answer this question we would first need to define a metric that could capture the quality of a pairing.

Dynamic intermediate models [14] such as mass-interaction models could be explored, in order to generate analogous but slightly different navigation behaviours between the audio and visual modalities, from the same gesture input. As an example, the input gestures could move one light simulated mass the position of which determines the target position for navigating the audio descriptor space, closely following the gesture input, but in the visual descriptor space, excite a system of several larger masses with high inertia, to create slowly moving image superpositions that continue to oscillate even after the sound navigation has stopped moving. This allows to design the interaction with the audio and visual modalities such that the temporal dynamics and information flow can be controlled and adapted separately, and the human sensory capacities of each modality can be taken into account (while dense superposition of sound grains creates audio texture, the flicker of constantly changing images might create sensory fatigue in humans).

Lastly, although my artistic inkling tends towards using still images for the visual corpus, an extension to using short image sequences taken from videos as basic unit of visual synthesis can be envisaged. This would pose interesting and complex research questions of cross-modal correspondences not only between the timbre of static sounds and images, but time-varying audio and visual stimuli.

5. CONCLUSION

CoCAVS explores new ways to probe artistically human multi-modal perception by creating inter-modal associations in gestural navigation in joint audio and image corpora. Thus, it augments timbre-based sonic exploration by colour and texture-based image exploration, sometimes with congruent inter-modal associations, following universal perceptive expectations [40], sometimes with contradictory ones. CoCAVS achieves this through a process of practice-based technological and artistic research leading to a software system usable by musicians, and an audio–visual performance piece.[1]

The CoCAVS approach of linking the aesthetic surface features of audio and images suggests the extension of a technique of appreciation of *musique concrète* to images: that of *reduced listening* coined by Pierre Schaeffer [24, 25], where, according to Chion, *"The sound is listened to for itself, disregarding the real or supposed origin, and the meaning that it can convey"* [5]. CoCAVS invites to apply this technique to visual perception, leading to what could be termed *reduced seeing*, where the viewer concentrates on the visual impression of colour, texture, balance, and where, adapting Chion's definition above to the visual domain, *The images are seen for themselves, disregarding the real or supposed origin, and the meaning that they can convey.*

The created piece revealed an important strand in my subconsious artistic aims when collecting audio and image corpora: My striving to find abstraction in the complexity of the real world such as acoustic scenes with a drone sound that could have been produced by a synthesiser, images of skies that appear like a colour gradient, or textures in nature that suddenly reveal their underlying abstract pattern.

Another telling characteristic of the project is that it is squarely opposed to all trends in modern computer graphics and sound synthesis, where everything is fluid, three-dimensional, parametric, and generative. Instead, just short sound grains and still images are used, the appearance of which is not given by parameters of a generative model, but by specifying the desired visual descriptors, modeled after the human sensory perception. And last, images are not generated by a disembodied artificial intelligence, but chosen by the human sensibility of the musician/AV performer and articulated with embodied physical gestures. As such it offers an alternative approach to animation and narrative, creating evolution from instants, and movement out of stillness.

6. ACKNOWLEDGMENTS

This work has been funded by the Arts, Sciences, Societies residency program 2021–2022 of the IMéRA Institute for Advanced Study, Aix–Marseille Université, and supported by Laboratoire Perception, Représentations, Image, Son, Musique (PRISM), CNRS, Aix-Marseille Université.

7. ETHICAL STANDARDS

No conflicts of interest (financial or non-financial) have been observed.

8. REFERENCES

[1] M. Adeli, J. Rouat, and S. Molotchnikoff. Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in human neuroscience*, 8:352, 2014.

[2] B.-L. Bartlett and C. Ellis. *Music Autoethnographies: Making Autoethnography Sing/Making Music Personal*. Australian Academic Press, 2009.

[3] F. Bevilacqua, O. Houix, N. Misdariis, P. Susini, and D. Schwarz. MIMES: outil de prototypage et d'exploration pour la création sonore. In *Conférence francophone sur l'Interaction Homme-Machine*, 2018.

[4] A. Chamberlain. Surfing with sound: An ethnography of the art of no-input mixing: Starting to understand risk, control and feedback in musical performance. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, AM'18, New York, NY, USA, 2018. Association for Computing Machinery.

[5] M. Chion. *Guide des objets sonores*. Buchet/Chastel, Paris, France, 1995.

[6] N. Collins. Audiovisual Concatenative Synthesis. In *International Computer Music Conference*, 2007.

[7] N. Collins and F. Olofsson. Klipp AV: Live algorithmic splicing and audiovisual event capture. *Computer Music Journal*, 30(2):8–18, 2006.

[8] N. Cook. *Analysing Musical Multimedia*. Oxford University Press, 1998.

[9] G. Daurer. Audiovisual perception. In S. Naumann and J. Thoben, editors, *See This Sound Compendium*. Ludwig Boltzmann Institute Media.Art.Research, Lentos Art Museum, Linz, Academy of Visual Arts, Leipzig, 2009. `http://www.see-this-sound.at/`.

[10] A. de Cheveigné and N. Henrich. Fundamental Frequency Estimation of Musical Sounds. *Journal of the Acoustical Society of America (JASA)*, 111:2416, 2002.

[11] O. Deroy and C. Spence. Crossmodal correspondences: Four challenges. *Multisensory research*, 29(1-3):29–48, 2016.

[12] O. Fraj, R. Ghozi, and M. Jaïdane-Saïdane. Audio texturedness indicator based on a direct and reverse short listening time analysis. *Multimedia Tools and Applications*, 76(24):26177–26200, Dec. 2017.

[13] F. Genton. L'image libérée ou le cinéma selon Hans Richter. In P.-L. T. (dir.), editor, *Peintres cinéastes*, pages 49–62. Ligeia 97-100 janvier-juin 2010, 2010.

[14] V. Goudard, H. Genevois, E. Ghomi, and B. Doval. Dynamic intermediate models for audiographic synthesis. In *Sound and Music Computing*, 2011.

[15] T. Grill and A. Flexer. Visualization of perceptual qualities in textural sounds. In *International Computer Music Conference (ICMC)*, 2012.

[16] B. Guellaï, A. Callin, F. Bevilacqua, D. Schwarz, A. Pitti, S. Boucenna, and M. Gratier. Sensus communis: Some perspectives on the origins of non-synchronous cross-sensory associations. *Frontiers in Psychology*, 10:523, 2019.

[17] J.-M. Lachaud. De l'usage du collage en art au xxe siècle. *Socio-anthropologie*, (8), 2000.

[18] I. Lallemand and D. Schwarz. Interaction-optimized sound database representation. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, Paris, France, Sept. 2011.

[19] L. McCarthy and S. Gibson. Scratch video and rave: The rise of the live visuals performer (1985–2000). In *Live Visuals*, pages 89–108. Routledge, 2022.

[20] J.-P. Moreau. *De la perception à la représentation - Analyser l'œuvre vidéomusicale*. CREArTe. EME éditions, 2021.

[21] V. Paredes, F. Bevilacqua, and J. Françoise. Polyspring : A python toolbox to manipulate 2-d sound database representations. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, Stockholm, 2023.

[22] G. Roma, O. Green, and P. Tremblay. Adaptive mapping of sound collections for data-driven musical interfaces. In *NIME*, 2019.

[23] C. Saitis, S. Weinzierl, K. von Kriegstein, S. Ystad, and C. Cuskley. Timbre semantics through the lens of crossmodal correspondences: a new way of asking old questions. In *International Symposium on Universal Acoustical Communication*, Sendai, Japan, 2018.

[24] P. Schaeffer. *Traité des objets musicaux*. Éditions du Seuil, Paris, France, 1edition, 1966.

[25] P. Schaeffer and G. Reibel. *Solfège de l'objet sonore*. ORTF, Paris, France, 1967. Reedited by INA Publications–GRM 1998.

[26] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi. MuBu & friends – assembling tools for content based real-time interactive audio processing in Max/MSP. In *Proc. ICMC*, Montreal, 2009.

[27] D. Schwarz. *Data-Driven Concatenative Sound Synthesis*. Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris, 2004.

[28] D. Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35(1):3–22, Mar. 2006. Special Issue on Audio Mosaicing.

[29] D. Schwarz. Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2):92–104, Mar. 2007. Special Section: Signal Processing for Sound Synthesis.

[30] D. Schwarz. The sound space as musical instrument: Playing corpus-based concatenative synthesis. In *Proceedings of the Conference for New Interfaces for Musical Expression (NIME)*, pages 250–253, Ann Arbor, MI, USA, May 2012.

[31] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton. Real-Time Corpus-Based Concatenative Synthesis with CataRT. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 279–282, Montreal, Canada, Sept. 2006.

[32] D. Schwarz, R. Cahen, and S. Britton. Principles and applications of interactive corpus-based concatenative synthesis. In *JIM*, GMEA, Albi, France, Mar. 2008.

[33] D. Schwarz, N. Schnell, and S. Gulluni. Scalability in content-based navigation of sound databases. In *Proc. ICMC*, Montreal, QC, Canada, 2009.

[34] S. Soraghan, F. Faire, A. Renaud, and B. Supper. A new timbre visualization technique based on semantic descriptors. *Computer Music Journal*, 42(1):23–36, 2018.

[35] X. Sun, X. Li, L. Ji, F. Han, H. Wang, Y. Liu, Y. Chen, Z. Lou, and Z. Li. An extended research of crossmodal correspondence between color and sound in psychology and cognitive ergonomics. *PeerJ*, 6:e4443, 2018.

[36] A. Tsiros. The dimensions and complexities of audio-visual association. In *Electronic Visualisation and the Arts (EVA)*, pages 149–156, 2013.

[37] A. Tsiros. The parallels between the study of crossmodal correspondence and the design of cross-sensory mappings. In *Electronic Visualisation and the Arts (EVA)*, pages 175–182, 2017.

[38] F. Tupin, M. Sigelle, and H. Maitre. Definition of a spatial entropy and its use for texture discrimination. In *International Conference on Image Processing*, volume 1, pages 725–728. IEEE, 2000.

[39] N. Warren and A. Çamcı. Latent Drummer: A New Abstraction for Modular Sequencers. In *NIME*, 2022.

[40] A. Wellek. Die Farbe-Ton-Forschung und ihr erster Kongreß. *Zeitschrift für Musikwissenschaft*, 9:582, 1927.