

AIMC 2024 (09/09 - 11/09)

The corpus' body. Embodied Interaction from Machine-Learning in Human-Machine Improvisation.

Pierre Saint-Germier¹ Clément Canonne¹ Marco Fiorini²

¹CNRS (STMS, IRCAM), ²Sorbonne-Université (STMS-IRCAM)

Published on: Aug 29, 2024

URL: <https://aimc2024.pubpub.org/pub/jylagyzp>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Introduction

Artificial musical improvisation is concerned with building agents capable of improvising music and interacting meaningfully with human performers. A major challenge is to endow such agents with sufficient musicianship for such interactions to be musically meaningful, which involves constructing appropriate generative models. However, research in Embodied Music Cognition [1] has shown that embodiment is essential to expressive and interactive properties of human collective music performance. This suggests that unless artificial agents are embodied in a significant sense, their behavior in collectively improvised performances will face serious limitations.

Various strategies are available to address the issue of embodiment. The most straightforward approach is to provide such agents with a robotic body. This comes however with difficulties and limitations, such as endowing robots with sufficiently fluid mechanical behavior to cover the expressive and emotional aspects of embodiment in music. Machine-Learning may provide an alternative *indirect* strategy if it can be shown that: (i) the data generated by embodied processes bear the mark of embodiment, (ii) the generative models constructed by machine learning from those data capture relevant aspects of embodiment, and (iii) the behavior of the agent exploiting such a model inherits some benefits of embodiment.

Our study proposes an empirical assessment of this indirect strategy, taking the machine-learning-based artificial improvisation software Somax2 [2] as a case study.

The improviser's two bodies

To isolate the relevant aspect of embodiment that may be reflected in corpus data and captured by machine learning, we propose to distinguish conceptually two dimensions of embodiment.

On the one hand, the musician's body *is a multimodal resource* that provides visual and auditory cues facilitating musical coordination. We shall refer to this feature as embodiment_{MR} . It allows for visible and predictable gestures to be the focus of joint attention and thus enhance coordination in collective musical performance. Additionally, it has been shown that the postures of performers provide a back-channel through which performers signal how they keep track of the behavior of partners, in much the same way that we use back-channeling in ordinary conversation [3]. On the other hand, the contingencies of the musician's body (e.g., the fact that a pianist has two hands of five fingers each) limit and shape the sort of musical signals that may be produced. This is notably the source of instrumental idiomat�icity and gestural expressivity in music [4]. Here embodiment plays the role of a *generative constraint* on musical performance, and we shall refer to that feature as embodiment_{GC} . The mechanical performance of music for [player piano](#) or [Disklavier](#) shows what a performance with no generative constraint from a human body sounds like and helps, by contrast, appreciate the constraining and shaping role of the human body in piano performance. In virtue of this limiting and

shaping effect, embodiment_{GC} allows listeners and co-improvisers to exploit low-level perceptual expectations, which are the basis for the perception and appreciation of musical expressivity [5][6], as well as coordination within collective improvisation [7].

A distinguishing feature of embodiment_{GC} is that it may be directly reflected in the musical signal, unlike embodiment_{MR}. The shaping effect of the body as a generative constraint leaves recognizable marks in the music itself, at various levels of musical structure, e.g. interval sizes, dynamic and melodic continuity for piano music, both in the synchronic and the diachronic organization of the musical material. As a result, it is conceivable that a machine-learning algorithm constructing a generative model from embodied_{GC} musical data may transmit some benefits of embodiment_{GC} (e.g., for coordination and expressivity) to its output.

The machine-learning-based artificial improvisation software Somax2, which we shall now describe, provided a useful tool to give an empirical assessment of this hypothesis.

Somax2

[Somax2](#), designed by the Music Representation team at IRCAM outputs stylistically coherent improvisations in audio or MIDI format, based on a generative model constructed by machine learning from a given corpus in audio or MIDI format, while interacting with a human improviser [2][8][9].

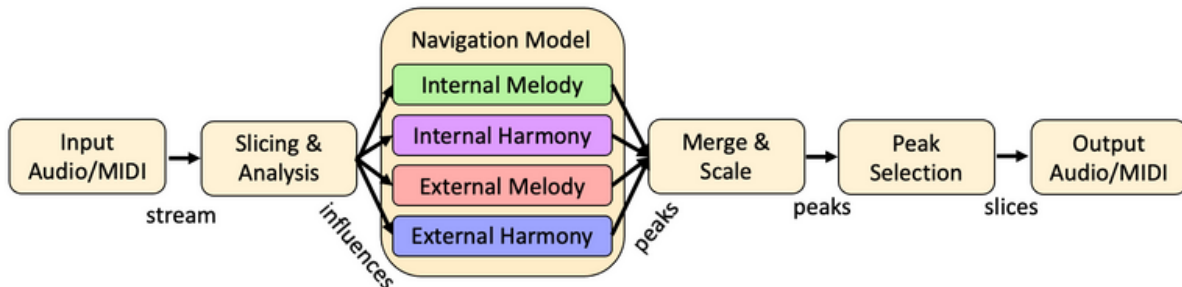


Figure 1

An overview of the steps through which Somax2 generates its audio or MIDI output at each given point in time, reacting to the incoming audio or MIDI influences from a live musician.

Source: *Somax2 User's Guide (2023)*.

Diverging from traditional generative methods, Somax2 constructs a model directly on top of a corpus of original musical data and generates a musical output by navigating through that corpus in a non-linear way.

More precisely (see [Figure 1](#)), Somax2 segments the given corpus into elementary units called *slices* and subjects each slice to detailed analysis based on various musical features, including harmony, melody, dynamics, etc. In real-time interactions with an external musician, Somax2 engages in a similar segmentation and multilayer analysis of the input stream and of its own output. In other words, Somax2 “listens” to its musician partner and to itself. These two channels of multilayered incoming information, called *influences*, are

compared in real time with the multilayered representation of the corpus, activating peaks on each layer every time a match is found between the incoming information and the corpus (See [Figure 2](#)). Each one of these peaks, resembling probability distributions, indicates potential output candidates within the original musical material that are coherent with the incoming information according to the layers of melody and harmony: peaks along the external melody and harmony layers record matches between the human performer's input and the corpus, while peaks along the internal melody and harmony layers record matches between Somax2's own output and the corpus. A sophisticated computation then scales and merges those four peaks to select a unique slice of the corpus to be played next. Importantly, this computation takes into account the previous peaks (represented in grey in [Figure 2](#)) so that Somax2 is not only sensitive to the current incoming information but keeps something like a short-term memory of the ongoing interaction.

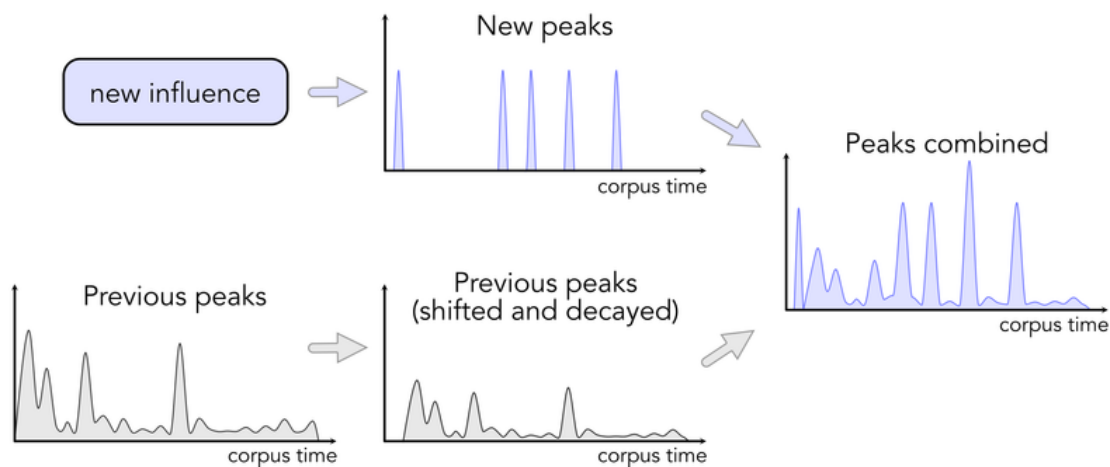


Figure 2

Peak shift, decay, and combination, as new influence is received. Source: *Somax2 User's Guide (2023)*.

Somax2's specific approach to the problem of human-machine co-improvisation makes it particularly suited to the investigation of our hypothesis about embodiment_{GC}. First, even though the computation of the next slice is only based on the four layers of internal/external melody and internal/external harmony, the idea of associating each state of the generative model to a concrete slice of the original corpus ensures that the output will preserve the more-fine grained aspects of the musical material that exceed melodic and harmonic analysis. As a result, the fine-grained marks of embodiment_{GC} in the synchronic organization of the music material, e.g., dynamic and intervallic relations between simultaneous notes, may be expected to be preserved. Second, the short-term memory endowed to Somax2 by the processes of peak shift and decay transmits the horizontal coherence of the corpus material to Somax2's output. For this reason, the marks of embodiment_{GC} in the diachronic organization of the music material, e.g., dynamic and melodic shapes, may also be expected to be preserved.

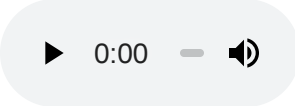
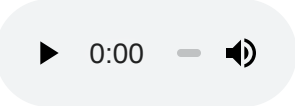
To provide empirical evidence that embodiment_{GC} may indeed be inherited from a corpus into the behavior of the Somax2 agent, the next step was to find a way to manipulate selectively and systematically embodiment_{GC} at the level of the corpus.

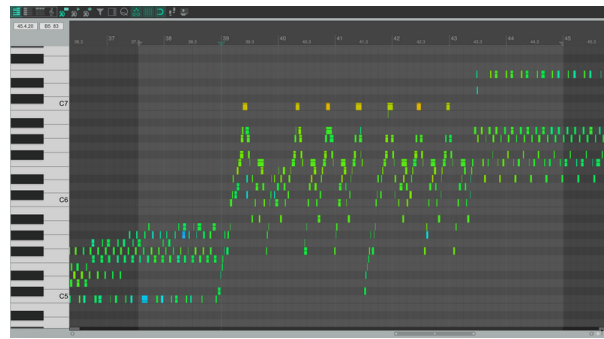
Manipulating embodiment_{GC} within a MIDI corpus

Even though embodiment_{GC} leaves recognizable marks on musical performances, finding a way to selectively manipulate the strength of that embodiment_{GC} is a difficult challenge. embodiment_{GC} acts as a significant shaping force on the musical material. As a result, it will most likely be correlated with other significant musical features, e.g., dynamic and melodic organization on both synchronic and diachronic dimensions, making it extremely difficult to manipulate embodiment_{GC} and leave all other significant musical features unchanged.

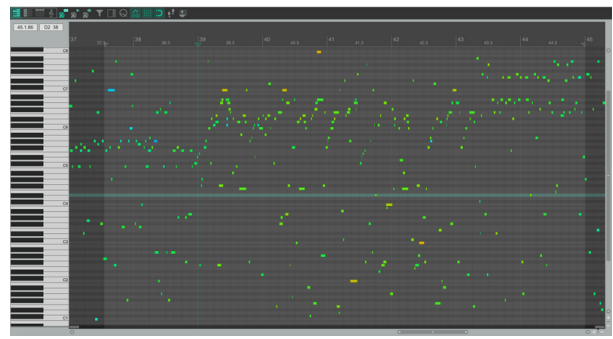
Our response to this challenge has been to opt for minimally invasive manipulations of embodiment_{GC} from a corpus of strongly embodied piano improvisations recorded in MIDI format. We considered two elementary operations. The first one, dubbed “Random Octaves”, consisted of randomizing the octave of each note event. The rationale for this manipulation is that the size of the human hand imposes a limit on the distance between simultaneous as well as adjacent notes. Operating a random octave jump on each note of a MIDI recording removes that limitation. The melodic shapes that reflect the path taken by a human hand are broken, but all the chromatic and dynamic relations are preserved. The second manipulation, dubbed “Random Dynamics”, consisted of randomizing the velocity of each note event in the corpus. The rationale for this manipulation is that the shape of the human hand imposes a limit on the difference in dynamics between simultaneous as well as adjacent notes. Randomizing the dynamics of each note of a MIDI recording removes that limitation. The dynamic shapes that reflect the touch of a human hand are broken, but all the pitch relations are preserved.

Since the two manipulations are orthogonal, we also considered their combination, dubbed “Random Dynamics and Octaves”, which gave us four experimental conditions, including the original corpus as a baseline. [Figure 3](#), [Figure 4](#), [Figure 5](#), and [Figure 6](#) show visualizations of those four conditions applied to the same fragment ([Audio 1](#)) of the corpus.

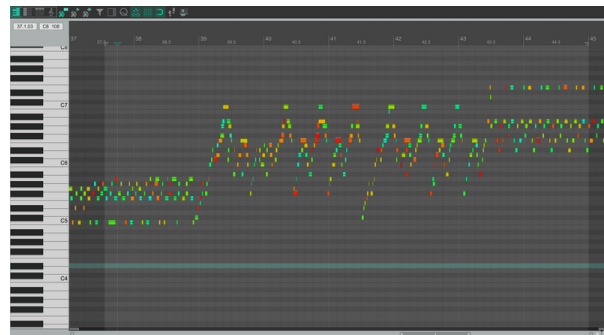
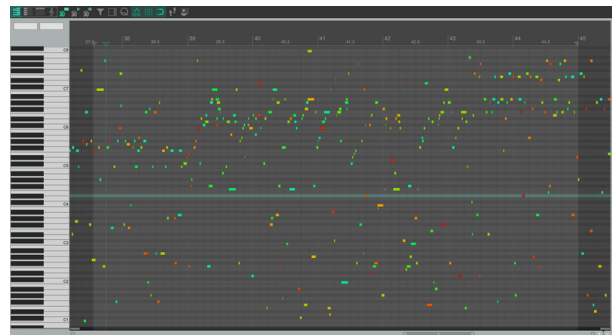
 <p>Audio 1 Excerpt from Miniature 4 of the A.M. corpus, without modification.</p>	 <p>Audio 2 Same excerpt, with random octaves.</p>
--	---

**Figure 3**

Piano roll notation for the above excerpt.

**Figure 4**

Piano roll notation for the above excerpt.

**Audio 3**Same excerpt, with
random dynamics.**Audio 4**Same excerpt, with
random octaves and
random dynamics.**Figure 5**Piano roll notation for the above excerpt. The
dynamics are indicated by the color of each
note event: the warmer the color, the higher
the dynamics.**Figure 6**

Piano roll notation for the above excerpt.

The goal of our first experiment was to check whether musically educated third-party listeners were able to detect by ear alone a change in embodiment for each manipulation.

Experiment 1

Material

We conducted audio and MIDI recordings of a corpus of piano improvisations by an internationally acclaimed performer Alexandros Markeas (A.M.) on a Yamaha C7 grand piano. The corpus was intended to reflect the

diversity of the musical material used by A.M. when they perform in the context of collective improvisation. We asked A.M. to record miniature pieces (approximately between 2 and 3 minutes), individually reflecting a particular aspect of A.M.'s improvisational material, but collectively approaching the extent and variety of their improvisation material. We asked A.M. to record as many miniatures as they needed to give a fair view of the range of their improvisational material. We excluded the miniatures where A.M. employed extended piano techniques that are not captured by the MIDI protocol (e.g., playing inside the case of the piano), which left us with a Corpus of seven miniatures (min=2'03; max=3'13).

We then isolated randomly chosen 15-second excerpts from each miniature of the Corpus and applied the three aforementioned manipulations to all of them. Each 32 excerpts were then converted from MIDI to Mp3 files, using the Steinberg clone virtual Synthesizer.

Participants

29 participants (age = 25.45; women: 17; men: 12) were recruited for this first study through the INSEAD-Sorbonne Université Behavioural Lab. Participants were screened based on their musical practice (a 5-year minimum; Mean musical practice = 10.86 years, SD = 6.70). Participants signed a written consent form and were compensated at a standard rate.

Procedure

Participants listened to each excerpt in random order. All excerpts were presented as generated by Artificial Intelligence. After each excerpt, participants had to rate, on a continuous scale, the extent to which they believed the music could have been improvised by a human pianist (from “Not at all” – 0 – to “Very much” – 10).

Results

A one-way ANOVA revealed a significant effect of Manipulation ($F=12.601$, $p<0.001$). As shown in [Figure 7](#), *post hoc* paired t-tests (adjusted for multiple comparisons using the Holm correction) showed that participants' ratings were significantly higher for “Original” ($M=6.775$, $SD=2.524$) than for “Random Dynamics” ($M=6.385$, $SD=2.660$) ($t=2.070$, $df=231$, $p=0.040$), “Random Octaves” ($M=5.976$, $SD=2.862$) ($t=3.433$, $df=231$, $p=0.001$) and for “Random Velocities and Octaves” ($M=5.795$, $SD=2.691$) ($t=4.605$, $df=231$, $p<0.001$). This suggests that the randomization of dynamics alone or octaves alone was sufficient to indeed reduce the marks of embodiment_{GC} in corpus data.

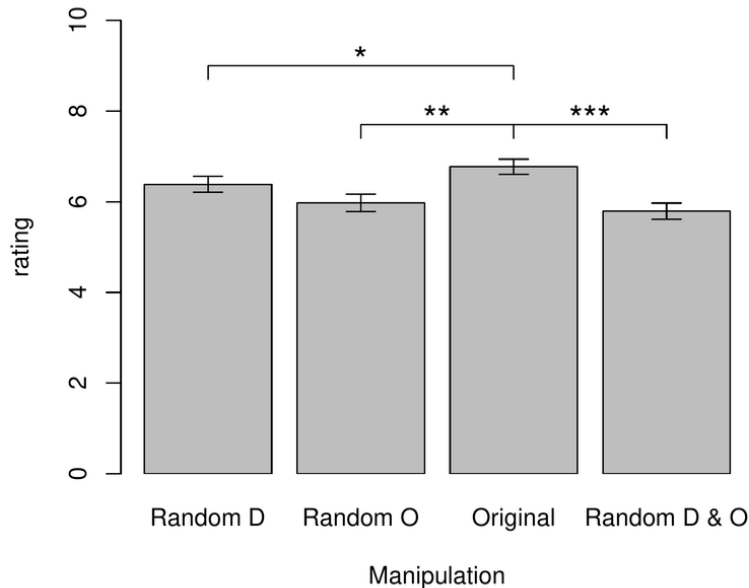


Figure 7

Mean ratings of the perception of human embodiment (0-10), depending on the manipulation of the MIDI recording (Random Dynamics, Random Octaves, Original, Random Octaves and Dynamics).

One might object that these results only show that our subjects were (to various degrees) more likely to ascribe a lower *human* embodiment_{GC} to the excerpts they heard when those excerpts underwent any one of our three manipulations. A low degree of *human* embodiment_{GC} is in principle compatible with a high degree of *non-human* embodiment_{GC} if a non-human body non-trivially constrains the music-making process. Perhaps some subjects who gave a low score of human embodiment_{GC} would have been able to imagine a different kind of body for which the given excerpt would have been nontrivially embodied_{GC}. However, the absence of widespread shared representations of such non-anthropomorphic piano-playing bodies in robotics and even in Science Fiction arguably makes it rather unlikely. Since Generative AIs are typically conceived as algorithmic rather than robotic agents, it is relatively safe to interpret a low score as a low score of embodiment_{GC} *tout court* in the context of this task.

Embodied_{GC} interaction by Machine Learning from embodied_{GC} data?

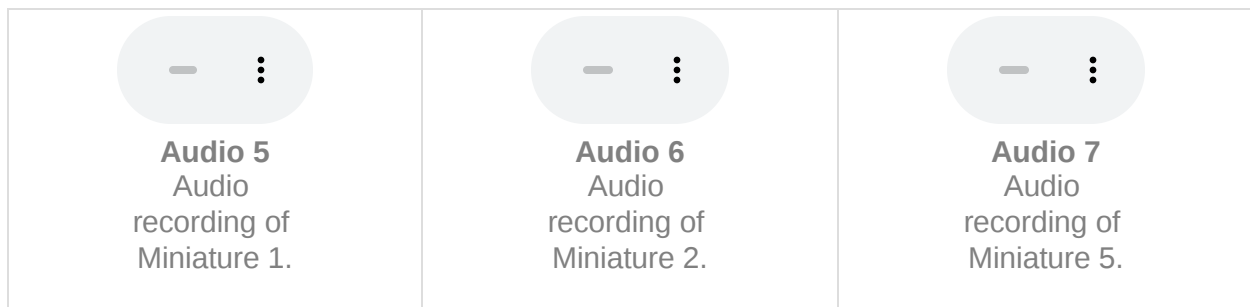
We have shown that erasing the traces of embodiment_{GC} is detectable by musically educated third-party listeners. But as we saw, embodiment_{GC} is supposed to facilitate the coordination and expressivity of collective

musical performance. If embodiment_{GC} is indeed transmitted from the corpus to the generative behavior of Somax2, changes in embodiment_{GC} at the level of the corpus may be expected to affect the experience of human improvisers interacting with Somax2. Such changes may also be expected to affect the perceived quality of the resulting collective performance from the standpoint of (musically educated) third-party listeners. This is what Experiments 3 and 4, respectively, were designed to assess.

Experiment 2

Material

We selected three miniatures from the Corpus collected for Experiment 1: Miniatures 1, 2, and 5, referred to as M1, M2, and M5 hereafter, reproduced below as [Audio 5](#), [Audio 6](#), and [Audio 7](#). The principle of selection was to ensure maximal acoustic and stylistic diversity between the musical material exemplified by each miniature. We applied the Random Octave and Random Dynamics operations to each track, which gave us an Extended Corpus of 9 tracks for Somax2.



The Somax2 player controls that modulate the computation of the next slide were kept constant for the whole experiment (i.e., between participants and across conditions). They were set to the values shown in [Figure 8](#) that were expected to give rise, overall, to the most fluid interaction with the human performers interacting with Somax2, given the diversity of the corpus material. (See [\[10\]](#) for a detailed description of Somax2's parameters.)

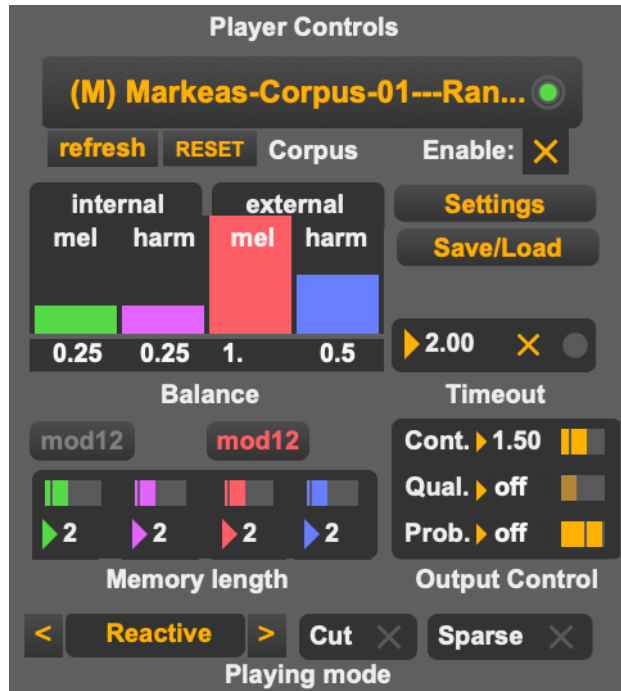


Figure 8

Somax2 player configuration for Experiment 2. Somax2 was used in *Reactive* playing mode, which means that it produces an output only in response to the input of its human partner. When the human partner stops giving inputs, Somax2 in *Reactive* mode stops producing outputs. The *Balance* settings fix relative weights for the layers of internal melody and harmony and external melody and harmony in the computation of the next slice to be played. Since the human partner played a melodic instrument, we gave more weight to the outer melodic layer. The Continuity value is set at 1.5, to maximize the chances that the navigation model will select consecutive slices, thus preserving as far as possible the consistency of the musical material in the corpus. The *Memory length*, measuring the size of Somax2's short-term memory was kept to its default value, i.e. 2. All the other parameters were set to their default value.

Participants

Ten professional musicians, recruited from the diverse Parisian Free Improvisation scene [11] participated in the experiment (mean age = 40.6; SD = 10.53; 7 male, 1 female, 1 non-binary, 1 did not provide the

information). The overall instrumentation was saxophone (N = 6), trumpet (N = 2), clarinet (N = 1) and euphonium (N = 1). All participants gave their informed written consent and were compensated at the standard rate for the employment of professional musicians in France.

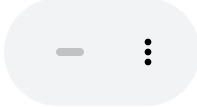
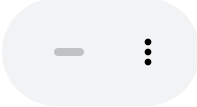
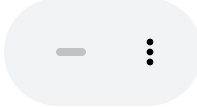
Procedure

Musicians were asked to perform 18 one-minute improvisations in duet with Somax2. For each performance, Somax2 was fed with one of the nine tracks of the aforementioned Extended Corpus. The MIDI information generated by Somax2 was then sent either toward a physical piano (Yamaha Upright U1 Disklavier), for half of the performances, or toward a virtual piano (Modartt Pianoteq 8, Upright U4 model) for the other half. After each performance, musicians were asked to provide ratings on 7-point Likert scales about 5 aspects of their experience when playing with Somax2: (a) the extent to which they felt *constrained* by Somax2; (b) the extent to which they felt *surprised* by Somax2; (c) the extent to which they felt *supported* by Somax2; (d) the extent to which they felt *stimulated* by Somax2; and (e) the extent to which they felt *immersed* in the performance while playing with Somax2. Importantly, musicians were placed in a separate studio booth, with no visual access to the booth containing the laptop with the Somax2 software, the Disklavier piano used by Somax2 for half of the performances, the sound engineer in charge of the recording session, and the scientist overseeing the Somax2 software. Musicians thus always heard Somax2 through their headphones and did not know whether the Disklavier or the virtual piano was used by Somax2.



Figure 9
Anonymized participant in the experimental setup.

Our study thus followed a 2×3 factorial design, with *Corpus Manipulation* (i.e., whether Somax2 used the “Original” corpus, the “Random Octaves” corpus, or the “Random Velocities” corpus to prompt its musical behavior) and *Output* (i.e., whether a physical Disklavier or a virtual piano is used as Somax’s output) as within-participants factors, and 5 dependent variables: *Constrained*, *Immersed*, *Stimulated*, *Supported*, and *Surprised*.

 Audio 8 Excerpt from a duet between Improviser 3 and Somax2 based on M5 with random octaves.	 Audio 9 Excerpt from a duet between Improviser 7 and Somax2 based on M1 with Random Dynamics.	 Audio 10 Excerpt from a duet between Improviser 10 and Somax2 based on M2.
---	--	---

Results

First, to assess whether our two experimental factors had any impact at all on our 5 dependent variables, we ran a MANOVA using the Stats package in R. A marginally significant MANOVA effect was obtained for *Corpus Manipulation* (Pillai's Trace=0.099, $F=1.761$, $p=0.067$). No significant effect of *Output* (Pillai's Trace=0.014, $F=0.485$, $p=0.787$) nor significant interaction between *Corpus Manipulation* and *Output* (Pillai's Trace=0.028, $F=0.476$, $p=0.905$) were found.

Second, given the results of our MANOVA, we analyzed the effect of Corpus on the participants' ratings for each one of our 5 dependent variables using a series of linear mixed regressions. For each one of our dependent variables, the following models were used, with "Original Corpus" as a base level:

- m_0 (null model): dependent variable $\sim 1 + (1 \mid \text{participant})$
- m_1 : dependent variable $\sim \text{Corpus} + (1 \mid \text{participant})$

The models were fitted with the function *lmer* from the R package *lme4* and compared using a likelihood ratio test.

For the "Constrained" dependent variable, the likelihood ratio test for model comparison was significant ($\chi^2=7.158$, $p=0.028$). Our model showed a significant negative effect of the "Random Octaves Corpus" on participant's ratings ($\beta=-0.614$, $z=0.238$, $p=0.011$). As shown in [Figure 10](#), participants thus felt less constrained when Somax2 used the "Random Octaves Corpus" than when it used the "Original Corpus".

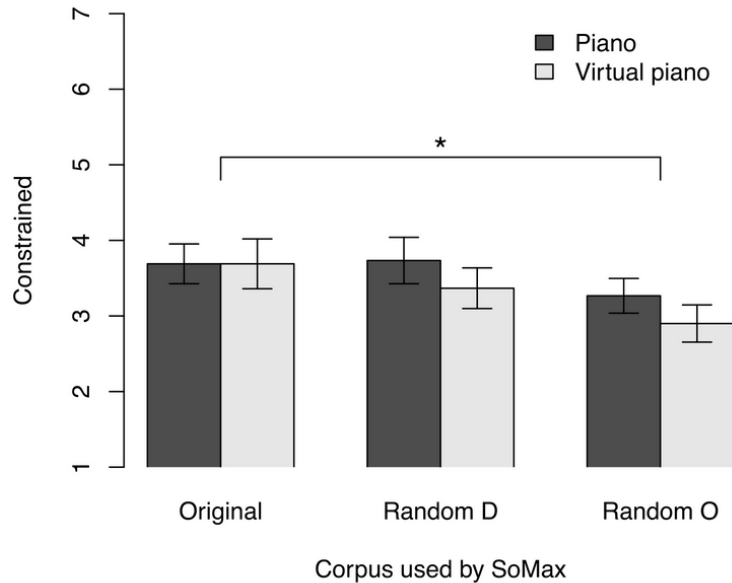


Figure 10

Mean ratings of the experience of feeling constrained (1-7) by the behavior of Somax2, depending on the modification applied to the Corpus (Original, Random Dynamics, or Random Octaves) and the nature of the Output (Piano, or Virtual Piano).

For the “Stimulated” dependent variable, the likelihood ratio test for model comparison was marginally significant ($\chi^2=4.742$, $p=0.093$). Our model showed a significant positive effect of the “Random Octaves Corpus” on participant’s ratings ($\beta=0.463$, $z=0.222$, $p = 0.038$). As shown in [Figure 11](#), participants thus felt more stimulated when Somax2 used the “Random Octaves Corpus” than when it used the “Original Corpus”.

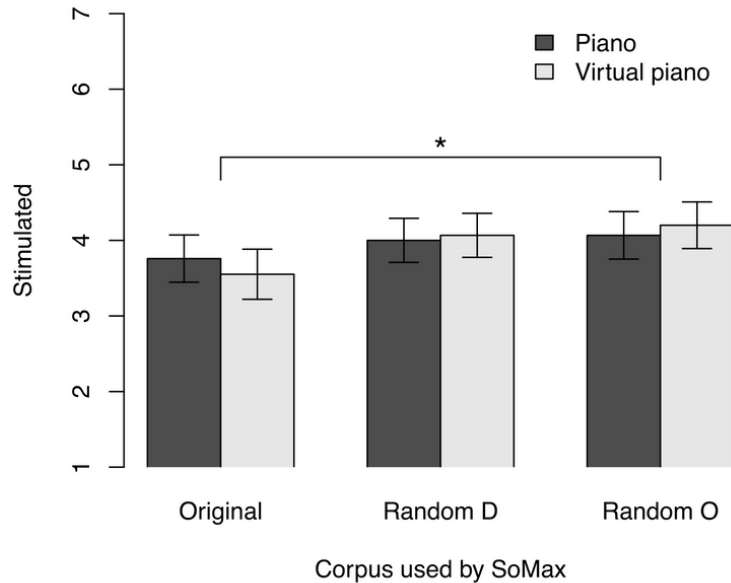


Figure 11
Mean ratings of the experience of feeling stimulated (1-7) by the behavior of Somax2, depending on the modification applied to the Corpus (Original, Random Dynamics, or Random Octaves) and the nature of the Output (Piano, or Virtual Piano).

For the “Immersed”, “Supported”, and “Surprised” dependent variables, the likelihood ratio tests for model comparison were not significant (resp. $\chi^2=3.188$, $p=0.203$; $\chi^2=0.426$, $p=0.808$; $\chi^2=0.411$, $p=0.814$).

The overall paucity of significant results may be explained as follows. To ensure the comparability of all the duets, we had to keep Somax2’s settings constant. However, those parameters are usually controlled in real-time by a human operator to make Somax2’s behavior more flexible and diverse in the course of performance. Setting the parameters once for all had the inevitable effect of making Somax2’s behavior more repetitive and stereotypical, hence limiting the chance of providing a positive experience to the human co-improviser on all dimensions.

We did find significant effects of the Random Octaves manipulation on two dimensions of the human performers’ experience. Given that it was more strongly recognized as a mark of disembodiment_{GC} in Experiment 1, it is plausible to interpret these effects as due to a lack of embodiment_{GC} on the part of Somax2. Assuming that feeling *less constrained* and *more stimulated* are markers of an overall positive experience, our data however suggest that interacting with a disembodied corpus had a small *positive* effect on the musicians’ experiences in the Random Octaves case, contrary to our expectations. This might be explained by the specifics of the population of musicians who took part in Experiment 2. Given their keen interest in the freer, most non-

idiomatic forms of collective improvisation, it is possible that their emphasis was more on the unexpectedness of the ongoing interaction rather than on tight coordination with Somax2. As a result, they might have preferred when Somax2 produced "weirder", less conventional outputs, which was more likely to happen when relying on the Random Octaves corpus. Alternatively, our pattern of results might be explained by the specifics of the original musical corpus used in our experiment: A.M. produced very dense improvisations, with a lot of thick chords. The Random Octaves might thus have had a twofold effect: on the one hand, it made the corpus feel more disembodied (as shown in Experiment 1); but on the other hand, it also created more space (with the events being more evenly distributed amongst the entire keyboard), which might have made it easier for the musicians to find their place in the overall sonic texture, and to develop their own ideas. In sum, the pattern of results observed in Experiment 2 could be explained by cultural reasons (the values favored by free improvisers) or interactional reasons (Somax2 leaving more space to its human partner). In any case, the embodiment_{GC} of the corpus seemed to make a difference, if not in the expected direction.

The experience of the human improviser is however only one perspective on the possible effects of our manipulations. Another relevant perspective is that of external hearers. Experiment 3 sought to investigate the potential effects of Corpus Manipulation on the appreciation of third-party listeners of the music produced by Somax2 in interaction with a human performer.

Experiment 3

Material

To obtain a comparable and ecological set of samples, only duo recordings from Experiment 2 in which Somax2 controlled the physical piano were used for Experiment 3. For each of the 10 improvisers recorded in Experiment 2, we randomly selected 3 tracks, representing each of our experimental conditions (Original, Random Octaves, Random Dynamics), so that each track was based on a different miniature (M1, M2, or M5). Finally, a 30-second excerpt was randomly extracted from each track, resulting in 30 musical stimuli.

Participants

28 participants (age = 25.32; women: 18; men: 10) were recruited for this study through the INSEAD-Sorbonne Université Behavioural Lab. Participants were screened based on their musical practice (a 5-year minimum; Mean musical practice = 11.43 years, SD = 5.61). Participants signed a written consent form and were compensated at a standard rate.

Procedure

Participants listened to the 30 excerpts in random order. All excerpts were presented as duets where Artificial Intelligence generated the piano part. After each excerpt, participants had to rate, on a continuous scale, the extent to which they found the duo improvisation to be successful (from "Not at all" – 0 – to "Very much" – 10).

Results

To assess the impact of our experimental manipulation on participants' ratings, the data were analyzed through a 1-way ANOVA, using the EZ package in R. Our statistical analysis revealed a significant effect of *Corpus Manipulation* ($F=26.708$, $p<0.001$). As shown in [Figure 12](#), *post hoc* paired t-tests (using the Holm correction for multiple comparisons) revealed that participants' ratings were significantly higher for "Original" ($M=5.971$, $SD=2.405$) than for "Random Octaves" ($M=5.016$, $SD=2.532$) ($t=4.930$, $df=279$, $p<0.001$) and for "Random Dynamics" ($M=4.607$, $SD=2.576$) ($t=7.314$, $df=279$, $p<0.001$). Participants' ratings were also significantly higher for "Random Octaves" than for "Random Dynamics" ($t=2.325$, $df=279$, $p=0.021$). In other words, participants were more likely to find the musical improvisation successful when Somax2 used the original embodied_{GC} Corpus.

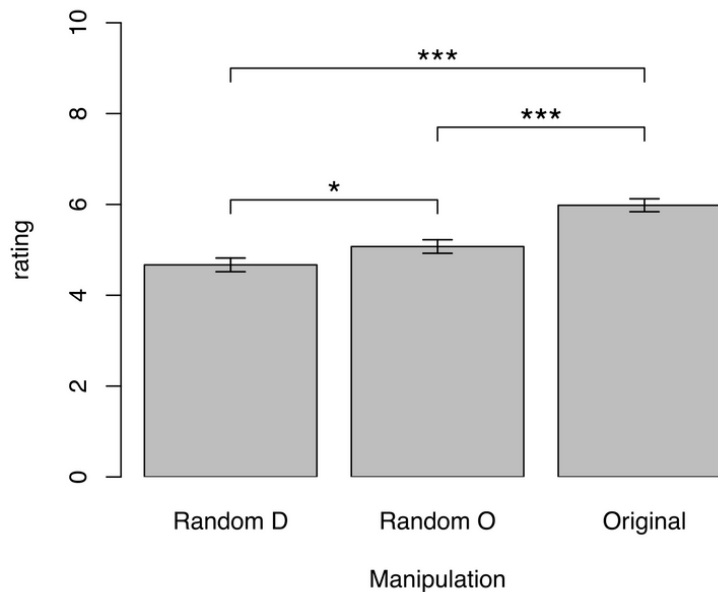


Figure 12
Mean ratings of judgments of success for the duo improvisation (1-10), depending on the manipulation applied to the Corpus (Random Dynamics, Random Octaves, or Original).

This interestingly contrasts with the perspective of the interacting improvisers investigated in Experiment 2, which revealed only a small positive effect of Corpus Manipulation in the direction of *disembodiment*_{GC}. Several explanations could be given to account for this divergence. First, although the listeners were screened for general musical practice, they were not necessarily familiar with the genre of freely improvised music, unlike the improvisers who took part in Experiment 2. The comparative weirdness of the improvisations generated from the modified tracks may have been appreciated less positively by the former for that reason.

Second, and more importantly, the metrics used in experiments 2 and 3 to ascribe a general valence to the experience were rather different and difficult to commensurate. Feeling less constrained and more stimulated are arguably signs of a positive music-making experience within an ensemble, but it says relatively little about the quality of the resulting music. Good improvised music may be obtained by performers who feel very constrained and little stimulated by their partners for example, because it forces them to find outstanding solutions to these problems[12]. We come back to this issue in the discussion below. What needs to be stressed here is that changes introduced in embodiment_{GC} at the level of the corpus used by Somax2 did impact external listeners' evaluations.

Discussion

The main question addressed by this study was whether embodiment, which seems *prima facie* neglected by software (vs robotic) approaches to artificial musical improvisation, may be indirectly captured by machine learning. This question was broken down into three hypotheses, that we assessed in turn: (i) the data generated by embodied processes bear the mark of embodiment, (ii) the generative model constructed by machine learning from such data captures relevant aspects of embodiment, such that (iii) the behavior of the agent exploiting such a model inherits some benefits of embodiment.

First, the isolation of the specific dimension of embodiment as a *generative constraint* (vs multimodal resource) provided a theoretical argument for (i), while Experiment 1 adduced empirical evidence for the auditory transparency of such marks. Second, the analysis of Somax2's design provided a theoretical argument in favor of (ii). Third, Experiments 2 and 3 showed that the experience of the musicians interacting with Somax2, and the success ratings of external listeners, respectively, were sensitive to the selective erasure of some marks of embodiment_{GC} , even if the sizes and directions of these effects were divergent: smaller and directed towards an enhancement of the experience in Experiment 2, larger and directed towards a deterioration of the perceived quality in Experiment 3.

The import of divergence regarding the confirmation of (iii) is not easy to evaluate. On the one hand, we observed an effect of the manipulation of the stronger mark of embodiment_{GC} in the corpus on the way Somax2's output affects the experience of co-improvisers and third-party listeners, which suggests that marks of embodiment_{GC} in the corpus *make a difference* to the output, and gives weight to the idea that some aspect of embodiment is transmitted by the machine-learning process. On the other hand, the nature of the observed effects raises a difficulty. A straightforward conclusion one may be tempted to draw from the embodied music cognition literature [1] is that the transmission of the effects of embodiment by machine learning should overall *benefit* the quality of both the experience of musicians and listeners. However, this inference may be too general to be applied indiscriminately to particular phenomena. The generative constraint of the body has both a negative aspect when one thinks of it as a constraint, and a positive aspect when one thinks of it as a shaping factor. So it may be expected that in some circumstances, the negative aspects outweigh the positive

ones. The suboptimal use of Somax2 in Experiment 2, with constant parameters across corpora and within performances, may very well have been such a circumstance. From this point of view, the fact that the manipulation of embodiment_{GC} made a difference overall can be seen as evidence in favor of (iii).

This being said, our study faces several limitations that force us to take the overall positive results in favor of the main hypothesis with several grains of salt. First, one may object that our manipulations of embodiment_{GC} were not selective enough. For instance, one might claim that the Random Octaves and Random Dynamics manipulation also diminish the overall aesthetic quality of the music. Then all the effects attributed to a weaker embodiment_{GC} of the corpus may be attributed to a corpus of weaker musical quality. The results of Experiment 3 from this point of view would be much less informative about (iii). This interpretation, however, would be hard to reconcile with the results of Experiment 2. More importantly, if one takes seriously the Embodied Music Cognition paradigm, such a correlation between marks of embodiment_{GC} and aesthetic properties is to be expected anyway, however surgical the manipulation. So this confounding factor is unavoidable in principle. The only option left to the experimentalist is to limit the risk by making the aesthetic impairment as small as possible. A less destructive alternative to our Random Octaves and Random Dynamics manipulation would have been to compare the corpus recorded by A. M. with a corpus of purely algorithmic music by an equally acclaimed composer. But then any observed effect would be equally attributable to the much greater stylistic differences between the improvised and algorithmic corpus.

Another limitation comes from the idiosyncratic musical genre, i.e., collective free improvisation, in which the study was conducted, and which limits the straightforward generalizability of our results. This limitation is however the counterpart of the advantages that this genre offers for this study. Contrary to first appearances, coordination in collective free improvisation is not random but obeys principles that are now well-studied[13][14]. This is also a genre for which Somax2 is routinely used in [professional musical performances](#). In addition, collective free improvisation in music may be seen as a paradigmatic example of a class of creative unscripted collective action that generalizes outside music to other performing arts such as dance and collective behavior found in sports and daily life[15].

Another concern may be that our results may not generalize beyond the idiosyncrasies of Somax2. We argued that the architecture of Somax2, and particularly, its reliance on a form of concatenative synthesis is essential to for the specific marks of embodiment_{GC} we manipulated to make a difference to the output. Since our aim was to give a proof of concept for the idea that a form of embodiment_{GC} may be obtained indirectly by machine-learning, such a limitation may not be a problem in itself. However, we chose those specific marks for methodological reasons. There are many other marks of embodiment_{GC} in musical signals, which makes it plausible that other architectures may be able to process them. This of course, remains an empirical hypothesis to be tested on a case-by-case basis.

Zooming out of the details of our study, one might consider the difference between the embodiment arguably obtained for algorithmic musical agents by the indirect route we have explored, and the physical embodiment afforded by robotics[16]. Robotic bodies offer both a multimodal resource for coordination and a generative constraint on the musical signal, unlike algorithmic agents, such as Somax2, which inherit, at best, the generative constraint reflected in the corpus used to train them. Furthermore, this generative constraint is only *indirectly simulated* by Somax, as it processes the marks of embodiment_{GC} in the corpus used to train it, whereas it is *causally* and *directly* imposed by the physical properties of robotic bodies.

The main advantage of simulated embodiment is that it comes at a lower cost, and allows for quick reconfigurations. It takes a few seconds to upload a new corpus in Somax2, and thus endow it with a new embodiment_{GC}. By contrast, it takes a lot of time and resources to design and construct a new robotic body. Furthermore, it requires considerable ingenuity to design robots with high *expressive* capacities, when it comes to the manipulation of musical instruments and musical sounds generally. On the contrary, the inheritance of embodiment_{GC} from highly expressive human bodies favors the transmission of rich expressive patterns to the outputs of Somax2. It might be objected here that roboticians can design non-anthropomorphic bodies, and thus extend the repertoire of embodiment beyond the limitations of the human body. Our study, by only relying on humanly embodied corpora, may suggest that this limitation is not overcome when embodiment is indirectly simulated by machine learning. It may be replied, however, that it is possible to manipulate humanly made corpora in the direction of *augmenting* embodiment, just like we manipulated it to diminish its marks. For example, by mixing corpora recorded from instruments with distinct instrumental idiomatilities one may synthesize generative constraints richer than those afforded by the human body.

Overall, it appears naive to conclude that algorithmic agents are by nature suffering from drastic limitations due to their lack of embodiment. If embodiment can be indirectly simulated in the way we suggested, then machine learning provides an alternative to robotics, when it comes to addressing the issue of embodiment in musical artificial intelligence.

Acknowledgments

This research is supported by the European Research Council (ERC) as part of the Raising Co-Creativity in Cyber-Human Musicianship (REACH) Project directed by Gérard Assayag (IRCAM), under the European Union's Horizon 2020 research (GA #883313) and has received help from the INSEAD business school for the experimental part. We warmly thank Jérémy Henriot for helping us with the recording of Experiment 2, and all the musicians who took part in the study.

Ethics Statement

The authors have no conflicts of interest to declare. The three experiments were approved by the INSEAD review board (protocol ID: 2023-16). This research is meant to contribute to the creative use of computers in

musical practices.

References

1. Lesaffre, M., Maes, P.-J., & Leman, M. (Eds.). (2017). *The Routledge Handbook of Embodied Music Interaction*. New York & London: Routledge. [↵](#)
2. Assayag, G., Bonnasse-Gahot, L., & Borg, J. (2024). Cocreative Interaction: Somax2 and the REACH Project. *Computer Music Journal*, 1(19). [↵](#)
3. Moran, N., Hadley, L. V., Bader, M., & Keller, P. E. (2015). Perception of 'Back-Channeling' Nonverbal Feedback in Musical Duo Improvisation. *PLOS ONE*, 10(6), 1–13. <https://doi.org/10.1371/journal.pone.0130070> [↵](#)
4. Souza, J. D. (2017). *Music at Hand*. New York: Oxford University Press. [↵](#)
5. Meyer, L. B. (1956). *Emotion and Meaning in Music*. [Chicago]: University of Chicago Press. [↵](#)
6. Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42(2), 594–609. [↵](#)
7. Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, 8–9, 998–1003. <https://doi.org/doi:10.1016/j.neunet.2010.06.002> [↵](#)
8. Borg, J. (2021). The Somax 2 Theoretical Model. *STMS-IRCAM, Technical Report*. [↵](#)
9. Borg, J. (2021). The Somax 2 Software Architecture Rev. 0.2.0. *STMS-IRCAM, Technical Report*. [↵](#)
10. Somax2 User's Guide. (2024). *STMS-IRCAM, Technical Report*. Retrieved from <https://github.com/DYCI2/Somax2/blob/master/Somax2%20User's%20Guide.pdf> [↵](#)
11. Roueff, O. (2006). L'invention d'une «scène» musicale, ou le travail du réseau: La programmation d'un club de musiques improvisées entre radicalisation et consécration (1991-2001). *Sociologie de l'Art*, (1), 43–76. [↵](#)
12. Golvet, A., Goupil, L., Saint-Germier, P., Matuszewski, B., Assayag, G., Nika, J., & Canonne, C. (2021). With, against, or without? Familiarity and copresence increase interactional dissensus and relational plasticity in freely improvising duos. *Psychology of Aesthetics, Creativity, and the Arts*. [↵](#)
13. MacDonald, R. A., & Wilson, G. B. (2020). *The art of becoming: How group improvisation works*. Oxford University Press. [↵](#)

14. Saint-Germier, P., & Canonne, C. (2022). Coordinating free improvisation: An integrative framework for the study of collective improvisation. *Musicae Scientiae*, 26(3), 455–475. [↵](#)
15. Coste, A., Bardy, B. G., & Marin, L. (2019). Towards an embodied signature of improvisation skills. *Frontiers in Psychology*, 10, 473994. [↵](#)
16. Weinberg, G., Bretan, M., Hoffman, G., & Driscoll, S. (2020). *Robotic Musicianship. Embodied Artificial Creativity and Mechatronic Musical Expression*. Springer. [↵](#)