



HAL
open science

Embodied sound design

Stefano Delle Monache, Davide Rocchesso, Frédéric Bevilacqua, Guillaume Lemaitre, Stefano Baldan, Andrea Cera

► **To cite this version:**

Stefano Delle Monache, Davide Rocchesso, Frédéric Bevilacqua, Guillaume Lemaitre, Stefano Baldan, et al.. Embodied sound design. *International Journal of Human-Computer Studies*, 2018, 118, pp.47-59. 10.1016/j.ijhcs.2018.05.007 . hal-04280735

HAL Id: hal-04280735

<https://hal.science/hal-04280735v1>

Submitted on 13 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Embodied Sound Design

Stefano Delle Monache^{a,*}, Davide Rocchesso^b, Frédéric Bevilacqua^c, Guillaume Lemaitre^c, Stefano Baldan^a, Andrea Cera^d

^a*Iuav University of Venice, Department of Architecture and Arts, Dorsoduro 2196, 30123, Venezia, Italy.*

^b*University of Palermo, Department of Mathematics and Computer Science, via Archirafi 34, 90123, 90133 Palermo, Italy.*

^c*Ircam - Institute for Research and Coordination in Acoustics/Music, 1, place Igor-Stravinsky 75004 Paris, France.*

^d*Independent sound designer, 2, Largo Trieste 36034 Malo - Vicenza, Italy*

Abstract

Embodied sound design is a process of sound creation that involves the designer's vocal apparatus and gestures. The possibilities of vocal sketching were investigated by means of an art installation. An artist–designer interpreted several vocal self-portraits and rendered the corresponding synthetic sketches by using physics-based and concatenative sound synthesis. Both synthesis techniques afforded a broad range of artificial sound objects, from concrete to abstract, all derived from natural vocalisations. The vocal-to-synthetic transformation process was then automated in SEeD, a tool allowing to set and play interactively with physics- or corpus-based sound models. The voice-driven process and tool, developed and evaluated through design exercises, show how an embodied sound sketching system can work in supporting the externalisation of sonic concepts.

Keywords: sound synthesis, conceptual design, sound design tool

1. Introduction

2 Interacting with and through representations is a key aspect of designers' activity.
3 A rich body of literature studied the role of sketches, drawings, and static forms of rep-

*Corresponding author. M. +39 347 6414957, T. +39 041 257 1852
Email addresses: stefano.dellemonache@iuav.it (Stefano Delle Monache),
davide.rocchesso@unipa.it (Davide Rocchesso), frederic.bevilacqua@ircam.fr (Frédéric Bevilacqua), Guillaume.Lemaitre@ircam.fr, GuillaumeJLemaitre@gmail.com (Guillaume Lemaitre), stefanobaldan@iuav.it (Stefano Baldan), andreawax@yahoo.it (Andrea Cera)

4 representations in the design domain (Schön, 1984; Purcell & Gero, 1998; Goldschmidt,
5 2014). A rather recent approach to the understanding of the design process, especially
6 in its early stages, has been focusing on the role of multi-modality and the contribu-
7 tion of non-verbal channels as key means of communication, kinaesthetic thinking, and
8 more generally of doing design (Tholander et al., 2008; Tversky et al., 2009).

9 Explanations, i.e. representations, emerge as multi-modal models from the con-
10 tinuous interplay between talk and action, and through the concurrent manipulation of
11 sketches and diagrams, physical props and artefacts. Designers are fluent in combining
12 utterances, drawings, props, and especially gestures to create and annotate models of
13 a situation, systems, processes and configurations (Kang et al., 2015). Active engage-
14 ment and performative action complement static forms of representation (e.g., white
15 boards, drawings and diagrams), and allow to convey complex spatio-dynamic proper-
16 ties, motion, trajectories, and time-based events. As visuospatial forms of communica-
17 tion can represent concepts more directly than verbal descriptions, they are suitable to
18 catch and express the dynamics of designs, behaviours and relations unfolding in time
19 and space.

20 Within the theoretical framework of distributed and embodied cognition, design
21 research has given attention to the role of gestures in visuospatial communication, as
22 peculiar cognitive artefacts through which designers represent structural and functional
23 information, think and collaborate (Becvar et al., 2008; Cash & Maier, 2016). Ulti-
24 mately, comprehending the rich dynamics of embodied sketching is crucial for the de-
25 velopment of appropriate technological systems that can support non-verbal displays
26 in conceptual design (Visser & Maher, 2011; Eris et al., 2014).

27 Within these premises, this contribution tackles the specific domain of sound de-
28 sign, that is the creative process of making sonic intentions audible (Pauletto et al.,
29 2016; Susini et al., 2014; Franinović & Serafin, 2013). This definition applies in
30 many different contexts, ranging from industrial products to computer games, where
31 the sound designer is called to give objects a “voice”, or a specific audible character,
32 sometimes following function (as in the sound for an electric car engine), sometimes
33 following form (as in audiovisual composition) (Susini et al., 2014)."

34 We propose an embodied account of the sound design process, which places the

35 bodily experience (i.e., communication of sonic concepts through vocal and gestural
36 imitations) at the center of the sound creation process. Vocalisations and gestures are
37 the primary cognitive artefacts available to sound designers to explain concepts and
38 sketch sonic representations cooperatively (Rocchesso et al., 2015; Delle Monache &
39 Rocchesso, 2016). Vocal sketching is a fast prototyping technique, aimed at the early
40 stage of sound design, which exploits the voice as means to portraying and imitating
41 non-speech sounds. The seminal research workshop by ? demonstrated the poten-
42 tial of such methodology. Yet, fully understanding how humans explain time-based
43 processes, such as sonic concepts, through vocal and gestural imitations is essential to
44 provide foundational training and communication support to sound practitioners in col-
45 laborative design scenarios (Rocchesso et al., 2016). Indeed, research in sound design
46 thinking is largely unexplored, and the expertise of professionals rather seems to be
47 based on individual paths in which music education, computer science, psychoacous-
48 tics and ultimately design intertwine to form the tacit knowledge of the practice (Özcan
49 & van Egmond, 2009).

50 Nykänen et al. (2015) framed the use of sketching in sound design, within the
51 “seeing–moving–seeing” model proposed by Schön & Wiggins (1992). However, a
52 critical reading of the article stresses how the conceptual stage in practice mostly relies
53 on verbal descriptions, and on the selection of advanced sound proposals. The reason
54 is not only the lack of a vocabulary on sound, shared with stakeholders, but also of
55 a more general attitude to sketch-thinking, being the medium, i.e. sound, normally a
56 sound recording which does not afford immediate manipulations. In their Volvo case
57 study, involving a sound design team and a product development team, Nykänen et al.
58 report that they had to make explicit the provisional characteristics that a sound sketch
59 would show. This situation is typical of current professional practices, which have not
60 embraced an embodied attitude to sound design. As a consequence, there is lack of
61 genuine cooperation in sound design processes, even between peers.

62 In embodied sound design, spatial forms of communication and non-verbal ut-
63 terances intertwine in the formation and explanation of auditory objects (Kubovy &
64 Schutz, 2010). Despite their ephemerality, vocal and gestural sketches have a repre-
65 sentational stability over time, showing a coherent mapping between the representing

66 world and the represented world, and providing material displays that can be built upon
67 and recalled during the stream of a discourse (Becvar et al., 2008; Delle Monache &
68 Rocchesso, 2016; Lemaitre et al., 2016a; Scurto et al., 2016). Connections with Fo-
69 ley artistry are easily found here (Pauletto, 2017), as the sound generation is inherently
70 embodied, but embodied sonic sketching aims at exploratory actions that could be com-
71 pared to scribbling with pencil on paper.

72 In this contribution, we propose that a tool for voice-driven sound synthesis would
73 automate the bridging between the representing and the represented world, by provid-
74 ing synthetic sound models that could be set, played and shared as instances of vocal
75 utterances. Target users of the tool are those professionals who work creatively with
76 sound – in product design, game design, branding, or audiovisual productions – and
77 who need to interact with stakeholders during the sound creation process. Although
78 continuous gestural interaction is envisioned as part of the creative process enabled
79 by the proposed tool, in this article we focus on the use of the voice as a means for
80 sketching and controlling synthetic sounds.

81 The manuscript is organised as follows. In Section 2 we set the theoretical back-
82 ground, by stressing the sensory-motor nature of auditory experiences. Then, in Sec-
83 tion 3, we approach the problem of the embodied representation of sound from the
84 three-folded perspective of sound perception, production, and articulation of voice and
85 gestures. In Section 4, we describe an artistic audio-visual installation that we realised
86 to explore the flow of the embodied sound design process, from the internal sonic con-
87 cept to its synthetic rendering, via the translation into actions (i.e., vocal sketches). The
88 representing world, that is the vocal sketches, and the represented world, namely the
89 synthetic counterparts, were compared and experimentally assessed in terms of natural-
90 ness and concreteness of the representation at hand. Eventually, the artistic installation
91 and its evaluation were functional to the development of SEeD, a semi-automated sys-
92 tem for the conversion of continuous vocalisations into synthetic sounds. The system
93 architecture, and its rationale are described in detail in Section 5, together with the
94 development that was driven by some design workshop experiences. Finally, the con-
95 sistency of SEeD as a sketching tool is assessed in Section 6, where we report the
96 experimental evaluation through a sound design exercise: We asked three professional

97 sound designers to sketch sounds that are similar to a set of examples produced with
98 SEED itself.

99 **2. Theoretical framing**

100 As humans we inhabit an enacted world, and we experience sounds that are the
101 product of our own actions, actions of other living creatures, or physical processes of
102 various kinds. In most everyday situations, sounds can be associated with the actions
103 that produced them or with the actions taken in response.

104 According to ideomotor theories (see Shin et al. (2010) for a review and a dis-
105 cussion), cognitive representations of sensory stimuli (images, sounds, etc.) and the
106 actions that produce them are tightly bound in memory and interact bidirectionally.
107 Activating any element of such a sensory-motor representation may activate the whole
108 representation. For example, activating the sound of an action may activate the motor
109 plans producing that action: Simply hearing the sound of an action may trigger or prime
110 that action. Auditory-motor associations, in particular, are short-lived and can be eas-
111 ily reconfigured by rapid learning of the association between sounds and the gestures
112 that produce them (Lemaitre et al., 2015). As a consequence, any associations between
113 gestures and sounds can be readily created, reconfigured and shaped by design.

114 Humans find it easier to imagine what they have previously experienced through
115 perception-action loops. If we want to communicate a sonic concept to another person,
116 we often try to re-enact the sonic process, internally represented as a perception-action
117 ensemble. Through vocal imitations, it is our vocal apparatus that gives access to such
118 internal representations. The vocal motor program that recreates a sonic process can
119 be described as an embodied auditory motor representation. We dedicate Section 3 to
120 the problem of sound representation, from an embodied perspective.

121 We argue that sound design should address the sensory-motor nature of auditory
122 experiences to be maximally effective. Embodied sound design is a process of sound
123 creation that extensively involves the designer's body. This contribution proposes a
124 voice-based embodied sound design process, and a related interactive system that is
125 empowering the designer to span a vast sonic space.

126 To make the vision of embodied sound design concrete, we embraced an artistic
127 route and realised an interactive audiovisual installation, where an artist–designer (Dunne
128 & Gaver, 1997) transformed a set of vocalisations into synthetic sounds that are or-
129 ganically ascribable to specific human utterances. In practice, the artist–designer had
130 considerable freedom to interpret a set of recorded utterances, and to manually repro-
131 duce them with his own tool-box. This approach provided a subjective vision of how
132 voice-driven production of synthetic sounds may actually be achieved. At the same
133 time, he was restricted by choice of sonic materials and synthesis techniques, as ex-
134 pressed in Section 4. These are not limits to creativity, but constraints that make the
135 vision realisable and expressing certain experiential qualities (Löwgren & Stolterman,
136 2004).

137 Subjectivity is necessary here, as there is no “correct” solution to the problem of
138 translating a vocalisation to a synthetic sound (Tuuri et al., 2011). Instead, there is a
139 space of possibilities that the artist–designer can thoughtfully explore, thus providing
140 valuable information for the tool design process. Constraints are necessary as well,
141 if we want to exploit the artist–designer’s work for a preliminary evaluation of the
142 technologies that are being proposed for the sound sketching process. Eventually, as
143 described in Section 5, the artistic exercise informed the design of a tool that auto-
144 mates the transformation of vocal expressions, by selection and manipulation of sound
145 models.

146 **3. Sound approaches to sound**

147 In designing synthetic sounds, there is an unresolved dilemma that the designer or
148 artist has to face: How to approach sound and its representations (De Poli et al., 1991;
149 Roden, 2010). Should we design sounds as they appear to the senses, by manipulat-
150 ing their proximal characteristics? Or should we rather look at potential sources, at
151 physical systems that produce sound as a side effect of distal interactions? Can our
152 body help establishing bridges between distal (source-related) and proximal (sensory-
153 related) representations? Can the intimate space of vocal and gestural articulations
154 be used to drive sound synthesis? Giving answers to these questions corresponds to

155 choosing appropriate sound models and preparing a set of effective sound synthesis
156 tools. These developments can be informed by research findings in perception, pro-
157 duction, and articulation of sounds. An embodied approach to sound design should
158 exploit knowledge in these areas, especially referring to voice and gesture, to create
159 action-oriented ontologies (Leman, 2008).

160 3.1. Perception

161 When considering what people hear from their environment, it emerges that sounds
162 are mostly perceived as belonging to categories of the physical world (Gaver, 1993).
163 For example, people do not hear “a series of impulsive sounds with an initial low-
164 frequency impulse followed by a rapidly rising pitch”: they simply hear “water drip-
165 ping”. Research on categorical sound perception has shown that, when asked to sort the
166 sounds of a kitchen environment, listeners spontaneously create four main categories of
167 solid, electrical, gas, and liquid sounds, even though the sounds within these categories
168 may be acoustically different (Houix et al., 2012). Similar results were found when
169 listeners were asked to sort *imitations* of such sounds, confirming that vocal imitations
170 convey the identity of the sounds (Lemaitre et al., 2011).

171 Whereas people tend to associate sounds to events involving distal physical ob-
172 jects (*sounding objects*), when the task is to separate, distinguish, count, or compose
173 sounds (Kubovy & Schutz, 2010) the attention inevitably goes to proximal *auditory*
174 *objects* represented in the time-frequency plane. The most prominent elements of the
175 proximal signal may be selected by simplification and inversion of time-frequency rep-
176 resentations. This produces the so-called auditory sketches (Isnard et al., 2016), which
177 have been used to test how the recognisability of imitations compares with that of
178 sparse time-frequency sound representations (Lemaitre et al., 2016a) and to highlight
179 the most relevant morphological elements. Tonal components, noise, and transients
180 can be extracted from sound objects with Fourier-based techniques (Verma et al., 1997).
181 Low-frequency periodic phenomena are also perceptually very relevant and often come
182 as trains of transients. Research on morphologic features and on extraction of audio
183 primitives of vocal imitations is making progress (Marchetto & Peeters, 2015, 2017).

184 Recent research has shown that vocal imitations can be more effective than verbal-

185 isations at representing and communicating sounds when these are difficult to describe
186 with words (Lemaitre & Rocchesso, 2014). This indicates that vocal imitations can be
187 a useful tool for investigating sound perception, and shows that the voice is instrumen-
188 tal to embodied sound cognition. When using vocal imitations, it must be considered
189 that there is the human individual at the center of the scene, with her preferences, lim-
190 itations, and idiosyncrasies. This makes the couples sound/imitation highly subjective,
191 but ensures the highest level of embodiment of the sonic space. When using vocal im-
192 itations to drive perception-based synthesisers, the resulting perception/action sound
193 synthesis is tightly connected to embodied representations of sound, especially if voice
194 control can be properly individualised.

195 3.2. *Production*

196 3.2.1. *Physics-based modelling*

197 In everyday contemporary environments, sounds are either produced by loudspeakers
198 or by various physical phenomena, such as mechanical contacts, or fluid-dynamic
199 processes. Leaving aside arbitrary electronic signals played via loudspeakers, an eco-
200 logical approach to sound synthesis may look at the sources and try to mimic the phys-
201 ical behaviour of sounding objects. Physics-based modelling of everyday sounds can
202 rely on detailed simulation of basic physical phenomena, and introduce simplifications
203 and abstractions for complex physical phenomena. Much of the physical-modelling
204 literature focused its attention to the properties of resonating objects, whose detailed
205 models are fed with patterned and filtered bursts of noise (Aramaki & Kronland-Martinet,
206 2006; van den Doel et al., 2001). Conversely, the Sound Design Toolkit (Baldan et al.,
207 2017) focuses on dynamic nonlinear interactions. It is based on a bottom-up hierarchy
208 that represents the dependencies between low-level models and temporally-patterned
209 textures and processes, organised into four classes: solids, liquids, gasses, and ma-
210 chines. These classes are grounded on different physical mechanisms, and they mirror
211 the perceived categories of everyday sounds.

212 3.2.2. *Corpus-based modelling*

213 Following the correspondence between classical and quantum physics, one may
214 look for sound quanta and for a quantum description of the sonic world. This was
215 indeed recommended, long ago by Gabor (1947), for proximal acoustic signals, and
216 gave rise to granular synthesis and related techniques. As a matter of fact, many ev-
217 eryday sounds are the product of a large number of micro-interactions in the physical
218 distal world, and for these the statistical properties of the distributions of events are
219 just as relevant as the qualities of individual events. Corpus-based synthesis based on
220 databases of short sound samples (Schwarz, 2007) is a very effective means to repre-
221 sent these kinds of continuously-varying sounds, especially those exhibiting textural
222 properties (Strobl et al., 2006; Schnell, 2011).

223 3.3. *Articulation*

224 A human can use her body to produce articulations that bridge the distal with the
225 proximal, to link production with perception via a motor representation. In the realm
226 of sound, this is mainly done through voice and gesture.

227 3.3.1. *Voice*

228 Phonation, Turbulence, and Myoelasticity¹ are three important components of vo-
229 cal imitations that can be activated independently, and therefore be present simultane-
230 ously (Helgason, 2014). Another component may be transient/click/impulse or, alter-
231 natively, this may be aggregated with myoelasticity, as low-frequency oscillations of
232 articulatory mechanisms can often be viewed as sequences of discrete pulses. These
233 component features can be extracted automatically from audio with time-frequency
234 analysis and machine learning (Peeters et al., 2015). They can be made to correspond
235 to categories of sounds as they are perceived and as they are produced in the physical
236 world. Lemaitre et al. (2016b) showed that naïve imitators can accurately match the
237 relative pitch, temporal behaviour, and spectral centroid of their vocalisations to the

¹For the sake of this research, myoelastic vocal fold vibrations contribute only to the phonation compo-
nent. Instead, the myoelastic component includes fairly slow periodic myoelastic vibrations such as those
produced by the lips when they are pressed together while an airstream is passed through.

238 corresponding features of non-speech abstract sounds . So, vocal articulations can be
239 made to correspond to features of proximal acoustic signals.

240 3.3.2. *Gesture*

241 Lemaitre, Scurto and colleagues have studied and compared how people describe
242 sounds with gestures and vocalisations (Scurto et al., 2015, 2016; Lemaitre et al.,
243 2017). Whereas vocalisations reproduce all features of the imitated sounds as faithfully
244 as vocally possible, the gestures focus on one salient feature with metaphors based on
245 auditory-visual correspondences. Such metaphors (e.g. mapping pitch to a spatial po-
246 sition or rapidly shaking hands to represent noisiness) are consistently shared across
247 participants yet not necessarily explicit in a culture.

248 Taken together, voice and gesture are used by humans to articulate the distinc-
249 tive traits of sounds, for the purpose of communicating sound ideas to other humans.
250 Articulations effectively bridge the physical production of everyday sounds to their
251 perception and to the formation of mental sound images, which can be stored as per-
252 ception/action ensembles. In short, voice and gesture are our natural sound sketching
253 instruments. This perspective on embodied sound representation sets the framework
254 within which the installation first, and the tool later were conceived.

255 4. Envisioning sound design by vocal sketching

256 For effective sound design, voice and gesture should act as entry points to vast sonic
257 spaces, such as those provided by sound synthesis models. In order to explore how vo-
258 cal utterances may be automatically converted to synthetic sounds, we conceived an
259 artistic installation, called *S'i' fosse suono*², where sixteen brief vocal self-portraits
260 are arranged in the form of an audiovisual checkerboard (Cera et al., 2016), depicted in
261 figure 1. The recorded non-verbal vocal sounds were used as sketches for synthetic ren-
262 derings, using physics-based modelling and corpus-based synthesis as reference sound
263 modelling techniques.

²In its web version, and without multi-touch support, *S'i' fosse suono* is available at <http://skatvg.eu/SIFosse/>

264 The conversion from vocal sketches to synthetic sound was done by Andrea Cera,
265 mostly based on his experience as a professional sound artist–designer. The artist–
266 designer was instructed to make exclusive use of the two sound–modelling approaches
267 to produce the sound prototypes. Only basic editing operations, such as layering, and
268 amplitude envelope, were allowed. Sound processing such as reverb, compression,
269 equalisation and modulations were retained only for embellishment in the completion
270 of the sounds. The available palette of physics-based sound models included fric-
271 tion, crumpling, impact, fluid flow, air turbulences, electric motor, and combustion
272 engine. Conversely the sound databases for the corpus-based granular synthesis have
273 been populated with the original vocal recordings and with the corresponding physics-
274 based sound realisations. Andrea Cera deployed a set of sound descriptors, such as
275 pitch tracking, onset detection, envelope follower, and several statistical moments of
276 the spectrum to extract some basic profiles from the vocal sounds to drive the sound
277 synthesis. In other cases, he interpreted the sound morphology of the vocal record-
278 ings to control the synthesis parameters. The artist–designer wrote a description of the
279 sound design process for each audio self-portrait, and these records were collected³.

280 The constraints given to the artist–designer in terms of usable sound models, and his
281 use of some automatic feature extractors, make the eventual automation of the process
282 relatively easy to foresee. The art installation was conceived to envision how vocal
283 sketching may be used to design sounds with a vast timbral palette, as it is given by
284 versatile sound synthesis models.

285 The rationale for this process can be derived from two perspectives: situated ontol-
286 ogy of design (Gero & Kannengiesser, 2014), and embodied sound cognition (Leman,
287 2008).

288 Indeed, the creation of *S'i' fosse suono* can be situated in three worlds (Gero &
289 Kannengiesser, 2014) or stages:

- 290 1. **interpreted** world: Sixteen participants imagine a sonic self-portrait in terms of
291 perception-action associations;

³See the Appendix to the manuscript for further details on the sound design strategy in vocal to synthetic conversion, and design implications.



Figure 1: *S'i' fosse suono*, with two sliders superimposed for evaluation

- 292 2. **expected** world: Each participant sets a vocal motor program to perform vocal-
- 293 izations that translate sonic imagination into action;
- 294 3. **external** world: The vocalisations are communicated to the artist–designer, who
- 295 interprets them as blueprints for synthetic sound composition.

296 Human agents with different roles have been playing in worlds 1 and 2 (partici-

297 pants), and in world 3 (artist–designer). In the envisioned tool-mediated sound design

298 process, instead, the sound sketcher would be playing in worlds 1 and 2, that is imag-

299 ining a sound first, and attempting to externalise the corresponding vocal articulation

300 later, while the translation of vocal blueprints into new sounds of the external world

301 would be performed by a machine. *S'i' fosse suono* gives joint access to the expected

302 world (i.e., the participants’ vocalisations) and to the external world (i.e., the synthetic

303 conversions by the designer-machine), as the audio-visual checkerboard chooses ran-

304 domly if playing back the vocal utterance or one of its two synthetic renderings.

305 In the framework of embodied sound cognition and mediation technology (Leman,
 306 2008), stages 1 and 2 can be associated with a *first-person* perspective, where sounds
 307 take the status of intentional actions. Stage 3 is that of a *third-person* perspective,
 308 where phenomena can be measured and translated, either by a sound designer or by
 309 a machine. The installation is experienced from a *second-person* perspective, where
 310 the observer gets involved in a context of intersubjective communication, as depicted
 in the resulting triangulation in Figure 2. From a phenomenological standpoint, if the

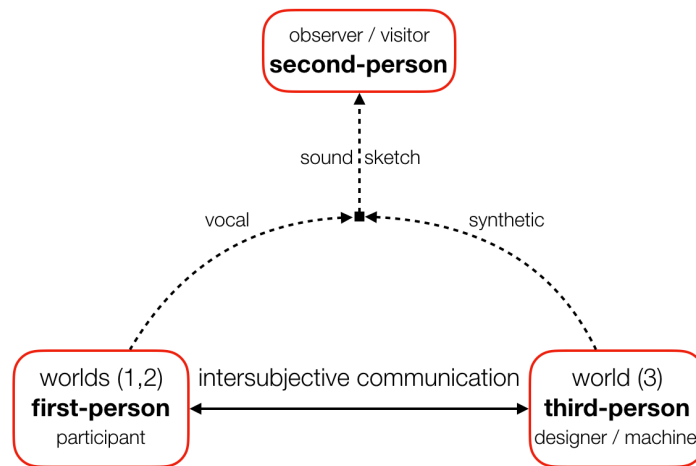


Figure 2: The observer forms an opinion on whether both vocal sounds and their synthetic translations represent plausible motor actions that can be ascribed to the visuo-spatial, intentional articulation of the mouth by the character-participant.

311
 312 observer experiences the sound sketch as made of actions with an intention, then the
 313 self-portrait communication act is found to be successful. If this *binding by causal-*
 314 *ity* (Schutz & Kubovy, 2009) occurs for both the vocal sketches and their synthetic
 315 translations, the effectiveness of the voice for sound sketching is demonstrated. Con-
 316 versely, the observer would experience the sound and the visual articulations as two
 317 separated, although synchronised events, not the result of an intention.

318 4.1. Reception

319 *S'i' fosse suono* was first exhibited at the ICT Conference of the European Commis-
 320 sion in Lisbon, Portugal, on October 20-22, 2015. In this and other public exhibitions

321 hundreds people interacted with the multi-touch screen installation. From informal
322 conversations with the interacting visitors, it emerged that the three sound realisations
323 were equally effective in binding the sensory information to spatial entities and tem-
324 poral events, i.e., at forming audio-visual objects (Kubovy & Schutz, 2010). The ac-
325 tion/sound association, that is the temporal simultaneity and spatial coincidence of the
326 visible and audible information, was found to be plausible and strong, which comes
327 at no surprise since the synthetic sounds were derived and causally consistent with a
328 recorded vocal utterance.

329 In the making of the installation, the artist–designer acted as a probe, to explore
330 the vast space of possible renditions of vocal utterances. In approaching the sound
331 conversion task, he embraced either an acousmatic or a concrete attitude, alternatively,
332 depending on whether the original vocal production was recognisable as an imitation of
333 an everyday sound phenomenon. In the acousmatic attitude (Chion, 1994), the conver-
334 sion strategy was abstracted from the physical information that could be derived from
335 the audiovisual recordings. The focus rather lies on the sound morphology *per se*. In
336 the concrete attitude, the nature of the reference recording prompted the artist–designer
337 to consider an everyday sound event.

338 In the practice, the two attitudes reflected the creative approach towards sound syn-
339 thesis. The concrete attitude called for a rather indexical use of the sound models, e.g.
340 vocal imitations of impacts or explosions were translated in corresponding synthetic
341 sound events, likewise. The abstract attitude instead stretches the possibilities of a
342 given sound model, by enhancing its sonic space⁴. More generally, this dual approach
343 to design is found during the conception of any product, in its visual, auditory, and
344 tactile properties (Özcan & Sonneveld, 2009). Although physical models seem to be
345 more suitable for concrete sound concepts, and granular synthesis more suitable for ab-
346 stract sound concepts, the two techniques were reported by the audience to be equally
347 effective in producing consistent and compelling sound realisations.

⁴The self-portraits in the fourth column from left, second and fourth rows from top in Figure 1 are good examples of vocal imitations prompting a concrete attitude. Rather, examples of abstract self-portraits are in the third column from left, first and second rows from top (<http://skatvg.eu/SIFosse/>).

348 *4.2. Naturalness and Concreteness of sound sketches*

349 Artist-designers can inform and inspire design, and the reactions from an audi-
350 ence exposed to their prototype realisations are valuable for steering project develop-
351 ments (Dunne & Gaver, 1997). But if the artist–designers are asked to work under
352 well-defined constraints, their prototypes can also be used to test some design assump-
353 tions, by formal experimentation.

354 In *S'i' fosse suono* the artist–designer used the two sound synthesis methods, intro-
355 duced and justified in Section 3: physics-based modelling and granular synthesis. In
356 order to see how their respective sonic spaces compare with each other and with the
357 space of vocal sketches, we ran an evaluation experiment, where participants positioned
358 each of the 48 audiovisuals (16 faces producing either vocal sounds or synthetic sounds
359 with one of the two synthesis methods) in a Naturalness vs. Concreteness space. In this
360 study, as opposed to other studies on musical and non-musical sounds (Dyck, 2016), a
361 sound is meant to be natural if it is perceived as coming from a human utterance; It is
362 concrete if it can be referred to a distal source.

363 *4.2.1. Apparatus*

364 The *S'i' fosse suono* installation was modified by adding a slider Natural ↔ Arti-
365 ficial and a slider Concrete ↔ Abstract, as in the rightmost column of figure 1. Apple
366 MacBook Pro laptops were used with Beyerdynamic DT770 Pro headphones. By using
367 the spacebar, the participant could play each audiovisual one by one, in random order.

368 *4.2.2. Participants and Task*

369 15 persons (six female, average age = 27.3 years, SD = 7.3) voluntarily participated
370 to the experiment, which did not last more than 15 minutes. They had to go through
371 the 48 stimuli (in randomised order) and, for each stimulus, rate its naturalness and
372 concreteness. The following explanation of the two scales was preliminarily given to
373 each participant: “An audiovisual is natural if the sound-image ensemble is a credible
374 human action. An audiovisual is artificial if the sound-image ensemble contains ele-
375 ments that have been clearly contrived by art. A sound is concrete if it is capable of
376 evoking a physical cause. A sound is abstract if it cannot be associated to a physical

377 generating event. For example, a person producing a bell noise would be artificial yet
378 concrete.”

379 4.2.3. Hypotheses

380 We expected that participants would be able to distinguish the sounds that are di-
381 rectly coming from a vocal production (natural) from those that are synthetic (artifi-
382 cial).

383 We expected that both physics-based and granular synthesis would be able to gen-
384 erate an ample range of sounds, some with an identifiable mechanical cause (concrete),
385 some very difficult to describe in physical terms (abstract). However, physics-based
386 models may be more biased toward concrete sounds, as compared to granular models.

387 Since the human voice can make sounds that are recognised as non-vocal (Lemaitre
388 & Rocchesso, 2014), it should be effective in producing both concrete and abstract
389 sounds, so vocal sketches should be rated in a wide range of concreteness.

390 4.2.4. Results

391 Figure 3 shows the mean ratings for the three families of sounds used in the instal-
392 lation.

393 Vocal sounds are perceived as more natural than both types of synthetic sounds
394 (vocals vs. physical models, $t_{15} = 20.0$, $p < .0001$). In our definition of naturalness,
395 this shows that listeners clearly perceived vocal sounds as produced by the voice, and
396 synthetic sounds as not produced by the voice. Both synthesis methods are perceived as
397 equivalently unnatural ($t_{15} = -1.8$, $p = .091$). Vocal sounds are also perceived as more
398 concrete than synthetic sounds (vocal vs. physical models, $t_{15} = 7.1$, $p < .0001$), and
399 both synthesis methods are perceived as equivalently abstract ($t_{15} = 0.57$, $p = 0.58$).
400 These findings are confirmed by non-parametric Friedman tests, indeed more appro-
401 priate for rating scales (naturalness: $\chi^2 = 24.13$, $p < .0001$, $df = 2$; concreteness:
402 $\chi^2 = 18.38$, $p < .0005$, $df = 2$).

403 Two-samples F-tests for equal variances show that both synthesis methods also
404 cover range sizes of concreteness that are not significantly different from each other
405 and from the range size of vocal sounds (vocal vs. physical models: $F_{15,15} = 0.46$,

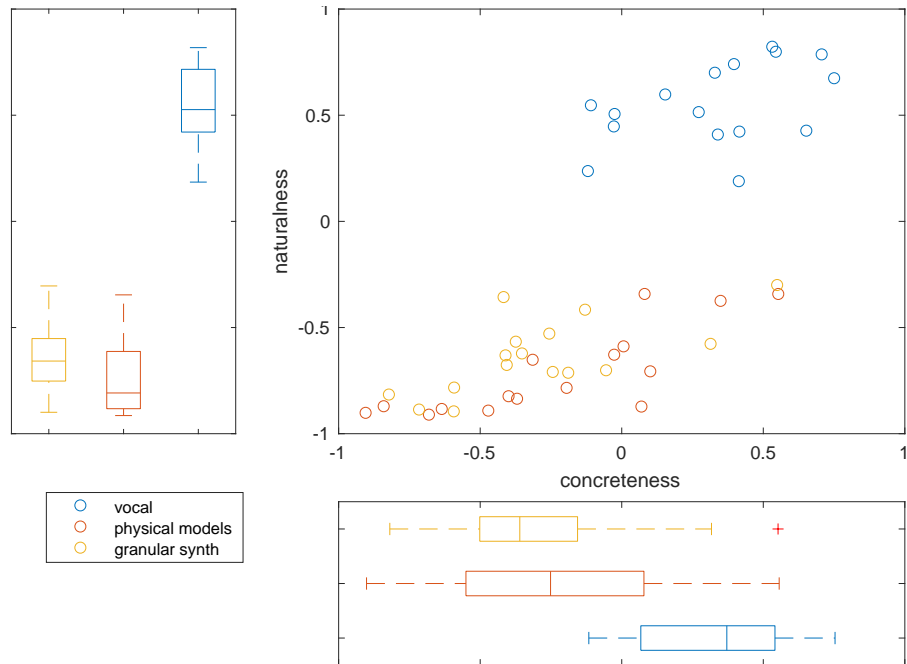


Figure 3: Scatterplot of the mean ratings of the 48 sounds of *S'i' fosse suono*, divided in the three groups of vocal, physics-based, and granular sounds

406 $p = .14$; physical models vs. granular synth: $F_{15,15} = 1.4$, $p = .50$). This suggests that
 407 the synthesis methods are as flexible as the voice to produce sounds that range from
 408 concrete to abstract (with a strong bias toward abstract sounds).

409 All participants were asked to comment freely after the test, and to express how
 410 difficult the task was. They generally found it easy to assess naturalness by thinking if
 411 they could have produced that sound as well, i.e. by a sort of embodied listening. Con-
 412 versely, several participants found it difficult to rate concreteness and reported some
 413 confusion between the attributes concrete and natural. This may explain the difference
 414 of the means of distributions of synthetic and vocal sounds along the concreteness axis.
 415 Nevertheless, they were able to distribute the examples of all three classes on a wide
 416 range of values of concreteness.

417 Overall, by confirming the hypotheses, the experiment also confirms the informal
 418 observations made in exhibitions and the impressions collected from the public, briefly

419 reported in Section 4.1.

420 In addition, the records by the artist-designer highlighted that sound renderings
421 were effectively produced by layering no more than three different sound models; that
422 onset, pitch, amplitude, and brightness were the most recurrent voice descriptors; and
423 that these were coupled to those synthesis parameters which could produce similar
424 timbral effects with minimum effort (i.e., corresponding variations in energy, pitch, and
425 density of the synthetic sounds). Thus, *S'i' fosse suono* represents a proof-of-concept,
426 whose assessment was used to specify the rationale of the sketching system. In the
427 Section 5, we first discuss the rationale of the system for embodied sound sketching,
428 then we describe its actual implementation together with further details on the control
429 strategies adopted.

430 **5. SEeD: Sound EmbodiEd Design**

431 In designing a tool to facilitate embodied sound design, we essentially tried to re-
432 place the artist–designer, as he was acting as a creative sound expert in the development
433 of *S'i' fosse suono*, with a semi-automated system that helps translating vocal and ges-
434 tural signals into synthetic sound. As synthesis techniques, we kept both physics-based
435 and corpus-based modelling (including granular and/or concatenative synthesis), as
436 they directly match the reported concrete or acousmatic attitudes of the artist–designer.
437 The effectiveness and versatility of both techniques have been experimentally assessed
438 with *S'i' fosse suono*.

439 *5.1. Design*

440 SEeD (Sound Embodied Design) is a system for sketching synthetic sound, based
441 on physics-based sound models and corpus-based synthesis. From the user’s perspec-
442 tive, and regardless of the underlying sound synthesis models, SEeD is structured into
443 two main modes: `set` and `play` (see Fig. 4).

444 In the `set` mode, the microphone captures a signal $u(t)$ of a vocal utterance and the
445 inertial measurement unit (IMU) captures movement signals (acceleration, orientation,
446 etc.) $z(t)$. Based on these signals, the system automatically proposes:

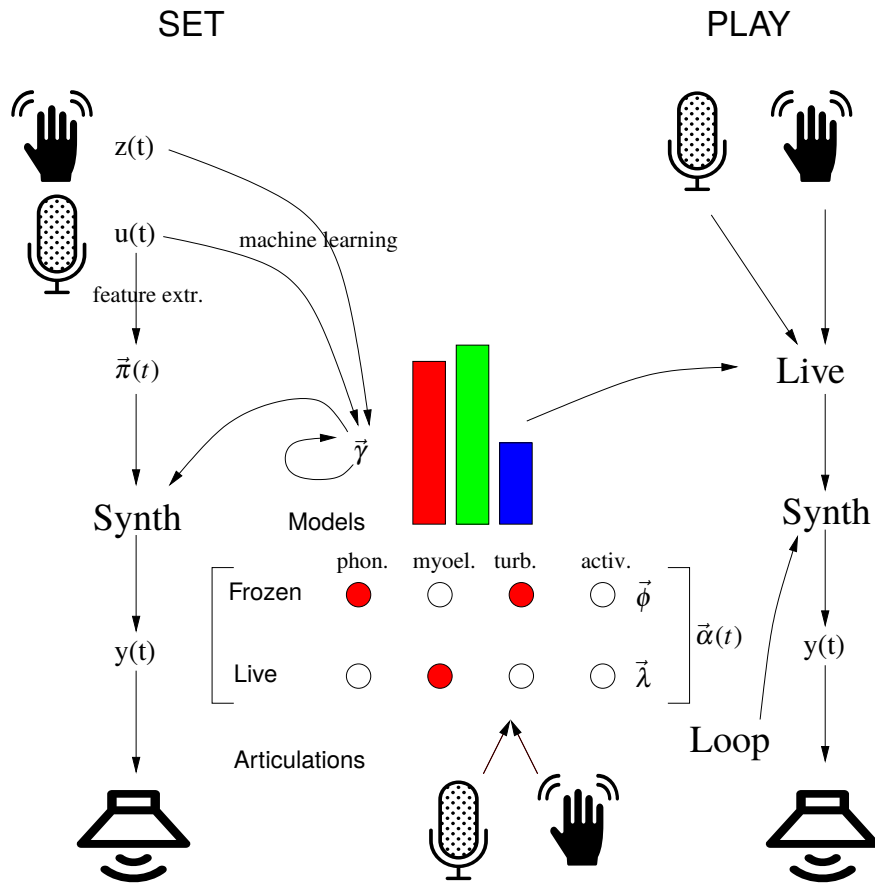


Figure 4: Sketch of SEED

- 447 1. a vector of weights $\vec{\gamma}$ (coloured bars in Fig. 4), giving the relative importance of
 448 different sound synthesis models (machine learning),
- 449 2. a vector of functions of time $\vec{\pi}$, with each element representing the temporal
 450 evolution of signal features (feature extraction).

451 Given the models weights and the features trajectories, the system returns a subset
 452 of its sound models and controls them to generate a synthetic sound similar to the
 453 input utterance. In line with the rationale of *S'i' fosse suono*, the set mode replaces
 454 that cognitive transition of the second-person perspective, in which the human agent,

455 the artist-designer, makes acousmatic or concrete hypotheses on the sound models and
456 corpora needed to return a synthetic impression of the given utterance. This operation
457 mode represents the moment in which the intersubjective communication between the
458 user and the machine is established. Indeed, the output $y(t)$ is an audio signal that goes
459 directly to the speakers and gives immediate feedback to the user, a sort of synthetic
460 echo to the proposed vocalisation.

461 Users can listen to such feedback and are given the possibility to change the model
462 weights $\vec{\gamma}$, to give more or less importance to a specific sound model. Each time the
463 user changes one of the model weights, a new feedback sound is produced. To make
464 the system relatively simple and usable even with a very limited graphical interface, it
465 is not possible to change values in $\vec{\pi}$ or to select a model which has not been included
466 in the $\vec{\gamma}$ vector, unless a new selection is made.

467 In fact, even though there are several available sound models, only the few (three)
468 that are ranked as most likely by the classifier are retained and their relative weights
469 γ_i get visualised through coloured sliders that are potentially modifiable. Indeed, the
470 experience of the artist-designer in *S'i' fosse suono* showed that three different sound
471 models at a time were enough to return the salient features of the vocal expressions. The
472 proposed system retains that economy of means. Labels will indicate to which model
473 each slider corresponds to but, in order to help the user with more stable configurations,
474 the relative order and color of models will be preserved. For example, “wind” will
475 always be red and be found on the left of “liquid”, which will always be blue. Notice
476 that this example works for both physics-based models of wind and liquids, and for
477 corpora of wind and liquid samples. Nothing prevents, however, to use arbitrary sound
478 corpora, either previously produced with physical models or simply having no relation
479 with the classes of physical phenomena.

480 The gestural part, coming from the reading of inertial sensors, provides gestural
481 signals $z(t)$ which may affect the choice of $\vec{\gamma}$. For example, a shaky gesture can give
482 more weight to a noisy model, whereas a continuous smooth gesture generally indicates
483 steady tonal sound (Scurto et al., 2016).

484 In play mode, $\vec{\gamma}$ is used to mix the output of sound models that will be control-
485 lable in real time through voice and gesture, according to the control layer of each

486 model. This modality does not require any visual interface. In the case of physical
487 models, each model is provided with a control layer that can perform a real-time map-
488 ping of vocal and gestural features into model parameters. In the case of corpus-based
489 synthesis, each sound sample is selected to match a particular set of sound descrip-
490 tors. Precisely, the voice is analysed resulting in a time-morphology vector of sound
491 descriptors (loudness, spectral centroid, noisiness, etc.) (Marchetto & Peeters, 2015).
492 The corpus-based synthesis corresponds thus to rendering a similar sound morphology
493 using the corpus samples.

494 In the play mode there is another submode called Loop, which requires both weights
495 $\vec{\gamma}$ and feature functions $\vec{\pi}$. For the sake of simplicity and effectiveness of control, the set
496 of feature functions is indeed partitioned into four articulatory controls $\vec{\alpha}(t)$: Phona-
497 tion, myoelasticity, turbulence, activity (Peeters et al., 2015). The synthetic sound
498 generation is looped, and through the $\vec{\lambda}$ selectors some elements of $\vec{\alpha}$ can be replaced
499 by live vocal and gestural control. Through the $\vec{\phi}$ selectors, some articulatory controls
500 can be frozen during looping.

501 As an example, if α_i is the turbulent component of the vocal utterance, controlling
502 α_i live allows the user to influence the turbulence-related synthesis parameters of the
503 model with his or her voice and gesture. More than one feature can be controlled live
504 at any given time, while keeping the others as they were recorded in the set mode, or
505 frozen. In this way, the human limitations in controlling multiple features (Lemaitre
506 et al., 2016b) can be overcome.

507 Gestures might also be used to temporally unfold the loop, controlling the reading
508 of the loop table, as in the mapping-by-demonstration method (Françoise, 2015). In
509 this case, the example gesture $z(t)$ that is recorded in the set stage is coupled with $u(t)$
510 by means of a Hidden Markov Model. This relation can then be used to regenerate
511 the sound descriptors while replaying the gesture. Performing the gesture with some
512 variations will generate variations in the sound descriptors, and consequently in the
513 final sound sketch.

514 *5.2. Implementation*

515 SEeD is essentially a modular Cycling'74 Max patch, which implements the dia-
516 gram in Figure 4 in most of its components⁵:

- 517 (1) `concat` (for corpus-based) and `physmod` (for physics-based) synthesis techniques
518 can be dynamically switched throughout `set` and `play` modes; once a mixture
519 of sound classes is set, its relative weights $\vec{\gamma}$ affect either the `physmod` sound
520 models or the `concat` sound corpora. In `play` mode, the `live` submode allows
521 to control the sound synthesis directly through vocalisations, while in `loop` sub-
522 mode the previously recorded stream $\vec{\pi}$ of audio descriptors is used to drive the
523 sound models. Additionally, in `loop` submode the stream $\vec{\pi}$ can be further re-
524 placed by a new recording, and yet without affecting the current `set` and weight-
525 ing.
- 526 (2) A Gaussian Mixture Model classifier (Françoise et al., 2014) is trained with the
527 user-provided vocal imitations of eight sound classes corresponding to the eight
528 sound models / corpora available, and used in the realisation of *S'i' fosse suono*:
529 blowing, car engine, crumpling, electric motor, hitting, liquid dripping, rubbing-
530 scraping, and shooting;
- 531 (3) During the `set` operation, the best three sound models and their relative contri-
532 butions (`weights`) are displayed. Eventually their balance in the mixture can be
533 adjusted manually on the GUI. Similarly, these weighting and tweaking apply to
534 the classes of sound samples used for the granular synthesis;
- 535 (4) An `articulation` control window allows to `freeze`, `loop`, or act `live` on ar-
536 ticulatory features. These features are integrated to give a high-level description
537 of vocalisations in terms of phonation, turbulence, myoelasticity, and general ac-
538 tivity. This layer allows to tailor to some extent the system responsiveness to
539 one's own vocal characteristics.

⁵See the accompanying video for the SEeD system at work.

540 The sound models used in the *physmod* section belong to the palette of physics-
541 based synthesisers available in the Sound Design Toolkit (SDT) (Baldan et al., 2017).
542 Voice descriptors include a pitch tracker, spectral characteristics (magnitude, centroid,
543 spread, skewness, kurtosis, flatness, flux, and onset), envelope follower, zero cross-
544 ing, and a detector of low-frequency vibrations. A subset of these descriptors is used
545 to provide articulatory controls (i.e., phonation, turbulence, myoelasticity) that can be
546 associated to the synthesis parameters. Such associations were designed by critically
547 looking at vocal input – sound output relationships emerged in the sound design process
548 of *S'i' fosse suono*. In particular, (i) Pitched vibrations in vocal activity are naturally
549 matched to control parameters affecting the emergence of pitched sounds. This is the
550 case of RPM in models of combustion engines and electric motors, and of bubble size
551 in the fluidflow model; (ii) Myoelastic articulations such as apico-alveolar or uvular
552 trills can be respectively associated to the parameters affecting the engine rumble or
553 the crushing energy in crumpling phenomena; (iii) Turbulent articulations are easily
554 associated to crumpling granularity or to explosions; (iv) The vocal activity provides
555 energy contours which are used to drive the throttle and the motor load in the engine
556 model or the wind speed in the air turbulence model. Using the *concat* engine (corpus-
557 based synthesis), the vocal contours expressed by audio descriptors are used to create
558 synthetic morphologies that are similar to the voice. As sound cannot be entirely de-
559 fined by the limited set of descriptors we use, the result depends also on the original
560 sound corpus. For example, a water sound corpus will produce sound morphologies
561 retaining some of the perceptual features of watery sounds.

562 5.3. Testing and development

563 The role of the artist–designer in *S'i' fosse suono* was crucial to envision the trans-
564 lation of the vocal sketches into synthetic sound, and how flexible this process should
565 be. Similarly, the development of SEED was further informed by interactions with
566 sound designers. Here we report and summarise the findings based on observations
567 and interviews with the sound practitioners that experienced the use of SEED in several
568 sound design workshops.

569 A preliminary release of SEeD was used in the *48-hours of sound design*⁶ in Château
570 La Coste, a workshop where five professional sound designers, varyingly active in the
571 fields of product sounds, animation movies, artistic installations, and auditory display,
572 were introduced to vocal sketching and asked to work with physics-based and corpus-
573 based sound models controlled by voice and gesture. In that embryonic version, the
574 system was actually made of two separate tools, one focused on the rather accurate
575 mimicking and synthesis (i.e., *physmod*), and the other focused on fostering the cre-
576 ative explorations of vocal utterances (i.e., *concat*). The complementary character of
577 the two tools emerged from the observed workflow, and it was further confirmed in
578 the debrief interviews with the sound designers, at the end of the workshop. In prac-
579 tice, *physmod* was found effective for the quick and rough production of sound ideas
580 through live vocal control, whereas *concat* was used at a later stage for creatively
581 shaping textural sounds, by means of loops and live gestures. For example, it was
582 suggested to integrate the switch from the *physmod* to the *concat* workflow by pop-
583 ulating the sound corpora with *physmod* sounds or classes pertaining to the available
584 sound models.

585 Voice and gesture-based interaction was found inherently fluent, facilitating adap-
586 tation and serendipity in the creation of raw sound materials. On the other side, the
587 sound designers stressed the inherent limitations of the sketching tool when coming to
588 mount the sound materials in a refined bundle: One main limitation was technical, that
589 is any successful sound design tool should afford a reliable integration in the workflow
590 of the sound designer's personal toolboxes. A second limitation is rather methodolog-
591 ical and refers to the fact that the proposed tool requires training and especially a new
592 approach towards sound creativity, i.e. the attitude to sketch-thinking with sound since
593 the very beginning of the project.

594 Embodied sound design tools consider the ambiguity of the sketch input as a re-
595 source, leaving the user potentially free to manage errors and conflicts later in the
596 process and even with other tools. As such, approaching the tools through the imita-
597 tive operation of established software for sound production is useless. It was reported

⁶<https://vimeo.com/169521601>.

598 how the embodied approach to sound design has the potential to modify the creative
599 process, by calling for a new methodology, and yet the necessary development of a
600 repertoire of practices and examples.

601 In a further iteration of the workshop on embodied practices for sonic sketching, the
602 participant industrial sound designers stressed how computer-mediated vocal sketch-
603 ing allowed them to experience positively a creative process which is far from their
604 everyday practice: Cooperation and action-reflection through contextual, live external-
605 isation of sound ideas improves team-building and communication. Finally, effective
606 user-centred, embodied sound design practices and tools have the potential to improve
607 sensibly the communication with customers and stakeholders.

608 Yet, despite the promising qualitative findings on the use of vocal sketching, we
609 wanted to have a clearer picture on the effectiveness of SEeD as technological support
610 to embodied sound design-thinking. For this purpose, we conceived a design exercise
611 in which we compared some SEeD-produced target sounds against the sketches repro-
612 duced by three sound designers, using their voice and the tool. The goal was too see if
613 the sketched representation is communicative of the originating sound. Ideally, voice-
614 driven configurations of synthetic sound models represent a more stable, yet dynamic
615 medium as compared to raw vocalisations and sound recordings. If the tool affords the
616 user to refer back to sounds (i.e., configurations of sound models), previously produced
617 by other peers, with a certain degree of reliability, then the bidirectional flow between
618 *perception* ↔ *articulation* ↔ *production* is established and mediated by the technol-
619 ogy. If so, the external sound representation exists, it is shared across internal mental
620 models and available to negotiation and collaborative practices.

621 **6. Evaluation**

622 The experimental evaluation of SEeD was set up around the following idea: To
623 provide some sound designers with several target sounds originally created with SEeD
624 and ask them to sketch the targets with the same tool, in a limited amount of time. The
625 goal was to observe whether the sketches would be closer to their corresponding targets
626 than to the other targets, by measuring the respective auditory distance. The reason for

627 using SEeD products as target sounds is that, in principle, a designer should be able to
628 reconstruct the target exactly. In practice, such perfect reconstruction is never achieved
629 in time- and tool-constrained sketching, and different sketchers may vary in their vocal-
630 sketching and tool-manipulation proficiency. This is a restricted form of evaluation, as
631 it only addresses the sketching effectiveness of the tool, and not its general usefulness
632 in a creative process.

633 *6.1. Experimental procedure*

634 *6.1.1. Setup*

635 The sound designers used a custom-made Cycling'74 Max v.7.3.1 user interface
636 of SEeD, which did not include the gesture-based interaction. The sound designers
637 were seated in a double-walled IAC sound isolated booth. The setup consisted of a
638 microphone (DPA d:fine omni), an audio interface (RME Fireface 400), a headphone
639 Beyerdynamic DT 250, and an Apple Mac Mini Intel Core 3 GHz, running MacOS
640 10.10.1 to record the sounds. The audio was recorded at a sampling rate of 44.1 kHz,
641 in 16 bits PCM Aiff files.

642 *6.1.2. Participants and procedure*

643 Three of the five sound designers who participated to the *48-hours of sound design*
644 were recalled six months later to test the new release of SEeD in a sketching task.

645 The user test was paced in four parts. First, each participant (P1, P2, P3) read the
646 instructions, the experimenter demonstrated the software and explained each different
647 sound synthesis model. Following, each participant trained the system with the vocal
648 examples of the eight classes of sound models available in the tool (and described
649 in Section 5.2). The participants spent 20 minutes of individual training, in order to
650 explore the tool with their voice, warm up and familiarise with the different sound
651 models. The experimenter was present to answer to questions. Before starting the
652 user test, the experimenter made a demonstration of the task. Each sound designer,
653 individually, was asked to reproduce 8 target sounds. The targets, 4 *physmod* and
654 4 *concat* sounds of a duration of 5 seconds each, had been previously created by
655 Andrea Cera with SEeD. In addition to the vocal control, both the designer of the

656 target sounds and the three participants were allowed to tweak manually the weights
657 of the sound classes, a reduced set of parameters of the sound models in `physmod`,
658 and the grain duration and the voice descriptors weights in `concat`. The target sounds
659 were presented in a random order, without disclosing how they had been generated
660 (i.e., `physmod` or `concat`). Essentially, the participants listened to the target and opted
661 for the most appropriate sound synthesis mode – `physmod` or `concat` – to sketch the
662 sound. The time slot to represent each target was constrained to 4 minutes.

663 6.1.3. Results

664 We analysed the results by isolating the final sketch for each participant and target
665 sound. We then computed the distance between each sketch and each target (as well as
666 the distances between the targets and between the sketches) for each sound designer.

667 The distance was adapted from the model of auditory distance created by Agus et al.
668 (2012). Originally, this model uses the time-frequency distribution of energy for each
669 sound, estimated using Spectro-Temporal Excitation Patterns (STEPs) that simulate
670 peripheral auditory filtering. Auditory distances are then computed by aligning pairs
671 of STEP time series using a dynamic time-warping algorithm. The cost of alignment
672 is used as the distance between two sounds. Here, we used auditory spectrograms
673 instead of STEP (Chi et al., 2005). To produce the auditory spectrogram, the acoustic
674 signal is analysed by a bank of constant-Q cochlear-like filters. The output of each
675 filter is processed by a hair cell model followed by a lateral inhibitory network, and is
676 finally rectified and integrated to produce the auditory spectrogram. Such a distance is
677 however sensitive to the duration of the sounds, and distances can only be compared
678 for sounds with the same duration. Therefore, signals were first zero-padded to the
679 duration of the longest sound. Sounds were normalised in amplitude.

680 Figure 5 represents the matrices of dissimilarities between the targets and the sketches,
681 as well as a hierarchical representation (dendrogram) of these distances, for each sound
682 designer (P1, P2, P3). For the sake of visualisation, the maximum value has been taken
683 from the matrices of all three participants, and the distances have been normalised to
684 such value. In each matrix $M = \begin{bmatrix} M_1 & M_2 \\ M'_2 & M_3 \end{bmatrix}$ the top-left 8-by-8 submatrix M_1 represents

685 the distances between target sounds, the bottom-right 8-by-8 submatrix M_3 represents
 686 the distances between the sketches, and the top-right 8-by-8 submatrix M_2 represents
 687 the distances between the sketches and the corresponding target sounds. Ideally, if the
 688 sketches would be identical to their target sounds then all the three submatrices would
 689 be identical. More loosely, if each sketch would be similar to its target then a dark
 690 diagonal would emerge in the top-right submatrix.

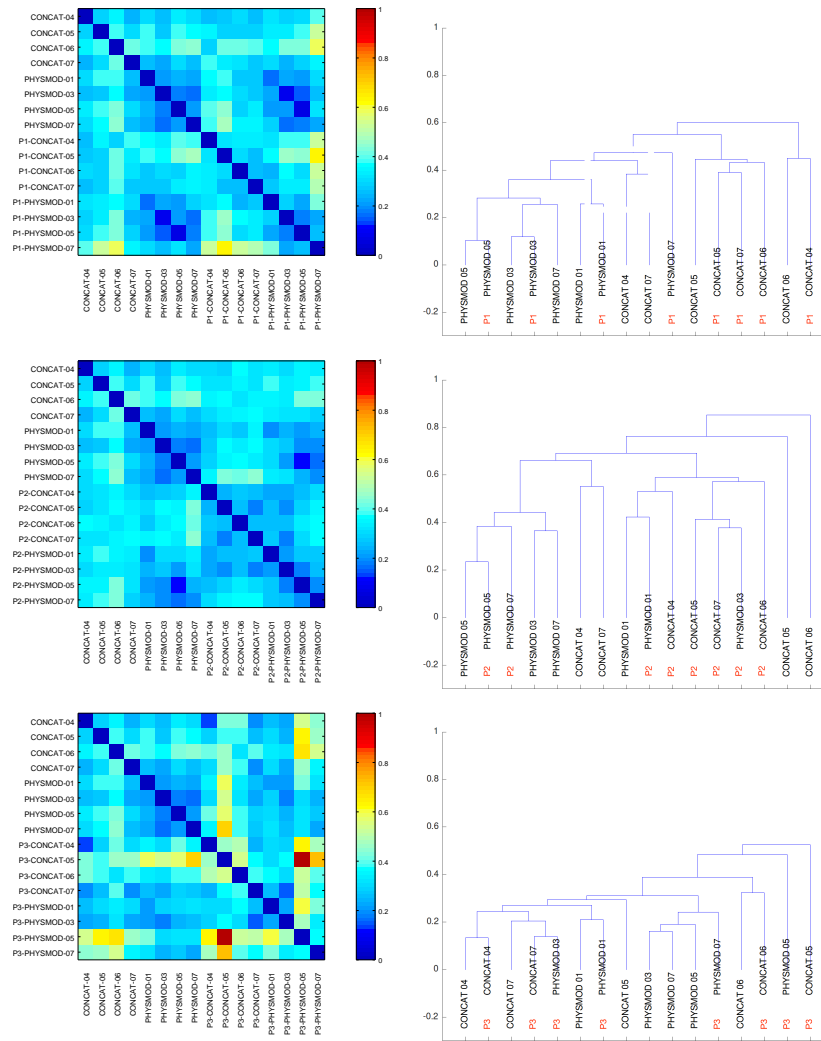


Figure 5: Left: Matrix of dissimilarities between the targets and the sketches. Right: Hierarchical representations of the distances. The results per each participant are arranged in rows (P1, P2, P3).

691 The visual inspection of the matrices shows that only a few sketches are actually
 692 close to their target. In particular, P1 created sketches that are close to their target for
 693 Physmod-05, Physmod-03, and Physmod-01; P2 created sketches that are close to
 694 their target for Physmod-05, and Physmod-01; P3 created sketches that are close to
 695 their target for Concat-04, and Physmod-01.

696 We can speculate that the closeness between the target and the sketch, when it
 697 occurs as a result of a tool-mediated replication, mark a conceptual space in which the
 698 mental models of the designer and the participant sketcher are shared and overlapping.

699 Table 1 shows the sound production mode used by the three participants to sketch
 700 each target sound. The choice of the production mode is a first indicator of the under-
 701 standing of the representing world embodied in the target sounds. In most of the cases,
 702 the sketchers used the same sound synthesis approach originally used in the targets. Yet
 703 of interest, sketches close to their targets could be achieved with different production
 704 modes as well.

	Physmod-01	Physmod-03	Physmod-05	Physmod-07	Concat-04	Concat-05	Concat-06	Concat-07
P1	physmod	physmod	physmod	physmod	concat	concat	concat	concat
P2	physmod	concat	physmod	physmod	physmod	concat	concat	concat
P3	concat	concat	physmod	physmod	concat	concat	concat	concat

Table 1: Participants' sketch production strategy for each target. The red texts highlight cases in which the sketched representation is made with a sound synthesis mode different from the one used for the creation of the target sound.

705 For instance, P3 represented the target Physmod-01 by using corpus-based sound
 706 synthesis (i.e., concat), effectively. P3 also preferred to use the concat mode to
 707 sketch the target Physmod-03. The resulting representation is also close to Concat-07,
 708 both target and its corresponding sketch. However, a closer inspection of P3 matrix and
 709 dendrogram shows how a certain ambiguity is already present in the relative distance
 710 between the targets Physmod-03 and Concat-07. P3 sketch of Physmod3 seems to
 711 preserve such a distance. Indeed, the same relative distance is preserved in P3 sketch
 712 of Concat-04, with respect to Concat-07 and Physmod-03, both targets and the cor-
 713 responding sketches.

714 A further similarity, or ambiguity, characterises the targets Physmod-03 and Physmod-07.
715 The corresponding sketches by P1 and P3 appear to preserve such a distance, both inter-
716 nally between sketches and externally with respect to the targets. The main difference
717 concerns the relative distance between the sketches, which reflects the diverse choice
718 of sound production mode by P1 and P3 for representing Physmod-03.

719 The analysis based on auditory distances is also useful to evaluate and compare
720 the performances of the three participants in the assigned task. By visual inspection
721 of the matrices in figure 5 we can argue that participant P1 was the most proficient, as
722 a partial dark diagonal in M_2 shows that there is auditory consistency between some
723 sketches and their corresponding target sounds. While bearing in mind that it is difficult
724 to generalise from only three analysed participants, we notice that participant P1 was
725 the most experienced between the three sound designers, and this can possibly justify
726 his highest proficiency.

727 The qualitative observations on the auditory distance matrices can be corroborated
728 by computation of the dissimilarity between the three submatrices of each participant.
729 A good measure of between-matrix dissimilarity is obtained via the so-called trace
730 norm $\|M\|$, which is given by the sum of absolute values of the eigenvalues of M .
731 For the three participants, the trace norms of their auditory distance submatrices are
732 reported in table 2. The column of $\|M_1 - M_3\|$ shows the capability of each participant
733 to achieve a set of sketches that has an internal metrical structure that is similar to that of
734 the target sounds (internal consistency). The two rightmost columns show the external
735 consistency, between the sets of sketches and the sets of target sounds. For participant
736 P1 the three matrices are more consistently similar to each other, thus showing that this
737 participant was the most successful in exploiting the tool to produce a set of sketches
738 that are both internally and externally consistent with the set of target sounds.

739 The SEeD tool supports to move back and forth from the vocal articulation to the
740 production of coherent synthetic impressions. The experiment shed light on the pos-
741 sibilities and limitations of the tool, in particular on the difficulties that designers may
742 meet when trying to represent a well-defined sound target with a sonic sketch. Further-
743 more, we ran *post-hoc* interviews with the participants and collected feedback on the
744 experience and the tool, in the light of future improvements.

Participant	$\ M_1 - M_3\ $	$\ M_2 - M_3\ $	$\ M_1 - M_2\ $
P1	1.80	1.78	1.80
P2	1.27	2.06	2.07
P3	2.90	2.15	1.98

Table 2: Degrees of dissimilarity between the submatrices of auditory distances.

745 *6.1.4. Post-experiment interviews*

746 The participants feedback on the sketching exercise and the tool is summarised be-
747 low. In general, the designers found the task quite hard. P2 reported that the complexity
748 of the sound morphology made it difficult to envisage the sound models and their con-
749 trol, thus suggesting that he opted for a re-production strategy based on acousmatic
750 approach rather than on the imitative behaviour. P1 stressed instead that the training
751 was essential for mastering the tool and turn sound ideas into vocalisations to drive the
752 synthetic representation.

753 Sketching by vocal imitation strategy with SEeD was found quite reliable, as the
754 participants reported that the system returned configurations of sound classes coherent
755 with the originating vocalisations (i.e., `set` mode). Yet, their major frustration derived
756 from the dynamic control of the sound models with the voice (i.e., `play` mode). P2
757 pointed out the effectiveness of the physics-based sound synthesisers, as they show a
758 rich and malleable user interface. Though, the manipulation of control parameters was
759 not always immediate.

760 The strengths and weaknesses of the physical models were found in their concrete-
761 ness. As their use was experienced by participants as more intuitive and immediate than
762 the corpus-based approach, the drawback of `physmod` is the apparently limited sound
763 palette. That was a contradictory observation by P1, which rather reflected a control
764 issue showed by the tool, that is the association between the articulatory streams and
765 the models parameters, in terms of width of the resulting sonic space. Conversely, the
766 participants stressed the creative potential of the `concat` synthesiser, especially in the
767 improvisation and exploration of textural sounds. The drawback of the abstractness of
768 this approach results in a less natural control. In particular, P1 reported the lack of the

769 pitch in the time-morphology vector of voice descriptors, which makes the exploration
770 of the sound corpora counter-intuitive. P1 suggested to harmonise the control strat-
771 egy of the two synthesis approaches. P3 reported the too many possibilities offered by
772 `concat` as a major difficulty preventing its immediate use.

773 Finally, the designers provided a feedback on the tool in general which reflected
774 their understanding and propensity towards an embodied approach to conceptual sound
775 design. Indeed, P1 recommended to even remove the labels of the sound classes in
776 order to avoid any possible influence on the creation process. On the other side, he
777 suggested to include pen-based interaction to manipulate the sound synthesis. P3 and
778 especially P2 showed a rather conservative attitude by suggesting the possibility to
779 select and weight the sound models manually.

780 The evaluation exercise shows that sketches created with SEeD can be tamed to
781 produce something predictable, and grounded in vocal motor skills and control. Simply
782 put in a visual analogy, we verified that the pencil (i.e., SEeD) allows to draw lines, and
783 that it can be used to produce something similar to a given drawing produced with the
784 same tool. In this respect, the experiences collected show two main loci of further
785 discussion and development.

786 One issue is technological and refers to the reliability and predictability of the tool.
787 Indeed, the `set` mode is individual centred and the classifier is trained to recognise the
788 imitations of a specific user. In this respect, as the user's intention and vocal motor
789 action unfold, the tool returns a synthetic representation coherently. The idiosyncrasy
790 shows up in `play` mode, wherein the customisation of the control layer is still heav-
791 ily constrained, making the exploration of the sonic space rather demanding. Further
792 development may include some kinds of adaptation of the control layer to the vocal
793 capabilities of the user. User's profiles could be stored and recalled with the classifier
794 training.

795 However, this possible design solution leads to the second locus of discussion,
796 which is methodological. Certainly user's profiles that are consistent with the individ-
797 ual centred `set` modes would partially solve the idiosyncrasy. Sketching (with SEeD),
798 that is design-thinking while making sonic representations, is not a sound selection task
799 whose final shape is left to the interpretation of the tool. Rather, it is an activity which

800 requires the fluency and expertise of the sketcher in order to access and empower im-
801 agery back and forth (?), during the conception of the sound design. When tools are
802 expressive, interaction becomes performative, yet not effortless. It requires practice
803 and training. In this respect, the further improvement of the tool goes together with
804 development of sketching practices of vocal scribbling. Ontologically-based studies of
805 vocal sketching protocols can reveal relevant information on the effectiveness of the
806 ideation process, and hence on the method and the tool (Gero & Kannengiesser, 2014;
807 Delle Monache & Rocchesso, 2016).

808 **7. Conclusion**

809 Sound design is an activity that is usually performed by experts who spend hours
810 manipulating the GUI elements of complex pieces of software. Such a practice is dis-
811 embodied from our vocal and gestural apparatus, which we naturally exploit to represent
812 and communicate sounds. But current practices may be disrupted by tools that make the
813 use of voice and gesture easy and direct. This contribution reports on the design pro-
814 cess that led to the implementation of SEeD, an embodied-sound-design tool. Before
815 putting a whole computational machinery (machine learning, sound synthesis, real-
816 time control) in a box, we developed and tested an artistic installation, where the role of
817 the machine was largely replaced by a sound artist–designer. Preliminary realisations
818 were also tested in professional settings, and this participatory design process con-
819 verged to a computational tool that gives users the freedom to sketch and refine sounds
820 using continuous vocalisations and gestures. Such freedom is indeed constrained by
821 the synthetic sonic space and control structure of the tool, just as choosing a set of
822 crayons and a certain paper would constrain a drawing act and give a material charac-
823 ter to a visual sketch. We expect that more tools for sonic sketching will be developed
824 in the future, aiming at immediacy of use and variety of results.

825 **8. Acknowledgments**

826 The work described in this paper is part of the project SkAT-VG, which received the
827 financial support of the Future and Emerging Technologies (FET) programme within

828 the Seventh Framework Programme for Research of the European Commission under
829 FET-Open grant number: 618067. Luca Ludovico and Davide Andrea Mauro con-
830 tributed to coding *S'i' fosse suono*. Most of the participants who donated their vocal
831 self-portraits are part of the theater group Cantiere Ca' Foscari, directed by Elisabetta
832 Brusa. The sound designers of the 48-hours of sound design in Château La Coste were
833 Simon Cacheux, Andrea Cera, Xavier Collet, Mathieu Pellerin, and Allister Sinclair.

834 **References**

- 835 Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recogni-
836 tion of musical sounds based on timbre. *The Journal of the Acoustical Society*
837 *of America*, 131, 4124–4133. URL: <http://dx.doi.org/10.1121/1.3701865>.
838 doi:10.1121/1.3701865. arXiv:<http://dx.doi.org/10.1121/1.3701865>.
- 839 Aramaki, M., & Kronland-Martinet, R. (2006). Analysis-synthesis of impact sounds
840 by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language*
841 *Processing*, 14, 695–705. doi:10.1109/TSA.2005.855831.
- 842 Baldan, S., Monache, S. D., & Rocchesso, D. (2017). The sound de-
843 sign toolkit. *SoftwareX*, 6, 255 – 260. URL: <http://www.sciencedirect.com/science/article/pii/S2352711017300195>. doi:<https://doi.org/10.1016/j.softx.2017.06.003>.
- 846 Becvar, A., Hollan, J., & Hutchins, E. (2008). Representational gestures as cognitive
847 artifacts for developing theories in a scientific laboratory. In *Resources, co-evolution*
848 *and artifacts* (pp. 117–143). Springer.
- 849 Cash, P., & Maier, A. (2016). Prototyping with your hands: the many roles of gesture
850 in the communication of design concepts. *Journal of Engineering Design*, 27, 118–
851 145.
- 852 Cera, A., Mauro, D. A., & Rocchesso, D. (2016). Sonic in(tro)spection by vocal sketch-
853 ing. In *Proc. XXI Colloquium on Musical Informatics*. Cagliari, Italy.

- 854 Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal anal-
855 ysis of complex sounds. *The Journal of the Acoustical Society of America*,
856 118, 887–906. URL: <http://dx.doi.org/10.1121/1.1945807>. doi:10.1121/
857 1.1945807. arXiv:<http://dx.doi.org/10.1121/1.1945807>.
- 858 Chion, M. (1994). *Audio-Vision: Sound on Screen*. New York: Columbia University
859 Press.
- 860 De Poli, G., Piccialli, A., & Roads, C. (Eds.) (1991). *Representations of Musical*
861 *Signals*. Cambridge, MA, USA: MIT Press.
- 862 Delle Monache, S., Polotti, P., & Rocchesso, D. (2010). A toolkit for explorations
863 in sonic interaction design. In *Proceedings of the 5th Audio Mostly Conference: A*
864 *Conference on Interaction with Sound AM '10* (pp. 1:1–1:7). New York, NY, USA:
865 ACM. URL: <http://doi.acm.org/10.1145/1859799.1859800>. doi:10.1145/
866 1859799.1859800.
- 867 Delle Monache, S., & Rocchesso, D. (2016). Understanding cooperative sound de-
868 sign through linkographic analysis. In *Proc. of the XXI CIM Colloquium on Music*
869 *Informatics* (pp. 25–32). Cagliari, Italy.
- 870 van den Doel, K., Kry, P. G., & Pai, D. K. (2001). Foleyautomatic: Physically-based
871 sound effects for interactive simulation and animation. In *Proceedings of the 28th*
872 *Annual Conference on Computer Graphics and Interactive Techniques SIGGRAPH*
873 *'01* (pp. 537–544). New York, NY, USA: ACM. URL: [http://doi.acm.org/10.](http://doi.acm.org/10.1145/383259.383322)
874 [1145/383259.383322](http://doi.acm.org/10.1145/383259.383322). doi:10.1145/383259.383322.
- 875 Dunne, A., & Gaver, W. W. (1997). The pillow: Artist-designers in the digital age.
876 In *CHI '97 Extended Abstracts on Human Factors in Computing Systems CHI EA*
877 *'97* (pp. 361–362). New York, NY, USA: ACM. URL: [http://doi.acm.org/10.](http://doi.acm.org/10.1145/1120212.1120434)
878 [1145/1120212.1120434](http://doi.acm.org/10.1145/1120212.1120434). doi:10.1145/1120212.1120434.
- 879 Dyck, J. (2016). Natural sounds and musical sounds: A dual distinction. *The Journal of*
880 *Aesthetics and Art Criticism*, 74, 291–302. URL: [http://dx.doi.org/10.1111/](http://dx.doi.org/10.1111/jaac.12286)
881 [jaac.12286](http://dx.doi.org/10.1111/jaac.12286). doi:10.1111/jaac.12286.

- 882 Eris, O., Martelaro, N., & Badke-Schaub, P. (2014). A comparative analysis of mul-
883 timodal communication during design sketching in co-located and distributed envi-
884 ronments. *Design Studies*, 35, 559–592.
- 885 Françoise, J. (2015). *Motion-sound Mapping By Demonstration*. Theses Université
886 Pierre et Marie Curie - Paris VI. URL: [https://tel.archives-ouvertes.fr/
887 tel-01206009](https://tel.archives-ouvertes.fr/tel-01206009).
- 888 Françoise, J., Schnell, N., Borghesi, R., & Bevilacqua, F. (2014). Probabilistic models
889 for designing motion and sound relationships. In *Proceedings of the International
890 Conference on New Interfaces for Musical Expression NIME* (pp. 287–292). URL:
891 <https://hal.archives-ouvertes.fr/hal-01061335/document>.
- 892 Franinović, K., & Serafin, S. (2013). *Sonic interaction design*. MIT Press.
- 893 Gabor, D. (1947). Acoustical quanta and the theory of hearing. *Nature*, 159, 591–594.
- 894 Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to
895 auditory event perception. *Ecological psychology*, 5, 1–29.
- 896 Gero, J. S., & Kannengiesser, U. (2014). The function-behaviour-structure ontology
897 of design. In A. Chakrabarti, & M. L. T. Blessing (Eds.), *An Anthology of The-
898 ories and Models of Design: Philosophy, Approaches and Empirical Explorations*
899 (pp. 263–283). London: Springer London. URL: [http://dx.doi.org/10.1007/
900 978-1-4471-6338-1_13](http://dx.doi.org/10.1007/978-1-4471-6338-1_13). doi:10.1007/978-1-4471-6338-1_13.
- 901 Goldschmidt, G. (2014). *Linkography: unfolding the design process*. MIT Press.
- 902 Helgason, P. (2014). Sound initiation and source types in human imitations of sounds.
903 In *Proc. Fonetik* (pp. 83–89). Stockholm, Sweden.
- 904 Houix, O., Lemaitre, G., Misdariis, N., Susini, P., & Urdapilleta, I. (2012). A lexical
905 analysis of environmental sound categories. *Journal of Experimental Psychology:
906 Applied*, 18, 52–80. doi:10.1037/a0026240.
- 907 Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. (2016). Auditory Sketches:
908 Very Sparse Representations of Sounds Are Still Recognizable. *PLoS ONE*,

909 11, e0150313+. URL: <https://hal.archives-ouvertes.fr/hal-01250175>.
910 doi:10.1371/journal.pone.0150313.

911 Kang, S., Tversky, B., & Black, J. B. (2015). Coordinating gesture, word, and diagram:
912 explanations for experts and novices. *Spatial Cognition & Computation*, 15, 1–26.

913 Kubovy, M., & Schutz, M. (2010). Audio-visual objects. *Review of Phi-*
914 *losophy and Psychology*, 1, 41–61. URL: [http://dx.doi.org/10.1007/](http://dx.doi.org/10.1007/s13164-009-0004-5)
915 [s13164-009-0004-5](http://dx.doi.org/10.1007/s13164-009-0004-5). doi:10.1007/s13164-009-0004-5.

916 Lemaitre, G., Dessein, A., Susini, P., & Aura, K. (2011). Vocal imitations and the iden-
917 tification of sound events. *Ecological Psychology*, 23, 267–307. URL: [http://dx.](http://dx.doi.org/10.1080/10407413.2011.617225)
918 [doi.org/10.1080/10407413.2011.617225](http://dx.doi.org/10.1080/10407413.2011.617225). doi:10.1080/10407413.2011.
919 617225. arXiv:<http://dx.doi.org/10.1080/10407413.2011.617225>.

920 Lemaitre, G., Heller, L. M., Navolio, N., & Zúñiga-Peñaranda, N. (2015). Priming
921 gestures with sounds. *PloS one*, 10, e0141791.

922 Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., & Susini, P. (2016a). Vocal im-
923 itations of non-vocal sounds. *PLoS ONE*, 11, e0168167. doi:10.1371/journal.
924 pone.0168167.

925 Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., & Susini, P. (2016b). Vocal imita-
926 tions of basic auditory features. *The Journal of the Acoustical Society of America*,
927 139, 290–300.

928 Lemaitre, G., & Rocchesso, D. (2014). On the effectiveness of vocal imitations and
929 verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135,
930 862–873. URL: [http://scitation.aip.org/content/asa/journal/jasa/](http://scitation.aip.org/content/asa/journal/jasa/135/2/10.1121/1.4861245)
931 [135/2/10.1121/1.4861245](http://scitation.aip.org/content/asa/journal/jasa/135/2/10.1121/1.4861245). doi:<http://dx.doi.org/10.1121/1.4861245>.

932 Lemaitre, G., Scurto, H., Françoise, J., Bevilacqua, F., Houix, O., & Susini, P. (2017).
933 Rising tones and rustling noises: Metaphors in gestural depictions of sounds. *PloS*
934 *one*, 12, e0181786.

- 995 Leman, M. (2008). *Embodied music cognition and mediation technology*. Cambridge,
996 MA: MIT Press.
- 997 Löwgren, J., & Stolterman, E. (2004). *Thoughtful Interaction Design: A Design Per-*
998 *spective on Information Technology*. The MIT Press.
- 999 Marchetto, E., & Peeters, G. (2015). A set of audio features for the morphological de-
1000 scription of vocal imitations. In *Proceedings of the 18th International Conference on*
1001 *Digital Audio Effects (DAFX-2015)* (pp. 207–214). Trondheim, Norway: Norwegian
1002 University of Science and Technology.
- 1003 Marchetto, E., & Peeters, G. (2017). Automatic recognition of sound categories from
1004 their vocal imitation using audio primitives automatically derived by si-plca and
1005 hmm. In *Proceedings of the International Symposium on Computer Music Multidis-*
1006 *ciplinary Research CMMR* (pp. 9–20). URL: [http://cmmr2017.inesctec.pt/
1007 wp-content/uploads/2017/09/1_CMMR_2017_paper_50.pdf](http://cmmr2017.inesctec.pt/wp-content/uploads/2017/09/1_CMMR_2017_paper_50.pdf).
- 1008 Nykänen, A., Wingstedt, J., Sundhage, J., & Mohlin, P. (2015). Sketching sounds–
1009 kinds of listening and their functions in designing. *Design Studies*, 39, 19–47.
- 1010 Özcan, E., & van Egmond, R. (2009). Product sound design: An inter-disciplinary
1011 approach? In *Undisciplined! Design Research Society Conference*. URL: [http:
1012 //shura.shu.ac.uk/531/](http://shura.shu.ac.uk/531/).
- 1013 Özcan, E., & Sonneveld, M. (2009). Embodied explorations of sound and touch in
1014 conceptual design. In *Design semantics of form and movement* (pp. 173–181).
1015 Taipei, Taiwan. URL: [https://www.northumbria.ac.uk/media/6278717/
1016 desform-2009-proceedings-v2.pdf#page=173](https://www.northumbria.ac.uk/media/6278717/desform-2009-proceedings-v2.pdf#page=173).
- 1017 Pualetto, S. (2017). The voice delivers the threats, foley delivers the punch: Embod-
1018 ied knowledge in foley artistry. In M. Mera, R. Sadoff, & B. Winters (Eds.), *The*
1019 *Routledge Companion to Screen Music and Sound* (pp. 338–348). Routledge.
- 1020 Pualetto, S., Cambridge, H., & Susini, P. (2016). Data sonification and sound
1021 design in interactive systems. *International Journal of Human-Computer Stud-*
1022 *ies*, 85, 1 – 3. URL: <http://www.sciencedirect.com/science/article/>

963 pii/S1071581915001299. doi:<http://dx.doi.org/10.1016/j.ijhcs.2015.>
964 08.005. Data Sonification and Sound Design in Interactive Systems.

965 Peeters, G., Françoise, J., Bevilacqua, F., Friberg, A., Marchetto, E., Brocal, G. M.,
966 Helgason, P., & Hellwagner, M. (2015). *Blind classifiers of imitations*. Deliverable
967 SkAT-VG project.

968 Purcell, A., & Gero, J. S. (1998). Drawings and the design process: A review of
969 protocol studies in design and other disciplines and related research in cognitive
970 psychology. *Design studies*, 19, 389–430.

971 Rocchesso, D., Lemaitre, G., Susini, P., Ternström, S., & Boussard, P. (2015). Sketch-
972 ing sound with voice and gesture. *interactions*, 22, 38–41. URL: [http://doi.acm.](http://doi.acm.org/10.1145/2685501)
973 [org/10.1145/2685501](http://doi.acm.org/10.1145/2685501). doi:10.1145/2685501.

974 Rocchesso, D., Mauro, D., & Delle Monache, S. (2016). miMic: The microphone as a
975 pencil. In *Proceedings of the TEI '16: Tenth International Conference on Tangible,*
976 *Embedded, and Embodied Interaction TEI '16* (pp. 357–364). New York, NY, USA:
977 ACM. URL: <http://doi.acm.org/10.1145/2839462.2839467>. doi:10.1145/
978 2839462.2839467.

979 Roden, D. (2010). Sonic art and the nature of sonic events. *Review of Phi-*
980 *losophy and Psychology*, 1, 141–156. URL: [http://dx.doi.org/10.1007/](http://dx.doi.org/10.1007/s13164-009-0002-7)
981 [s13164-009-0002-7](http://dx.doi.org/10.1007/s13164-009-0002-7). doi:10.1007/s13164-009-0002-7.

982 Schnell, N. (2011). *Real-Time Audio Mosaicing*, [http://imtr.ircam.fr/imtr/](http://imtr.ircam.fr/imtr/Real-Time_Audio_Mosaicing)
983 [Real-Time_Audio_Mosaicing](http://imtr.ircam.fr/imtr/Real-Time_Audio_Mosaicing). Technical Report IRCAM. URL: [http://imtr.](http://imtr.ircam.fr/imtr/Real-Time_Audio_Mosaicing)
984 [ircam.fr/imtr/Real-Time_Audio_Mosaicing](http://imtr.ircam.fr/imtr/Real-Time_Audio_Mosaicing).

985 Schön, D. A. (1984). *The reflective practitioner: How professionals think in action*.
986 Basic Books.

987 Schön, D. A., & Wiggins, G. (1992). Kinds of seeing and their functions in designing.
988 *Design studies*, 13, 135–156.

- 989 Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of*
990 *Experimental Psychology: Human Perception and Performance*, 35, 1791–1810.
- 991 Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE Signal Processing*
992 *Magazine*, 24, 92–104. doi:10.1109/MSP.2007.323274.
- 993 Scurto, H., Lemaitre, G., Françoise, J., Bevilacqua, F., Susini, P., & Voisin, F. (2016).
994 Embodying sounds: Building and analysis of a database of gestural and vocal imita-
995 tions. In *Gesture, Creativity, Multimodality. Proceedings of ISGS 7 : 7th Conference*
996 *of the International Society for Gesture Studies* (p. 269). The University of Texas at
997 Austin, Austin, TX: International Society for Gesture Studies.
- 998 Scurto, H., Lemaitre, G., Françoise, J., Voisin, F., Bevilacqua, F., & Susini, P. (2015).
999 Combining gestures and vocalizations to imitate sounds. *The Journal of the Acous-*
1000 *tical Society of America*, 138, 1780–1780.
- 1001 Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideo-
1002 motor theory. *Psychological Bulletin*, 136, 943–974.
- 1003 Strobl, G., Eckel, G., & Rocchesso, D. (2006). Sound texture modeling: A survey.
1004 In *Proceedings of the Sound and Music Computing (SMC) Conference* (pp. 61–65).
1005 Marseille, France.
- 1006 Susini, P., Houix, O., & Misdariis, N. (2014). Sound design: an applied, experimental
1007 framework to study the perception of everyday sounds. *The New Soundtrack*, 4,
1008 103–121.
- 1009 Tholander, J., Karlgren, K., Ramberg, R., & Sökjer, P. (2008). Where all the interac-
1010 tion is: Sketching in interaction design as an embodied practice. In *Proceedings of*
1011 *the 7th ACM Conference on Designing Interactive Systems DIS '08* (pp. 445–454).
1012 New York, NY, USA: ACM. URL: <http://doi.acm.org/10.1145/1394445.1394493>.
1013 doi:10.1145/1394445.1394493.
- 1014 Tuuri, K., Eerola, T., & Pirhonen, A. (2011). Design and evaluation of prosody-
1015 based non-speech audio feedback for physical training application. *Internat-*
1016 *ional Journal of Human-Computer Studies*, 69, 741 – 757. URL: <http://www>.

- 1017 sciencedirect.com/science/article/pii/S1071581911000760. doi:<http://dx.doi.org/10.1016/j.ijhcs.2011.06.004>.
- 1018
- 1019 Tversky, B., Heiser, J., Lee, P., & Daniel, M.-P. (2009). Explanations in gesture, dia-
1020 gram, and word. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial language*
1021 *and dialogue* chapter 9. (pp. 119–131). New York, NY: Oxford University Press.
- 1022 Verma, T., Levine, S., & Meng, T. (1997). Transient modeling synthesis: a flexi-
1023 ble analysis/synthesis tool for transient signals. In *Proceedings of the International*
1024 *Computer Music Conference (ICMC)* (pp. 48–51). Thessaloniki, Greece.
- 1025 Visser, W., & Maher, M. L. (2011). The role of gesture in designing. *Artificial Intelli-*
1026 *gence for Engineering Design, Analysis and Manufacturing*, 25, 213–220.