Embodied Sound Design

Stefano Delle Monache[a,*], DavideRocchesso[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

Interacting with and through representations is a key aspect of designers' activity. 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 (DavideRocchesso), 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)

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

Explanations, i.e. representations, emerge as multi-modal models from the continuous interplay between talk and action, and through the concurrent manipulation of sketches and diagrams, physical props and artefacts. Designers are fluent in combining utterances, drawings, props, and especially gestures to create and annotate models of a situation, systems, processes and configurations (Kang et al., 2015). Active engagement and performative action complement static forms of representation (e.g., white boards, drawings and diagrams), and allow to convey complex spatio-dynamic properties, motion, trajectories, and time-based events. As visuospatial forms of communication can represent concepts more directly than verbal descriptions, they are suitable to catch and express the dynamics of designs, behaviours and relations unfolding in time and space.

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

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

We propose an embodied account of the sound design process, which places the bodily experience (i.e., communication of sonic concepts through vocal and gestural imitations) at the center of the sound creation process. Vocalisations and gestures are the primary cognitive artefacts available to sound designers to explain concepts and sketch sonic representations cooperatively (Rocchesso et al., 2015; Delle Monache & Rocchesso, 2016). Vocal sketching is a fast prototyping technique, aimed at the early stage of sound design, which exploits the voice as means to portraying and imitating non-speech sounds. The seminal research workshop by **?** demonstrated the potential of such methodology. Yet, fully understanding how humans explain time-based processes, such as sonic concepts, through vocal and gestural imitations is essential to provide foundational training and communication support to sound practitioners in collaborative design scenarios (Rocchesso et al., 2016). Indeed, research in sound design thinking is largely unexplored, and the expertise of professionals rather seems to be based on individual paths in which music education, computer science, psychoacoustics and ultimately design intertwine to form the tacit knowledge of the practice (Özcan & van Egmond, 2009).

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

In embodied sound design, spatial forms of communication and non-verbal utterances intertwine in the formation and explanation of auditory objects (Kubovy & Schutz, 2010). Despite their ephemerality, vocal and gestural sketches have a representational stability over time, showing a coherent mapping between the representing world and the represented world, and providing material displays that can be built upon and recalled during the stream of a discourse (Becvar et al., 2008; Delle Monache & Rocchesso, 2016; Lemaitre et al., 2016a; Scurto et al., 2016). Connections with Foley artistry are easily found here (Pauletto, 2017), as the sound generation is inherently embodied, but embodied sonic sketching aims at exploratory actions that could be compared to scribbling with pencil on paper.

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

The manuscript is organised as follows. In Section 2 we set the theoretical background, by stressing the sensory-motor nature of auditory experiences. Then, in Section 3, we approach the problem of the embodied representation of sound from the three-folded perspective of sound perception, production, and articulation of voice and gestures. In Section 4, we describe an artistic audio-visual installation that we realised to explore the flow of the embodied sound design process, from the internal sonic concept to its synthetic rendering, via the translation into actions (i.e., vocal sketches). The representing world, that is the vocal sketches, and the represented world, namely the synthetic counterparts, were compared and experimentally assessed in terms of naturalness and concreteness of the representation at hand. Eventually, the artistic installation and its evaluation were functional to the development of SEeD, a semi-automated system for the conversion of continuous vocalisations into synthetic sounds. The system architecture, and its rationale are described in detail in Section 5, together with the development that was driven by some design workshop experiences. Finally, the consistency of SEeD as a sketching tool is assessed in Section 6, where we report the experimental evaluation through a sound design exercise: We asked three professional sound designers to sketch sounds that are similar to a set of examples produced with SEeD itself.

2. Theoretical framing

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

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

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

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

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

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

3. Sound approaches to sound

In designing synthetic sounds, there is an unresolved dilemma that the designer or artist has to face: How to approach sound and its representations (De Poli et al., 1991; Roden, 2010). Should we design sounds as they appear to the senses, by manipulating their proximal characteristics? Or should we rather look at potential sources, at physical systems that produce sound as a side effect of distal interactions? Can our body help establishing bridges between distal (source-related) and proximal (sensory-related) representations? Can the intimate space of vocal and gestural articulations be used to drive sound synthesis? Giving answers to these questions corresponds to choosing appropriate sound models and preparing a set of effective sound synthesis tools. These developments can be informed by research findings in perception, production, and articulation of sounds. An embodied approach to sound design should exploit knowledge in these areas, especially referring to voice and gesture, to create action-oriented ontologies (Leman, 2008).

3.1. Perception

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

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

Recent research has shown that vocal imitations can be more effective than verbalisations at representing and communicating sounds when these are difficult to describe with words (Lemaitre & Rocchesso, 2014). This indicates that vocal imitations can be a useful tool for investigating sound perception, and shows that the voice is instrumental to embodied sound cognition. When using vocal imitations, it must be considered that there is the human individual at the center of the scene, with her preferences, limitations, and idiosyncrasies. This makes the couples sound/imitation highly subjective, but ensures the highest level of embodiment of the sonic space. When using vocal imitations to drive perception-based synthesisers, the resulting perception/action sound synthesis is tightly connected to embodied representations of sound, especially if voice control can be properly individualised.

3.2. Production

3.2.1. Physics-based modelling

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

3.2.2. Corpus-based modelling

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

3.3. Articulation

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

3.3.1. Voice

Phonation, Turbulence, and Myoelasticity[1] are three important components of vocal imitations that can be activated independently, and therefore be present simultaneously (Helgason, 2014). Another component may be transient/click/impulse or, alternatively, this may be aggregated with myoelasticity, as low-frequency oscillations of articulatory mechanisms can often be viewed as sequences of discrete pulses. These component features can be extracted automatically from audio with time-frequency analysis and machine learning (Peeters et al., 2015). They can be made to correspond to categories of sounds as they are perceived and as they are produced in the physical world. Lemaitre et al. (2016b) showed that naïve imitators can accurately match the relative pitch, temporal behaviour, and spectral centroid of their vocalisations to the

[1] For the sake of this research, myoelastic vocal fold vibrations contribute only to the phonation component. 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.

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

3.3.2. Gesture

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

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

4. Envisioning sound design by vocal sketching

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

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

The conversion from vocal sketches to synthetic sound was done by Andrea Cera, mostly based on his experience as a professional sound artist–designer. The artist–designer was instructed to make exclusive use of the two sound–modelling approaches to produce the sound prototypes. Only basic editing operations, such as layering, and amplitude envelope, were allowed. Sound processing such as reverb, compression, equalisation and modulations were retained only for embellishment in the completion of the sounds. The available palette of physics-based sound models included friction, crumpling, impact, fluid flow, air turbulences, electric motor, and combustion engine. Conversely the sound databases for the corpus-based granular synthesis have been populated with the original vocal recordings and with the corresponding physics-based sound realisations. Andrea Cera deployed a set of sound descriptors, such as pitch tracking, onset detection, envelope follower, and several statistical moments of the spectrum to extract some basic profiles from the vocal sounds to drive the sound synthesis. In other cases, he interpreted the sound morphology of the vocal recordings to control the synthesis parameters. The artist–designer wrote a description of the sound design process for each audio self-portrait, and these records were collected[3].

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

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

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

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

[3] See 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

2. **expected** world: Each participant sets a vocal motor program to perform vocalizations that translate sonic imagination into action;

3. **external** world: The vocalisations are communicated to the artist–designer, who interprets them as blueprints for synthetic sound composition.

Human agents with different roles have been playing in worlds 1 and 2 (participants), and in world 3 (artist–designer). In the envisioned tool-mediated sound design process, instead, the sound sketcher would be playing in worlds 1 and 2, that is imagining a sound first, and attempting to externalise the corresponding vocal articulation later, while the translation of vocal blueprints into new sounds of the external world would be performed by a machine. *S'i' fosse suono* gives joint access to the expected world (i.e., the participants' vocalisations) and to the external world (i.e., the synthetic conversions by the designer-machine), as the audio-visual checkerboard chooses randomly if playing back the vocal utterance or one of its two synthetic renderings.

In the framework of embodied sound cognition and mediation technology (Leman, 2008), stages 1 and 2 can be associated with a *first-person* perspective, where sounds take the status of intentional actions. Stage 3 is that of a *third-person* perspective, where phenomena can be measured and translated, either by a sound designer or by a machine. The installation is experienced from a *second-person* perspective, where 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

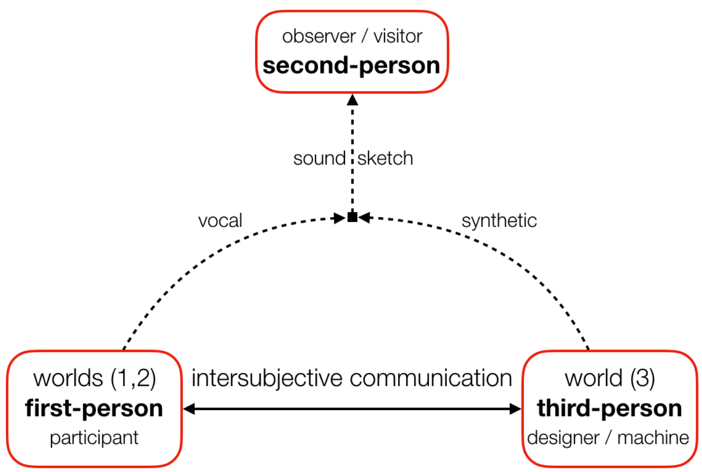

Figure 2: The observer forms an opinion on whether both vocal sounds and their synthetic translations represent plausible motor actions that can be ascribed to the visuo-spatial, intentional articulation of the mouth by the character-participant.

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

4.1. Reception

S'i' fosse suono was first exhibited at the ICT Conference of the European Commission in Lisbon, Portugal, on October 20-22, 2015. In this and other public exhibitions hundreds people interacted with the multi-touch screen installation. From informal conversations with the interacting visitors, it emerged that the three sound realisations were equally effective in binding the sensory information to spatial entities and temporal events, i.e., at forming audio-visual objects (Kubovy & Schutz, 2010). The action/sound association, that is the temporal simultaneity and spatial coincidence of the visible and audible information, was found to be plausible and strong, which comes at no surprise since the synthetic sounds were derived and causally consistent with a recorded vocal utterance.

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

In the practice, the two attitudes reflected the creative approach towards sound synthesis. The concrete attitude called for a rather indexical use of the sound models, e.g. vocal imitations of impacts or explosions were translated in corresponding synthetic sound events, likewise. The abstract attitude instead stretches the possibilities of a given sound model, by enhancing its sonic space[4]. More generally, this dual approach to design is found during the conception of any product, in its visual, auditory, and tactile properties (Özcan & Sonneveld, 2009). Although physical models seem to be more suitable for concrete sound concepts, and granular synthesis more suitable for abstract sound concepts, the two techniques were reported by the audience to be equally effective in producing consistent and compelling sound realisations.

[4]The 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/`).

4.2. Naturalness and Concreteness of sound sketches

Artist-designers can inform and inspire design, and the reactions from an audience exposed to their prototype realisations are valuable for steering project developments (Dunne & Gaver, 1997). But if the artist–designers are asked to work under well-defined constraints, their prototypes can also be used to test some design assumptions, by formal experimentation.

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

4.2.1. Apparatus

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

4.2.2. Participants and Task

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

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

4.2.3. *Hypotheses*

We expected that participants would be able to distinguish the sounds that are directly coming from a vocal production (natural) from those that are synthetic (artificial).

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

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

4.2.4. *Results*

Figure 3 shows the mean ratings for the three families of sounds used in the installation.

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

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

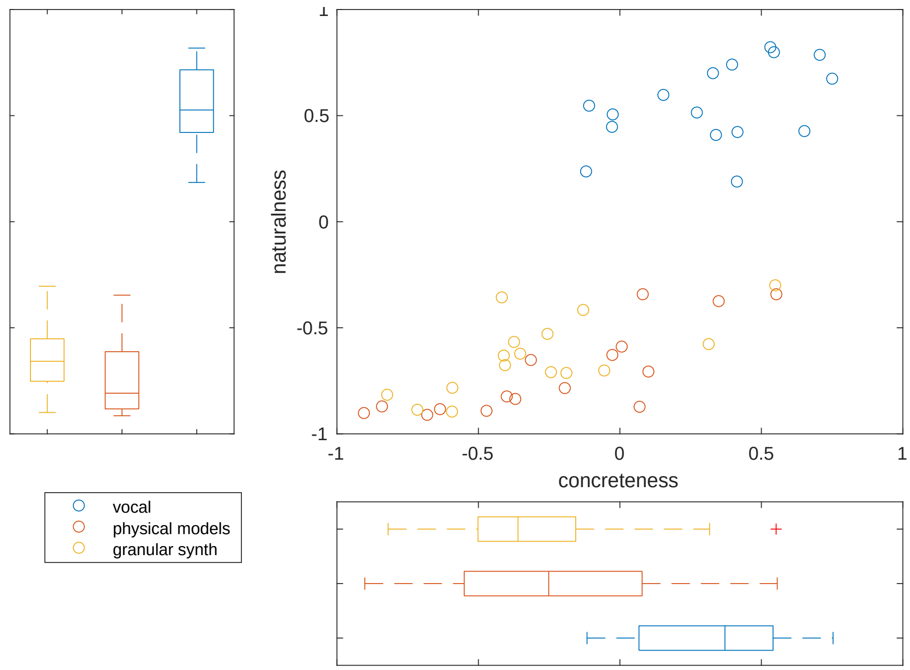

Figure 3: Scatterplot of the mean ratings of the 48 sounds of *S'i' fosse suono*, divided in the three groups of vocal, physics-based, and granular sounds

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

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

Overall, by confirming the hypotheses, the experiment also confirms the informal observations made in exhibitions and the impressions collected from the public, briefly reported in Section 4.1.

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

5. SEeD: Sound EmbodiEd Design

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

5.1. Design

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

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

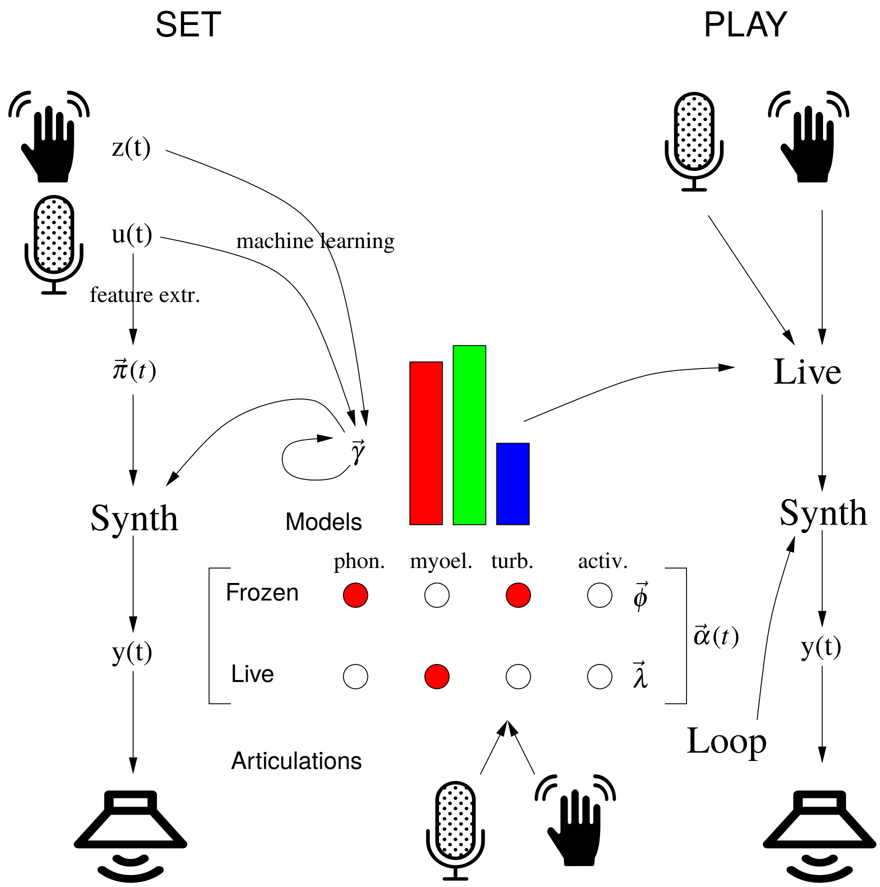

Figure 4: Sketch of SEeD

1. a vector of weights $\vec{\gamma}$ (coloured bars in Fig. 4), giving the relative importance of different sound synthesis models (machine learning),

2. a vector of functions of time $\vec{\pi}$, with each element representing the temporal evolution of signal features (feature extraction).

Given the models weights and the features trajectories, the system returns a subset of its sound models and controls them to generate a synthetic sound similar to the input utterance. In line with the rationale of *S'i' fosse suono*, the `set` mode replaces that cognitive transition of the second-person perspective, in which the human agent, the artist-designer, makes acousmatic or concrete hypotheses on the sound models and corpora needed to return a synthetic impression of the given utterance. This operation mode represents the moment in which the intersubjective communication between the user and the machine is established. Indeed, the output $y(t)$ is an audio signal that goes directly to the speakers and gives immediate feedback to the user, a sort of synthetic echo to the proposed vocalisation.

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

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

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

In `play` mode, $\vec{\gamma}$ is used to mix the output of sound models that will be controllable in real time through voice and gesture, according to the control layer of each model. This modality does not require any visual interface. In the case of physical models, each model is provided with a control layer that can perform a real-time mapping of vocal and gestural features into model parameters. In the case of corpus-based synthesis, each sound sample is selected to match a particular set of sound descriptors. Precisely, the voice is analysed resulting in a time-morphology vector of sound descriptors (loudness, spectral centroid, noisiness, etc.) (Marchetto & Peeters, 2015). The corpus-based synthesis corresponds thus to rendering a similar sound morphology using the corpus samples.

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

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

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

5.2. Implementation

SEeD is essentially a modular Cycling’74 Max patch, which implements the diagram in Figure 4 in most of its components[5]:

(1) `concat` (for corpus-based) and `physmod` (for physics-based) synthesis techniques can be dynamically switched throughout `set` and `play` modes; once a mixture of sound classes is set, its relative weights $\vec{\gamma}$ affect either the physmod sound models or the concat sound corpora. In `play` mode, the `live` submode allows to control the sound synthesis directly through vocalisations, while in `loop` submode the previously recorded stream $\vec{\pi}$ of audio descriptors is used to drive the sound models. Additionally, in `loop` submode the stream $\vec{\pi}$ can be further replaced by a new recording, and yet without affecting the current `set` and weighting.

(2) A Gaussian Mixture Model classifier (Françoise et al., 2014) is trained with the user-provided vocal imitations of eight sound classes corresponding to the eight sound models / corpora available, and used in the realisation of *S’i’ fosse suono*: blowing, car engine, crumpling, electric motor, hitting, liquid dripping, rubbing–scraping, and shooting;

(3) During the `set` operation, the best three sound models and their relative contributions (`weights`) are displayed. Eventually their balance in the mixture can be adjusted manually on the GUI. Similarly, these weighting and tweaking apply to the classes of sound samples used for the granular synthesis;

(4) An `articulation` control window allows to `freeze`, `loop`, or act `live` on articulatory features. These features are integrated to give a high-level description of vocalisations in terms of phonation, turbulence, myoelasticity, and general activity. This layer allows to tailor to some extent the system responsiveness to one’s own vocal characteristics.

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

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

5.3. Testing and development

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

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

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

Embodied sound design tools consider the ambiguity of the sketch input as a resource, leaving the user potentially free to manage errors and conflicts later in the process and even with other tools. As such, approaching the tools through the imitative operation of established software for sound production is useless. It was reported

[6] `https://vimeo.com/169521601`.

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

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

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

6. Evaluation

The experimental evaluation of SEeD was set up around the following idea: To provide some sound designers with several target sounds originally created with SEeD and ask them to sketch the targets with the same tool, in a limited amount of time. The goal was to observe whether the sketches would be closer to their corresponding targets than to the other targets, by measuring the respective auditory distance. The reason for using SEeD products as target sounds is that, in principle, a designer should be able to reconstruct the target exactly. In practice, such perfect reconstruction is never achieved in time- and tool-constrained sketching, and different sketchers may vary in their vocal-sketching and tool-manipulation proficiency. This is a restricted form of evaluation, as it only addresses the sketching effectiveness of the tool, and not its general usefulness in a creative process.

6.1. Experimental procedure

6.1.1. Setup

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

6.1.2. Participants and procedure

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

The user test was paced in four parts. First, each participant (P1, P2, P3) read the instructions, the experimenter demonstrated the software and explained each different sound synthesis model. Following, each participant trained the system with the vocal examples of the eight classes of sound models available in the tool (and described in Section 5.2). The participants spent 20 minutes of individual training, in order to explore the tool with their voice, warm up and familiarise with the different sound models. The experimenter was present to answer to questions. Before starting the user test, the experimenter made a demonstration of the task. Each sound designer, individually, was asked to reproduce 8 target sounds. The targets, 4 `physmod` and 4 `concat` sounds of a duration of 5 seconds each, had been previously created by Andrea Cera with SEeD. In addition to the vocal control, both the designer of the target sounds and the three participants were allowed to tweak manually the weights of the sound classes, a reduced set of parameters of the sound models in `physmod`, and the grain duration and the voice descriptors weights in `concat`. The target sounds were presented in a random order, without disclosing how they had been generated (i.e., `physmod` or `concat`). Essentially, the participants listened to the target and opted for the most appropriate sound synthesis mode – `physmod` or `concat` – to sketch the sound. The time slot to represent each target was constrained to 4 minutes.

6.1.3. Results

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

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

Figure 5 represents the matrices of dissimilarities between the targets and the sketches, as well as a hierarchical representation (dendrogram) of these distances, for each sound designer (P1, P2, P3). For the sake of visualisation, the maximum value has been taken from the matrices of all three participants, and the distances have been normalised to 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 the distances between target sounds, the bottom-right 8-by-8 submatrix M_3 represents the distances between the sketches, and the top-right 8-by-8 submatrix M_2 represents the distances between the sketches and the corresponding target sounds. Ideally, if the sketches would be identical to their target sounds then all the three submatrices would be identical. More loosely, if each sketch would be similar to its target then a dark diagonal would emerge in the top-right submatrix.

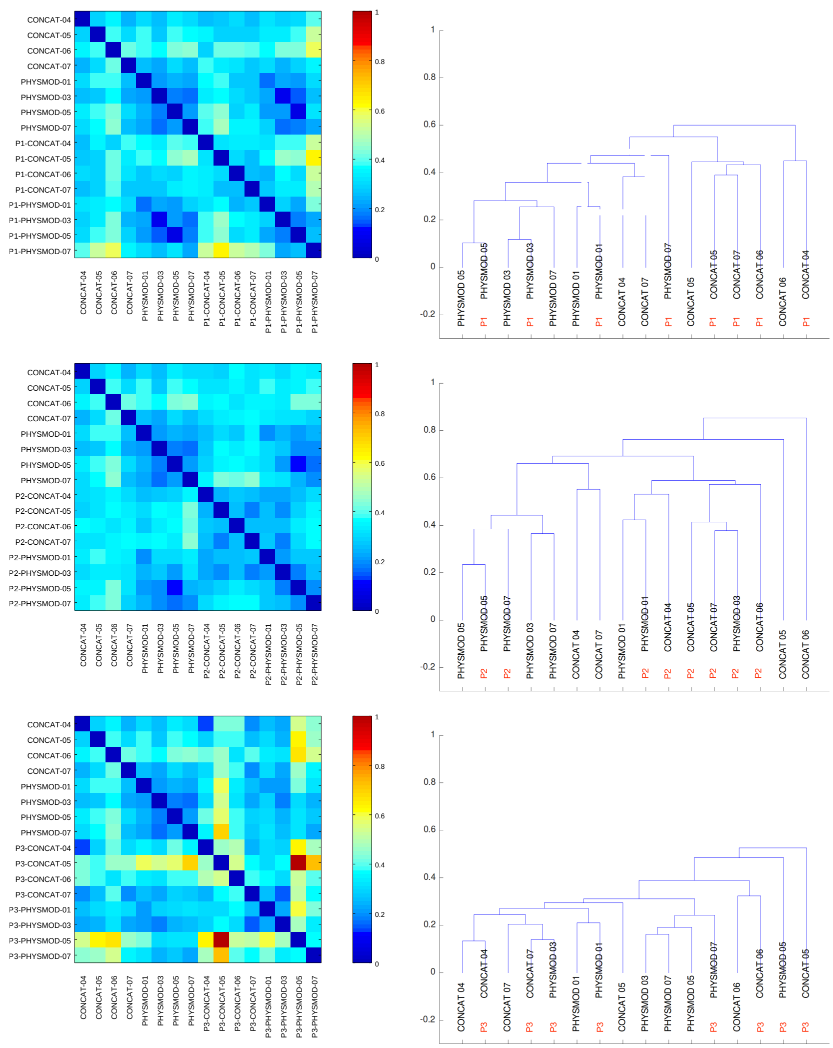

Figure 5: Left: Matrix of dissimilarities between the targets and the sketches. Right: Hierarchical representations of the distances. The results per each participant are arranged in rows (P1, P2, P3).

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

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

Table 1 shows the sound production mode used by the three participants to sketch each target sound. The choice of the production mode is a first indicator of the understanding of the representing world embodied in the target sounds. In most of the cases, the sketchers used the same sound synthesis approach originally used in the targets. Yet of interest, sketches close to their targets could be achieved with different production 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.

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

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

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

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

The SEeD tool supports to move back and forth from the vocal articulation to the production of coherent synthetic impressions. The experiment shed light on the possibilities and limitations of the tool, in particular on the difficulties that designers may meet when trying to represent a well-defined sound target with a sonic sketch. Furthermore, we ran *post-hoc* interviews with the participants and collected feedback on the 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.

6.1.4. Post-experiment interviews

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

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

The strengths and weaknesses of the physical models were found in their concreteness. As their use was experienced by participants as more intuitive and immediate than the corpus-based approach, the drawback of `physmod` is the apparently limited sound palette. That was a contradictory observation by P1, which rather reflected a control issue showed by the tool, that is the association between the articulatory streams and the models parameters, in terms of width of the resulting sonic space. Conversely, the participants stressed the creative potential of the `concat` synthesiser, especially in the improvisation and exploration of textural sounds. The drawback of the abstractness of this approach results in a less natural control. In particular, P1 reported the lack of the pitch in the time-morphology vector of voice descriptors, which makes the exploration of the sound corpora counter-intuitive. P1 suggested to harmonise the control strategy of the two synthesis approaches. P3 reported the too many possibilities offered by `concat` as a major difficulty preventing its immediate use.

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

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

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

However, this possible design solution leads to the second locus of discussion, which is methodological. Certainly user's profiles that are consistent with the individual centred `set` modes would partially solve the idiosyncrasy. Sketching (with SEeD), that is design-thinking while making sonic representations, is not a sound selection task whose final shape is left to the interpretation of the tool. Rather, it is an activity which requires the fluency and expertise of the sketcher in order to access and empower imagery back and forth (**?**), during the conception of the sound design. When tools are expressive, interaction becomes performative, yet not effortless. It requires practice and training. In this respect, the further improvement of the tool goes together with development of sketching practices of vocal scribbling. Ontologically-based studies of vocal sketching protocols can reveal relevant information on the effectiveness of the ideation process, and hence on the method and the tool (Gero & Kannengiesser, 2014; Delle Monache & Rocchesso, 2016).

7. Conclusion

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

8. Acknowledgments

The work described in this paper is part of the project SkAT-VG, which received the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission under FET-Open grant number: 618067. Luca Ludovico and Davide Andrea Mauro contributed to coding *S'i' fosse suono*. Most of the participants who donated their vocal self-portraits are part of the theater group Cantiere Ca' Foscari, directed by Elisabetta Brusa. The sound designers of the 48-hours of sound design in Château La Coste were Simon Cacheux, Andrea Cera, Xavier Collet, Mathieu Pellerin, and Allister Sinclair.

References

Agus, T. R., Suied, C., Thorpe, S. J., & Pressnitzer, D. (2012). Fast recognition of musical sounds based on timbre. *The Journal of the Acoustical Society of America*, *131*, 4124–4133. URL: `http://dx.doi.org/10.1121/1.3701865`. doi:`10.1121/1.3701865`. `arXiv:http://dx.doi.org/10.1121/1.3701865`.

Aramaki, M., & Kronland-Martinet, R. (2006). Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 695–705. doi:`10.1109/TSA.2005.855831`.

Baldan, S., Monache, S. D., & Rocchesso, D. (2017). The sound design 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`.

Becvar, A., Hollan, J., & Hutchins, E. (2008). Representational gestures as cognitive artifacts for developing theories in a scientific laboratory. In *Resources, co-evolution and artifacts* (pp. 117–143). Springer.

Cash, P., & Maier, A. (2016). Prototyping with your hands: the many roles of gesture in the communication of design concepts. *Journal of Engineering Design*, *27*, 118–145.

Cera, A., Mauro, D. A., & Rocchesso, D. (2016). Sonic in(tro)spection by vocal sketching. In *Proc. XXI Colloquium on Musical Informatics*. Cagliari, Italy.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*, 887–906. URL: http://dx.doi.org/10.1121/1.1945807. doi:10.1121/1.1945807. arXiv:http://dx.doi.org/10.1121/1.1945807.

Chion, M. (1994). *Audio-Vision: Sound on Screen*. New York: Columbia University Press.

De Poli, G., Piccialli, A., & Roads, C. (Eds.) (1991). *Representations of Musical Signals*. Cambridge, MA, USA: MIT Press.

Delle Monache, S., Polotti, P., & Rocchesso, D. (2010). A toolkit for explorations in sonic interaction design. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound* AM '10 (pp. 1:1–1:7). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1859799.1859800. doi:10.1145/1859799.1859800.

Delle Monache, S., & Rocchesso, D. (2016). Understanding cooperative sound design through linkographic analysis. In *Proc. of the XXI CIM Colloquium on Music Informatics* (pp. 25–32). Cagliari, Italy.

van den Doel, K., Kry, P. G., & Pai, D. K. (2001). Foleyautomatic: Physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* SIGGRAPH '01 (pp. 537–544). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/383259.383322. doi:10.1145/383259.383322.

Dunne, A., & Gaver, W. W. (1997). The pillow: Artist-designers in the digital age. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems* CHI EA '97 (pp. 361–362). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1120212.1120434. doi:10.1145/1120212.1120434.

Dyck, J. (2016). Natural sounds and musical sounds: A dual distinction. *The Journal of Aesthetics and Art Criticism*, *74*, 291–302. URL: http://dx.doi.org/10.1111/jaac.12286. doi:10.1111/jaac.12286.

Eris, O., Martelaro, N., & Badke-Schaub, P. (2014). A comparative analysis of multimodal communication during design sketching in co-located and distributed environments. *Design Studies*, *35*, 559–592.

Françoise, J. (2015). *Motion-sound Mapping By Demonstration*. Theses Université Pierre et Marie Curie - Paris VI. URL: `https://tel.archives-ouvertes.fr/tel-01206009`.

Françoise, J., Schnell, N., Borghesi, R., & Bevilacqua, F. (2014). Probabilistic models for designing motion and sound relationships. In *Proceedings of the International Conference on New Interfaces for Musical Expression* NIME (pp. 287–292). URL: `https://hal.archives-ouvertes.fr/hal-01061335/document`.

Franinović, K., & Serafin, S. (2013). *Sonic interaction design*. MIT Press.

Gabor, D. (1947). Acoustical quanta and the theory of hearing. *Nature*, *159*, 591–594.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, *5*, 1–29.

Gero, J. S., & Kannengiesser, U. (2014). The function-behaviour-structure ontology of design. In A. Chakrabarti, & M. L. T. Blessing (Eds.), *An Anthology of Theories and Models of Design: Philosophy, Approaches and Empirical Explorations* (pp. 263–283). London: Springer London. URL: `http://dx.doi.org/10.1007/978-1-4471-6338-1_13`. doi:`10.1007/978-1-4471-6338-1_13`.

Goldschmidt, G. (2014). *Linkography: unfolding the design process*. MIT Press.

Helgason, P. (2014). Sound initiation and source types in human imitations of sounds. In *Proc. Fonetik* (pp. 83–89). Stockholm, Sweden.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., & Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, *18*, 52–80. doi:`10.1037/a0026240`.

Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. (2016). Auditory Sketches: Very Sparse Representations of Sounds Are Still Recognizable. *PLoS ONE*, *11*, e0150313+. URL: https://hal.archives-ouvertes.fr/hal-01250175. doi:10.1371/journal.pone.0150313.

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

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

Lemaitre, G., Dessein, A., Susini, P., & Aura, K. (2011). Vocal imitations and the identification of sound events. *Ecological Psychology*, *23*, 267–307. URL: http://dx.doi.org/10.1080/10407413.2011.617225. doi:10.1080/10407413.2011.617225. arXiv:http://dx.doi.org/10.1080/10407413.2011.617225.

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

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

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

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

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

Leman, M. (2008). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.

Löwgren, J., & Stolterman, E. (2004). *Thoughtful Interaction Design: A Design Perspective on Information Technology*. The MIT Press.

Marchetto, E., & Peeters, G. (2015). A set of audio features for the morphological description of vocal imitations. In *Proceedings of the 18th International Conference on Digital Audio Effects (DAFX-2015)* (pp. 207–214). Trondheim, Norway: Norwegian University of Science and Technology.

Marchetto, E., & Peeters, G. (2017). Automatic recognition of sound categories from their vocal imitation using audio primitives automatically derived by si-plca and hmm. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research* CMMR (pp. 9–20). URL: http://cmmr2017.inesctec.pt/wp-content/uploads/2017/09/1_CMMR_2017_paper_50.pdf.

Nykänen, A., Wingstedt, J., Sundhage, J., & Mohlin, P. (2015). Sketching sounds–kinds of listening and their functions in designing. *Design Studies*, *39*, 19–47.

Özcan, E., & van Egmond, R. (2009). Product sound design: An inter-disciplinary approach? In *Undisciplined! Design Research Society Conference*. URL: http://shura.shu.ac.uk/531/.

Özcan, E., & Sonneveld, M. (2009). Embodied explorations of sound and touch in conceptual design. In *Design semantics of form and movement* (pp. 173–181). Taipei, Taiwan. URL: https://www.northumbria.ac.uk/media/6278717/desform-2009-proceedings-v2.pdf#page=173.

Pauletto, S. (2017). The voice delivers the threats, foley delivers the punch: Embodied knowledge in foley artistry. In M. Mera, R. Sadoff, & B. Winters (Eds.), *The Routledge Companion to Screen Music and Sound* (pp. 338–348). Routledge.

Pauletto, S., Cambridge, H., & Susini, P. (2016). Data sonification and sound design in interactive systems. *International Journal of Human-Computer Studies*, *85*, 1 – 3. URL: http://www.sciencedirect.com/science/article/

pii/S1071581915001299. doi:http://dx.doi.org/10.1016/j.ijhcs.2015.08.005. Data Sonification and Sound Design in Interactive Systems.

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

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

Rocchesso, D., Lemaitre, G., Susini, P., Ternström, S., & Boussard, P. (2015). Sketching sound with voice and gesture. *interactions*, *22*, 38–41. URL: http://doi.acm.org/10.1145/2685501. doi:10.1145/2685501.

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

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

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

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

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

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1791–1810.

Schwarz, D. (2007). Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, *24*, 92–104. doi:10.1109/MSP.2007.323274.

Scurto, H., Lemaitre, G., Françoise, J., Bevilacqua, F., Susini, P., & Voisin, F. (2016). Embodying sounds: Building and analysis of a database of gestural and vocal imitations. In *Gesture, Creativity, Multimodality. Procedings of ISGS 7 : 7^{th} Conference of the International Society for Gesture Studies* (p. 269). The University of Texas at Austin, Austin, TX: International Society for Gesture Studies.

Scurto, H., Lemaitre, G., Françoise, J., Voisin, F., Bevilacqua, F., & Susini, P. (2015). Combining gestures and vocalizations to imitate sounds. *The Journal of the Acoustical Society of America*, *138*, 1780–1780.

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, *136*, 943–974.

Strobl, G., Eckel, G., & Rocchesso, D. (2006). Sound texture modeling: A survey. In *Proceedings of the Sound and Music Computing (SMC) Conference* (pp. 61–65). Marseille, France.

Susini, P., Houix, O., & Misdariis, N. (2014). Sound design: an applied, experimental framework to study the perception of everyday sounds. *The New Soundtrack*, *4*, 103–121.

Tholander, J., Karlgren, K., Ramberg, R., & Sökjer, P. (2008). Where all the interaction is: Sketching in interaction design as an embodied practice. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems* DIS '08 (pp. 445–454). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1394445.1394493. doi:10.1145/1394445.1394493.

Tuuri, K., Eerola, T., & Pirhonen, A. (2011). Design and evaluation of prosody-based non-speech audio feedback for physical training application. *International Journal of Human-Computer Studies*, *69*, 741 – 757. URL: http://www.

sciencedirect.com/science/article/pii/S1071581911000760. doi:http://dx.doi.org/10.1016/j.ijhcs.2011.06.004.

Tversky, B., Heiser, J., Lee, P., & Daniel, M.-P. (2009). Explanations in gesture, diagram, and word. In K. Coventry, T. Tenbrink, & J. Bateman (Eds.), *Spatial language and dialogue* chapter 9. (pp. 119–131). New York, NY: Oxford University Press.

Verma, T., Levine, S., & Meng, T. (1997). Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 48–51). Thessaloniki, Greece.

Visser, W., & Maher, M. L. (2011). The role of gesture in designing. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *25*, 213–220.