Sonic in(tro)spection by vocal sketching

Andrea Cera
Iuav University of Venice

Davide Andrea Mauro
Iuav University of Venice

Davide Rocchesso
Iuav University of Venice
roc@iuav.it

ABSTRACT

How can the art practice of self-representation be ported to sonic arts? In *S'i' fosse suono*, brief sonic self-portraits are arranged in the form of an audiovisual checkerboard. The recorded non-verbal vocal sounds were used as sketches for synthetic renderings, using two seemingly distant sound modeling techniques. Through this piece, the authors elaborate on the ideas of self-portrait, vocal sketching, and sketching in sound design. The artistic exploration gives insights on how vocal utterances may be automatically converted to synthetic sounds, and ultimately how designers may effectively sketch in the domain of sound.

1. VOCAL SKETCHING

Vocal sketching is the act of communicating a sonic concept using the voice. Assuming an idea of sound exists in the mind of a person, sketching is a representational act that uses the most direct means for sound production.

An old jazz adage says "If you can't sing it, you can't play it". Given the constraints and possibilities of the voice organ, an interesting question is whether the way we imagine sounds is affected by our sketching practice and abilities. In the context of modern theories of embodiment, such as the ideomotor theory or the theory of event coding, the internal representations come in the form of perception-action ensembles [1], and we can imagine what we have previously experienced through perception-action loops. Given the effectiveness of sound communication by vocal imitation [2], it seems that our sonic imagery, while transcending the vocal acoustic space, can be effectively and compactly represented via vocal gestures. These vocal sketches can be exploited as proxies for non-utterable sounds or as entry points to vast sonic spaces.

Similarly to drawing, vocal sketching can be instrumental to establishing a positive-feedback loop that let sonic ideas emerge and take form, through a continuous confrontation between internal and external representation. Such a loop can be entirely controlled by the designer or, in a participatory framework, it can be open to other participants, whose sketches can contribute to a reflective process [3].

2. ON ABSTRACT AND CONCRETE SOUNDS

An underlying axiom of early computer music research and practice was that "there are no theoretical limitations" as "any perceivable sound can be generated" [4]. This view of the computer as a white canvas for sound and music was a promoting factor for abstraction in sound synthesis and composition. On the other hand, when the raw audio material to be painted on canvas was coming from recordings, the limitless view promoted an acousmatic, reduced way of listening [5]. Are these computer-music approaches utopian attitudes or rather feasible approaches to sound design and composition?

Quoting Giorgio Morandi (1890-1964), a painter who used to fill his white canvases with only bottles, vases, and carafes, "nothing is more abstract than reality" and "to achieve understanding it is necessary not to see many things, but to look hard at what you do see". Which sounds can we imagine? We can imagine, and possibly imitate, those sounds that we can (directly or indirectly) produce, and those of the environment that have the potential to trigger an action.

Action-sound associations are plastic [6] and, as such, can be designed. However, experiments in dimensionality reduction and categorization have highlighted the anchoring role of basic mechanisms of physical sound production, such as gas turbulence, fluid dynamics, impact, friction, rotary machines, etc. [7]. Similarly, the sounds that can be vocally produced can be decomposed and analyzed in terms of fundamental mechanisms such as turbulence, myoelastic vibration, impact, phonation. So, there seems to be a lexicon of basic elements for sound composition, a range of not many things that can be outlined, colored, and arranged on the white canvas in infinitely many ways. In this respect, doing sound design may mean to compose sound using this lexicon of physically-grounded phenomena, thus keeping a direct link to sound imagination.

3. SONIC SELF-PORTRAITS

The self-portrait is a recurrent exercise in visual arts of all times, and across all techniques, styles, and movements. How can we translate self-representation to sonic arts? How can we exploit sound synthesis in perception-action loops? Some attempts have been previously made to devise audio-visual tools that stimulate and enhance self-expression, and to produce visual-sonic self-portraits. For example, Polotti and Goina [8] realized a system based on correspondences between elementary sonic and movement units, and asked participants to express themselves through movement-generated sound.

The EU Project SkAT-VG aims at developing tools and methods to exploit the innate vocal sketching abilities of humans, in the early stage of the sound design process [9]. In this context, the perception-action loop is including interpretation and synthesis, and tools for effective sonic introspection are going to be available. However, in an exploration stage, sound-designer expertise is sought to envision meaningful translations of vocal utterances into designed, artificial sounds, a sort of inspection into the possibilities of vocal sketching. From a science and technology perspective, the artist/designer acts as an exploratory probe: He inspects the utterances and uses them as proxies to imagined sound spaces. The resulting designed sounds let us foresee what an automatic translation process might be aimed at, so te become an effective sonic extension of the voice.

In *S'i' fosse suono* (lit., if I were sound), we asked sixteen persons (including the second and third authors) to represent themselves with a brief non-verbal vocal sound. They were audio-visually recorded and their performances were used as sketches for two different synthetic renderings, one using physics-based sound modeling, and the other using synthesis by recomposition of vocal grains. The resulting materials were used to compose an audiovisual interactive checkerboard, which is a proof of concept of vocal sketching and its exploitation in sound design and sonic arts. By producing a sonic self-portrait, each participant was requested to make imagination audible. The sound designer (Andrea Cera) was expected to understand, interpret, and transform such representational act. Two sets of constraints were given to the sound designer for such transformation, so that two different synthetic sound renderings were produced for each recorded vocal production:

1. use physics-based sound models of fundamental sound-generating mechanisms, as made available by the Sound Design Toolkit (SDT) [10];

2. manipulate the audio buffer containing the recorded vocalization, using granular techniques (MuBu) [11].

4. DESIGN PROCESS

If framed in a situated ontology of design [12], the creation of *S'i' fosse suono* can be described as the sequence:

1. **interpreted** world: The sixteen participants (Figure 1) imagined a sonic self-representation in terms of perception-action associations and concepts;

2. **expected** world: Each participant set a motor program for acting sonically by means of the voice, thus translating imagination into action;

3. **external** world: The utterances were communicated to the sound designer, who interpreted them as blueprints for synthetic sound composition.

Stage 2 is reached from stage 1 via focusing, i.e. taking some aspects of the interpreted world and using them as goals for the expected world. Stage 3 is the effect on the external world achieved via goal-driven action. Such sequence may be looped, in such a way that the interpreted world (and its perception-action associations) may be modified after new experiences and interpretation of the external world. In this case, however, we present the result of an open-loop process, where the sound designer interprets and affects the external world. The sequence is also indicative of what a tool-mediated sound design process may produce, where the sound designer may be involved in stages 1 and 2, while the translation of vocal blueprints into new sounds of the external world may be performed by a machine. In the proposed artistic installation, instead, human agents with different roles have been involved for stages 1-2 (participants) and for stage 3 (sound designer). The expected world and the external world are made jointly accessible as the audio-visual checkerboard chooses randomly, upon being clicked in one face box, if playing back the vocal utterance or one of its two renderings in synthetic sound. The constraints given to the sound designer in terms of usable sound models, and his use of some automatic feature extractors, make the automation of the external world easier to foresee.

In the framework of embodied music cognition and mediation technology [13], stages 1 and 2 can be associated with a first-person perspective, where "moving sonic forms take the status of actions to which intentionality can be attributed". Stage 3 is that of a third-person perspective, where phenomena get somehow measured and translated, either by a human observer (in this case, the sound designer) or by a machine. The experience of the installation is that of a second-person perspective, where the observer gets involved "with physical energy in a context of intersubjective communication".

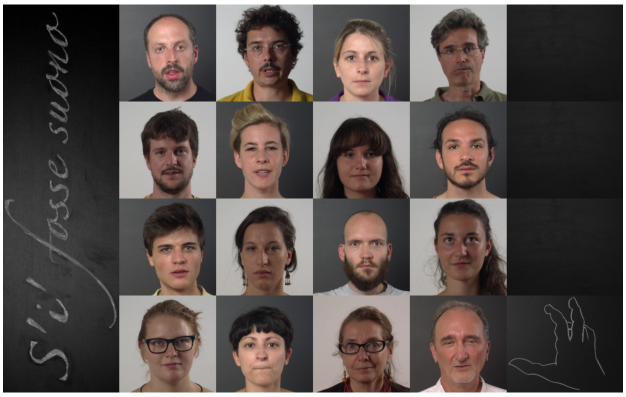

Figure 1. *S'i' fosse suono*

5. SOUND DESIGN

The sound design-by-transformation process is here presented in some detail by two examples, one based on physical models, and the other based on audio manipulation. The vocal production is that of the participant portrayed in row 4 and column 2 of Figure 1. In articulatory terms, her sonic self-portrait can be described as a train of labial myoelastic pulses superimposed to a steady phonation. The voice-driven sound design process is divided into three stages (physical models):

1. **Analysis**. Two streams are automatically extracted from audio: (A) discrete onsets (Ableton Live) and (B) continuous pitch (SDT pitch extractor).

2. **Synthesis**. Two processes are executed: (X) Stream A drives the SDT bubble sound model (Figure 2, left) and (Y) Stream B drives the gear ratio and the RPM parameters of the SDT dc motor model.

3. **Rendering**. The audio outputs of processes X and Y are layered. Process X is processed through a convolution reverb, following a hand-drawn automation curve.

In the case of audio buffer manipulation, the analysis stage is unchanged, and the second stage is replaced by

2. **Synthesis**. Using MuBu granulator (Figure 2, right): Stream A drives duration and position, in such a way that each onset places the granulator head in proximity of one of the vocal events and then back to a point where there is only the sustained note,

and the third stage is reduced to reverberation.

6. THE PERSPECTIVE OF THE SOUND DESIGNER

When the first author (sound designer) started working on the collection of the sixteen recordings, he immediately realized that he needed a perspective from which to observe and judge the designed sounds and their relations with the original audio-visual counterparts. The driving idea of bad imitation [14] was considered to be a valuable strategy for this case. This paradigm is based on the deliberate production of a discrepancy between one imitated object and one imitating agent, which brings novelty and surprise in the imitation process. It relies on bounces between the internal world and the external world, driven by processes of perception (listening) and action (reproducing). The bad-imitation paradigm can be embedded in the programming of an interactive behavior [15], but it can also be exploited during the preparatory stage of a creative production.

In *S'i' fosse suono*, bad imitations were used to limit the scope of analysis of the original recording by embracing one of two mutually exclusive attitudes, defined by the sound designer as a) acousmatic, or b) concrete:

a. Acousmatic approach to the recorded utterances. The sound designer was trying to abstract from physical information that could be derived from visual moving images, without searching neither for possible referent physical phenomena nor for articulatory details of the vocal apparatus. The original sound was approached as if it was artificial, trying to imagine which would be the control signals that would have led to its synthetic production. The SDT sound models were selected according to the proximity of their sound output to the hypothetical sound output of such an ideal synthesizer. Once the choice was made, the designer forced himself to stick with it. For example, participant in row 2 and column 3 (Figure 1) produced a complex and articulated vocal expression, using ingressive air streams and her mouth as a filter. The designed synthetic sounds, both in the SDT and in the MuBu version, bear significant differences from the uttered spectral signatures. However, these discrepancies are offset by the parallelism between the temporal evolutions of the imitated and imitating morphologies;

b. Concrete approach to the recorded utterances. The sound designer was prompted, by the nature of the reference recording, to consider a (natural or mechanical) everyday sound event. As a consequence, there was no other choice than using the SDT sound models meant to imitate such events (explosions, motors, wind, etc.). In this case, a weaker adherence between the imitated and imitating sounds, both in terms of spectral signatures and of temporal evolution, was accepted due to the physical metaphor overriding the morphological similarities.

The choice between these two attitudes (acousmatic vs. concrete) was also suggested by two ways in which the participants enacted their self-representation: some of them tended to simply explore their voice, while some others seemed to look for specific sounds (wind, far out explosion, motor, etc.), as if to sonically embellish their image. Similar tendencies of self-embellishment are found in the widespread practice of taking visual selfies [16].

It is interesting to notice that what has been described as an acousmatic approach, is indeed referring to a hypothetical synthesizer, an imaginary sound-producing device that just confirms that "there can never be any sound 'as such', and that sounds are always events of and in a field of relationships" [17]. Even under a self-imposed discipline, the sound designer could not escape the irreducible nature of listening.

7. RECEPTION AND DEVELOPMENTS

S'i' fosse suono was realized in the form of interactive installation, with a graphic framework (Figure 1), audio-visual recordings, and synthetic soundfiles held together in a Processing program. The piece was first exhibited at the ICT Conference of the European Commission in Lisbon, Portugal, on october 20-22, 2015. Since then, it has been used many times to elicit an intuitive understanding of the concept behind the SkAT-VG project. A multi-touch screen supported a playful engagement with the piece [1]. Many visitors, while initially triggering single utterances with circumspection, quickly showed a mixture of surprise and pleasure and started playing with the checkerboard with rapid sequences of finger taps. The overlapping events produce sorts of audio-visual arpeggios, whose actual outcome can never be predicted, as each face gesture is randomly coupled to the original or to one of the two synthetic sounds. In a second-person perspective [13], if the observer succeeds in re-enacting the sound sketch as made of actions

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

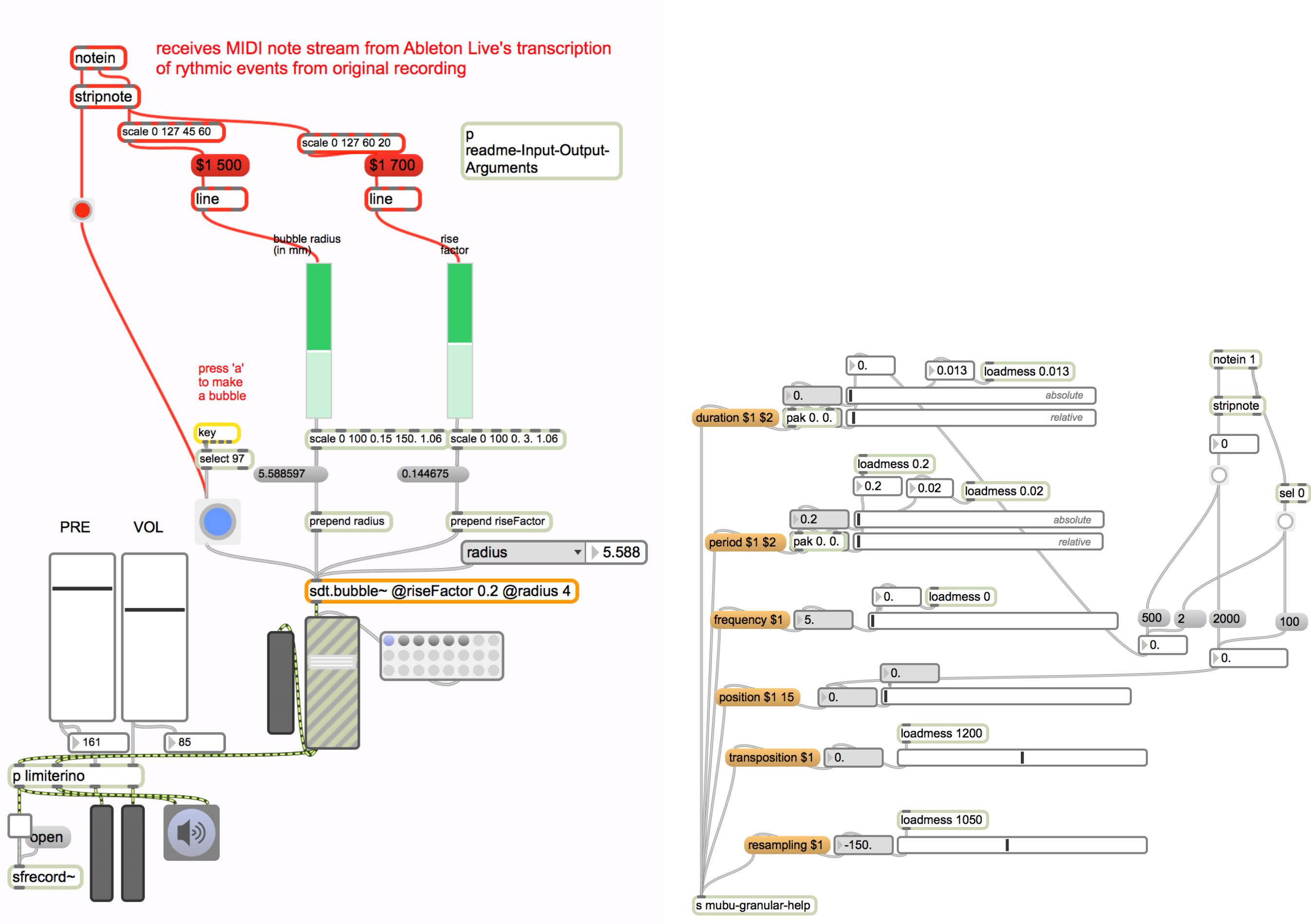

Figure 2. Stream of bubbles with the SDT (left) and granulation with MuBu (right)

with an intention then the communication act is found to be successful. Since this was found to happen for both the vocal sketches and their synthetic translations, it means that plausibility is preserved in the translation stage, and there is evidence of the effectiveness of the voice for sound sketching

"Audio-visual objects are constructs of the mind" [18], the result of a process that produces the most plausible binding of sensory information to objects and events. From the perspective of the observer, the three sound realizations result equally effective at eliciting such binding. The fact that the synthetic sounds are derived from a vocal utterance, which was interpreted as an imitation in acousmatic or concrete terms, makes the association plausible and strong.

This work shows both the concreteness of vocal sound materials and the versatility of sound models that refer to fundamental sound production mechanisms. The role of the sound designer was that of a probe, to explore the vast space of possible interpreted renditions of vocal utterances. By inspecting the sound design process, some indications were derived for future automation of the rendering process, towards a more effective use of the voice as a sound sketching tool. The two embraced attitudes, the acousmatic and the concrete, were largely dependent on the nature of the original vocal production in its recognizability as an imitation of an everyday sound phenomenon. The leading synthesis technique was tightly following such attitude. Nevertheless, in most cases physical models or audio-buffer manipulation were equally effective in producing consistent and compelling sound realizations. The automation of the voice-to-synth conversion processes, as envisioned by the SkAT-VG project, is encouraging us to explore iteration and plasticity in perception-action loops of sound design processes [19]. It will be interesting to see how new forms of self-representation will emerge.

8. ACKNOWLEDGMENTS

Luca Ludovico (University of Milan) coded the web version of *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. We acknowledge 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: SkAT-VG (Sketching Audio Technologies using Vocalizations and Gestures, skatvg.eu).

9. REFERENCES

[1] B. Hommel, “The theory of event coding (TEC) as embodied-cognition framework,” *Frontiers in Psychology*, vol. 6, no. 1318, 2015.

[2] G. Lemaitre and D. Rocchesso, “On the effectiveness of vocal imitations and verbal descriptions of sounds,” *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[3] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen, “User sketches: A quick, inexpensive, and effective way to elicit more reflective user feedback,” in *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*, NordiCHI ’06, (New York, NY, USA), pp. 105–114, ACM, 2006.

[4] M. V. Mathews, “The digital computer as a musical instrument,” *Science*, vol. 142, no. 3592, pp. 553–557, 1963.

[5] M. Chion, *Audio-Vision: Sound on Screen*. New York: Columbia University Press, 1994.

[6] G. Lemaitre, L. M. Heller, N. Navolio, and N. Zúñiga-Peñaranda, “Priming gestures with sounds,” *PLoS ONE*, vol. 10, no. 11, 2015.

[7] G. Lemaitre, A. Dessein, P. Susini, and K. Aura, “Vocal imitations and the identification of sound events,” *Ecological Psychology*, vol. 23, no. 4, pp. 267–307, 2011.

[8] P. Polotti and M. Goina, “EGGS in action,” in *Proceedings of the International Conference on New Interfaces for Musical Expression*, (Oslo, Norway), 2011.

[9] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard, “Sketching sound with voice and gesture,” *interactions*, vol. 22, pp. 38–41, January 2015.

[10] S. Delle Monache, P. Polotti, and D. Rocchesso, “A toolkit for explorations in sonic interaction design,” in *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, AM ’10, (New York, NY, USA), pp. 1:1–1:7, ACM, 2010.

[11] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi, “MuBu & friends - assembling tools for content based real-time interactive audio processing in Max/MSP,” in *Proceedings of the International Computer Music Conference*, (Montreal, Canada), 2009.

[12] J. S. Gero and U. Kannengiesser, “The function-behaviour-structure ontology of design,” in *An Anthology of Theories and Models of Design: Philosophy, Approaches and Empirical Explorations* (A. Chakrabarti and M. L. T. Blessing, eds.), pp. 263–283, London: Springer London, 2014.

[13] M. Leman, *Embodied music cognition and mediation technology*. Mit Press, 2008.

[14] A. Cera, “Écoutes et mauvaises imitations,” in *In actu: de l’expérimental dans l’art* (E. During, L. Jeanpierre, C. Kihm, , and D. Zabunyan, eds.), Djon, France: Les presses du réel, 2009.

[15] A. Cera, “Loops, games and playful things,” *Contemporary Music Review*, vol. 32, no. 1, pp. 29–39, 2013.

[16] F. Souza, D. de Las Casas, V. Flores, S. Youn, M. Cha, D. Quercia, and V. Almeida, “Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram,” in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN ’15, (New York, NY, USA), pp. 221–231, ACM, 2015.

[17] A. Di Scipio, “The politics of sound and the biopolitics of music: Weaving together sound-making, irreducible listening, and the physical and cultural environment,” *Organised Sound*, vol. 20, pp. 278–289, 12 2015.

[18] M. Kubovy and M. Schutz, “Audio-visual objects,” *Review of Philosophy and Psychology*, vol. 1, no. 1, pp. 41–61, 2010.

[19] D. Rocchesso, D. A. Mauro, and S. Delle Monache, “mimic: The microphone as a pencil,” in *Proceedings of the TEI ’16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI ’16, (New York, NY, USA), pp. 357–364, ACM, 2016.