# A Proposal for the Interactive Sonification of the Human Face

Davide Bonafede, Luca A. Ludovico and Giorgio Presti
*Dipartimento di Informatica "Giovanni Degli Antoni",*
*Università degli Studi di Milano, Via Celoria, 18, 20133 Milan, Italy*
{*luca.ludovico, giorgio.presti*}*@unimi.it*

Keywords: Interactive Sonification, Human Face, Face Tracking, Facial Expressions, Sonification Space.

Abstract: In the context of a broader project addressing the sound description of the aspect and expressions of the human face, a prototype has been implemented and presented during a scientific dissemination initiative. The final goal is to employ face-tracking and sound-synthesis techniques in order to strengthen or even replace visual communication, in particular sonifying facial expressions for visually impaired people. In this work we present the implementation of a proof of concept and some early results.

## 1 INTRODUCTION

Sonification is a technique for communicating information which employs sound to describe data and interactions without using vocal signals in order to facilitate their interpretation (Hermann et al., 2011).

According to (Hermann, 2008), sonification has to meet some fundamental requirements:

- Sound must reflect properties or objective relationships of data in input;
- Data transformation must be systematic; namely, a precise definition of how sound changes in function of data has to emerge;
- Sonification must be reproducible, i.e. sounds resulting from the same data and interactions must be structurally identical;
- The system must ultimately be designed to be used with different input data as well as with repetitions of the same input data.

The translation of information into sound can be realized through a number of different techniques, depending on the characteristics of input data, the approach adopted to process them, and the expected results.

In literature, multiple strategies are presented. Starting from the simplest technique, it is worth citing *audification*, namely a way to interpret a one-dimensional time series as a sound to be reproduced as audio in order to deliver information to listeners. *Auditory icons* are non-verbal sounds familiar to listeners, since they can be encountered in everyday life, and their meaning intuitively follows from listeners' experiences. An example is the rumbling sound produced by a moka pot when coffee is ready. *Earcons* are short sound events whose properties are associated with the parameters of the data to be delivered. The crucial difference from auditory icons is the lack of a real relationship between the idea to convey and the associated sound. An example is the acoustic signal used on planes to make passengers interact with flight attendants. So-called *parameter-mapping sonification* translates the characteristics of a data set into sound, mapping them onto acoustic parameters, such as pitch, velocity, timbre, brightness, etc. By raising the number of sound attributes, the multidimensionality of the sonification increases, too. Possible applications include sonic plots, e.g. audible skydiving altimeters. Finally, *model-based sonification* focuses on how acoustic events are generated in response to user interactions with a physical model imposed on data, thus creating an interactive system capable of generating acoustic signals.

In order to frame our proposal in a clear way, we will adopt a graphical tool introduced in (Ludovico and Presti, 2016) and shown in Figure 1: the *sonification space*. Such a representation aims at providing the expected sonic outcome of the sonification at a glance. The horizontal axis of the diagram is associated with the temporal granularity, and the vertical axis represents the level of abstraction[1] of the delivered sound. In the sonification space, two element types can be placed: the *main feature* (i.e. how the sonification appears as a whole), and *data bindings* (i.e. how single sonification parameters are mapped onto

[1] In this context, a low level of abstraction implies simple sounds without any kind of universally-accepted meaning, as opposed to a high level, which implies complex and culture-dependent meaningful sounds.



Bonafede, D., Ludovico, L. and Presti, G.
A Proposal for the Interactive Sonification of the Human Face.
In *Proceedings of the 2nd International Conference on Computer-Human Interaction Research and Applications (CHIRA 2018)*, pages 163-169
ISBN: 978-989-758-328-5

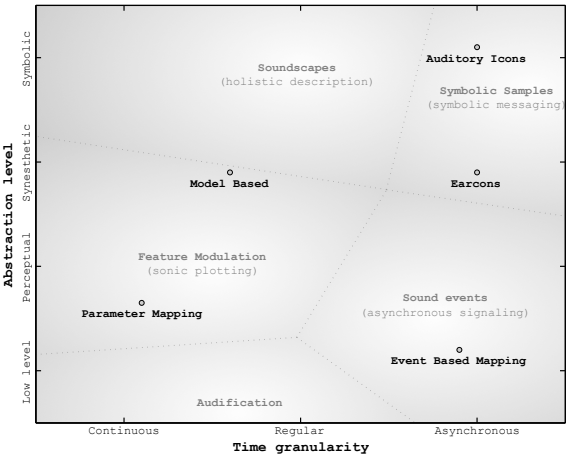


Figure 1: The sonification space proposed in (Ludovico and Presti, 2016), showing the position of some traditional sonification techniques.

sound).

For the sake of clarity, you can consider this space as divided into several areas which cluster sonification techniques and approaches (see Figure 1). For example, soundscapes can be found in the area containing the most abstract and continuous results, whereas simple triggered alarms are hosted in the low-abstraction asynchronous area. Even if some clusters can be identified, the sonification space was designed to be continuous, and some forms of sonification can raise legitimate doubts about their exact position. Nevertheless, the sonification space is just a support tool to represent complex sonifications in a straightforward way.

The paper is structured as follows: Section 2 presents the state of the art; Section 3 provides implementation details; Section 4 discusses our main design choices; Section 5 describes the framework in which an early experimentation took place; and, finally, Section 6 provides the road map for future work.

## 2 STATE OF THE ART ABOUT HUMAN FACE SONIFICATION

Scientific literature contains some relevant works dealing with human-face sonification. In this section we will review the most promising approaches and highlight the main differences from our proposal.

Reference (Guizatdinova and Guo, 2003) describes the development and usability evaluation of a sonification system for alternative access to facial expressions through so-called *eARmoticons*, namely sound entities that should evoke the emotional feelings of a listener like those a visual image could produce. The goal is to facilitate communication and interpretation of visual images for people with special needs. The research discusses an experimental evaluation conducted on volunteers to understand the sensitivity and the response of users to eARmoticons.

Another work shows a visual creativity tool to automatically recognize facial expressions and track facial muscle movements in real time in order to produce sounds (Valenti et al., 2010). The feature vector thus obtained is used as input to a Bayesian network which classifies facial expressions into several categories (e.g., angry, disgusted, happy, etc.). Classification results are used along with the feature vector to generate a combination of sounds that change in real time depending on facial expressions.

In contrast with our proposal, discussed in detail in Section 3, the mentioned techniques aim to infer emotions from raw data and drive sonification consequently, even when they claim to apply a direct mapping of parameters.

The research reported in (Funk et al., 2005) proposes face-detection and optic-flow algorithms to associate facial movements with sound synthesis in a topographically specific fashion. Gesture-to-sound mappings are oriented to musical performance, also known as *musification*: facial traits and expressions are used as a controller that triggers MIDI events to be sent to synthesis modules. Possible applications proposed by the authors include: the sonification of facial movements as a form of biofeedback to increase awareness of habitual behavior patterns; a support tool for physiotherapy in the treatment of facial paralysis or forms of brain damage which compromise face or facial expression recognition capabilities; an assistive technology for the blind. The mentioned use cases may be in common with ours, but the present proposal does not aim to produce a musical performance as the *main feature*, even if it uses musification for a particular *data binding*.

## 3 IMPLEMENTATION DETAILS

In this section we will provide details about our proposal, concerning the data to be sonified, the algorithms adopted to map facial data onto sound features, and the technologies employed in the implementation.

### 3.1 Adopted Technologies

The prototype is mainly based on three components: (i) a face tracking module, (ii) a set of Pure Data pa-

tches, and (iii) the Open Sound Control (OSC) protocol to make the mentioned modules communicate.

The definition of *face tracking* embraces those techniques and algorithmic approaches that determine if an image contains a face and track its position. In this field, a fundamental algorithm is the Active Appearance Model (AAM), which is able to create a statistic model of a deformable figure (Cootes et al., 2001). This approach has been profitably tested also in the field of face recognition (Edwards et al., 1998). Once the facial model is available, it is necessary to strengthen it against factors such as light conditions, facial expressions, and head orientation. A more advanced way to obtain facial tracking is to use a three-dimensional face prototype. We used a non-rigid model: a 3D mask is calculated from the first frame where the face is detected, and then it is continuously updated and adapted to expressions and orientation.

*Pure Data* is an open-source visual programming language for multimedia. Its main distribution was developed by Miller Puckette in the 1990s in order to create interactive computer music and multimedia works (Puckette et al., 1996). Our application is based on a set of Pure Data patches receiving data from the webcam and implementing the algorithms for their sonification. The face-tracking module and Pure Data patches are currently running in a low-budget laptop computer, but the final goal is to embed this system into a wearable device.

Information exchange among different components is based on a standard format. *OSC* is a music-oriented protocol widely adopted in computers and multimedia devices to support communication between synthesizers, computing units, and other devices (Wright, 2005). OSC applications include real-time sound and media processing environments, web interactivity tools, software synthesizers, distributed music systems, and so on.

## 3.2 Structure of the Application

The application can be roughly divided into two parts, as shown in Figure 2:

1. a *video-capture module*, that catches moving images from a webcam, tries to identify the presence of faces and tracks them through a three-dimensional model;
2. a *computation module*, that processes the generated data flow, so as to create, manage, and manipulate sounds accordingly.

Figure 2 intuitively conveys an important concept: tracking and sonification activities achieve a reduction of data dimensionality. Under this perspective, stereo mixing corresponds to information

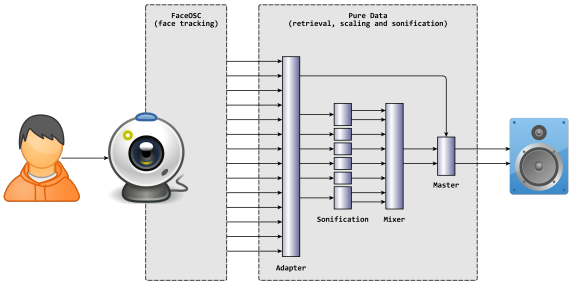


Figure 2: The general schema of the application.

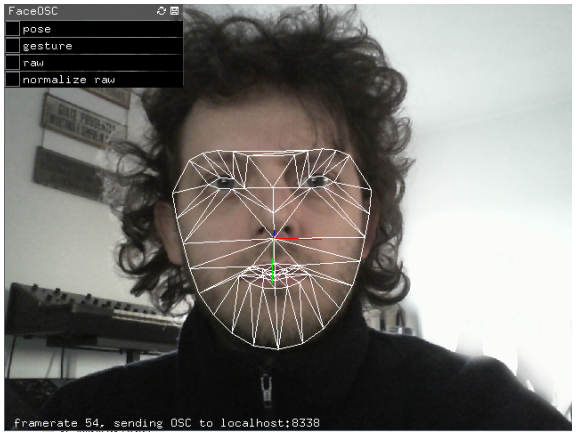


Figure 3: A man's face with a superimposed 3D mask.

multiplexing, since many data mapping results are mixed into a stereo soundscape.

The former part, called *FaceOSC*,[2] is a C++ open-source project developed on top of the APIs written by Jason Saragih. It manages face parameters and their representation in terms of OSC messages. Apart from an entry-level computer, the only hardware device involved is the webcam, that typically allows to manage crucial parameters like saturation, auto-focus, exposure, and frame rate.

When the module is running, it calculates the key facial parameters of the user in front of the webcam, identifying also dark areas and shadows. The three-dimensional model thus obtained is able to follow the movements of the subject in an adaptive way. This operation is performed by evaluating the variations of side lengths for the triangles that constitute the 3D mask (see Figure 3).

The second part of the application, dealing with computations and sonification, was entirely implemented in Pure Data. It can be further subdivided into 4 logical blocks:

1. the *adapter* module, with a specialized submodule for each group of facial parameters (eyes,

[2] https://github.com/kylemcdonald/ofxFaceTracker/ releases

mouth and head orientation). It splits the aggregated information from the video-capture module into specialized data about single facial characteristics, also scaling values to the allowed intervals;

2. the *sonification* module, where data are transformed into sound. This central aspect will be detailed in Section 4;
3. the *mixer* module, where sound levels are managed independently, like tracks in a traditional audio mixer;
4. the *master volume* module, which adjusts the overall sonification volume, turning sound output off when no face is recognized and reactivating it when a face is tracked.

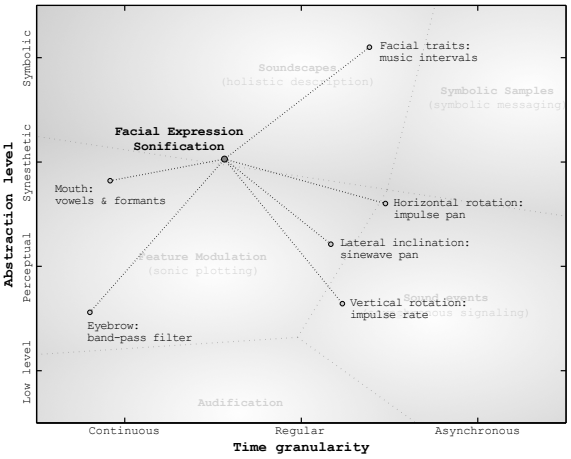


Figure 4: Position of our proposal within the sonification space.

# 4 SONIFICATION DETAILS AND DESIGN CHOICES

The sonification block is the core step of the proposal where facial data are transformed into sound according to many possible algorithms. The global project will explore the effectiveness of different sonification strategies, whereas, in this preliminary work, we are arbitrarily selecting some of the possibilities.

A key point is that data about facial traits and expressions will be captured and sonified as they are, without trying to interpret them (e.g., inferring the underlying emotions), since *a priori* interpretations may be considered biased and subjective. In accordance with (Tanveer et al., 2012), we believe that a visual-to-auditory domain shift is a better method than emotion inference due to:

(i) complexities in expression-to-emotion mapping, (ii) the problem of capturing a multitude of possible emotions coming from a limited number of facial movements, (iii) the difficulty to correctly predict emotions due to the lack of ground truth data, and (iv) cultural differences in expressing emotional nuances.

For these reasons, we prefer to delegate interpretations and inferences to the end user.

The proposed approach takes into account both static and dynamic aspects of face recognition. Figure 4 shows their logical position within the sonification space.

Concerning static parameters, when a face is first recognized, its traits[3] are evaluated in order to discriminate a face from another. The sonification of these characteristics generates a music background for the whole duration of the experience referring to that face. Depending on static facial traits, such parameters are mapped onto sets of notes from different musical scales, with variable envelopes and timbres (e.g., sine or sawtooth waveforms processed by low-pass filters). The resulting sound background, asynchronous with respect to face expressions, represents a soundscape, logically located in the upper-right part of the sonification space. This strategy should help users to discern between different faces, without the purpose of being unambiguous: similar somatic features should result in similar sonifications.

The dynamic parameters we aim to sonify are linked to the variations produced by eyebrows, mouth, and head position in space.

The eyebrow height modulates the central frequency of a band-pass filter applied to a white noise. Since the mapping is a direct one, this part of the sonification presents a continuous time behavior and a quite low degree of abstraction.

Concerning the mouth, its width defines the fundamental pitch of a frequency-modulated wave, while the oral aperture controls both the modulation index and the cut-off frequency of 4 cascading low-pass filters applied to the overall wave. The slope of the filter curve is directly proportional to the number of filters used on the same sound event. This choice responds to an attempt to describe the resonance of the oral cavity when acoustic waves are reflected inside it. Mouth sonification partially falls into *model-based sonification* and partially into *parameter mapping sonification*. In terms of position in the sonification space, it is located in the area of continuous time and synaesthetic abstraction, since it aims to recall mouth shape via vowel sounds.

[3]For *traits* we intend those features that are independent from specific facial expressions, such as the relative position of mouth, jaws and eyes.

Concerning head position in space, vertical rotation is described through a pitched sound repeatedly generated, with an increasing reproduction speed (and, consequently, an increasing pitch) as the face turns towards the zenith, in order to suggest the sense of elevation. Horizontal rotation is represented through an impulsive sound which is activated in the left channel when the face rotates to the left, and in the right channel vice versa, with a playback rate directly proportional to the rotation amount. Finally, lateral inclination is represented by a sine wave, whose source is located in the stereo image depending on the direction detected. All the mapping techniques concerning head position and orientation are activated only when the head is not in a frontal "idle" state (i.e. only when neither tilt nor rotation are null). These observations, along with the adoption of modulated sound events for sonification, push the mentioned data bindings towards the asynchronous area, in some cases more abstract and "synaesthetic" due to the adoption of panning.

The overall sound result of the sonification behaves much like a model-based one, sometimes resembling a soundscape, sometimes recalling a sonic plot; nevertheless, it is quite responsive to changes and characterized by a complex sound texture, which suggests to place it within the sonification space as shown in Fig.4.

Finally, it is worth underlining that the present work focuses on the conceptual framework, without taking into account the pleasantness of the sonification (a non-trivial issue discussed, e.g., in (Hermann et al., 2015) and (Susini et al., 2012)), nor the issues of information overload and audio channel saturation, particularly relevant for visually impaired people. Please note that we are not presenting a musical instrument nor a gesture-controlled synthesizer, especially since some facial expressions derive from autonomic reflexes.

The ultimate goal of the global project will be to let the final user customize the experience by selecting his/her own samples as the source material to be modulated by the sonification engine.

## 5 MeetMeTonight 2016

An early implementation of the prototype was presented at the stand of the LIM (Laboratory of Music Informatics, University of Milan) in occasion of the 4th edition of the European Researchers' Night,[4] specifically during the Milanese initiative called "MeetMeTonight 2016" held at the Giardini Indro Montanelli of Milan on September 30 - October 1, 2016.[5] This free-entrance public event aimed to bring university research outside traditional academic environments, addressing a wide audience ranging from primary school children to elder people. It was the occasion to achieve an early validation of our approach thanks to the presence of thousands of heterogeneous visitors.

The research theme chosen by the LIM for the 2016 edition was sonification, namely the possibilities offered by sound as an alternative communication channel to deliver non-audio information. A number of experiences were proposed to a general audience, including some historical examples of sonification (e.g., meteorological data, global warming, pollution levels in Milan, etc.) and a step sequencer driven by the detection of visitors' position inside the stand.

In this context, an area was dedicated to our face sonification experience. This space was equipped with a front webcam to catch the user's face, two loudspeakers for audio output, a background screen (not visible to the user) to let the audience watch the ongoing computation, and a computer running the required software hidden "behind the scenes".

During this experience, the sonification prototype described above was tested by a large number of people, very heterogeneous in age, education level, technological knowledge, culture and nationality, etc. (see Figure 5). Moreover, the duration of the event, spanning over 2 days, 10 hours per day, allowed to test very different lightning conditions.

Figure 5: People around the installation at MeetMeTonight 2016.

Concerning the *video-capture module*, the application proved to effectively recognize the presence of a face, without false positives, but it showed three cases of wrong tracking: the eyebrows of users wea-

[4] http://ec.europa.eu/research/researchersnight/

[5] http://www.meetmetonight.it/

Figure 6: Different facial expressions caught by the webcam and sonified accordingly during a demo session at MeetMeTonight 2016.

ring glasses, the mouths of bearded faces, and – unexpectedly – also in case of overexposure. These problems, experienced by 30 users out of 150 approx., are pushing us to improve the algorithms in use for this module. In general terms, performances were not affected by age nor gender differences. The *computation module*, on the other hand, proved to be very robust, and it did not produce glitches nor system failures.

At MeetMeTonight we did not plan a formal test phase about the sonification itself. In any case, as a side note, we realized that most visitors (both active users and passive audience) remained positively impressed: reactions ranged from amusement and surprise in hearing a sound description of their own face to the desire to better understand the underlying processes and algorithms.

Users experienced the application in an unsupervised environment, autonomously experimenting with different facial expressions (see Figure 6). The LIM staff was mainly consulted *a posteriori*, answering user requests for technical explanations and providing insights about the sonification intuitively produced.

The ability quickly developed by most people to recognize, even if approximately, facial structures and movements suggests the effectiveness of the mapping adopted in our sonification, but more formal and extended tests will be conducted when sonification techniques have been fine-tuned.

## 6 CONCLUSION AND FUTURE WORK

In this paper we have presented an approach towards face sonification and an implementation based on webcam face tracking and Pure Data processing. Both static and dynamic face characteristics have been considered and sonified accordingly. The goal of such a prototype, presented during a public event, was to explore the potential of technologies applied to the field of face sonification.

Future work will consist in the improvement of feature selection and sonification strategies, the administration of extensive tests under controlled conditions, and, consequently, the implementation and testing of a prototype in a real-world scenario. The test protocol we are planning to use will include a training phase, where users have the opportunity to unveil sonification strategies and relationships with their own movements, followed by a recognition task, consisting in the evaluation of other people's expressions by listening the corresponding sonifications.

We aim to investigate the applicability of this prototype to assist visually impaired people, in accordance with the recognition of sonified objects in real environments that has been proposed, e.g., in (Ribeiro et al., 2012). In this case, the idea is to adopt sonification to allow the recognition of human faces, to track facial expressions, and consequently to let the user infer emotions and reactions.

## REFERENCES

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998). Face recognition using active appearance models. In *European conference on computer vision*, pages 581–595. Springer.

Funk, M., Kuwabara, K., and Lyons, M. J. (2005). Sonification of facial actions for musical expression. In *Proceedings of the 2005 conference on New interfaces for musical expression*, pages 127–131. National University of Singapore.

Guizatdinova, I. and Guo, Z. (2003). Sonification of facial

expressions. In *New Interaction Techniques '03*, pages 44–51.

Hermann, T. (2008). Taxonomy and definitions for sonification and auditory display. In *Proceedings of the 14th International Conference on Auditory Display, Paris, France June 24 - 27, 2008*. International Community for Auditory Display.

Hermann, T., Hildebrandt, T., Langeslag, P., and Rinderle-Ma, S. (2015). Optimizing aesthetics and precision in sonification for peripheral process-monitoring. In *Proceedings of the 21st International Conference on Auditory Display (ICAD-2015)*, pages 317–318.

Hermann, T., Hunt, A., and Neuhoff, J. G. (2011). *The sonification handbook*. Logos Verlag Berlin.

Ludovico, L. A. and Presti, G. (2016). The sonification space: A reference system for sonification tasks. *International Journal of Human-Computer Studies*, 85:72–77.

Puckette, M. et al. (1996). Pure data: another integrated computer music environment. *Proceedings of the second intercollege computer music concerts*, pages 37–41.

Ribeiro, F., Florêncio, D., Chou, P. A., and Zhang, Z. (2012). Auditory augmented reality: Object sonification for the visually impaired. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, pages 319–324. IEEE.

Susini, P., Misdariis, N., Lemaitre, G., and Houix, O. (2012). Naturalness influences the perceived usability and pleasantness of an interfaces sonic feedback. *Journal on Multimodal User Interfaces*, 5(3-4):175–186.

Tanveer, M. I., Anam, A., Rahman, A., Ghosh, S., and Yeasin, M. (2012). FEPS: A sensory substitution system for the blind to perceive facial expressions. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 207–208. ACM.

Valenti, R., Jaimes, A., and Sebe, N. (2010). Sonify your face: facial expressions for sound generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1363–1372. ACM.

Wright, M. (2005). Open sound control: an enabling technology for musical networking. *Organised Sound*, 10(3):193–200.