FP7-ICT-2013-C FET-Future Emerging Technologies-618067

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D1.1.1.bis
Extra Periodic Report

First Author	Davide Andrea Mauro
Responsible Partner	IUAV
Status-Version:	Final-1.0
Date:	October 15, 2015
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	Extra Periodic Report
Date of delivery to the EC:	15/10/2015

Workpackage responsible for the Deliverable	WP1
Editor(s):	Davide A. Mauro, Davide Rocchesso
Contributor(s):	Davide A. Mauro, Davide Rocchesso, Stefano Delle Monache, Sten Ternström, Guillaume Lemaitre, Olivier Houix, Patrick Susini, Nicolas Misdariis, Geoffroy Peeters, Patrick Boussard, Hélène Lachambre
Reviewer(s):	Davide A. Mauro
Approved by:	All Partners

Abstract	This Extra Periodic Report addresses the technical aspects of the project as a follow up of its first year.
Keyword List:	periodic report

Disclaimer:

This document contains material, which is the copyright of certain SkAT-VG contractors, and may not be reproduced or copied without permission. All SkAT-VG consortium partners have agreed to the full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

The SkAT-VG Consortium consists of the following entities:

#	Participant Name	Short-Name	Role	Country
1	Università Iuav di Venezia	IUAV	Co-ordinator	Italy
2	Institut de Recherche et de Coordination Acoustique/Musique	IRCAM	Contractor	France
3	Kungliga Tekniska Högskolan	KTH	Contractor	Sweden
4	Genesis SA	GENESIS	Contractor	France

The information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Document Revision History

Deliverable D1.1.1

Version	Date	Description	Author
First draft	16/07/2015	Import from template	DAM
New core	17/08/2015	New core report with response to reviewers	ROC
Ircam	21/08/2015	Contributions from Ircam	GL
Final	14/10/2015	Version to be submitted	ROC

Project Periodic Report

Grant Agreement number: FP7-618067

Project acronym: SkAT-VG

Project title: Sketching Audio Technologies using Vocalizations and Gestures

Funding Scheme: FP7-ICT-2013-C FET-Future Emerging Technologies

Date of latest version of Annex I against which the assessment will be made: 30/08/2013

Periodic report: P1bis (1st extended)

Period covered: from 01 January 2015 to 23 October 2015

Name, title and organisation of the scientific representative of the project's coordinator:

Prof. Davide Rocchesso,
Dipartimento di Culture del Progetto,
Università Iuav di Venezia
Tel: +39 041 257 1852
E-mail: roc@iuav.it

Project website address: http://www.skatvg.eu

Declaration by the Scientific Representative of the Project Coordinator

I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;

The project (tick as appropriate):

☑ has fully achieved its objectives and technical goals for the period;

□ has achieved most of its objectives and technical goals for the period with relatively minor deviations.

□ has failed to achieve critical objectives and or is not at all on schedule.

The public website, if applicable

☑ is up to date

□ is not up to date

To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project (Section 4) and if applicable with the certificate on financial statement.

All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported under Section 2.4 (Project Management) in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: Davide Rocchesso

Date:

Signature of scientific representative of the Coordinator: ....

Table of Contents

1 Publishable summary **8**

1.1 Context and Objectives . . . 9

1.2 Description of Work . . . 9

1.3 Main Results . . . 10

1.4 Expected Final Results and their Potential Impact and Use . . . 11

2 Core of the report for the period **13**

2.1 Response to Reviewers . . . 13

2.2 Project Objectives for the Period . . . 28

2.2.1 Structure of the work packages . . . 28

2.2.2 Context of the Reporting Period within the Work Plan . . . 32

2.2.3 Project Objectives for the second year . . . 33

2.3 Work Progress and Achievements during the Period . . . 34

2.3.1 Work Package 1: Project coordination . . . 37

2.3.2 Work Package 2: Case studies . . . 37

2.3.3 Work Package 3: Phonetic listening and world-event representation . . . 39

2.3.4 Work Package 4: Perception and cognition of vocalizations and expressive gestures . . . 42

2.3.5 Work Package 5: Automatic imitation recognition . . . 51

2.3.6 Work Package 6: Imitation-driven sound synthesis . . . 55

2.3.7 Work Package 7: Sonic Interaction Design . . . 60

2.4 Project Management during the period . . . 71

3 Deliverables and Milestones tables **75**

3.1 Deliverables . . . 75

3.2 Milestones . . . 75

4 Explanation of the Use of Resources and Financial Statements **78**

5 List of Publications, Networking, and Dissemination Activities **83**

6 Ethical issues **87**

7 Relations with other projects **92**

Index of Figures

1 Timing of the different work packages as foreseen in the DoW. Orange arrows indicate interaction between WPs. 30
2 Interaction between the work packages. 31
3 Activities of Year 2 as extracted from the Redmine project management tool. . 35
4 Principle of pilot studies conducted in Task 4.3. 49
5 S'i' fosse suono. 63
6 Person months of Years 1 and 2 as compared to the overall planned effort over three years. 81
7 Costs per partner for Years 1 and 2 as compared to the overall planned costs over three years. "p": Personnel, "o": Other Direct Costs. 82
8 IRCAM Acknowledge of receipt of the IRCAM database by the Commission Nationale de l'Informatique et des Libertés (CNIL). According to this letter, no answer within two months indicates that the database is accepted. 88
9 IRCAM Acknowledge of receipt of the IRCAM database of video recordings (2015) by the Commission Nationale de l'Informatique et des Libertés (CNIL) . 89
10 IRCAM Consent forms and questionnaire for subjects. 90
11 KTH consent form. 90
12 Reply from EPN-Stockholm 91

List of Acronyms and Abbreviations

DoW Description of Work

EC European Commission

PM Person Month

WP Work Package

GA Grant Agreement

CA Consortium Agreement

M Milestone

Mo Month

Q Quarter

1 Publishable summary

Sketching Audio Technologies using Vocalizations and Gestures
www.skatvg.eu

Pleasant, yet functional. The fusion of these two adjectives describes most of the work of designers, in any domain. In the aural domain, designers aim at giving a pleasant and functional 'voice' to the objects that will populate future soundscapes. Improved safety, health, and quality of life are the possible benefits for society at large.

The idea of SkAT-VG is to exploit the most natural of sound design tools: human voice and gestures. Humans have surprising capabilities in communicating sound, especially in interactive contexts, but a thorough understanding of how this happens and how these capabilities could be exploited requires an ambitious research plan. In the project, the Iuav University of Venice develops design methods and tools based on vocal and gestural sketching. The French company Genesis provides an industrial framework and application contexts. The possibilities and limitations of the human voice as a sound sketching apparatus are charted by partner KTH in Stockholm, where thousands of utterances are being collected, annotated, and classified in relation to the physical phenomena they are supposed to mimic. This work is done together with the Ircam institute in Paris, where three research teams are involved in studying the features (Sound Analysis-Synthesis), understanding the human response (Perception and Sound Design), and exploiting the non-acoustic component (Sound Music Movement Interaction) of imitative or evocative vocalizations.

The four institutional partners of SkAT-VG are also interacting with professional and academic stakeholders in order to define the scope of sound design in future interactive contexts. In the near future, designers might sketch novel responsive sounds, for a car or for a coffee machine, by using the whole expressive potential of their voice and body.

1.1 Context and Objectives

Sketches are materials for the development and communication of ideas in the early stages of any design process. Sketching is commonly done with paper and pencil in visual design, but it is far from being straightforward when it comes to drafting the sonic behavior of objects. SkAT-VG aims at understanding and exploiting vocalizations and gestures, which are the most natural analogues to hand and pencil to communicate sound in action.

SkAT-VG aims at achieving three main objectives, which will lead to improved understanding, classification, and exploitation of human vocalizations and gestures:

1. **Understand.** To extend existing knowledge in perception and production of vocal imitations and expressive gestures;

2. **Classify.** To develop automatic classifiers of vocal and gestural imitations, based on what is imitated, by integrating signal analysis with the physio-mechanics of vocal production;

3. **Design.** To explore the effectiveness of vocal and gestural sketching in sonic interaction design, by exploiting automatic classification for selection and parameterization of sound synthesis models.

These objectives define the three SkAT-VG Milestones:

M1 Accumulation of a large enough database of recorded, sorted, and labeled imitations (Mo12-15);

M2 Automatic classifiers of vocal and gestural imitations into categories of imitated sounds (Mo22-25);

M3 Integrated sketching tools (Mo36-37).

1.2 Description of Work

Most of the work in the first year of the project was aimed at achieving M1. In order to gain a better understanding of non-verbal communication through voice and gesture, the project partners KTH and IRCAM coordinated a campaign of recordings, measurements, and experiments with human participants. At KTH attention was more focused on extracting the primitives of vocal production, as they emerge from non linguistic tasks. A special representation for articulators was devised and used to annotate the recordings. Conversely, IRCAM moved from everyday sounds to see how people organize them into perceptual spaces and use these internal representations to communicate by means of vocal and gestural imitations. Expert and naïve performers were asked to produce vocalizations that vary along elementary auditory features, and the results showed that the fidelity of the imitations is moderate for isolated acoustic features yet the identification of complex sound events is effectively conveyed by vocal imitations.

Milestone M2 is being achieved by the end of second year. In particular, IRCAM has devised new features and developed original signal-based classifiers for vocal and gestural imitations. KTH has shown that articulatory features can be automatically extracted from audio by application of auditory receptive fields. New scientific knowledge is emerging out of the convergence of three scientific perspectives: Production (Articulatory descriptors); Perception (Sound categories); Signals: (Signal primitives).

Progress towards M3 has been centered around interactions with sound design professionals and workshops on vocal sketching (GENESIS, IUAV). These activities have been exploiting the new scientific knowledge through design exercises and prototype systems for vocal and gestural sketching. Two major software frameworks have been produced and are being distributed, for physics-based sound synthesis and for the integration of audio modules in a sound design workflow.

1.3 Main Results

The main achievements during the first 22 months of SkAT-VG have been:

An exploratory database of audio recordings of various imitations, indexed and annotated by action primitives and attributes of origin [Hel14]. Recordings include diverse collected media, bespoke pilot recordings of three subjects, and an indexed set of sounds from the literature [New04]. Deliverable D2.2.1;

A database of recordings of skilled imitators, acquired with a structured and strict protocol, presented in more details in D2.2.2 and D3.3.1;

Scientific evidence that vocal imitations communicate sounds more effectively than verbalizations [LR14];

A corpus of 52 referent sounds that are unambiguously identified (through an identification experiment) as belonging to three families and 26 categories: Included in Deliverable D4.4.1 (Mo21);

A setup for audio-visual recording of imitations of the referent sounds and pilot studies aimed at defining the recording protocol: Included in Deliverable D4.4.1 (Mo21);

A database of about 8000 recordings of vocal and gestural imitations, including audio recordings, high speed video recordings, depth camera and body tracking, and data from inertial measurement units fixed on participants' wrists: Included in Deliverable D4.4.1 (Mo21);

A set of strategies for analysis and classification of vocal imitations, based on morphology [MP15] and on temporal dynamics: Included in Deliverable D5.5.1 (Mo23);

A novel movement representation, based on wavelet analysis and particle filtering, partially implemented under the MuBu framework: Included in Deliverable D5.5.1 (Mo23);

A set of analysis tools for voice analysis using Receptive Fields [LF15a][LF15b]: Included in Deliverable D5.5.1 (Mo23);

A software framework named SkAT Studio, for the integration of modules within custom audio processing workflows: Included in Deliverable D6.6.1 (Mo24);

Physics-based sound model of vehicle motor sounds [BLDB15]: Included in Deliverable D6.6.1 (Mo24);

A new software architecture and public release of the Sound Design Toolkit (SDT) framework, integrated with all the sound models necessary to synthesize the timbral families emerged from the 26 perceptually discriminable categories of sounds: Included in Deliverable D6.6.1 (Mo24);

A set of interviews with sound designers from France, Italy, UK, and Japan: To be included in Deliverable D7.7.2 (Mo36);

Four workshops on vocal sketching [DBMR14, DRBM15]: To be included in Deliverable D7.7.2 (Mo36);

Sketch-a-Scratch, a tool for multisensory texture exploration at the tip of the pen [DRP14, RDP15]: To be included in Deliverable D7.7.1 (Mo36);

miMic, a demonstrator based on the metaphor of the microphone as pencil, it is a system architecture that, through a microphone augmented with inertial sensors, can empower the user with a wide sound palette that can be directly controlled by voice and gesture [RDA16]: To be included in Deliverable D7.7.1 (Mo36);

MIMES, a family of 3D printed, interactive objects to support the vocal and gestural sketching of expressive sounds interactively: To be included in Deliverable D7.7.1 (Mo36);

"S'i' Fosse Suono" installation, an interactive video mosaic of audio self-portraits, developed in cooperation with the sound designer Andrea Cera, to exploit the concepts and tools being developed in the project, and to communicate the long-term vision of SkAT-VG: To be included in Deliverable D7.7.1 (Mo36).

Enlargement of the SkAT-VG community by establishing contacts and networking with interested external stakeholders, designers, scientists, and professionals (section 5 of this report);

Dissemination through public initiatives (World Voice Day 2014, ICT 2015), seminars, press releases, and articles (Interactions, The New Soundtrack, ASA Press Room);

1.4 Expected Final Results and their Potential Impact and Use

The long-term vision of the SkAT-VG project is to introduce non-verbal vocalizations and expressive manual gestures at every stage of the design process, from early sketches to the final evaluation of the sound quality of products, wherever the sonic behavior of objects is relevant for their use and aesthetics.

At the end of SkAT-VG, a designer will be able to use vocal and manual gestures to create synthetic, model-based sounds. As it should happen with sketches at the early stage of any design process, the sound models can be exchanged between collaborators, edited and refined. The vision is that technologies that enable fast and intuitive prototyping, refinement, and evaluation of product sounds will greatly facilitate the process of industrial design, boost the creativity of sound designers, and improve the quality of our sound environment.

If successful, the project will determine significant advances in Europe in the design practices for a variety of products, such as films and multimedia shows (sound effects), games (sound-mediated sense of agency), everyday products (sonic affordances and aesthetics), environments (soundscapes), human-machine interfaces, and vehicles. The concepts developed during the project are also believed to be applicable to areas other than sonic sketching. In fact, once SkAT-VG would have developed the appropriate tools to infer the user's intentions, it will be possible to consider vocal imitations and manual gestures more generally for expressive and intuitive human-computer interactions (including new interfaces for musical instruments and creative interfaces for non-professionals) in different fields of applications. Moreover, as it often happens when the expressive abilities of humans are exploited through technology, new unforeseeable applications and activities may spontaneously emerge.

2 Core of the report for the period

This Addendum to the First Periodic Report serves the purpose of an additional technical review, recommended after the first review meeting, that was held in Venice on January 30, 2015. Since it was recommended that we should show evidence of the suggested refocus of activities, we start the core part of this Addendum by giving detailed replies to the critical comments that were given in the Technical Review Report.

2.1 Response to Reviewers

These notes summarize the actions taken to address the observations made by the reviewers in their Technical Review Report dated February 14, 2015.

Issue:

1.a – However, the reviewers are left with the concern that the first year of work shows an overall lack of ambition. At this point no major scientific achievement beyond the state of the art has been obtained, nor does any such achievements appear to be seriously envisaged after year one. Due to the FET nature of the project, which calls for risk-taking, bold scientific initiatives, such a low profile may turn out to be an issue for the following periods.

Reply:

In the second year the scientific structure of the project is emerging out of the convergence of three directions of investigation. Essentially, human vocal imitations may be represented from these perspectives:

Production: Articulatory descriptors (phonation, myoelastic, turbulent);

Perception: Sound categories, based on sound-production reference mechanism or on sound morphology;

Signals: Primitives emerging from machine learning.

This bold combination of efforts is improving the understanding of human non-verbal voice communication much beyond the state of the art, and defining the possibilities and the limits of the voice-processing tools that will be developed in the project and beyond.

The measurements and analyses of voice production define a radically new research path in phonetics, which extends beyond the realm of verbal communication. The experimental work on perception has already been acknowledged by the scientific community as defining the current scientific understanding about vocal and gestural sound communication. The work on machine learning is twofold: On the one hand it is necessary to provide innovative use cases with suitable (and not necessarily innovative) vocal/gestural recognizers; On the other hand, the specific (and little-studied) nature of vocal imitations is addressed by using innovative methods aimed at extracting vocal primitives (see sec. 2.3.5, page 52).

...

Overall, the project stands on the aforementioned three legs (production, perception, and signals), whose mutual confrontation gives a solid scientific foundation and advances the state of the art. Such progress would have not been possible without the combination of skills and interests that are found in the SkAT-VG consortium. The three legs have not been mentioned before, as far as I can tell.

Issue: 1.c#1 – The reviewers think that there is a risk of achieving only conventional research, if the members follow only "closed-world" approaches. Notably, the project should *refocus* on the core theme of "sound design" *per se*, that is: the creation, invention of sounds by sound designers using voice and gesture. While it is understandable that the consortium chose in a first phase to study vocal imitations, there is a risk of this research boiling down to more conventional areas such as "sound retrieval" or "sound control". This refocus could take the form of creating case studies involving more clearly and upfront sound design, as well as clearer definitions of sound models, clearer identification of key scientific and technical issues and core technological bottlenecks and a shared basic methodology.

Reply: While voice-driven retrieval and control have been active research topics in sound and music computing for a few years, no one has even tried to combine and exploit knowledge on production and perception of voice sounds to design new sounds. So, it is true that sound design is the core theme of the project, and the consortium has been improving its understanding of sound design processes mainly through workshops. In the second year, one sound design workshop was run at the University of York and another is going to take place at Aalborg University Copenhagen by the end of the year. The definition and development of sound models has also made progress, and a set of models is now available to cover the classes emerging from perception studies. The availability of these models allow to nail down the design workshops by extensive use of sound synthesis. While the earliest workshops focused on the practice of vocal sketching for sound communication and definition in design tasks, in the second and third year the design sessions are using the proposed tools to go from sketch to prototype sounds, to be demonstrated in applications. Sound design professionals are increasingly being involved in the project activities.

Issue: 1.c#2 – The interviews with sound designers should be exploited more effectively and could be used as a mechanism to provide the re-focus suggested in recommendation #1. This recommendation is NOT to engage in traditional user studies in order to ascertain an average user viewpoint or use case or solve a particular small interface concern, but rather to find a small number (possibly just one) of expert sound designers to act as "visionary practitioners" and "lead users" for the project. The input from such a user might provide valuable inspiration and visions for the future of sound design, which in turn could be used to ensure a more daring scope for the project. The way this input is integrated in the flow of the project could then in turn form the basis for a stronger shared methodology and work process.

Reply: The main results of the interviews with sound designers have been presented to the other SkAT-VG partners during project meetings, and a summarizing article is in preparation. Among the interviewees, the well-known sound designer and composer Andrea Cera stands up for his curiosity and imagination, grounded on solid product sound design experience. Starting in the summer of 2015, Andrea Cera has been formally hired by IUAV, thus becoming part of the project consortium. He is contributing by (i) using the proposed sound models to "translate" vocal sketches into designed synthetic sounds (e.g., installation *S'i' fosse suono*), (ii) testing the models, tools and demonstrations (Sound Design Toolkit, SkAT Studio, miMic), and (iii) interfacing with industries (especially car manufacturers) to test some of the SkAT-VG results in product sound design. A one-day or two-days workshop with professionals from industry is going to be held in Aix-en-Provence in spring 2016. In the first stage, the workshop will be run as a sound design competition based on sketching. In the second stage, the designers will be asked to work cooperatively at evolving the selected sound sketch into a refined sound prototype.

Issue: 1.c#3 – Interactions with other projects should be made clearer. In particular we will need to ensure that no double funding of work is taking place.

Reply: A section on "Relations with other projects" was added to the final version of the First Periodic Report and is further expanded in section 7 of this intermediate report.

Issue: 1.c#4 – The consortium should clarify, how multimodality will be addressed scientifically. In particular, which models will be used or studied, how voice and gesture can be combined productively to enhance the efficiency of interaction.

Reply: We have collected high-quality audio-visual recordings of both expert performers and lay persons. In particular, WP4 has collected a large database of imitations performed by participants specifically instructed to imitate sounds with vocalizations and gestures. There were two main objectives: i. Understand what types of gestures imitators use and how gestural imitations interact with vocal imitations; ii. Develop algorithms that capture the relevant gestural features. The first objective was addressed by a qualitative analysis of the database of imitations (D4.4.1) and a controlled experimental study. The main conclusion is that gestures are not as precise as vocalizations to describe the acoustic parameters of the referent sounds, but may have a metaphorical function to indicate some aspects of the referent sounds that the voice is unable to convey. The interactive systems will take these findings into account: the voice and gesture typically convey complementary information, which might not be precisely synchronized. To reach the second objective, gestural features were defined, extracted, and submitted to statistical analyses, which confirmed the hypotheses. The new features were used to train a classifier that recognizes if a gesture imitates a noisy or stable sound. More details about these studies are reported in Section 2.3.4 of this report (page 49).

...

The miMic and MIMES demonstrators both exploit voice and gesture jointly. Simple strategies of sensor fusion are adopted, that take the experimental results into account. For example, in the sound models used in miMic some of the parameters are mapped both to audio and to motion features, so that if shaking gestures are applied these gestures override the vocal control in determining the noisiness of the resulting sound.

Issue:

1.c#5 – WP7 (sonic interaction design) has only formal interactions with the rest of the project at the end of the project, which is an issue for integrating user input in the developed prototypes. Visionary use cases should be integrated more clearly with prototype development.

Reply:

It is true that the interactions highlighted in the Gantt chart of the DoW (Figure 1) only occur in the second part of year three. However, dependencies of WP7 on WP2 and WP4 were planned to occur much earlier and were depicted in the PERT chart of Figure 2 of the DoW. These interactions did occur in the first 18 months of the project: WP7 workshop activities on vocal and gestural sketching were strongly informed, and continuously updated by WP2 and WP4 findings, in order to ground design exercises in phonetics and auditory perception of vocal imitations. Conversely, the motor and machines sound design scenario, explored in WP7, was reflected in WP4 experiments on the categories of vocal imitations, and in the annotations of vocal imitations being performed in WP2. Similarly, the iterative development of the mock-ups and demonstrators SkAT Studio, miMic, and MIMES reflected the acquisition of new scientific and design knowledge. Interactions between WP7 and WP5 are starting to occur in the second part of year two, with the availability of the first prototypes of automatic classifiers. Moreover, some assumptions emerging from the interviews with sound designers have been confirmed by the WP4 experimental results, and have been taken into account as guidelines in the development of the prototypes. That is the case, of using the voice to select a timbral family, i..e. the synthetic model representing a sound category, and exploiting gestures to refine the control. A long term collaboration with the sound designer Andrea Cera is aimed at collecting feedback and ideas for the improvement of the prototypes. New possible collaborations for industrial use cases with a French manufacturer are under negotiation and a collaboration with French sound designer Xavier Collet has began, whereas forthcoming workshops will possibly include specific design briefs provided by other companies.

Issue:

1.c#6 – A substantial dissemination strategy is missing. Dissemination is not only about papers, but also about prototypes, which could be made available to some audiences, possibly with focused advertisement campaigns using social media to find suitable expert visionary users.

Reply: Dissemination of scientific results is being pursued using conventional academic channels, which are scientific conferences and journals. Specific attention has been given to more general audiences and lay public which may be informed about the project from scientific magazines, special events, or through social networks. For these reasons, along with the official website (http://www.skatvg.eu/), a number of accounts on different services has been setup:

- Twitter (https://twitter.com/SkATVG);
- Vimeo (https://vimeo.com/skatvg);
- Freesound (http://www.freesound.org/people/skat_vg/);
- Soundcloud (https://soundcloud.com/skat-vg).

Dissemination of software products is done using different approaches, ranging from the free software approach taken with the Sound Design Toolkit (where all sources are available via GitHub) to the distribution of Matlab-produced binaries as it has been envisioned for the prediction models of articulatory parameters using the auditory receptive fields toolbox. The project website is the place were all the links to software are being collected, each new release being advertised via Twitter. The demonstrators are also made widely accessible. In particular, the miMic demonstrator is being developed as a platform product, all the constructive details and documentation being available in the buildinprogress web platform.
The database of vocal and gestural imitations collected at Ircam will be published as a journal paper including the collection of data, the statistical analysis and, potentially, the identification experiments (see D4.4.1), accompanied with a publicly available website with the audio data. Due to its over-terabyte size, the consortium is still searching for a convenient publication strategy for the video database.

Issue: 1.c#7.a – The annotation system described in D3.3.1 and related to WP3 seems a bit complex. To address its reliability, it would be nice to see:

- if different annotators agree with their annotations;
- if the same annotator, when asked to re-annotate the same file after some time (e.g., weeks) from his/her first annotation, still produces a similar annotation.

Reply: A reannotation test was performed to assess the reliability of the annotation. The reannotation test was performed on 40 imitations with a time lag of 6-8 weeks between the original annotation and reannotation. The test indicates that most of the articulatory parameters were correctly reproduced in reannotation, indicating that the annotation system was reliable. Still, some discrepancies were found concerning the use of specific values. In light of these discrepancies, the use of some value oppositions was abandoned (e.g. aryepiglottal vs. ventricular, modal vs. pressed), and in some cases all occurrences of a particular annotation were reviewed using unified criteria for their representation (e.g. dorsal place distinctions and click sounds). The reannotation formed part of the general aim of increasing the reliability of the annotation as well as ensuring uniformity in annotation conventions across the whole database.

Issue: 1.c#7.b – Still concerning the annotation system, in case it will be used to produce features to be used by the classifiers to be studied in WP5, it will be important to be able to generate those features automatically, or at least to minimize the time needed to generate them, in case a semi-automatic generation procedure will be used.

Reply: From the full multi-layered annotation system, KTH has derived a sub-set of three layers (phonation, myoelastic and turbulent) that can be estimated automatically by the prediction models of articulatory parameters using the auditory receptive fields toolbox.

Issue: 1.c#8.a – As a preliminary step, it could be important to see if there are some features that do not change significantly from one recording to the other one, when considering different recordings of the same imitation, done by the same imitator.

Reply: In the experiments performed in WP4, users u were asked to imitate a referent sound p belonging to a sound category c. For each p, they had several trials t to imitate the sound using two modalities m (vocal only and vocal+gesture). For a given u, p and m, the variation over the recordings (trials t) is small according to the audio features we have chosen. However, the variations are larger over m.

Issue: 1.c#8.b – In case of a positive answer to 1.c#8.a, it could be interesting to see if only small changes on the values of some features still occur if one asks the same imitator to re-do the same imitation after some time (e.g., weeks) from his/her first set of recordings. Since a preliminary set of recordings has been already obtained, this could be easily achieved by contacting again the imitators already contacted during the first year of the project, or their subset.

Issue: 1.c#8.c – Moreover, it could be interesting to see if one imitator, after being asked to imitate the same sound several times, "learns" in a sense how to produce the imitation itself. This could be detected, e.g., studying if his/her last recordings are more "similar" to each other than the first recordings (this could happen, for instance, if the values of a subset of features were in a sense invariant in the last recordings).

Issue: 1.c#8.d – Similarly, it could be interesting to see if, when comparing the imitations of the same sound produced by different imitators, the values of some features are similar for the different imitators.

Reply: Our goal is precisely to find audio features o that are invariant over users u for a given prototype p. Currently, we found that several strategies can be used by users u to imitate a given prototype p. Our current research is to determine this set of strategies and to check whether the strategy used depends more on the referent sound to be imitated p than on the user u. In the ideal case we would like $p(o|p,u) = p(o|p)$. Fortunately, the set of strategies used does not seem very extended. Our current direction in terms of research (sound basis decomposition) aims precisely at finding the strategy/sound basis, decompose the imitations over this in order to develop a user-dependent model.

Issue: 1.c#8.e – In case of a positive answer to 1.c#8.d, it could be nice to identify (if they exist) features that typically differ in value when considering different sounds.

Reply: IRCAM (WP4+WP5) has conducted statistical analyses of the databases of imitations (Issues 1.c#8.a, d and e). It consists in comparing a number of features for each of the recorded imitations: pitch, pitch strength, zero-crossing rate, noisiness, modulation, spectral centroid, etc. These analyses have particularly focused on the (dis)similarities between imitations of the same reference stimuli done by the same speaker (the different trials) or by different speakers. The results have shown that the strategies to imitate a given sound are very consistent across participants, and have occasionally singled out a few outlier participants using individual strategies. They have also shown a clear discrimination between voiced imitations (used for abstract sounds or sounds produced by an engine) and unvoiced imitations (systematically used for mechanical interactions), as well as between impulsive and longer imitations, stable and non stable imitations, etc. This knowledge helped for the development of descriptors and automatic classifiers. More details can be found in section 2.3.4 of this report page 46.

We are not planning to have imitators come a second time to redo their imitations. Instead, the plan is to study how imitators adjust their imitations when they are provided with a feedback: the response of a fellow human listener ("I do not see what you mean"), or the sounds generated by e.g. the SkAT-VG tools. This study is planned for the third year of the project (see D4.4.1).

Issue: 1.c#8.f – Although the imitations were made by experts, it could be nice to be able to assess the quality of the imitations themselves, possibly by removing from the database unsuccessful imitations. These could be detected, e.g, by

- asking the experts about how they perceived their own imitation or the ones performed by the other experts; or
- having the imitations judged/recognized by some public, by compiling a questionnaire; or
- making the imitators listen to their own imitations after some time (e.g., weeks) from their recording, and asking them to identify the imitated sound. An incorrect classification at this stage by the imitator would be a sign that the imitation was not good, or at least that it was associated with a class of sounds difficult to be identified.

Reply: The initial plan was to sort *the whole* database into successful and unsuccessful imitations (i.e. imitations that allow recognition). However, we have decided to use a different strategy:

- Conduct statistical analyses on the *whole* database of imitations (see Issue1.c#8.a, d, and e).
- Conduct experimental identification studies with human subjects on *sub-selections* of the database, selected from the previous point.

This choice is justified in detail in Section 2.3.4 (page 45). In short, WP5 informed us that sorted data are of no real use. Conversely, having a large dataset is instrumental to getting good statistics. This is however a deviation from the initial plan that has been discussed and approved by the consortium at the project meeting in Aix-en-Provence, on august 2015.
Furthermore, the results of the statistical analyses have allowed us to isolate participants that follow a unique imitation strategy (in comparison to the global trend shared across participants).

Issue: 1.c#8.g – Given the potentially high number of classes to be recognized and possible similarities among the classes, the multiclass classification problem to be solved when identifying different sounds could be, among various options,

- solved in a hierarchical way, by grouping similar classes into one class, then identifying the correct class inside the group;
- solved by producing, as an output of the trained classifier, not only the best predicted class, but also a second/third best predicted class.

Reply: The two proposed solutions have been taken into account. However, the organization of categories appeared more complex than initially thought. First, since each sound category is only represented by two prototypes, it appeared that people focused more on the imitation of the exact prototype than on the category. This leads to splitting the categories into two. In the opposite, it appeared that similar strategies are used for categories from the same family (e.g., using voiced sounds for machine family, unvoiced for interaction family) which clearly leads to potential grouping and to hierarchical classification. However, machine and interaction sounds also belong to a parallel abstract description. Apart from manual grouping/splitting of the categories in order to perform hierarchical classification, automatic grouping can of course be performed in order to find the best topology. Also, the actions taken to tackle issues 1.c#8.a/d (dataset analysis) have pointed out the existence of similarities between imitations of different stimuli.

The second proposed solution has indeed been adopted. To satisfy the use case development, the classifier now provides a ranking list of results, ordered by decreasing likelihood.

Issue: 1.c#8.h – It would be nice to have a final system that could adapt its classifications to the final user, e.g., by accepting online suggestions by him/her. For instance, in case of an incorrect classification by the system, detected by the user, the system could take into account this feedback from the user by changing its parameters in a suitable way, thus performing a sort of online learning.

Reply: The SkAT-VG project will not produce one single "final system". Instead, by reaching Milestone 3 (Integrated Sketching Tools), the project will provide a set of instruments for sound designers. Among these, there will be a few different kinds of classification, that may result in one or more sound models that can be controlled to produce a sonic sketch. The development of the miMic demonstrator has already highlighted some of the different possibilities. Online learning is certainly one interesting direction that we are going to explore. At the beginning of WP5, user adaptation has been studied through the use of relevance-feedback: the retrieval process is done iteratively by adapting the search weight (weight of the acoustical dimension) to user feedback. However, this is not currently included it the prototypes, but it may be in the future. User adaptation in terms of model adaptation (such as in speech recognition) will be studied later within the project. It is also planned that the prototype MIMES implements the "interactive machine learning" scenario, where users can record their own imitations into the system.

Issue: 1.c#8.i – In case of a large number of classes to be recognized, it would be interesting to study, possibly but not necessarily even theoretically, if the number of examples used to train the classifier is large enough to prevent overfitting, taking into account also the complexity (e.g., number of parameters, VC dimension) of the classifier itself.

Reply: A deep understanding of the classification performance limits poses difficulties which are beyond the scope of the project. As pointed out in Task 5.2, we found that automatic classification on the whole set of 26 categories is not coherent with the content of the dataset. For each classifier, the number of categories will thus be limited to 10, at most, reducing the complexity of the classifier itself. Moreover, analytic studies of the dataset (see issues 1.c#8.a and 1.c#8.d) and analytic listening (Task 5.1) concluded that some categories are imitated in very similar ways; this leads to category confusion, hence limits to the optimal recognition performances are known to exist.

The suggested approach will be taken into account during the development of the classifiers. The machine learning tools exploited in the near future will be, among others, the Support Vector Machines (SVMs). These are well-founded into statistical learning theory, and some results are available in terms of fat-shattering dimension of the training set (we will rely on RBF kernels, thus VC dimension bounds are ineffective). Moreover, model selection by hyper-parameter tuning will be driven by measures on the number of support vectors and by structural risk minimization. As general remark, we point out that in our approach the classifiers are always evaluated by crossvalidation, thus giving insights in the generalization capabilities of the classifiers themselves.

Issue: 1.c#9 – We recommend an additional review to take place 10/12 months after the first review, showing evidence of the recommended refocus of activities.

Reply: The additional review is taking place on October 22-23, 2015 in Lisbon, attached to project participation to ICT 2015. This actually comes after less than 9 months from the first review, and it forced us to anticipate some of the deliverables (D3.3.2, D5.5.1, D6.6.1) in draft form. A refined version of these deliverables will be sent in due time (respectively, months 24, 23, and 24), possibly taking advantage of interaction with the reviewers. The other remaining deliverable to be provided before the review meeting is D4.4.1, due on month 21.

This section is explaining, in the form of a response to reviewers, how the activities have been refocused.

Issue: 2.a – However, some work packages show limited interactions with the rest of the project. For instance, WP7 (sonic interaction design) has only interactions with the rest of the project at the end of the project, which is an issue for integrating user input into the development of prototypes. The multi-modality aspect of the project (e.g. how voice and gesture may complement each other, what technical difficulties, approaches are foreseen for multi-modal integration) has not yet been seriously considered. In addition to this, there is no clear and strong dissemination strategy yet.

Reply: See responses to issues 1.c#4, 1.c#5 and 1.c#6. Moreover, stronger interactions have been established between IRCAM and KTH, and between IUAV and GENESIS. For the first bilateral cooperation, IRCAM and KTH have been collaborating to create the database of referent sounds and imitations. KTH is currently annotating the IRCAM database of imitations. Furthermore, a new collaborative study has started in June 2015. This work addresses the question: "Does speakers' native language constrain their non-linguistic imitative vocalizations?" The details of this study are reported in Section 2.3.4 of this report, page 50. For the second bilateral cooperation, two coordination meetings took place in February 2015, with participants from IUAV, IRCAM and GENESIS. A joint paper entitled *Physically informed car engine sound synthesis for virtual and augmented environments* has been jointly written by IUAV and GENESIS authors, and presented at the SIVE 2015 workshop, part of the IEEE Virtual Reality conference. In march 2015 Stefano Baldan (IUAV) was at GENESIS for a research visit.

WP7 has been collecting knowledge and pieces of software being developed in other WPs, integrating them in mock-ups, and testing them as design hypotheses: A set of SkAT-VG tools and architectures are emerging through demonstrations and workshops (S'i fosse suono, miMic, MIMES, SkAT Studio). The main results of the interviews have been provided by WP7 to the other WPs. Some WP6 sound generators have already been integrated in WP7 applications, and the classification tools from WP5 are in the process of being integrated.

Issue: 2.b.1 – However, some questions are left unclear such as the methodology or rationales for choosing the so-called "primitives", the selection of sound categories to imitate, and the resulting claims that can be made concerning the "coverage" of the database. The sounds in the databases are considered in their entirety, i.e. independently of how they could evolve in time with respect to constraints, as foreseen in the main scenario.

Reply: The rationale behind the selection of referent sounds is fully documented in Deliverable 4.4.1. In short, it was constrained by three criteria. First, it was essential that the selection of sounds provides a good coverage of the different application fields of product sound design. To reach such a good coverage, we adopted three families of referent sounds (three points of view): sounds of machines and products, basic mechanical interactions, and abstract sounds.

The second criterion required that it was also crucial that the selection of referent sounds be balanced in perceptually and cognitively relevant categories. Therefore, the three families were further organized in taxonomies of mutually-exclusive categories. Third, we selected referent sounds that we assumed could elicit a large and balanced variety of articulatory mechanisms. This criterion was met by carefully selecting (KTH - IRCAM collaboration) the categories within each family so that to provide as much variety and as few redundancy as possible as regards the articulatory mechanisms potentially elicited by the referent sounds.

...

The sounds here are not interactive, i.e. they do not react to users' inputs. They last from a few milliseconds to about ten seconds. Some of these have a complex temporal evolution (i.e., a printer printing out several pages). We chose to avoid complex auditory scenes involving a sequencing and layering of events. We reasoned that the human voice is better suited to produce "isolated" sounds than complex sounds. This further led us to consider that we should enable the users of SkAT-VG tools to combine and compose different imitations to create complex scenes.

Issue:

2.b.2 – The main case study foreseen (sound design) was taken as a starting point for the whole project, but is still not clearly defined: Is this process envisioned to be performed in real-time or off-line? How can the sound model (behaviour in time: i.e. the way the sound evolves in time according to the context) be specified? And would such a process be done with editing tools in a second phase or also using vocal and gesture?

Reply:

The project focuses on the sketching phase of the sound design process, the phase that is the least supported by existing sound processing and editing tools. In sketching, many ideas have to be made concrete and comparable in a very short time. Sketching, in any design area, is inherently performative, and it has elements of vagueness that open to discussion and interpretation. That is why the project aims at real-time sound synthesis controlled by voice and gesture in a very direct way. Of course, when sketches are turned to complex and refined prototypes it is necessary to tweak a multitude of parameters, and editing tools may be needed as well. SkAT-VG is mainly using physics-based sound models because they are structurally close to how humans describe sounds through sound-producing phenomena. Moreover, it is scientifically interesting to look for correspondences between how sounds are produced by natural or artificial mechanisms, and how these are mimicked using voice and gesture. However, other kinds of sound models, namely concatenative and granular models, are used as well in demonstrators, thus showing that sound sketching is not tied to a specific category of sound synthesis methods.

Issue:

2.b.3 – The decision to use discrete annotations for vocal articulations could be made clearer. This choice was discussed in section 2.2.3 of the periodic report, but it is still not very clear why "discrete parameters are necessary as input to machine learning".

Reply:

The comment was made in connection with machine learning of articulatory states. The point made there was that time-continuous parameters were not superior or advantageous to discrete parameters when it comes to identifying articulatory states in machine learning. The comment was not intended to imply that time-continuous parameters could not be used as input, but rather that they do not provide any advantages over discrete parameters in the context of articulatory analysis.

Issue: 2.b.4 – In WP4 an ontology of sound types was designed by the consortium (though this word is not used). Some explanations could be provided about how was this ontology designed, and its coverage, i.e. which guarantee of exhaustivity can be provided to users? Here, it could also be interesting to study how vocalizations are used to imitate sounds in the animal world (in case there are publications on this topic).

Reply: The rationale for the selection of referent sounds has been detailed in response to Issue 2.b.1.

We have chosen not to use any animal sounds, because (industrial) sound designers usually do not design such sounds (though Foley artists might do so). However, we are aware of a series of articles that have studied how people perceive imitations of animal sounds (onomatopoeia mostly) [LEP+82, LEW+83, LHE+84]. The results show that listeners better recognize the imitations than the real sounds. But this is probably very specific to onomatopoeias of animal sounds.

Issue: 2.b.5 – WP5. 6 MM were claimed. In this WP, most tasks did not start yet, except for 5.1. However, a number of questions could have been addressed: What new knowledge has been obtained at this point? What are the new ideas being investigated with regard to gesture analysis?

Reply: At the time this issue was written, WP5 had just started. At that time, given that the dataset of vocal/gesture imitations was not available, it was not possible to derive knowledge from it. For this reason, works concentrated on development of generic methods for time-series classification and relevance feedback. At this point of the project, WP5 has started Tasks 5.1 and 5.2. Different sets of audio descriptors have been developed that allowed (Task 5.1) a deep description of the dataset, and (Task 5.2) the development of the blind classification system. The deep description of the dataset has allowed to highlight that various strategies are used to produce the imitations. An extensive listening of the dataset has been completed that will allow to compile a sort of dictionary of the most common imitation strategies. This knowledge should drive the design of tailored descriptors, able to clearly measure the content of each imitation.

About gesture analysis, the analysis of the database brought us new insights to model movement primitives based on their frequency contents. We developed a new approach in using continuous wavelet transform (CWT) and particle filtering. The former proved better than more conventional STFT to describe the gesture phenomena; Particle filtering is being tested to postprocess the CWT. The CWT method has already been implemented as a real-time tool and allows to find and track gesture primitives based on their frequency content. This will enable a deeper gesture description and will be used in the tools for real-time gesture analysis.

Issue: 2.b.6 – WP7. The activities of in this WP appear sound, but it is unclear how they are connected to the rest of the project. A set of interviews were conducted but have not been published. What is the output of this process? How are interviews analysed and fed back into the work of the consortium? What is the emerging balance between everyday existing sounds versus invented new sounds? Through the design of prototypes and interfaces, a number of interesting scenarios have been proposed, for example: using voice to sketch, and gesture to finalize. How are such ideas brought back into the work of the other partners? In the second year it will be of great importance to connect this work with the rest of the project and to identify strong visionary expert case studies.

Reply: The conclusions output from the interviews have been presented to the other SkAT-VG partners during project meetings. They will also be delivered by the end of October in a more formal way, in a written and publishable document.

Some potential case studies have been identified, which will take place during year 3: Genesis has been (and is currently) trying to find a case study and/or testers for SkAT Studio. Several preliminary contacts with industrial stakeholders could not produce any followup, due to confidentiality issues. However, an agreement is about to be found with the acoustic team of a French company for them to be testers for SkAT Studio. Also, a workshop with at least five sound designers will be held during spring 2016.

Issue: 2.d – Some doubts are cast concerning the scientific impact being at the expected level of a FET project. The overall recommendation is to focus on sound design and non closed-world experiments. A stronger shared methodology can be a way to facilitate such a re-focus, as well as clear and federative case-study which integrates contributions from all partners (even if it does not integrate all of them).

Reply: The responses to issues 1.a, 1.c.#1, 2.b.2 are also relevant for this issue. From the scientific viewpoint, the consortium shares the ambitious goal of providing a coherent and integrated description of vocal imitations primitives from a three-folded perspective, that is "Production - WP3, Perception - WP4, and Signals - WP5". The findings being achieved are innovative, as reliable, quantifiable connections and correlations are emerging between perception, articulation, and audio prototypes (i.e., signal primitives) of vocal imitations. The progressive availability of synthetic sound models (i.e., timbral families), to cover the categories of sound primitives that humans can imitate, sets the concrete link to the main focus and research in sound design. The development of the Sound Design Toolkit, SkAT Studio, abstractions from demonstrators – miMic, Mimes – and forthcoming prototypes, now available for use in workshops and case studies, provide sound design tools, sonic "paper and pencil" to support the exploration of the use of voice and gesture in sonic sketch-thinking practices. Hence, the consortium efforts integrate towards the development of a sound design methodology which has been proposed, but never investigated systematically nor fully evaluated.

Issue: 4.a – The first version of the periodic report shows that 9PM have already been spent in management by IUAV out of 12 foreseen, which is excessive (see comments in WP1).

Reply: Issue addressed in final version of the First Periodic Report.

Issue: 4.b – However, the KTH partner may seem to be involved in research that is not clearly related to the project. More links between partners could be envisaged, to ensure that this work, of an understandably more academic nature, can benefit to the project.

Reply: The incorporation of auditory receptive fields (ARF) as a feature-detection mechanism is possible thanks to a close collaboration with another KTH department, Computational Biology (CB). Prof. Tony Lindeberg of CB has, together with Anders Friberg, adapted a more general theory of receptive fields to the auditory perception case. The KTH person-months in WP5 concern only the adaptation of this ARF branch to the purposes of the SkAT-VG project, i.e. detecting voice-related features in the audio signal, and the collaboration with the IRCAM team on providing automatic classifiers.

Issue: 5.a – All first year deliverables are labelled public, but appear to not be publicly available. In a similar manner the various prototypes are not made public in a uniform way. As a result the web site does not provide as much information as could be desired. In general, more pro-active initiatives are expected, to disseminate the all types of results (not only the papers).

Reply: The approved public deliverables are available through the project web site. See also response to issue 1.c#6.

Issue: 5.d – Interviews with sound designers have been conducted and will be reported later in the project. However these interviews could be better exploited to provide insights to the consortium throughout the duration of the project.

Reply: See replies to issues 1.c#2 and 2.b.6.

Issue: 6 – The ethical issues should be reported in a dedicated Section in the periodic reports. Copies of the ethical approvals and of the consent forms should be provided to the EC prior to the activity and sho¡uld be attached in annex to the periodic reports.

Reply: The issue was addressed in a specific section of the final version of the First Periodic Report. In this report, the section has been expanded (see sec. 6) to include new documents clearing up all ethical issues related to experiments and recordings performed at IRCAM and KTH.

2.2 Project Objectives for the Period

The SkAT-VG project has the following three main objectives:

DoW:

1. To extend existing knowledge in perception and production of vocal imitations and expressive gestures;
2. To develop automatic classifiers of vocal and gestural imitations, based on what is imitated, by integrating signal analysis with the physio-mechanics of vocal production;
3. To explore the effectiveness of vocal and gestural sketching in sonic interaction design, by exploiting automatic classification for selection and parameterization of sound synthesis models.

As by the DoW, the objectives for the **second year** are summarized as

Analyzing imitations and automatically extracting meaningful representations (features, primitives) from imitation audio-visual signals;

Predicting the categories (classification) of imitated sound sources;

Defining the timbral families in terms of sound synthesis models;

Defining sonic interactive scenarios and applications.

Together, the above objectives should lead to M2, scheduled for the end of the second year "Automatic classifiers of vocal and gestural imitations into categories of imitated sounds".

2.2.1 Structure of the work packages

The work plan is divided into work packages (WPs), which follow the logical phases of the implementation of the project, and include consortium management (in WP1) and assessment of progress and results (distributed among WPs and particularly relevant in WP7). Dissemination and exploitation are not described as a separate work package but rather distributed into the research work packages. Work packages are further decomposed into tasks, whose verifiable outcomes are called deliverables. The structure of WPs and Tasks, as well as their foreseen unfolding in time, is represented in Figure 1.

The development of the project over three years foresees three Milestones.

DoW:

Milestone M1 (Accumulation of a large enough database of recorded, sorted, and labeled imitations) represents the point where prior studies are integrated with new experimental results to provide enough accumulated data to effectively start the beginning of WP5 (automatic imitation recognition).

Milestone M2 (First implementation of automatic classifiers of vocal and gestural imitations) is the point where the first tools for segmentation and classification of imitations, informed by the knowledge in vocal and gestural production, will be available. This will boost the tasks in WP6 and WP7.

Milestone M3 (Integrated sketching tools) represents the final achievement of the project, where a tool to convert vocal and gestural sketches into instances of sound models will be available and evaluated. Also, the utility of vocal sketching in real-world design contexts will be manifest at this point.

The first period was mainly devoted to the initial fundamental scientific investigations on the perception and production of imitations, as well as to the analysis of requirements and definition of scenarios for applications of vocal and gestural sketching. These studies led to M1 at the end of Year 1. Milestone M2 is being achieved by the end of Year 2 thanks to the studies on machine learning, and to the extensive analysis of vocal and gestural imitations. Moreover, major decisions on how to turn technologies into tools and applications (milestone M3) have been taken in the second reporting period.

The temporal organization of activities over the three years is represented in the PERT chart of Figure 2, where activities are assigned to branches, and the nodes represent significant stages of the project, including Milestones. This representation provides a concise view of the tasks and work packages as they should overlap and unroll in time. Each elliptical node summarizes a period of the project (range of months) and it may contain a Milestone. The arcs are labeled with work package or task numbers.

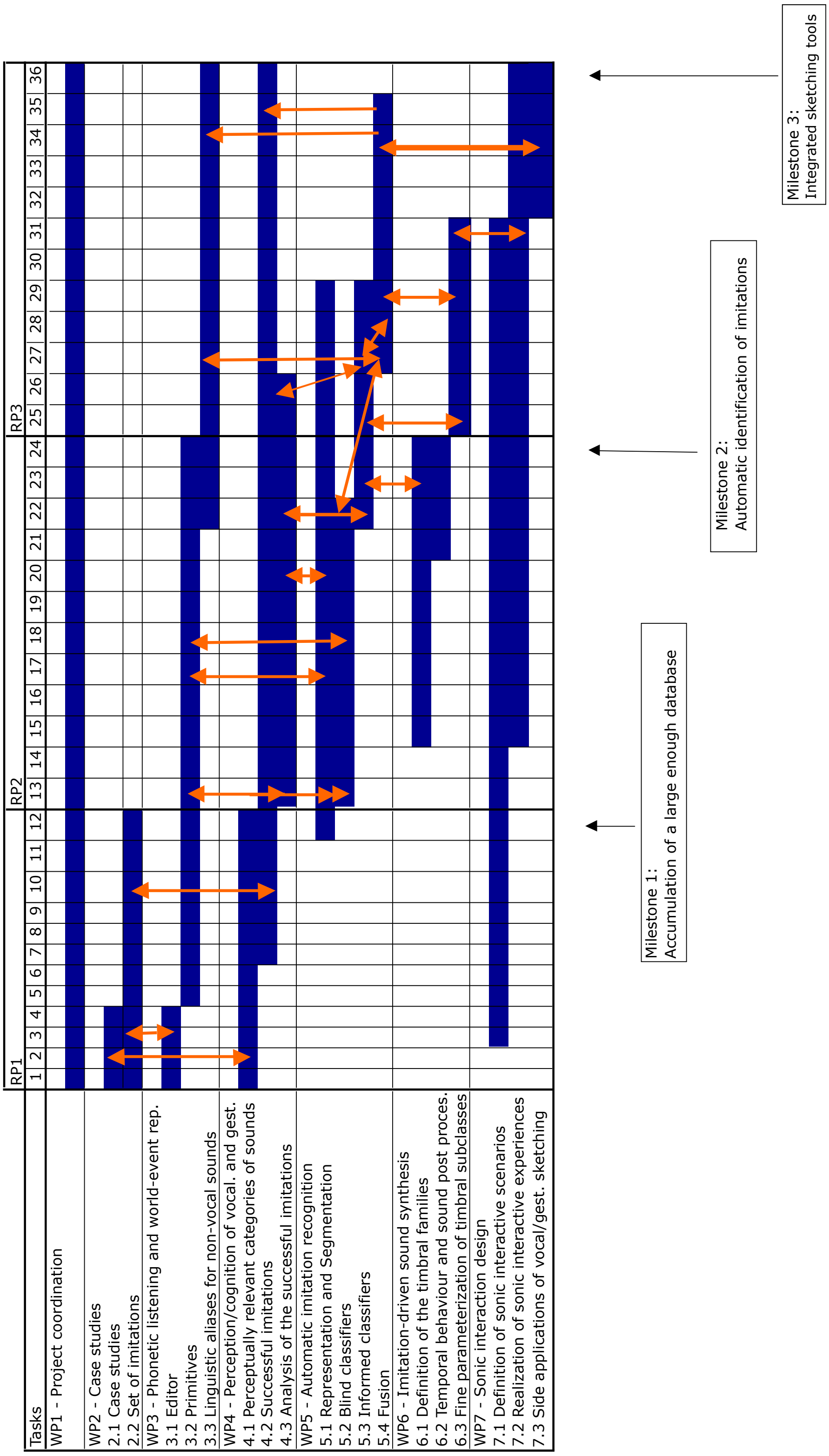


Figure 1: Timing of the different work packages as foreseen in the DoW. Orange arrows indicate interaction between WPs.

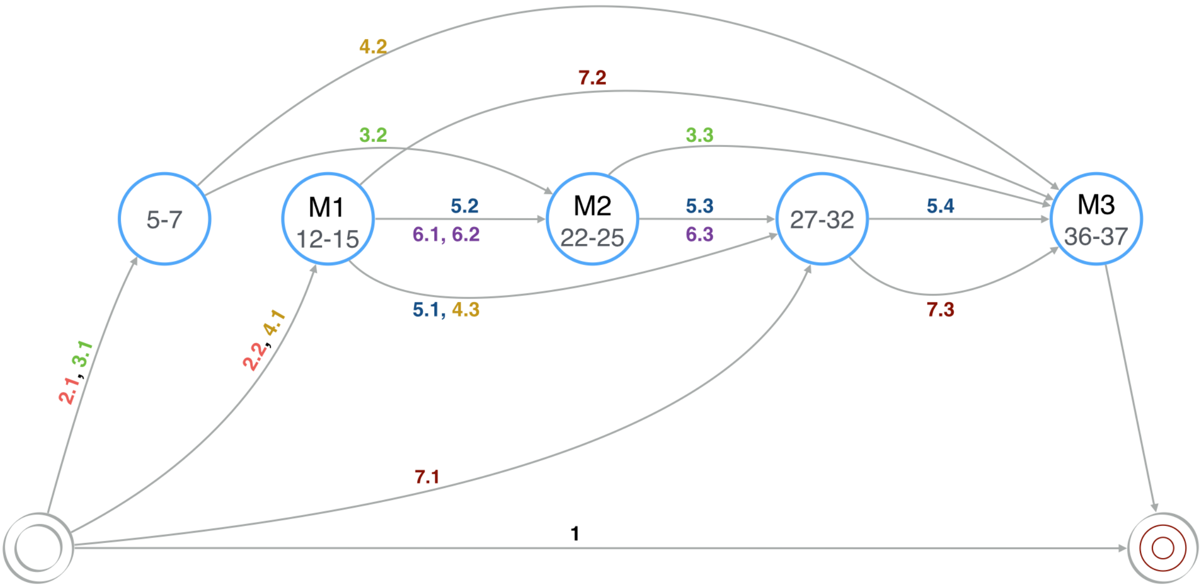


Figure 2: Interaction between the work packages.

2.2.2 Context of the Reporting Period within the Work Plan

Most of the activities in the Reporting Period are finalized at achieving Milestone M2:

> WP3 (Phonetic listening and world-event representations): Task 3.2 (primitives). The annotation of a database of 52 imitations from 4 improvisational actors has been finalized. These data were designed to cover a range of articulatory mechanisms as big as possible and were very successful in this respect. WP5 is now using these articulatory data to develop informed classifiers for articulatory analysis (Task 5.3). In addition, the KTH team is now in the process of making articulatory annotations of the dataset recorded at IRCAM, aiming to make annotations of at least four imitators. These data will provide further input for the informed classification in Task 5.3.
>
> WP4 (Perception and cognition of vocalizations and expressive gestures): Tasks 4.2 (successful imitations) and 4.3 (analysis of successful imitations). A database of 52 referent sounds, and a database of about 8000 audio video recordings of vocal and gestural imitations of the referent sounds has been collected in Task 4.2. The material in Task 4.2 has been passed to Task 3.2 for supplementary articulatory analysis and labeling. Task 5.2 (blind classifiers) uses the large database of recordings of imitations to develop and train automatic classifiers.
>
> WP5 (Automatic imitation recognition): Task 5.1 (representation and segmentation) and Task 5.2 (blind classifiers). IRCAM finished a first version of the classifier of vocal imitations into abstract, interaction and machine categories (Task 5.1 and 5.2). It has been trained on the database produced in WP4. This will be ported to real-time environment for integration into (WP6) to drive the synthesis. IRCAM also developed a gesture analysis framework to compute movement primitives. They have been used to analyse the database created in WP4. KTH is finishing a system for the automatic recognition of articulatory mechanisms, based on the database annotated in WP3. This will allow the start of the Task 5.3.
>
> WP6 (Imitation-driven sound synthesis): Task 6.1 (definition of the timbral families) builds on the categories provided by Task 4.1 and provides sound synthesis models. Task 6.2 (temporal behavior) faces the problem of combining and controlling the sound models, based on the experiences accumulated in WP7.
>
> WP7 (Sonic interaction design): WP7, which spans the whole project, is continuously exchanging information with the other WPs. In particular, workshop activities continue to take place as part of Tasks 7.1 and Task 7.2 (respectively, definition and realization of sonic interaction scenarios). Exercises and basic training exploit the system of voice production attributes of Task 2.1, and the perceptually relevant categories of sounds of Task 4.1. The problem of controlling the sound model results from the interplay between WP7, WP6, and WP5.

Impact-generating activities

> **DoW:** The SkAT-VG project has the potential to have an impact in science (understanding how vocal and gestural imitations are produced and perceived), technology (design tools, auditory displays, and sonic interactive artefacts), and society (designing a better soundscape for human beings and technological artefacts).

After ten months from the First Reporting Period, SkAT-VG has been stratifying layers of new scientific knowledge, which actually represent the new state of the art on the understanding how vocal and gestural imitations are produced and perceived: The classification on vocal production and articulation mechanisms is a foundation work, which goes well beyond the alphabet system of phonetic notation (IPA); the emerging ontology of perceptually relevant sound categories represents the current state of the art on auditory perception of vocal imitations. Despite being yet in an embryonal state, the collection of new basic sound models, arranged in timbral families, represents an important step further, towards the design of physics-based sound synthesis tools, beyond the lab simulation of everyday sound phenomena. These results are summarised in the installation-demo "S'i fosse suono", developed in collaboration with the sound designer Andrea Cera. This proof-of-concept provides a glance on the ideal SkAT-VG methodology and tools, envisioned by the project in the long term.

SkAT-VG has been expanding the network of external stakeholders and expert sound designers, in order to tailor the design of sketching tools. In addition, educational research in workshop activities is involving students and professionals towards the development and systematization of sonic sketching exercises and practices. Finally, SkAT-VG dissemination strategy through press releases, publications in scientific magazines and journals, and public events has been strengthening the attention toward the SkAT-VG project, both in scientific communities and in the general public.

2.2.3 Project Objectives for the second year

The objectives for the second year of SkAT-VG are given in Table 1. It includes the declared Milestone and Deliverables as well as other objectives indicated in the DoW. Elements of work of the second year that will contribute to future Milestone and Deliverables have been grayed out in the table.

Milestones	Deliverables	Other
M2 - Automatic classifiers of vocal and gestural imitations into categories of imitated sounds	D3.3.2 - Final comprehensive annotation of the database of imitations	
	D4.4.1 - A large set of vocal and gestural imitations	Publications
	D5.5.1 - Blind classifiers of imitations	Tools
	D6.6.1 - Automatic system for the generation of sound sketches	Tools
M3 - Integrated sketching tools	D4.4.2 - An analysis of how vocal and gesture primitives are sequenced	Publications
	D5.5.2 - Informed classifiers of imitations	Tools
	D6.6.2 - Front-end applications for interactive sound prototyping	Tools
	D7.7.1 - Interactive prototypes realized with the SkAT-VG tool	Demonstrations
	D7.7.2 - Applications of vocal sketching	Workshops
		Publications

Table 1: Objectives of SkAT-VG for the second year.

2.3 Work Progress and Achievements during the Period

The realization of SkAT-VG tasks and activities is represented in Figure 3, which has been automatically generated by the project-management tool.

Table 2 concisely shows the progress of WPs and Tasks in Years 1 and 2, with the corresponding deviations from the planned progress. The deviations in each task are explained in the rest of this section. Table 3 assigns the Year 2 deviations to individual partners of the Consortium.

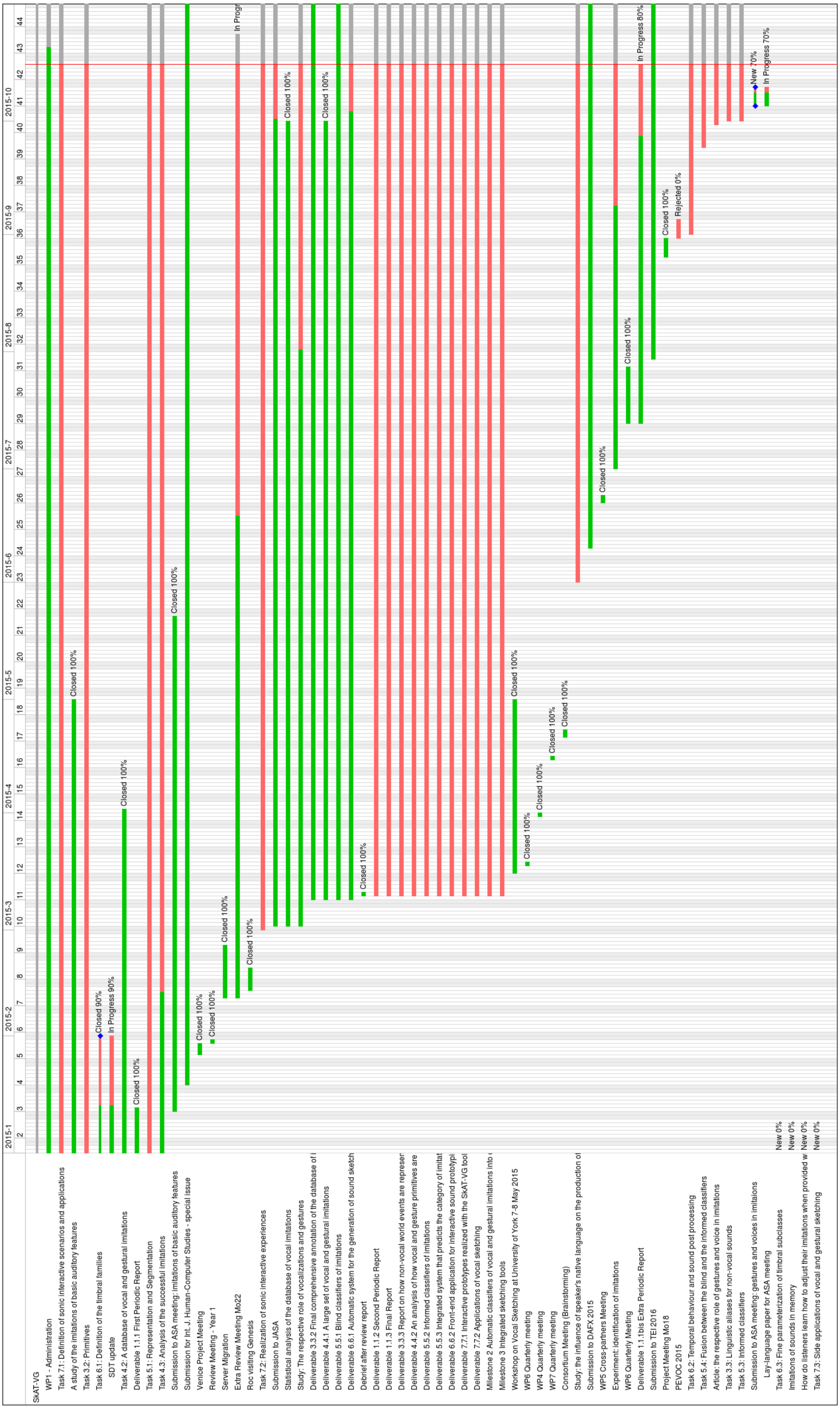


Figure 3: Activities of Year 2 as extracted from the Redmine project management tool.

WP	Description	Task	Progress	Deviation
WP1	Coordination		AtS	No
WP2	Case studies	T2.1	Completed	No
		T2.2	Completed	Minor slight delay
WP3	Production	T3.1	Completed	No
		T3.2	Almost done	Yes discrete annotation, more PM
		T3.3	Started	No
WP4	Perception	T4.1	Completed	No
		T4.2	Almost done	Yes: Statistical analysis of the whole database, annotation of only a subpart
		T4.3	In progress	No
WP5	Machine Learning	T5.1	In progress	No
		T5.2	Completed	No
		T5.3	Started	No
WP6	Synthesis	T6.1	Almost done	No
		T6.2	Started	No
WP7	Design	T7.1	In Progress	Minor anticipations
		T7.2	In progress	No
		T7.3	Not Started	No

Table 2: Task progress and deviations.

Partner	WPs	Deviation
IUAV	WP7	Anticipation: automatic clustering, extension of sound synthesis palette
IRCAM	WP4	Modification: statistical analysis of the whole database, annotation of a subpart
KTH	WP2	Extension: additional subjects have been recorded in year 2
	WP3	Modification and Delay: discrete instead of time-continuous annotations; more PMs (transferred from WP2) needed for annotation
GENESIS	WP7	Anticipation: sound synthesis and control of vehicle sounds

Table 3: Deviations per partner.

2.3.1 Work Package 1: Project coordination

DoW: to ensure financial and administrative management of the project; to develop a spirit of co-operation between the partners; to ensure consensus management and information circulation among the partners, to ensure project reporting and interface with the Project Officer; to co-ordinate and control project activities to keep it within the objectives, to ensure quality management of the project.

Within WP1, resources are dedicated to manage the communication inside the project consortium with specific tools (e.g. Redmine, ownCloud) and towards the European Commission, to prepare and conduct project meetings and reviews, to prepare the minutes (see Section 2.4), to manage the fund transfers towards the partners, to monitor and report on the execution of the financial plan (more details in Section 4). A GANTT of the activities of Year 2, as extracted by the Redmine project management tool, is shown in Figure 3.

The Project Management during the period is explained in section 2.4.

Summary:

No Significant Deviations from Annex I

All Objectives achieved according to Schedule

No Corrective Actions Required

2.3.2 Work Package 2: Case studies

In WP2 the Consortium created an exploratory database of imitation case studies (Task 2.1). Then, a controlled procedure was devised for making a database of the productions of skilled imitators (Task 2.2).

Further details are available in the First Periodic Report.

Task 2.1: Case studies

DoW: Case studies will be collected from commercial recordings. Their quality will probably not be adequate, but they will be inspirational and reveal what skilled imitators can do. Then, a list of “action/sound primitives” will be defined: the classes of basic mechanical interactions that subjects can imitate. They will have to fulfill three requirements: being the simplest imitable sounding interactions, being combinable to form the sound events of Task 4.1 (IRCAM), and covering the timbral families and scenarios of Tasks 6.1 and 7.1 (IUAV). The emphasis on “what the voice can do” imposes an approach based on the source filter model of sound production. Combining potential sources (trains of impulses, noises) and filters (from low- to high-Q resonators) a priori suggests a number of classes (impacts, whistles, bubbles, etc), but a precise set will be defined and limited at the output of Task 2.1.

Progress
Work on this task was described in the First Periodic Report.

No Deviations from Annex I
Task 2.1 is completed.

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 2.2: Set of imitations

DoW: High-quality recordings of imitations of the primitives will be made using skilled imitators, in a digital video format with both hi-fi audio (airborne and contact microphones), and video (frontal and profile views of mouth and hands). The video and contact microphone signals will assist the phonetic transcription in WP3. Task 2.2 will take care of using mechanical sounds, as well as sounds generated by IUAV in Task 6.1, as a the substrate of the imitations. This will ensure the compatibility with WP6 and WP7. Such classes will be analyzed one at a time in WP3.

Progress
Work on this task was described in the First Periodic Report.

Deviations from Annex I
For WP2, there are no deviations from Annex I, other than minor technical modifications of the setup.

Objectives achieved according to Schedule
The KTH recordings of skilled imitators were completed in early months of the second year of the project. The final annotation of these recordings is completed and they are now being used as input for machine learning in WP5. A set of recordings of non-professionals made at IRCAM are now also be being annotated by the KTH team and will be added to the pool of data used for developing the informed articulatory classifiers in WP5.

The objective of the recordings in T2.2 is to provide WP5 with a sufficient amount of data to develop the automatic informed recognition. While a quantifiable target for the amount of data can not be given (since it is unknown beforehand how much input data WP5 actually requires), the KTH team expects that the existing recordings of KTH (professionals) and IRCAM (non-professionals) will be sufficient. We therefore regard Task 2.2 as completed.

Corrective Actions Required
No Corrective Actions Required

2.3.3 Work Package 3: Phonetic listening and world-event representation

This WP seeks to document how real-world events are imitated at the phonetic level. Recordings of skilled imitators were analyzed and annotated by phoneticians. The resulting labelling, paired with the audio, then forms the input data to the machine-learning developed in WP5.

Task 3.1: Editor

DoW: Implementation of a data format and a graphical parameter editor for performing manual transcription of the selected articulatory parameters, from audio/video files to multitrack data files.

Progress
Work on this task was described in the First Periodic Report.

No deviations from Annex I
The task is completed.

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 3.2: Primitives

DoW: Time-continuous phonological transcriptions of a number of imitations of the action primitives obtained in WP2. This will initially be done manually, using spectrograms, audio and video. These transcriptions will be in the form of continuous traces of manually estimated articulatory parameters (APs), connecting sequences of landmark points of recognizable phonetic configurations, i.e.

phonemes. AP values represent the degree of activation of such phonetic sources as phonation, frication, plosives; and modifiers such as vowel resonances and stops. AP values and phonemes will form the vocal primitives of the project. In unclear cases, the imitators will be subjected to direct articulatory measurements. The vocal primitives will be stored synchronously with the original audio, accompanied by annotations of gestural primitives, and processed by IRCAM in WP4 and WP5. For running a perceptual-feedback validation of the transcription work, it would be valuable, though not strictly necessary, to have an articulatory synthesizer that is driven to reproduce the vocal sounds interactively during the manual transcription. It is not clear that the current state-of-the-art in articulatory synthesis is good enough for this task. The existing TADA system from Haskins Laboratories will be tested as a possibility.

Progress

As discussed in the first periodic report, it was found that time-continuous representations of articulatory parameters would be less practical than discrete representations. For some articulatory parameters, time-continuous annotations would be difficult to implement and unlikely to yield any benefits in classification over time-discrete annotations. Therefore, for the articulatory annotation the conventional paradigm of discrete phonetic attributes was chosen, but supplemented, where necessary, with an optional degree of scaling for position or intensity. The revised annotation scheme including some examples, is described in the First Periodic Report (D3.3.1).

Work in WP3 since the first Review meeting has been focused on 4 main points. The first was to complete the articulatory annotation of the 4 professional actors recorded at KTH. This annotation has now been finalized. This annotation work has aimed at increasing the reliability of the annotation as well as ensuring uniformity in annotation conventions across the data base. This work included a reannotation test, as recommended by the reviewers at the first review meeting.

Second, WP3 has been working with WP5 to decide on the articulatory classes used as classifiers in WP5. A primary aim of WP3 is to provide WP5 with data for machine learning that are annotated with respect articulation. This annotation is rich in articulatory detail to ensure that all aspects that may be significant for conveying referent sounds through imitation are covered. The KTH team analysed the articulatory data to arrive at a set of articulatory parameters that were amenable to analysis using Auditory Receptive Fields (ARF). Three contrastive articulatory features were identified as highly significant for imitation. First, the presence vs. absence of vocal fold phonation. Second, the presence vs. absence of slow myoelastic vibrations (not produced by the vocal folds). And third, the presence or absence of turbulence in the signal. These three contrastive parameters are used as a starting point for the articulatory machine learning performed in WP5. The choice of these 3 parameters was based in part on the analysis of the correlation between articulatory settings and the basic category of the referent sounds performed in Task 2.1. This analysis established a strong correlation between vocal fold phonation and referent sounds related to engines/motors and animals. Slow myoelastic vibrations (like trilling with the tongue tip) were also associated with engines/motors and animals, but were used to convey a lower pitch. However, vocal tract turbulence of different kinds correlated more strongly with basic mechanical interactions in gases and solids. From the point of view of WP5, it was considered feasible to base the initial ARF machine learning on these three binary contrastive articulatory parameters.

Third, WP3 has been working to extract appropriate data sets for use in WP5. The articulatory parameters used in the initial machine learning in WP5 make use of five of the eight articulatory annotation tiers in the database: Vocal fold phonation, Aryepiglottal vibration, Nasality, Lip manner and Tongue manner. On the basis of various combinations of these tiers (and the values within each tier), two data sets were generated for each articulatory parameter; for example, a data set containing slow myoelastic vibrations, as well as a (representative) data set containing occurrences in which slow myoelastic vibrations were absent.

Meetings and Events

Bilateral IRCAM - KTH meeting, for the concurrent development of WP2, WP3, and WP4: July 23, 2015, in Paris.

Deviations from Annex I

The choice of a system of discrete rather than time-continuous annotations is a deviation from Annex I. The annotations have been made not only of the action primitives, but also of more compound imitations.

Status Relative to Schedule

At present the annotation work is on schedule. The final comprehensive annotation of the four professional subjects recorded at KTH is already completed. The annotation of at least 4 non-professional IRCAM subjects is in progress and is expected to be completed by Mo24.

Corrective Actions

The person-power assigned to the annotation task was doubled.

Task 3.3: Linguistic aliases for non-vocal sounds

DoW: When the target sound is very far from what is physically possible with the voice, sound symbolism in the form of onomatopoeias (invented sound-words) is what people generally use. KTH will consider how vocally inaccessible sounds might be specified using word-like aliases or semi-symbolic sounds. Because these are usually transcribed differently in different languages, SkAT-VG may initially need to define a new unambiguous phonetic representation, specific to the SkAT-VG system, that would have to be learned by its users. The International Phonetic Alphabet will be used as a starting point. By asking prospective SkAT-VG users to adopt a well-considered convention for such linguistic aliases, the system, being largely phoneme-based, could be trained or even constrained to map certain sound-words to non-vocal real-world events. A later refinement of the system (not in SkAT-VG) might ultimately make it more language-specific.

Progress
Task 3.3 examines the use of onomatopoetic words as input to the SkAT-VG system. In the initial phase of Task 3.3 we conduct workshops in collaboration with WP7 that examine the potential uses of onomatopoetic input in the system. Initially, the focus of the workshops is on gathering information on the use of onomatopoeia in sound sketching, especially in sketching sounds that are difficult to convey faithfully using vocalizations. The results from the workshops will help us decide which categories of sounds the onomatopoetic input should encompass. In the second phase of Task 3.3, building on the results from the preceding workshops, we explore to what extent it is feasible to use onomatopoeia as input in the system. Also, we assess the importance of language specific onomatopoeia, i.e., whether the system can utilize a "universal" set of onomatopoetic word forms or whether the input should be language-specific and thus reflect the conventions of sound-to-word conversion of individual languages. Lastly, we consider the eventual implementation of the onomatopoetic component in the SkAT-VG system.

Meetings and Events
Bilateral KTH - IUAV Skype meeting, for the concurrent development of WP3, and WP7, with respect to the planning and schedule of the joint SkAT-VG workshop on vocal sketching at Aalborg University in Copenhagen (November 23-27, 2015): October 6, 2015.

Deviations from Annex I
No deviations.

Status Relative to Schedule
The work on the task has just started.

Corrective Actions
None.

2.3.4 Work Package 4: Perception and cognition of vocalizations and expressive gestures

Overall, WP4 has three main objectives. First, WP4 aims at studying how people produce and perceive vocal and gestural imitations with *experimental studies*. The second objective is to provide the project (WP6 and WP7 in particular) with *datasets and new insights* on how vocal and gestural imitations can be practically used in the context of sound design. The third objective is to use vocal and gestural imitations as new tools to *investigate sound perception* *and cognition in general*

During the first half of the project, WP4 mainly addressed the first and second objectives. The studies and recording sessions conducted by WP4 resulted in a better understanding of how imitators produce vocal and gestural imitations, and what information users perceive from these imitations. The results of the studies have provided the project with databases of vocal and gestural imitations, and led the project to better define how vocalizations and gestures should be used in the SkAT-VG tools.

In terms of tasks, WP4 completed Task 4.1, nearly completed Task 4.2 during the reporting period, and made progress toward Task 4.3.

In WP4, the Consortium collected and selected a database of 52 referent sounds, sorted in perceptually relevant categories (Task 4.1, see First Periodic Report).

The main achievement of WP4 in the reporting period has been the completion of Task 4.2: WP4 collected a database of about 8000 audio/video recordings of vocal and gestural imitations of these sounds. This database has been manually sorted to eliminate technically flawed recordings, resulting in about 6000 usable recordings. Statistical analyses of the acoustic features were conducted on the whole database of vocal imitations, the video recordings were qualitatively analyzed, and identification experiments are currently being developed on a subpart of the database. Importantly, these studies introduce another notion of "sketching", based on a sparsified representation of the sound signals, and compare human imitations with sparsified sketches.

In the reporting period, IRCAM also conducted a series of pilot studies (Task 4.3). A first study analyzed how two experts (professional singers specialized in extended vocal techniques) and two lay participants imitated different sets of sounds varying along elementary auditory features: tempo, sharpness, pitch, and onset. A second study investigated the respective roles of voice and gestures during imitation. The results of this study have identified different gestural strategies (from descriptive to metaphorical, e.g. using shaky gestures to express noisiness, conveying one piece of information with the voice and another piece of information with gestures), and developed the computation of gestural features (based on a wavelet representation) used to train gesture classifiers.

Overall, these studies have resulted in a deeper understanding of what happens when an imitator imitates a reference sounds with voice and gestures. We have learned that the vocalizations *sound similar* to the referent sounds to the listener, despite a *moderate accuracy* in precisely reproducing basic acoustic features. However this similarity is sufficient for the listeners to recover a good deal of information about the *sound source*. Furthermore, these results have also taught us that imitators use gestures in a very different way than they use vocalizations: gestures in the air are not very precise at reproducing the trajectory of acoustic features, but they convey very important information to the receiver by symbolic means.

WP4 has planned new exciting studies for the next period. In particular, IRCAM has started to investigate the *semantics* of the imitations. IRCAM will also collaborate with KTH to assess the constraints that speakers' native languages impose on non-linguistic vocalizations. Our first insights suggest that these constraints are much more limited for non-linguistic than linguistic vocalizations, which is something unexpected according to the current theories of speech production. Another exciting future work will be to further assess the bimodal aspects of sound imitation, and whether the *combination* of voice and gestures is critical for the perception of imitations. After the deep characterization of the phenomenon of imitations, the next and final step of WP4 will consist of studying imitations of sounds *in memory* (instead of referent sounds played back). This final step will test the bold idea that imitations of sounds can become a tool to study sound perception and cognition in general.

Task 4.1: Perceptually relevant categories of sounds

DoW: Perceptually discriminable categories of sounds will be defined. The task will be based on existing knowledge, and complement it by recording sounds, and conducting categorization and discrimination experiments to define different categories of sounds in terms of interaction, temporal and timbral properties. A strong interaction with IUAVwill occur, to define the relevant temporal and timbral families of Task 6.1. Eventually, the outcome of Task 4.1 will consist of a large set of sounds, sorted in perceptually relevant categories. These categories will consist in combinations of the mechanical action primitives studied by KTH in Task 2.1.

Progress
Work on this task was described in the First Periodic Report.

No Deviations from Annex I
Task 4.1 is completed.

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 4.2: Successful imitations

DoW: Successful imitations will be sorted out by conducting identification experiments. Imitations of the perceptually relevant categories of sounds (Task 4.1) will be recorded focusing on different modalities: vocal and gesture. Particular attention will be drawn on expressive gestures related to temporal evolution of timbral properties based on the list of action/sound primitives defined in Task 2.1, and complementary characteristics found in Task 2.2. The outcome will be a large set of vocal and gestural imitations that successfully convey the different categories of sounds, and a set of gesture primitives in addition to the set of vocal primitives obtained in Task T2.2. This database will form the set of examples (vocal and gesture) required in Task 5.2. Task 4.2 will also study imitations of sounds generated by IUAVin Task 6.1. A specific methodology will be developed to handle the large number of sounds required by WP5: Tasks 4.1 and 4.2 therefore go in parallel with Tasks 2.1 and 2.2.

Progress

Task 4.2 is now nearly completed. The core of Task 4.2 consisted in recording a large database of audio/video recordings of vocal and gestural imitations. Then, statistical analyses of the whole set of audio recordings, qualitative analyses of the video recordings, and identification experiments using a sampling of the database were (or are being) conducted.

Database of vocal and gestural imitations The recording setup involved the synchronization of several audio and video devices: a microphone, a webcam, a fast HD camera (GoPro), a depth camera (Microsoft Kinect), and inertial measurement units fixed on participants' wrists (Ircam's "Musical objects"). In December 2014, IRCAM started the recording of vocal and gestural imitations of the 52 referent sounds selected in Task 4.1, based on pilot studies and discussions with KTH. The pilot studies allowed to refine the recording protocol.

The recording sessions were completed in April 2015. Fifty participants took part and produced a total of 7929 imitations (4410 in the V condition, 3519 in the V+G condition), making an average of 1.5 imitation per participant and per referent sound. This makes a total of about one terabyte worth of data. The data were first manually screened by listening to each audio track and watching the videos. At this stage, 536 imitations (7%) were rejected because of the poor quality (resulting in a total of 7393 files, 4062 in the V condition, 3331 in the V+G condition). Most issues resulted from participants stopping the recording before the end of their imitation. Twenty-one participants produced at least one good audio recording for each referent sound in the V condition, 29 in the V+G condition.

Some other data files were also missing because of technical issues during recordings. In the V condition, there were 3586 imitations with both the audio and GoPro data files. In the V+G condition, there were 2726 imitations for which the audio, GoPro, Kinect, and MO files were recorded properly. In total, this results in 6312 usable imitations (i.e. with all data collected; 20% of the imitations missed at least one of the data files).

The whole database (in the form of a hard drive) was physically delivered to GENESIS and IUAV on April 20, 2015; and to KTH on June 12, 2015, via regular mail (a first drive had been previously damaged during shipping).

Analysis and sorting of the database The initial plan (as written in the DoW) was in effect to sort *the whole* database into successful and unsuccessful imitations (i.e. imitations that allow recognition). This could be done by having individual participants (e.g. the imitators themselves) coming to the lab as suggested by the reviewers, but the size of the database (about 7000 files) makes this practically very difficult. One solution was to use crowdsourcing tools, such as Amazon's Mechanical Turk. There were two reasons for sorting the whole database: 1. Provide WP5 with "good exemplars" for training the classifiers; 2. Provide WP4 with identification data in order to study what makes an identification successful in Task 4.3. However, after much pondering, we realized that such a strategy was not optimal:

- Discussions with WP5 informed us that they preferred more exemplars than good exemplars

- We realized that the results obtained with crowd sourcing methods are of poor quality [CP15].

So we considered the following facts. First, the main advantage of a huge database is that it allows to run automatic statistical analyses on large samples, and thus obtain solid statistical data. Second, its main drawback is that it makes any kind of study using human subjects very daunting, if not unmanageable. Therefore we decided to replace the goal of sorting the *whole* database by two other studies:

- Conduct statistical analyses on the *whole* database of imitations
- Conduct experimental identification studies with human subjects on a *sample* of the database, selected from the previous point. The main advantage is that we will able to select referent sounds to fit our research hypotheses and get good quality results.

This is however a deviation from the initial plan that has been formally discussed and approved by the Consortium in august 2015 during the project meeting in Aix-en-Provence.

Statistical analyses of the whole database IRCAM has conducted statistical analyses on the databases of imitations. The main steps of this task were:

- Computation of audio descriptors found in WP5 on the signals, related to:
  - general characteristics: noisiness, pitch strength, zero-crossing rate, absolute duration;
  - continuity/repetitiveness: number of non-silent regions, ratio between silent and non-silent portions of the signal;
  - evolution in time: signal slopes and modulations, both in amplitude and frequency.

  These descriptors characterize what the voice is capable of doing [LJH+15]; they are computed and summarized over time.
- Apply machine learning tools, such as hierarchical clustering and principal component analysis, to enhance the main cues of the dataset.
- Visualize in a meaningful way the obtained statistics, finding conclusions.

Based on these computations, we can visualize the descriptors values across speakers and across referent sounds. This analysis gives a fairly good overview of what is in the databases, hence helping to single out idiosyncratic behaviors.

The results show a very clear distinction between tonal and noisy imitations. Imitators produced tonal imitations when they imitated machines (especially those with engines, motors, and rotating parts) and abstract sounds, and produced noisy imitations when they imitated interactions (that mainly are broadband noisy signals). Imitators were able to reproduce the temporal patterns like repetition and the impulsive profiles of referent sounds. In particular, in the case of abstract sounds, the referent sounds with an upward profile were well accentuated by imitators who produced imitations with a marked upward profile. Overall these results show that imitators were able to reproduce the main aspects of the referent sounds ("tonalness" and temporal profiles), and that these aspects are well captured by the descriptors developed in WP5.

In addition, we found that the strategies were remarkably consistent across imitators. We could not cluster imitators based on their imitations. Only occasionally we found a few outlier imitators that somehow diverged from the main shared strategy.

Outcomes from these analyses are also used to select relevant subsets of sounds, which will be then used in follow-up experiments. The results of the different steps are included in D4.4.1 and we are planning a publication of these data together with the database.

Qualitative analyses of video recordings IRCAM also conducted qualitative analyses of the video recordings, by watching the whole database of imitations and identifying different strategies for the gestural imitations. This analysis was used to draw the hypotheses that we used to study the respective roles of vocalizations and gestures with more formal experiments.

Identification experiments This work addresses the question of the "success" of an imitation. So far [LR14], we have addressed this question by using a forced-choice task (N-response classification task), in which participants listen to an imitation and select the best-matched referent sound among a list of N possible alternative. The success of the imitation is then defined as the identification accuracy, i.e. the number of hits.

Two remarks can be made. First, the hit rate for the imitations alone has little meaning, since it is completely dependent on the task (the set of alternatives). This measure is only meaningful if compared to some reference. [LR14] used verbal descriptions of the sounds as such a comparison. Second, the task actually amounts in comparing a sound (the imitation) to other sounds (the referent sounds). Therefore, the minimal interpretation of the results is that imitations *sound similar* to the reference sounds. These results actually do not tell us whether listeners can recover the source of a referent sounds by listening to its imitation, which one of the main hypotheses of the SkAT-VG project.

We are therefore working on a new study with the following specifications:

- Yes/no tasks. For each imitation, the participants see a potential label, and simply indicates whether they think the label correctly describes the sound.
- We use a number of comparison stimuli. The best possible stimuli are the reference sounds themselves. Then we also use degraded versions of the reference sounds as comparisons. We have compared classical degradations known to impinge on sound identification: stop-band filtering, temporal envelope applied to a white noise, spectral envelope made stationary, phase scrambling, etc. We have finally chosen to use "auditory sketches": resynthesis of the referent sounds based on sparsified representations of the referent sounds [SDPD13]. The nice thing about this is that it is scalable and we can control how different features are degraded (what can the voice do? [LJH$^+$15]).

The results of the study will therefore be, for a selection of imitations, the number of correct identifications of the corresponding category labels. In fact, the classification scores will be analyzed in the context of the Signal Detection Theory to account for the potential biases of the participants toward certain responses. This number will be compared to the different comparisons used in the experiment: the reference sound itself, the filtered version, etc. The most interesting result will be the comparison with the automatic imitator: are human speakers smarter than just systematically following basic acoustic properties of the reference sounds?

Do they adapt to each sound by picking up the most relevant features and reproduce them within the constraints of their vocal apparatus?

Discussion between Ircam and composer Georges Aperghis (Golden Lion for Lifetime Achievement, La Biennale di Venezia, 2015) will be conducted about the idea of automatic imitation, for a musical piece that will be created in 2017.

Meetings and Events

Quarterly meeting through video conferencing: January 28, 2015 (#4 together with project meeting in Venice); March 31, 2015 (#5); July 21, 2015 (#6);

Bilateral IRCAM - KTH meeting, for the concurrent development of WP2, WP3, and WP4: July 23, 2015, in Paris.

Deviations from Annex I:
Instead of conducting identification experiments on the *whole* database of imitations, identification experiments are conducted on a *sample* of the database of imitations and statistical analyses of the whole database.

All objectives achieved according the Schedule

No Corrective Actions Required

Task 4.3: Analysis of the successful imitations

DoW: Integration of the results of WP3 and WP4 aim at analyzing what makes an imitation successful. Identifying which sound features cannot be rendered by the human voice and gesture will inform Task 3.3 and Task 5.3 about the sounds that require linguistic aliases. The vocal and gesture primitives identified respectively by KTH in WP3 and IRCAM in Task 4.2 will be used to analyze the imitations, with a focus on the temporal combination. Multimodal analysis will inform on the pertinent gesture characteristics that can complement vocal imitation. The outcome of Task 4.3 will therefore inform two other tasks: 1. It will help WP5 refining its classifiers by providing it with correspondence rules between imitations and imitated sounds. 2. It will inform IUAVin Task 6.2 about the aspects that are important for the fine tuning of the timbral families.

Progress

Two studies have been completed within Task 4.3: the study of how expert and lay speakers imitate basic auditory features, and the study of the respective roles of vocalizations and gestures during imitations. A new study about the influence of the speakers' native language is also in progress.

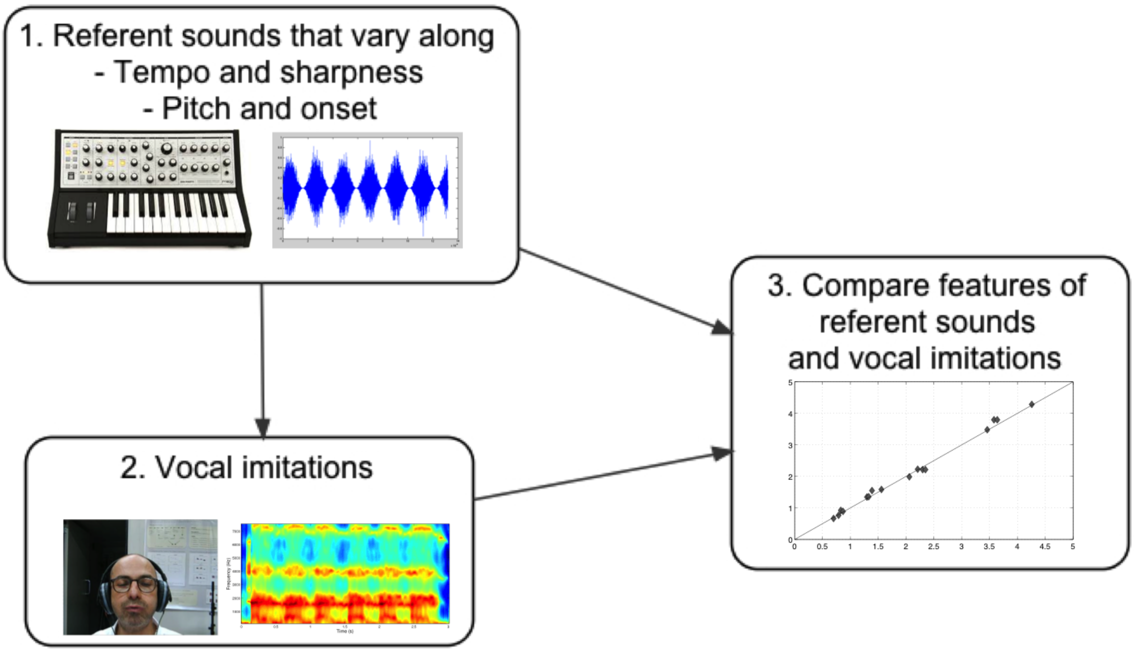

1. Referent sounds that vary along
- Tempo and sharpness
- Pitch and onset
2. Vocal imitations
3. Compare features of referent sounds and vocal imitations

Figure 4: Principle of pilot studies conducted in Task 4.3.

Vocal imitations of basic auditory features IRCAM conducted a series of pilot studies already in Year 1, to prepare the analysis of vocal and gestural imitations. These studies analyzed how two experts (professional singers specialized in extended vocal techniques) and two lay participants imitated different sets of synthetic referent sounds varying along elementary auditory features: tempo and pitch (i.e. musical features), and sharpness and onset (i.e. timbral features, see Figure 4). A feature comparison of referent sounds and vocal imitations is revealing that speakers were more precise to imitate musical than timbral features. For the timbral features, the analyses are showing that participants relied on different strategies. Recording sessions were also videotaped. Even though the speakers were not instructed to produce any gesture, analysis of the videos shows that they did produced gestures: they used their hands to highlight and reinforce some aspects of the imitations, and used movements of the torso and the head to indicate the beginning of an imitation. This suggests that analysis of gestural imitations in Task 4.3 will have to distinguish between voluntary and ancillary gestures. This study has been submitted to the Journal of the Acoustical Society of America and is currently under revision.

The role of voice and gestures during imitations During Year 2, IRCAM studied the role of gestures during imitations of sounds. We first qualitatively studied the database of imitations and drew three hypotheses:

- Vocal imitations are more accurate than gestural imitations of rhythmical sequences.

- Imitators use very often “shaky gestures” to express that the reference sound has a noisy texture
- When imitators imitate sounds made of different layers, they use different strategies, one of which consists of conveying one layer with the voice and one with the gestures.

We tested these hypotheses in a subsequent experimental study. We designed a specific set of referent sounds and recorded vocal and gestural imitations of these sounds. Then we designed a set of gestural measurements (with WP5) based on the wavelet representation of acceleration data. These gestural features were submitted to statistical analyses that confirmed that the data were in good agreement with the hypotheses. Finally, we used these new features to train a classifier that recognizes if a gesture imitates a noisy or stable sound. IRCAM is currently drafting a manuscript for submission to PLOS ONE or Frontiers in Psychology

Influence of the speaker’s native language This work addresses the question: “Does speakers’ native language constrain their non-linguistic imitative vocalizations?” This question can be refined as: “can we observe articulatory mechanisms that are specific to a given language in the non-linguistic vocalizations of speakers of a language in which these articulatory mechanisms are usually not present? Are speakers not any longer constrained by their native language as soon as their vocal utterances are not linguistic? How do the native language’s constraints compare to individual differences of ability ?”.

Based on discussions with KTH at Ircam in June 2015, the following plan was established:

- The starting point will be a table that lists the different tokens of the International Phonetic Alphabet (with a focus on consonants) and checks their existence in French, Swedish, English, Italian and Mandarin Chinese (PH);
- Next to that will be tally of how often we observed these different tokens in the vocal imitations recorded in Paris and Stockholm;
- Based on this, ad hoc hypotheses will be formulated;
- If necessary, recordings will be redone in Paris;
- These recordings will be annotated at KTH.

This work is potentially publishable, either in a linguistic journal or in an acoustics journal.

Meetings and Events

There were no meetings specific to Task 4.3 except the quarterly meetings.

No Deviations from Annex I

All objectives achieved according the Schedule

No Corrective Actions Required

2.3.5 Work Package 5: Automatic imitation recognition

WP5 deals with the automatic estimation of the categories of vocal imitations from the analysis of the audio signal. In this part we summarize the progress achieved in WP5 from its start in December 2014 until September 2015. Task 5.1 (representation and segmentation) and 5.2 (blind classifier of imitation) are under finalization. Task 5.3 (informed classifier) and 5.4 (fusion classifier) are about to start.

Task 5.1: Representation and Segmentation

DoW: The front-end application developed in WP5 will extract two types of meaningful representations of the imitation signals: Sound representations: low-level features (based on spectral moments, modulation spectrum, etc.) and perceptual features (loudness, pitch, sharpness, roughness, etc.); Gesture representations: gesture primitives obtained from WP4. These representations will be used to predict the high-level representations (vocal primitives) of WP3. This learning will be performed using the examples and analyses provided by WP3 and WP4. The mapping from sound to vocal primitives will be done by KTH. Imitations also involve the complex combination of vocal and gestural primitives. The front-end application of WP5 will segment the imitations into sequences of meaningful multimodal elements.

Progress

Five lines of work have been followed at Ircam along the Period.

Study on classification by DTW alignment and relevance feedback. Initial research work concerned the modeling of audio feature sequences over time using Dynamic Time Warping. A prototype has been developed to test this approach, which integrates *user relevance feedback*: alignment costs weighting is adapted by iterative user choices. In this first phase, automatic classification of imitations by usual audio features has proved to be not effective. Activities have thus focused on finding audio signals representations suitable for machine learning, as planned in Task 5.2-4.

Development of new audio features for the morphological categories. Three sets of descriptors have been developed for the imitations of the family of Abstract referent sounds: Local Trend (LTd), Global Trend (GTd) and Morphological (Md). All the three exploit the same low-level features: spectral centroid, spectral spread, spectral rolloff and pitch; moreover, the lowest formant and the frequency of short-time spectrum peak have been added. LTd are built by computing the local derivative of the underlying feature, thus expressing the short-term variations. GTd, instead, point out the long-term cues of the features and express them by a finite alphabet. LTd and GTd rely on Hidden Markov Models (HMMs) for their explicit temporal modeling. Md do not need this kind of modeling: they express the presence of the relevant cues typical of the Abstract family. An article, resuming application of LTd, GTd and Md, will be presented at the DAFx-2015 conference in early December [MP15]. Details about automatic learning are given in Task 5.2, while a deeper description of the whole system is provided in D5.5.1.

Development of audio features for dataset analysis. Task 5.1 has then advanced by providing a newly developed set of Statistics descriptors (Sd), which have been exploited in Task 4.2 (with results in D4.4.1). Sd are inspired by the Md, but tailored to the whole dataset. Among the others, a feature of *modulation* has been developed, which can be applied to both the temporal and spectral representations of an underlying signal. Sd have also been used for classification: details are given in Task 5.2 and in D5.5.1.

Development of a basis. The latest advances in Task 5.1 are toward the design of high-level descriptors for the imitations. To have a better understanding of the variety of signals in the dataset, a subset of it has been listened: this has given advices about how to design relevant descriptors. The analyses are being summarized into the form of a dictionary of signal bases, such that all the imitations of the Dataset could, in principle, be described as a combination (even with superposition) of one or more basis.

Future work: automatic learning of sound basis. Right after the manual compilation of signal bases, advanced machine learning techniques are planned to be applied to the same problem. The comparison between manual and automatic found bases would be scientifically relevant, possibly giving insights about more compelling descriptors for Task 5.2-4.

Based on the multimodal database built in WP4, a novel movement representation based on Continuous Wavelet Analysis was developed at IRCAM, and implemented under the MuBu framework. This representation was justified from a qualitative analysis of the database that clearly shows the importance of taking into account the frequency content of the gestures. In particular, we found that a time-frequency representation was necessary. The Short-time Fourier transform (STFT) was compared to the Continuous Wavelet Analysis (using Morlet basis). For the frequency range needed, there was a clear advantage to use the Continuous Wavelet Analysis to define various frequency components of the movements. Such components were defined as "gesture primitives" that can be superimposed in time and scales (corresponding to log frequencies). We also developed a method to track the temporal evolution of these primitives using particle filtering. In particular we implemented a class of methods that allows for tracking several primitives simultaneously called "track-before-detect". The gestural part of the database from WP4 is currently being analysed based on this methodology.

KTH investigated the possibility for recognition of imitated sounds to be based on Audio analysis using Receptive Fields (ARF) [LF15a][LF15b]. For SkAT-VG, a first level of analysis will facilitate both extraction of voice parameters (pitch and formant trajectories, noise resonances, etc.) as well as features that can be used for classification of voice articulation types. The following analysis methods are being developed:

- Basic time-frequency representation (spectrogram), invariant under transposition in log frequency;
- Enhancement of partials in the time-frequency plane, improving the analysis of time-varying pitch;

- Enhancement of formants;
- Enhancement of harmonic fundamental;
- Onset/offset enhancement.

IUAV has been working at the extraction of principal components from a database of 152 short excerpts of vocal imitations by Fred Newman [New04], and at the organization of a bi-dimensional sonic space by clustering [RM14, MR15].

Task 5.2: Blind classifiers

> **DoW:** Classifiers that predict the categories of imitated sound sources (from WP4) directly from low-level multimodal (sound and gesture) representations.

Progress

Before the actual classification, some studies regarding the dataset contents have been conducted.

Early studies. The dataset built by WP4 is based on three families of referent sounds: Abstract, Machines and Interactions, each with its own set of reference stimuli. This, initially, originated the idea of training a single automatic classifier, able to recognize all the 26 categories in the dataset. As expected, early results found that the three families are not orthogonal: a given imitation can, in principle, belong to three categories, one from each family. The current approach is thus of building three distinct classifiers, one per family.

Classification into Abstract family. As anticipated in Task 5.1, the LTd and the GTd descriptors are modeled by Hidden Markov Models (HMMs), respectively continuous and discrete. In both cases the hidden states of the models are defined according to four basic signal *shapes*, which have been selected as a resume of the Abstract family contents: *silence*, *up*, *down* and *stable*. Recognition performances have been improved by moving to more specific signal descriptors, such as the Md set. The modeling of temporal evolution of the signals is, in this case, embedded into the descriptors; HMMs have thus been replaced by k-Nearest Neighbor (kNN) algorithm. Md descriptors provide, up to the time of writing, the best classification performances. Exploitation of automatic learning tools for Md is currently being done.

Classification into Interaction and Machine families. Blind classification on these families is currently under development. Specific descriptors are being developed, using an approach similar to Md set; main cues of the imitations are being highlighted, and used as basis for descriptors design. As temporal evolution of imitations will probably be embedded in descriptors, there is no need for temporal modeling by HMMs; we are thus moving toward state-of-the-art classifiers such as Support Vector Machines.

About the gesture analysis, the database from WP4 is currently being analyzed using the gesture primitives defined in T.5.1. As expected, the recognition of sound categories using gesture features only is very challenging. The most promising approach is to use clustering, which allows for defining different movement strategies.

A simple classification tree was developed at IUAV for demonstrating model selection by imitation in miMic (see 2.3.7). This is going to be replaced by the more accurate and complete classifier being developed in this Workpackage. Also in the miMic demonstrator, individual-centered selection has been demonstrated by using the classification by examples provided by IRCAM as part of its MuBu objects (see table 10).

Task 5.3: Informed classifiers

DoW: Classifiers that predict the categories of imitated sound sources (from WP4) from the high-level phonetic representations (from WP3) (instead of the low-level multimodal representations). Since high-level representations are used to model the categories, the statistical models will remain tractable by a human.

Progress
Work on this task has not started yet.

Task 5.4: Fusion

DoW: Fusion between the blind and the informed classifiers. The performances of the two approaches (blind and informed) will be compared. From the results, fusion techniques will be developed to take advantage from both approaches (robustness versus precision). For each task, modules will be adapted in order to deal with the linguistic aliases (Task 3.3).

Progress
Work on this task has not started yet.

Meetings and Events
WP5 has been actively discussed in all the official project meetings (Paris, Stockholm, Venice, Aix-en-Provence) and internal meetings. WP5 has also been a central topic in the brain-storming meeting of Venice (04/2015). A two-day meeting in Paris (06/2015) was organized as part of WP5. The goals were:

- to further establish the specifications of the annotated audio data (raw-audio-data and type of annotations) to be submitted by the other WPs;

- to establish the specifications of the inputs and outputs of the modules to be developed;
- to check that the outputs of WP5 correspond to the expected inputs of the Use Cases;
- to check that the various tasks of WP5 are complementary and can be achieved within the project duration.

No Deviations from Annex I

All Objectives achieved according to Schedule

No Corrective Actions Required

2.3.6 Work Package 6: Imitation-driven sound synthesis

Objectives of WP6 are to provide salient timbral families of sound synthesis tools, which cover the ontology of perceptually relevant sound categories, as being defined in WP4 (T6.1), to calibrate the synthesis tools for effective vocal control and temporal behaviour, including sound post processing operations (T6.2), and to develop software architectures and interfaces for sound design (T6.3) which integrate the classifier, developed in WP5, the sound synthesis and post processing sections, based on the feedbacks collected through continuous exchanges with WP7 experimental activities. From the deliverables viewpoint, the achievement of these objectives is distilled in D.6.6.1 "Automatic system for the generation of sound sketches". The deliverable D.6.6.1 (P), being anticipated at Mo 22, includes the first implementation and release of the software resulting from the accomplishment of the tasks. Provided that T6.3 is planned to officially start at Mo 25, it is clear that the nature of deliverable D.6.6.1 reflects continual progresses, with iterative development and refinement of the software. Deliverable D.6.6.2 (R) "Front-end application for interactive prototyping", planned for Mo 30, will cover the systematized description of the emerging SkAT-VG tools, and control strategies for sound design applications, as being devised in WP7 activity. WP6 activities were planned to start at month 13, but they were partially anticipated to the first year.

In practice, the progress of the tasks in the second year, towards the achievement of the automatic system for the generation of sound sketches concerned:

- **T6.1**: i) the expansion and optimization of the physics-based sound synthesis algorithms derived from the Sound Design Toolkit, to complete the low-level taxonomy and provide the "vocabulary" needed for the definition of the timbral families; ii) the selection and organization of sets of sound models, and their relevant parameter space, in timbral families. This activity completes the task.
- **T6.2**: i) the development of audio descriptors as MaxMSP externals, tailored to the vocal control of the timbral families, ii) the development of post processing modules, such as reverb and pitch shift, functional to the implementation of specific timbral families, iii) the development of SkAT Studio architecture and modules, the proposed architecture was originally sketched by Genesis in WP7, starting from the analysis of

the interviews with professional sound designers, and its refinement is functional to the accomplishment of T6.3.

Task 6.1: Definition of the timbral families

> **DoW:** Physics-based sound synthesis makes an extensive work of parameterization necessary, to define the different timbral families of each model. The specification of appropriate timbral families of the SkAT-VG system will be the main activity of this task. In particular, they will have to match the categories sorted out by WP3 and WP4, and will define the subclasses onto which the outputs of WP5 will be mapped.

Progress

The conceptual framework behind the structure of the Sound Design Toolkit (SDT), based on an ecological approach to sound synthesis, makes an excellent starting point for the activities of WP6. Sound categories defined in WP4 can be grouped into three main families:

1 Abstract sounds,

2 Machine sounds,

3 Basic mechanical interactions.

For the latter family, many categories have an almost direct relationship with the low-level sound synthesis models, as proposed in the SDT taxonomy. Machine sounds can be further divided in other three groups:

a Electric appliances,

b Combustion engines,

c Other complex mechanical interactions.

The latter group can be rendered with higher level sound textures and events, composed by a combination of lower level basic mechanical interactions, many of them already present in the SDT taxonomy. Electric appliances and vehicle sounds are so complex to require *ad hoc* synthesis models.

During year 1 a complete rewrite of the Sound Design Toolkit was started, and the work continued in year 2. The software architecture is now composed of a core framework, plus a set of externals and user interface patches for Max and PureData. The core framework is developed in ANSI C and designed to be portable across different operating systems (Windows, Mac OS X, Linux). It exposes a C API which gives access to all the implemented synthesis, analysis and processing algorithms, allowing their use in a wide variety of applications. The maintained externals are nothing but wrappers for the algorithms in the core framework, and are organized in patches with an intuitive user interface.

The following physically informed sound synthesis algorithms were added in year 1:

- Aerodynamic noise against solid surfaces,
- Aerodynamic resonance in hollow cavities,
- Karman vortexes across thin objects,
- Liquid sounds: dripping, filling, gushing etc.,
- Combustion engine sounds.

In year 2 these models have been refined, and the following ones were implemented from scratch:

- Electric motors,
- Explosions.

Moreover, the following models already belonging to previous versions of the Sound Design Toolkit were radically redesigned and ported inside the new software architecture:

- Inertial mass,
- Modal resonator,
- Impact interactor,
- Friction interactor,
- Crumpling and breaking,
- Rolling,
- Scraping.

The current state of development of the Sound Design Toolkit allows the definition of the corresponding timbral families for all of the 26 sound categories defined in WP4, as displayed in table 4. An extensive work of parametrization of the SDT models is still in progress and will produce a set of patchers, one for each timbral family, containing the appropriate synthesis algorithms together with the parameter ranges and combinations needed to reproduce the corresponding sound categories. By the end of year 2, in compliance with the project DoW, this set of patchers will be available in the GitHub repository[1] containing the official distribution of the Sound Design Toolkit.

Meetings and Events

- WP6 first quarterly meeting (via Skype), 18th March 2015

[1] https://github.com/SkAT-VG/

Abstract sounds	
Up	Conventional synthesis techniques (additive, subtractive, AM, FM...)
Down	
Up/Down	
Impulse	
Repetition	
Stable	
Machines	
Alarms	Conventional synthesis
Buttons and switches	Impact
Doors closing	Impact
Filing and sawing	Scraping
Fridge hums	Electric motor
Mixers and blenders	Electric motor
Printers fax and xerox	Electric motor + rolling + impact
Windshield wipers	Electric motor + friction
Vehicles exterior revs up	Combustion engine
Vehicles interior accelerating	Combustion engine
Mechanical interactions	
Blowing	Aerodynamic noise
Whipping	Aerodynamic noise + explosion
Shooting	Explosion
Crumpling	Crumpling
Rolling	Rolling
Rubbing and scraping	Friction + scraping
Hitting and tapping	Impact
Dripping and trickling	Fluid flow
Filling	Fluid flow
Gushing	Fluid flow

Table 4: The 26 sound categories that are emerging from experiments are listed on the left column and coupled with the SDT sound models that are being used to represent them. The sounds are grouped in three large classes.

- Genesis - IUAV meeting in Aix-en-Provence, followed by the joint participation to SIVE/IEEEVR for the presentation of the joint article, 23rd – 25th March 2015
- WP6 second quarterly meeting (via Skype), 16th July 2015

No Deviations from Annex I

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 6.2: Temporal behaviour and sound post processing

> **DoW:** Once the roughly parameterized synthesis algorithm will be selected, it will be necessary to establish (i) how it behaves in time, (ii) how and if the output sound must be post-processed and, (iii) if a sequence of subclasses must be used, how to pass from one timbre to the other. This task is essentially about calibration and control of time-varying parameterizations.

Progress
T6.2 is essentially focusing on calibration, vocal control and temporal behaviour of sound synthesis algorithms. In the second year, the task is progressing along three main tracks:

1. The extension of the Sound Design Toolkit with sound analysis tools for the extraction of timbral audio descriptors from vocal imitations. The implemented algorithms include a fundamental frequency estimator, an extractor of statistical spectral moments (centroid, spread, skewness, kurtosis), a signal envelope follower, an onset detector and other low level indicators such as signal zero crossing rate, spectral flatness and spectral flux. The design, use and further processing of these descriptors is specifically oriented towards the extraction of meaningful information from vocal signals, useful to dynamically control the available synthesis algorithms through vocalization.

2. The development of post processing modules functional to the implementation of specific timbral families. A maximally-diffusive reverberation algorithm has been implemented and included in the Sound Design Toolkit. It is an important building block for thickening and propagating micro-textures in physics-based sound synthesis, which will be presented at the DAFx-2015 conference in early december [RBD15]. A time domain pitch shifter has also been implemented, using a well-known algorithm [Zoe02]. Although not yet used in any SDT synthesis model, it could be used in the future as a post-processing tool, to simulate Doppler effect for moving sound sources.

3. The current state of SkAT Studio is framed in the development of a stable environment suitable for the control of the temporal behavior of sound synthesis. SkAT Studio was proposed as general framework to allow each partner to include quickly its developments as loadable modules. The major improvements concerned the stability of the whole architecture, and the addition of a few functionalities to meet the design requirements, emerged in the early experimentations of the tool. Essentially, the early sketch proposed in the first reporting period is evolving to set the ground for the accomplishment of T6.3, regarding the definition of user interfaces for the intuitive exploration of voice-driven sonic spaces and sound design. A true Model/View/Controller architecture has been set up for the coding of SkAT Studio. This had some consequences on the saving process, particularly on the format of a backup. Keeping the use of SkAT Studio unchanged, it improves the stability of the software. Two new functionalities have been proposed:
   - a "recorder" module, which allows to record the output of a synthesis;
   - an "undo/redo" functionality, useful in the building of a framework.

No Deviations from Annex I

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 6.3: Fine parameterization of timbral subclasses

DoW: From sketch to prototype. It is necessary to give the designer the possibility of processing and refining the sounds automatically generated on the basis of the results of Tasks 6.1 and 6.2. This task involves the definition and development of an interface that a designer can intuitively use for the exploration of a neighborhood of the default sound (the sketch). Such a sound will represent a sort of reference point (landmark) that the designer will be able to repeatedly play with, towards the definition of a prototype. In particular, the interface will allow the designer to vary a set of perceptually meaningful physics-based parameters and a set of controllers of the temporal behaviour of the sound. This task will be performed in cooperation between IUAV and GENESIS.

Progress
Work on this task has not started yet.

2.3.7 Work Package 7: Sonic Interaction Design

During the first reporting period, WP7 design activities were mainly aimed at exploring the underlying assumptions for the communication of sonic concepts, as being investigated by

WP2, WP3, and WP4. The aim was to establish a shared, methodological connection between scientific and design research. In the second year, the analysis of the workshop experiences collected led to the systematization and distillation of the scientific contributions of the other active WPs into well-defined introductory design exercises. Therefore, the research-through-design activities of WP7 made progress on:

- understanding how to exploit vocal sketches for design communication,
- the analysis of the cognitive and perceptual aspects of the early stage of the sound design process,
- the development of a methodology of sketch-thinking practices for sound design.

A first demonstrator – miMic –, and a proof-of-concept in the form of installation – S'i' fosse suono – were developed, whereas the SkAT Studio framework underwent several improvements to embody the desired features, progressively emerging from scientific and design researches. A second demonstrator – MIMES – was specifically developed to highlight the role of gesture in sound definition and manipulation. Tighter collaborations with professional sound designers were established, whereas the workshops planning has been focusing on concrete sound design cases.

Task 7.1: Definition of sonic interactive scenarios and applications

DoW: The aim will be to coordinate all the basic research efforts, with the definition and realization of concrete experiences that will serve as test benches of the SkAT-VG tool, as described in task 7.2. The task will consist of user studies and will be benefit form GENESIS experience in industrial sound. It will be active along most of the project duration. It will involve an intensive exchange with almost all of the other WPs and partners, and in particular WP2, WP3 and WP4 to distill relevant cases. The monitoring role of this task will be important to focus research toward applicable results, to be assessed in a number of design experiences.

Progress

Task 7.1 represents the SkAT-VG sketch book and sets the playground for the design exploration of the other WPs work. In other words, scientific contributions are nurtured from the design perspective, and mature sound design hypotheses are selected and investigated through workshops and mock-ups in Task 7.2.

In the first reporting period, GENESIS conducted several interviews with sound designers from France, Italy, UK and Japan, in order to get perspective on how vocal imitations may have a role or can impact the workflow of sound designers. The text of the interviews was provided to the Consortium. The analytical summary will set the basis for the preparation of a publication in collaboration with IUAV . GENESIS hired Xavier Collet to make a summary of all interviews, which is due by mid-october.

Genesis organized two new meetings in Aix-en-Provence with sound designers:

- Norio Kubo spent one day in a work session about sound design. First he presented the tools he uses for sound design (LEA, Adobe Audition, ...). Then, SkAT Studio was presented, in its current state, with the existing scenarios. Norio Kubo showed much interest on the use of SkAT Studio, especially for the use of gesture for specifying a pattern to apply to a texture.

- A full interview was done with Xavier Collet, a French sound designer who works on video games and movies sound tracks. As usual, the interview was fully video recorded. This designer already uses his voice as a control parameter for the production of sounds: for example, he uses the level of the voice to control the cut-off frequency of a filter. He was therefore very interested by SkAT in general, and the SkAT Studio tool in particular.

IUAV has been refining the design of the sound models, the timbral families being developed in WP6, through demonstrations and prototypes. The expansion of the Sound Design Toolkit [DPR10] was exploited for the live soundtrack of the theatrical play "Bodas de Sangre", within the collaboration with the PhD School of Computer Science of Ca' Foscari University of Venice. The sound models were further refined, in terms of exploration of the sound design space, in the design of the demo "miMic", and of the installation "S'i' fosse suono".

IUAV developed a first demonstrator, "miMic", an augmented microphone for vocal and gestural sketching. Vocalizations are classified and interpreted as instances of sound models, which the user can play with by vocal and gestural control. The physical device is based on a modified microphone, with embedded inertial sensors and buttons. Sound models can be selected by vocal imitations that are automatically classified, and each model is mapped to vocal and gestural features for real-time control. miMic will be exploited as experimental tool in forthcoming workshop activities. A paper on miMic has been accepted for presentation at the ACM Conference on Tangible, Embedded, and Embodied Interaction (TEI), Eindhoven, February 2016.

IRCAM has been developing "MIMES", a family of interactive objects that allow for interactively designing expressive sounds, by means of gesture and voice.

IUAV developed the installation "S'i' fosse suono", in collaboration with the sound designer Andrea Cera. The installation consists of a multi-touch video mosaic of interactive vocal and synthesis-converted self-portraits (see Figure 5). It represents a proof-of-concept of the SkAT-VG system, in which the articulation of vocal utterances is interpreted and converted in coherent synthetic sounds. A HTML 5 implementation is almost finished (`http://www.lim.di.unimi.it/subsites/sifosse/`) thanks to a collaboration with Luca A. Ludovico of the University of Milan. The proposal for an article on "S'i' fosse suono" for Volume 26 ("lend me your ears! - sound and reception") of Leonardo Music Journal was sent to the editor in chief, who responded very favourably.

IUAV organized a two days workshop on vocal sketching at the Department of Theatre, Film and Television of the University of York, England. The workshop was aimed at refining and structuring the introductory exercises devised in previous editions, while starting to set the

Figure 5: S'i' fosse suono.

ground for evaluating the effectiveness of vocal sketching compared to other forms of sound creation in the conceptual stage of the design process.

Meetings and Events

GENESIS organized meetings with

- Norio Kubo, Aix-en-Provence, April. Norio Kubo's sound design process and methods.
- Xavier Collet, Aix-en-Provence, September 24th: (French) sound designer for multimedia (art, video games, web sites, movies). Full interview of the designer, presentation of SkAT-VG and of the developed tools.

  WP7 meeting IUAV - GENESIS - IRCAM Marseille, February 13th, 2015;

  WP7 quarterly meeting (via Skype), April 15th, 2015;

  WP7 Brainstorming cross-partners meeting IUAV - GENESIS - IRCAM Venice, April 21st-22nd, 2015;

  WP7 quarterly meeting (via Skype), July 17th, 2015;

  Electroacoustic direction and live performance with the Sound Design Toolkit, for the theatrical play "Bodas de Sangre", in collaboration with the acting company "Cantiere Ca' Foscari", and the PhD School of Computer Science of the Ca' Foscari University of Venice, Teatro S. Marta, Venice, May 26-27th, July 7th, 2015.

Deviations from Annex I
No deviations from Annex I for the reporting period. The PMs used by IUAV in the first reporting period, to anticipate the work on the Sound Design Toolkit and on sound synthesis models, which may well be ascribed to WP6, have been rebalanced in WP6 activities.

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 7.2: Realization of sonic interactive experiences, using the SkAT-VG tool

> **DoW:** Implementation by IUAV of sonically augmented mock-ups, selected from the applications emerging from Task 7.1. The definition and realization of real design sessions will aim at providing user tests of the capabilities of the SkAT-VG tool. In a series of workshops, designers will be invited to solve some design problems under a set of constraints. The analysis of the workshops, both in terms of reports and prototypes, will provide the assessment bench of the results of the project. The most compelling design processes will be documented with short movies, thus giving immediate evidence of the effectiveness of the SkAT-VG system. In order to face the risks discussed in Section v), the definition and the integration of auxiliary tools, based on other criteria than imitation (i.e. the linguistic aliases developed in Task 3.3), could be considered in the context of the design workshops.

Progress
In Task 7.2, a bottom-up process is centered on devising and supporting the emergence of relevant sketch-thinking behaviours in sound design, and provide guidance for the design and use of SkAT-VG tools. Work progress in Task 7.2 focused on refining the outcomes gained from a first set of workshops, aimed at probing the effectiveness of vocal imitations in acting out sound design ideas. The collected information have been crystallized in well-defined introductory exercises on the representation of concepts in the aural domain, through vocal sketches. A workshop format on product sound sketching has been devised and is being tested as framework to investigate more systematically the sound designer's behavior, and to unfold the process of sound sketch-thinking. In parallel, the contributions from the other WPs, and the observations emerged through workshops and internal design sessions led to the design and development of miMic, the augmented microphone for vocal and gestural sketching. In this respect, goal of this demonstrator is to investigate a system architecture that would seamlessly support the ubiquity of sketching, from analog representations to synthetic displays. The development of this tool is instrumental to advance research-through-design activities, especially regarding the assessment of the fluidity and effectiveness of individual and collaborative sketching workflows.

The audiovisual installation "S'i' fosse suono", developed in collaboration with sound designer Andrea Cera, is aimed at communicating the essence of the SkAT-VG project, and explores the effectiveness of voice-driven synthetic sketches in a complementary way. In a sense, the sound designer embodied the hypothetical SkAT-VG system, by interpreting creative vocal utterances provided by laypersons, selecting the most appropriate sound models (i.e., timbral families), and providing a coherent synthetic version of the vocal sketch.

Growing the practice of vocal sketching: The workshops are structured in RtD activities aimed at a three-folded objective, that is i) ground the design exercises and their evaluation in phonetics (i.e., elicitation and articulation) and auditory perception of vocal imitations; ii) study the sound designer's behavior; iii) collect and refine exercises in structured sketching practices for sound design and education. In 2015, IUAV held a two days workshop on sonic sketching, organized with Dr. Sandra Pauletto at the Department of Theatre, Film and Television (TFTV) of the University of York. TFTV, Music and Computer science postgraduate students participated in the workshop that explored strategies and practices for the creation of new sounds for objects, scenarios and data sets. A two days workshop format runs as follows: the first day is devoted to the development of analytical skills, and to the practice of using voice to communicate sonic interactions. In the second day, teams are introduced to the basics of product sound quality and sound synthesis (Sound Design Toolkit), and engage in the exploration of given design scenarios.

Day 1 - Morning - On action-sound relationship:

- Ear-cleaning exercises introduce the discussion about everyday listening and ecological perception;

- Foley-oriented "physical" sound synthesis exercises foster the discovery of sound affordances in the manipulation of physical objects to produce given target sounds;

- Sound-action analysis and exercise explore the relationship between function, objective, action and sound. Sonic interactions are decomposed in tasks (objectives) and action primitives, in order to exploit the rich repertoire of everyday actions, for design purposes.

Day 1 - Afternoon - On the use of voice to communicate sound:

- Vocalization techniques for the production of basic sound effects [New04] have been systematized according to the ontology of perceptually relevant sound categories, investigated in WP4, and within the framework of source types and voice production attributes, proposed in WP3. The improvisational and playful playground is kept in foreground, yet the collective practice stimulates analytical skills and fosters the participants' iterative exploration and control of the initiation mechanisms;

- The (vocal) imitation game is a collaborative guessing game based on a card deck of verbal description of mechanical interactions and machines, the goal is to guess the sonic concept mimicked only with the voice. The design research is focused on collecting evidence of emerging strategies in the elicitation and communication of vocal imitations, outside the laboratory. The competitive playground pushes participants to learn and improve quickly to mimic sonic concepts, and exploit iconic, metaphoric or ancillary gestures to support and fine-tune the control on the voice articulation, especially when the first imitation attempts do not work;

- The exercise on acousmatic narratives requires to design the sonification of complex processes of fictional machines, according to given sound design themes (e.g., magic, horror, futuristic, etc.), and produce polyphonic, collaborative sketches that are presented in form of live performances.

Day 2 - Morning: On product sound:

- Introduction to product sound quality and exercise, the relationship between object, function, action, and sound is analyzed according to a comparative approach, in order to emphasize the concept of sound quality;

- The automotive scenario, selected in the SkAT-VG project, is explored in the second exercise wherein the teamwork focuses on the production of vocal sketches of motor sounds for various vehicles with a futuristic streamlining. A video editor and a microphone are exploited to produce video-prototypes with sonic overlays (e.g., `https://vimeo.com/128886746`);

- Introduction to the Sound Design Toolkit, the workshop is the natural venue wherein demonstrators and mock-ups of SkAT-VG tools are introduced and tested in sound design assignments. Timbral families and control modules through voice and gestures are made progressively available in the form of MaxMSP patches, and are refined iteratively towards functional prototypes.

Day 2- Afternoon - Design session:

- Anatomy of sketching, starting from the studies in the visual domain and interaction design, the exercises done in the first day and a half are summarized and re-assessed in terms of design-thinking process. Sketching is described and analyzed from the cognitive and perceptual viewpoints, in order to stress the inherent characteristics of this peculiar design-thinking process by means of representations;

- A proper design assignment, based on a given brief, is used to assess the overall effectiveness of the workshop activities, and observe the sketching process in a real design context. In the last workshop, teams were asked to create a system to make access to health data intuitive, easy and informative, to extend the device for social sharing information, and to demonstrate its functions and capabilities through videoprototyping.

The workshop in York was the occasion to film the whole design process of the teams. Due to some logistic problems it was not possible to have a sufficient number of participants to assess the effectiveness of vocal sketching compared to other forms of sound creations, in the conceptual stage of the sound design process. However, the next workshop will have a large number of participants. The research objective is to observe and compare a) the time required to produce sketches; b) the number of sketches produced; c) the duration of the verbal thinking phase; d) the degree of refinement and quality of outcomes, under three constrained working conditions, that is

1. one team is allowed to use only vocalisations in sketching and video prototyping;

2. one team is allowed to use only vocalisation-driven sound synthesis in sketching and video prototyping;

3. one team is allowed to use any technique, approach and tool (Foley, sound recording and processing, etc.), except vocalisations and voice-driven sound synthesis.

We asked the participants to fill a questionnaire to evaluate the workshop and vocal sketching. As general observations, participants were eager of discovering their natural abilities in communicating through vocal imitations, and virtuoso iterations became soon evident in the warming up exercises of the first day. On the other hand, we observed a strong reticence to vocal sketching in the design assignments. Enacting the use of the voice to develop and explore design ideas from scratch is hard, being the decision-making process mostly based on verbal discussion, whereas non-verbal vocalizations and utterances come into play once a sonic idea is selected. Mastering the human voice is difficult, and this severely affects the design outcome, in terms of aesthetic quality and number of iterations. Especially, it is difficult to detach from representational strategies based on naïve metaphors and stereotypes. However, once the teams managed to engage in vocal production, they produced quite refined presentations in a very short time. The expressive use of manual gestures during the actual sketching was very limited, being the designers mainly focused on the quality of the recordings in front of the microphone. Instead, they made a creative use of several tools, pipes and boxes, to augment their vocal sounds. These observations were taken in account in the development of miMic. A forthcoming one-week workshop will take place at Medialogy, Copenhagen, in late November 2015. The workshop program will focus on three days of sketching activities, and two days of prototyping and evaluation. 25 postgraduate students in Sound and Music Computing will participate. We will propose again the experimental conditions, as initially planned for the workshop in York.

miMic, the augmented microphone: miMic is a system architecture that, through the augmented microphone, can empower the user with a wide sound palette that can be directly controlled by voice and gesture (`http://buildinprogress.media.mit.edu/projects/2385/steps`). The idea of augmenting the microphone with an Inertial Measurement Unit (IMU), to capture manual gestures and two switches to select and play the timbral families emerged during workshops and internal body-storming session (see minutes of the IUAV – GENESIS – IRCAM meeting, held in Marseille, in February 2015, and minutes of the brainstorming cross-partners meeting, held in Venice in April 2015). The experience gained through workshops, brainstorming sessions, tests with partial realizations, and further discussions, led us to foresee a sound design process structured into four stages, as described in Table 5. In stage 0 of our sound design funnel, by selecting one sound model or a mixture of sound models, the designer is effectively defining a sonic concept. At stage 1, by vocally mimicking a selected sound mechanism, she gets acquainted with the available sound space, simply by vocal mirroring. Exploration is extended at stage 2, as soon as the user realizes that control is not limited to a restricted set of vocalizations. One can go beyond mirroring and let creative uses of the voice explore new neighborhoods of the sonic space that is made available by a given sound model. While stages 0 to 2 can all be performed through the miMic without any visual display, stage 3 possibly requires the manipulation of each single model parameter, made available as a GUI element (typically, a virtual slider). miMic can still

n.	Stage	Mode	Tool
0	Select	Select	miMic
1	Mimic	Play	miMic
2	Explore		miMic
3	Refine	Play + Tune	miMic + GUI

Table 5: The four stages of sound design, using miMic.

be supportive in this stage, as real-time vocal-gestural control can be applied while changing parameters one by one. Stage 3 is indeed the resolving iteration, where a sound sketch is hard-lined into a sound prototype. The system architecture, developed as MaxMSP patch, is the result of a tight interplay between WP5, WP6, and WP7. The system architecture will be used in the forthcoming workshop in Copenhagen.

MIMES: MIMES is a family of interactive objects that allow for interactively designing expressive sounds. MIMES exploit the user's movements as well as the user's voice. First, the user records a "vocalization" to propose a basic sound "morphology". The user also records a movement and/or an action on the object. The system can then make an association between the proposed sound morphology and the movement/action. The user can replay the initial sound morphology by re-performing the action/movement. Most importantly, the user can further modify the sound morphology by altering the initial movement. In other words, the user can interactively sculpt the initial vocalisation. Various sound synthesis models can also be chosen (manually or automatically), which further extend the sound palette of the final result. The MIMES as physical objects are 3D printed with a neutral material (ABS), and equipped with different wireless sensors (a triple-axis gyro, a triple-axis accelerometer and a triple-axis magnetometer). The object contains also a force-sensitive-resistor (FSR) and two piezo sensors that are connected to the main board through I2C using a Teensy 3.0 development board. These physical objects were designed during a previous project (ANR Legos) to study sensori-motor learning. These interactive objects are being further developed within SkAT-VG, adding the voice component, and within a completely different research context and goals.

S'i' fosse suono, audiovisual installation: this is the first outcome of the collaboration between IUAV and the sound designer Andrea Cera. S'i' fosse suono, shown in figure 5, is a collection of audio self-portraits, displayed as interactive tiles of a video mosaic (`http://buildinprogress.media.mit.edu/projects/2553/steps`). Each tile, representing the video recording of a self-portrait, can be explored through touch on an interactive display. The audiovisual installation is aimed at communicating the essence of the SkAT-VG project, by means of a proof-of-concept of the ideal SkAT-VG system: by selecting through touch a video tile, the user can actually watch the performance and vocal articulation of the self-portrait, produced by a person (ideally the sound designer), and listen and compare the original vocal utterance with two synthetic versions, either produced with the sound models (i.e., timbral families) available in the Sound Design Toolkit, or with granular synthesis from MuBu objects (see table 10). Participants were asked to produce a short utterance which would describe and represent themselves, according to the following brief:

- If you were a sound, which sound would you be? Think about a sound which would represent yourself;
- It can be abstract, environmental, or related to one or more objects;
- Make it with your voice;
- Words, singing, animal calls, or imitation of musical instruments are not permitted;
- The duration should be between 3 and 5 seconds.

Beyond the SkAT-VG project communication goal, the making of S'i' fosse suono reveals two important research objectives: in practice, by acting as SkAT-VG system, the sound designer Andrea Cera had to interpret the vocal imitations, selecting the most appropriate sound models which embody the main characteristics of the utterances, choose and extract the most meaningful features from the vocal sounds, and articulate in time the profiles obtained, to manipulate the salient control parameters of the sound models. This process, despite being done manually, mainly overlaps with the four stages devised in the design of miMic (stage 0, select; stage 1, play; stage 2, explore; stage 3, refine). The making of S'i' fosse suono has been carefully documented, with the aims of analyzing retrospectively the sound-design thinking and collect important feedbacks regarding the refinement of the sound models, in terms of sound quality and control spaces. The audiovisual installation will be shown at the ICT Conference in Lisbon, together with the two demos miMic and MIMES.

Meetings and Events
WP7 Brainstorming cross-partners meeting IUAV - GENESIS - IRCAM Venice, April 21st-22nd, 2015.

No Deviations from Annex I

All Objectives achieved according to Schedule

No Corrective Actions Required

Task 7.3: Side applications of vocal and gestural sketching

DoW: The task, led by GENESIS will explore possible applications of SkAT-VG in areas such as sound effects for movies, real-time interaction in games, and sound information retrieval. Although these areas are not in the central focus of the project, the partners have the skills and interests that allow testing the SkAT-VG findings and technologies in a wide range of applications.

Progress

Work on this task has not started yet. However, the interviews with sound designers, done by GENESIS highlighted a set of side-applications which would possibly have a relevant immediate impact, in terms of market innovation. In particular, requests of tools based on sound information retrieval to navigate databases by means of vocal utterances, and transform sound by means of drawing manipulations on its visual representations have been taken in account. The development of these tools will start once the basic pieces of SkAT-VG software will be robust enough to be included in these side-applications. Side-applications based on non-verbal interaction represent the pacemaker towards a more visionary use of voice to sketch sounds.

Two potential partners have been identified for the experimentation and evaluation of the vocal sketching tools: a French manufacturer (still under negotiation) and Xavier Collet.

2.4 Project Management during the period

As in Year 1, project management has been relatively easy in Year 2 as well, due to the clear complementarity of project partners in terms of areas of activity. The points of interaction and exchange between partners have been easy to define and finalized to achieve specific research goals.

Project management and coordination are described in WP1, which

> **DoW:** has functions of financial and administrative management. Within WP1, resources are dedicated to manage the communication inside the project consortium and towards the European Commission, to prepare and conduct project meetings and reviews, to prepare the minutes, to manage the fund transfers towards the partners, to monitor and report on the execution of the financial plan. Resources are also dedicated to quality control, to assure that the development process follows the quality rules for the project. The measurable success factors for all other Work Packages are monitored in WP1.

The objectives of WP1

> **DoW:** To ensure financial and administrative management of the project. To develop a spirit of co-operation between the partners. To ensure consensus management and information circulation among the partners. To ensure project reporting and interface with the Project Officer. To co-ordinate and control project activities to keep it within the objectives. To ensure quality management of the project.

have been achieved. No deviations from the workplan for Y2 were deemed nor observed.

Management

The management structure described in Section 2.1.1 of the DoW, and implemented as reported in the First Periodic Report, is unchanged in Y2. The roles are:

Coordination Team IUAV

Project Manager Davide Rocchesso

Scientific Assistant Stefano Delle Monache

Administrative Assistant Ilaria Rosa

Project Committee :

IUAV Local Manager Davide Rocchesso

IRCAM Local Manager Patrick Susini

KTH Local Manager Sten Ternström

GENESIS Local Manager Patrick Boussard

Work Package Leaders :

WP1 - **IUAV** Davide Andrea Mauro

WP2 - **KTH** Sten Ternström

WP3 - **KTH** Pétur Helgason

WP4 - **IRCAM** Guillaume Lemaitre

WP5 - **IRCAM** Geoffrey Peeters

WP6 - **IUAV** Stefano Baldan

WP7 - **GENESIS** Patrick Boussard

Quality assurance and risk management

According to what is specified in Section 1.3.1 of the DoW, and whenever it is possible, preventive actions are conceived at the project design stage to reduce the identified failure risks related to each WP. Contingency plans may be activated when a risk actually occurs. For these reasons, some works have been anticipated and a couple of deviations (one in Year 2) have been approved, as reported in table 2.

Collaboration infrastructure and access to documents

The following elements of a collaboration infrastructure have been established since the beginning of the project:

Web sites and Social Networks `http://skatvg.iuav.it/` mirrored on `http://skatvg.eu/`, `https://twitter.com/SkATVG`, `https://vimeo.com/skatvg`. The website is intended as a showcase for the project itself. It encompasses a section for presenting the project, one section for presenting the partners involved in the project, and a number of sections to keep track of what is happening in the context of the project. Recent news and updates are "tweeted" to reach a wider audience. Accounts have been set on Freesound (`http://www.freesound.org/people/skat_vg/`) and Soundcloud (`https://soundcloud.com/skat-vg`) for sharing sounds.

Mailing lists General list `skat-vg@ircam.fr` used extensively as an efficient discussion platform and for general organizational purposes. Managed at IRCAM;

Coordinator list `skat-vg@iuav.it` used for internal purposes by the Coordination Team. Managed at IUAV.

Redmine project management tool https://redmine.skatvg.iuav.it Redmine is a flexible project management web application, which provides several useful features:

Issues This is the core functionality of the application. Everything regarded as important for the project can be raised as an "issue" (with various types and definitions) and the evolution can be managed, assigned to specific people, and monitored.

Documents and Files These repositories contain documents that can be shared between partners and be directly linked in the Issues.

Wiki A Wiki is used as a collaborative workplace to share information, e.g. Lists of similar and relevant technologies, Relevant Literature, Calls for Conferences, and so on.

Others (GANTT, Calendar, Activity) These facilities are used for specific purposes such as automatically producing GANTT charts (see Figure 3,) sharing events on a calendar, keeping track of the overall project activity.

SVN repository https://skatvg.iuav.it/svn/skatvg_svn: A revision control system for collaboratively writing documents, publications, and reports (especially in LaTeX) and for collaborative software development. The repository is readable from the Redmine page and it exposes the directories: • Budget • Code • Data • Documents • Misc

GitHub repository https://github.com/SkAT-VG: GitHub is a Web-based Git repository hosting service, which offers all of the distributed revision control and source code management (SCM). It is the main public resource where to find the software outcomes of the project. It allows the researcher from the consortium to update the code while permitting to the general public to download the source code.

Build in Progress http://buildinprogress.media.mit.edu has been chosen to document the design and development process of some SkAT-VG prototypes.

Teleconferencing A number of solutions have been evaluated, with a decision to use Skype. Conference calling is an effective tool to reach consensus about technical issues among multiple partners. Minutes are kept in the project's Redmine installation.

Starting from the beginning, the project had planned intra- and inter-WPs group calls, where members updated each other on the development status of the individual WPs and on plans for the successive periods. In most cases more focused follow-up discussions were held by some of the partners.

Meetings and exchanges

Since January 2015, in its second year the project arranged the following meetings and research visits:

General Meetings

- Venice Project and Review Meeting: 27-30th January, 2015. Minutes of the meeting are available on Redmine.

- Aix-en-Provence Project Meeting Mo18: 26-28th August, 2015. Minutes of the meeting are available on Redmine.

Cross-partner Meetings

- Venice: Genesis-IUAV-IRCAM-KTH. 21-22nd April 2015. Minutes of the meeting are available on Redmine. Focus of the meeting was on vocal production and analysis.
- Paris: IRCAM-IUAV-GENESIS-KTH, 22-24th June 2015. Minutes of the meeting are available on Redmine. Focus of the meeting was on WP5.

Teleconferences:

- WP2: WP Formally ended on Mo12;
- WP3: 6th October 2015, with WP7;
- WP4: 31st March 2015, 21st July 2015;
- WP5: 23rd February 2015;
- WP6: 18th March 2015, 16th July 2015;
- WP7: 15th April 2015, 21st July 2015.

Minutes from these meetings are available on Redmine.

Research Visits:

- Davide Rocchesso (IUAV) at Genesis: 16-18th February, 2015.
- Stefano Baldan (IUAV) at Genesis: 23rd March, 2015.

3 Deliverables and Milestones tables

3.1 Deliverables

All Year 1 and 2 deliverables (excluding the periodic reports) are summarized in Table 6. The year 2 deliverables are described as follows:

D3.3.2 Final comprehensive annotation of the database of imitations.

D4.4.1 A large set of vocal and gestural imitations. It includes the database of referent sounds used to elicit the imitations, and a statistical analysis of the database of imitations.

D5.5.1 Blind classifiers of imitations. The deliverable describes the work performed by IRCAM for automatic recognition of sound categories from vocal imitations and for automatic recognition from gesture, and KTH for automatic recognition of articulatory mechanism from vocal.

D6.6.1 Automatic system for the generation of sound sketches. Distribution of the softwares Sound Design Toolkit and SkAT Studio.

3.2 Milestones

Milestone M1, "Accumulation of a large enough database of recorded, sorted, and labeled imitations", was achieved after Year 1.

Some work aimed at achieving Milestone M2, "Automatic classifiers of vocal and gestural imitations into categories of imitated sounds", was anticipated to the first year of the project, and its achievement can be considered to be complete by the end of Year 2. In particular, IRCAM has been doing and is still doing work to assess the success of vocal and gestural imitation in conveying the imitated referent sounds, conducted statistical analyses of the large database of imitations, and developed classifiers and descriptors. KTH has extended the annotation work and developed classifiers based on auditory receptive fields. IUAV has started developing prototype applications that make use of machine learning for sound model selection.

Activities towards the definition and implementation of "Integrated sketching tools", namely Milestone M3, have also been done during the second year of the project. The work of GENESIS centered around interactions with several sound design professionals and development of a tool (SkAT Studio) that addresses some of the needs of such stakeholders. IUAV has been organizing, conducting, and analyzing two design workshops on the use of vocal sketching for product design. In these workshops, the new sound models and prototype tools are being exploited and exposed to user testing for further refinement.

Del. no.	Deliverable name	WP no.	Lead beneficiary	Nature	Dissemination level	Delivery date from Annex 1 (Mo)	Delivered Yes/No	Actual / Forecast delivery date (Mo)	Comments
2.2.1	Explorative collection of imitated sounds	2	KTH	R	PU	4	Yes	5	
2.2.2	Extensive set of recorded imitations	2	KTH	R	PU	12	Yes	12	
3.3.1	Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives	3	KTH	R	PU	12	Yes	12 / 14	draft / update
3.3.2	Final comprehensive annotation of the database of imitations	3	KTH	R	PU	24	Yes	22 / 24	draft / update
4.4.1	A large set of vocal and gestural imitations	4	IRCAM	R	PU	21	Yes	21 / 21	
5.5.1	Blind classifiers of imitations	5	IRCAM	R	PU	23	Yes	22 / 23	draft / update
6.6.1	Automatic system for the generation of sound sketches	6	IUAV	P	PU	24	Yes	22 / 24	draft / update

Table 6: Deliverables for Years 1 and 2 of SkAT-VG. The approved deliverables, which are publicly available from the SkAT-VG website, are grayed-out in the table.

Milestone no.	**Milestone name**	**WP no.**	**Lead beneficiary**	**Delivery date from Annex 1 (Mo)**	**Delivered Yes/No**	**Actual / Forecast delivery date (Mo)**	**Comments**
M1	Accumulation of a large enough database of recorded, sorted, and labeled imitations	2, 3, 4	KTH	12	Yes	12/15	draft/update (Figure 2)
M2	Automatic classifiers of vocal and gestural imitations into categories of imitated sounds	3, 4, 5, 6	IRCAM	24	Yes	22/25	draft/update (Figure 2)

Table 7: Milestones for Years 1 and 2 of SkAT-VG.

4 Explanation of the Use of Resources and Financial Statements

No financial statements are due for the reporting period. Nevertheless, the following tables detail the major costs that occurred in the second year, in a form consistent with Section 2.4 of the DoW. The figures are approximate, as they have been computed at the beginning of October 2015. Overall, the **estimated costs** are consistent with the planned budget.

Description	Resources	Cost
	IUAV	
Personnel	Research positions D.A. Mauro, researcher, 01/01/15 - ; S. Baldan, researcher, 01/01/15 - ; S. Delle Monache, assistant professor, 01/01/15 -; A. Cera, sound designer, 01/10/2015 -)	84168
	Management I. Rosa, administration collaborator, 01/01/15 - ; D. Rocchesso	26750
Equipment	Computers	5300
	Software licenses	1000
	Audiovisual equipment	2000
	Fabrication costs for prototypes	150
	Other	300
Travel	10 Conference trips S. Baldan: IEEE VR 2015 (Arles - Aix en Provence, France), 03/2015; A. Del Piccolo: Summer School (Lille, France) 05/2015; S. Baldan, S. Delle Monache ICAD 2015 (Graz, Austria) 07/2015; A. Del Piccolo: SMC Conference 2015(Maynooth, Ireland) 07/2015; D. Rocchesso: AudioMostly (Thessalonika, Greece) 10/2015; D. Rocchesso, S. Delle Monache, S. Baldan: ICT Conference 2015 and Extra Review Meeting (Lisbona, Portugal) 10/2015; 12/2015 S. Baldan: DAFX 2015 (Trondheim, Norway)	9000
	2 Participations to Project Meetings D. Rocchesso, S. Delle Monache: Project Meeting (Aix en Provence, France) 08/2015	1100
	4 Cross-partner and Review Meetings Review meeting (Venice, Italy), 01/15; D. Rocchesso Iuav-Genesis-Ircam: (Marseille, France), 02/2015; Iuav-Genesis-Ircam-KTH (Venice, Italy) 04/2015; D. Rocchesso Iuav-Ircam (Paris, France) 06/2015;	4300
	2 Research workshops D. Rocchesso, S. Delle Monache (York, Uk), 05/2015; D. Rocchesso, S. Delle Monache, S. Baldan (Copenaghen, Denmark) 11/2015	4500
	1 Research Missions S. Delle Monache: Starts Symposium (Bruxelles, Belgium) 06/2015	700
Dissemination	Publications and promotion	4100
Total direct costs		143368

Description	Resources	Cost
	IRCAM	
Personnel	Research positions (G. Lemaitre, F. Voisin, E. Marchetto, J. Francoise, G. Meseguer, researchers; R. Dubelski, S. Levesque, expert participants in experiments)	140079
	Permanent research and management positions (F. Bevilacqua, O. Houix, N. Misdariis, G. Peeters, P. Susini)	49000
	Master internship (H. Scurto)	706
	Rewards for participants in experiments	100
Equipment	Computers	1487
	Audiovisual equipment	75
Travel	Organization of WP5 Meeting Paris, France, 6/15	428
	9 Participations to Project Meetings	5284
	Conference trip	2544
Total direct costs		199703

Description	Resources	Cost
	KTH	
Personnel	Research positions Sten Ternström, professor, 01/01/15 - ; Anders Friberg, associate professor, 01/01/15 - ; Pétur Helgason, 01/01/15 - ; Tony Lindeberg, 01/01/15 - 31/01/15; Glaucia Salomão, 01/01/15 - 31/12/15	122232
Equipment	Computers	1421
	Materials	205
	Ethical vetting fee	532
Travel	Meetings Project and Review Meeting (Venice), 01/15; Project Meeting (Aix), 08/15; ICT + Review (Lisbon), 10/15, Ircam visit (Paris)	6793
Total direct costs		131183

Description	Resources	Cost
GENESIS		
Personnel	Research positions Patrick Boussard, researcher, 01/01/15 - ; Hélène Lachambre, researcher, 01/01/15 - ; Guillaume Stempfel, researcher, 01/01/15 - ; Stéphane Molla, 01/01/15 -	89894
	Management Patrick Boussard, 01/01/15 -	5041
Equipment	Computers	2188
	Software licenses	548
	Audiovisual equipment	2997
Travel	Participations to Project Meetings P. Boussard, H. Lachambre, G. Stempfel: Project/Review Meeting (Venice, Italy), 01/15; P. Boussard, H. Lachambre, G. Stempfel: Project meeting (Aix-en-Provence, France), 08/15	5560
	Cross-partner meetings H. Lachambre: IUAV-Genesis (Venice, Italy), 04/15; WP7 Meeting (Aix en Provence, France), 03/15; H. Lachambre: WP5 meeting (Paris, France), 06/15; SkAT presentation to potential partners (Paris, France), 06/15	2796
	Conferences H. Lachambre, P. Boussard: SIVE/IEEEVR (Arles, France), 03/15; H. Lachambre: Gretsi (Lyon, France), 09/15	4577
	Interviews	1185
Total direct cost		114787

Table 8 and Figure 6 report the overview of Person-Months status (cumulative). Overall, the table gives account of a homogeneous development of the scientific work.

The only Workpackage that has been finished in the reporting period is WP2, with an effort (20.28PM) smaller than the expected (28PM).

PM Claimed	WP1	WP2	WP3	WP4	WP5	WP6	WP7	TOTAL
IUAV	6.06/12	3.38/4	3.38/3	3.37/3	2.95/2	19.99/30	14.88/15	54.02/69
IRCAM	1.76/3	2/2	2/4	34.76/48	21.47/28	1.5/4	2.6/4	66.09/93
KTH	1.6/3	10.5/16	16/32	0/2	5.6/8	0/3	0/4	33.7/68
GENESIS	1/4	4.4/6	0/0	5.6/6	4/6	4/8	14.5/20	33.5/50
Total	10.42/22	20.28/28	21.38/39	43.73/59	34.02/44	25.49/45	31.98/43	187.3/280

Table 8: Effort in Person-Months, per Workpackage and per partner. Years 1 and 2.

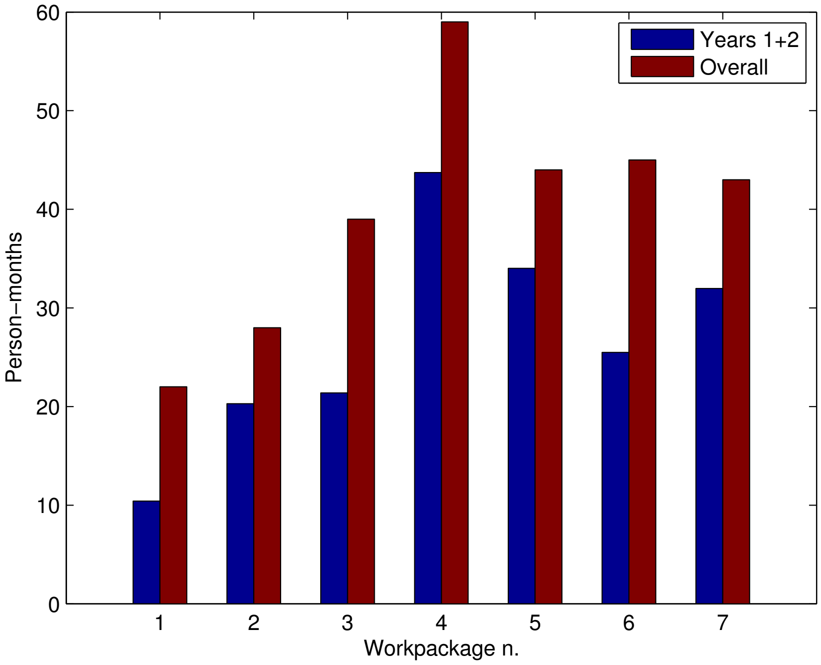


Figure 6: Person months of Years 1 and 2 as compared to the overall planned effort over three years.

Table 9 and Figure 7 report the overview of Personnel and Other direct costs.

Costs	Personnel	Other	TOTAL
IUAV	255787/394800	57877/117000	313664/511800
IRCAM	321222/487753	17764/91200	338986/578953
KTH	229151/445600	18106/80000	247257/525600
Genesis	210709/264550	32172/32666	242881/297216
Total	1016869/1592703	125919/320866	1142788/1913569

Table 9: Personnel and Other Direct Costs, per partner: Years 1 and 2 / Whole Project.

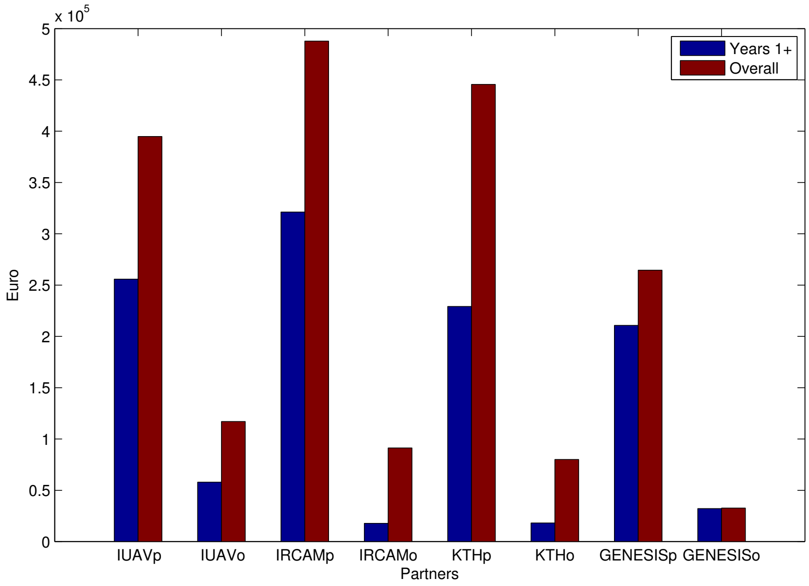


Figure 7: Costs per partner for Years 1 and 2 as compared to the overall planned costs over three years. “p”: Personnel, “o”: Other Direct Costs.

5 List of Publications, Networking, and Dissemination Activities

Publications:

1. "Sonic Introspection by Vocal Sketching" A. Cera, D. A. Mauro, and D. Rocchesso. Accepted article proposal for Volume 26 ("lend me your ears! - sound and reception") of Leonardo Music Journal, 2016

2. "miMic: The microphone as a pencil" D. Rocchesso, D. Mauro, and S. Delle Monache. Accepted at the 10th International Conference on Tangible, Embedded and Embodied Interaction (TEI), Eindhoven (The Netherlands), 14-17 February 2016.

3. "Combining gestures and vocalizations during sound imitation" H. Scurto, G., J. Françoise, P. Susini, F. Bevilacqua. In preparation for PlosONE or Frontiers in Psychology, 2015.

4. "Identification of vocal imitations" G. Lemaitre, O. Houix, N. Misdariis and P. Susini. In preparation fo JASA, 2015.

5. "Gesture Analysis Using the Continuous Wavelet Transform" J.Françoise, F. Bevilacqua et al.. In preparation, 2015.

6. "Vocal imitations of basic acoustic features" G. Lemaitre, A. Jabbari, N. Misdariis, O. Houix, and P.Susini. Submitted to the Journal of the Acoustical Society of America, 2015.

7. "Vocal imitations of basic auditory features" G. Lemaitre, A. Jabbari, O.Houix, N. Misdariis, P. Susini, The Journal of the Acoustical Society of America, vol. 137 (4), p. 2268, 2015. (Proceedings of the meeting of Acoustical Society of America, Pittsburgh, PA).

8. "Combining gestures and vocalizations during sound imitation" H. Scurto, G. Lemaitre, J. Françoise, P. Susini, F. Bevilacqua. To appear in the Proceedings the of the meeting of Acoustical Society of America, Jacksonville, FL, 2015.

9. "A set of audio features for the morphological description of vocal imitations" E. Marchetto and G. Peeters. Digital Audio Effects Conference, Trondheim, Norway, 2015.

10. "Reverberation still in business: Thickening and propagating micro-textures in physics-based sound modeling" D. Rocchesso, S. Baldan, and S. Delle Monache. Digital Audio Effects Conference, Trondheim, Norway, 2015.

11. "Analyzing and organizing the sonic space of vocal imitation" D.A Mauro, and D. Rocchesso. Audio Mostly 2015, Thessaloniki (Greece), 07-09 October, 2015.

12. "Multisensory texture exploration at the tip of the pen" D. Rocchesso, S. Delle Monache and S. Papetti. International Journal of Human-Computer Studies, in press, 2015.

13. "To "Sketch a Scratch" " A. Del Piccolo, S. Delle Monache, D. Rocchesso, S. Papetti, D.A. Mauro. 12th Sound & Music Computing conference (SMC), Maynooth (Ireland), July 26 - August 01, 2015.

14. "Growing the practice of vocal sketching" S. Delle Monache, D. Rocchesso, S. Baldan, D.A. Mauro. 21st International Conference on Auditory Display (ICAD–2015), Graz (Austria), 07-10 July, 2015.

15. "Sketching sound with voice and gesture." D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, and P. Boussard. interactions 22:1, 2015, pp. 38-41.

16. "Non-Verbal Imitations as a Sketching Tool for Sound Design." G. Lemaitre, P. Susini, D. Rocchesso, C. Lambourg, P. Boussard, In Mitsuko Aramaki et al., editors, Lecture Notes in Computer Sciences : Sound, Music, and Motion. Springer, Berlin/Heidelberg, Germany, 2014, pp. 558-574.

17. "Bauhaus legacy in Research through Design: the case of Basic Sonic Interaction Design." S. Delle Monache, D. Rocchesso. International Journal of Design [Online] 8:3, 28 December 2014.

18. "Self-organizing the space of vocal imitations" D. Rocchesso, and D.A. Mauro. XX Colloquio di Informatica Musicale, Rome (Italy), 20-22 October, 2014.

19. "His engine's voice: towards a vocal sketching tool for synthetic engine sounds" S. Baldan , S. Delle Monache, and L. Comanducci. XX Colloquio di Informatica Musicale, Rome (Italy), 20-22 October, 2014.

20. "Sounding objects in Europe", D. Rocchesso. The New Soundtrack, Volume 4, Issue 2, pp 157–164, September, 2014.

21. "A design exploration on the effectiveness of vocal imitations" S. Delle Monache , S. Baldan D.A. Mauro, and D. Rocchesso. 40th International Computer Music Conference (ICMC) joint with the 11th Sound and Music Computing conference (SMC), Athens (Greece), 14-20 September, 2014.

22. "2001-2016: Oggetti sonanti in Europa." D. Rocchesso, Associazione Italiana di Acustica 41° Convegno Nazionale, Pisa (Italy), 17-19 June, 2014.

23. "Sound initiation and source types in human imitations of sounds", P. Helgason. In proceedings of FONETIK. pp 83–89, Stockholm (Sweden), 09-11 June, 2014.

24. "Might as Well Jump: Sound Affects Muscle Activation in Skateboarding" P. Cesari, I. Camponogara, S. Papetti, D. Rocchesso, and F. Fontana. PLOS ONE, Volume 9, Issue 3, 9 March 2014.

25. "On the effectiveness of vocal imitations and verbal descriptions of sounds." G. Lemaitre, and D. Rocchesso. Journal of Acoustical Society of America 135(2):862–873, 2014.

26. “Sketch a Scratch.” S. Delle Monache, D. Rocchesso, and S. Papetti. 8th International Conference on Tangible, Embedded and Embodied Interaction (TEI), Munich (Germany), 16-19 February, 2014.

Networking:

1. Tony Lindeberg, professor, KTH, Dept. of Computational Biology, `http://www.csc.kth.se/~tony/`.

2. Luca Andrea Ludovico, assistant professor, Università degli Studi di Milano, `http://www.ludovico.net/`.

3. Xavier Collet, sound designer, `http://xaviercollet.com`.

4. Fernando Ocaña, Creative Director at Semcon Hybrid Design Studios, Sweden, `http://www.semcon.com/en/Services/Design/`.

5. Isabelle Ballet, Audio Director at Ubisoft, video game publisher, France, `https://www.ubisoft.com/en-US/studio/paris.aspx`.

6. Norio Kubo, Director and Product Sound Designer at Yokohama Acoustics Institute, Inc., Yokohama, Japan, `http://yokohama-onkyo.jp/`.

7. Andrea Cera, independent Sound Designer, Italy `http://andrea.cera.free.fr/`.

8. Mathieu Pellerin, freelance Sound Designer, France, `http://www.mathieupellerin.com/`.

9. Jean-François Sciabica, scientist and engineer in sound perception at PSA Peugeot-Citröen, France.

10. Andy Farnell, sound designer, researcher in procedural audio at Queen Mary University of London, England, `http://obiwannabe.co.uk/`.

11. Adam Stark, freelance sound designer and software developer, England, `http://www.adamstark.co.uk/`.

12. Christian Heinrichs, PhD candidate in Electronic Engineering and Computer Science at Queen Mary University of London, England.

13. Richard Kronland-Martinet, Research Director at CNRS/LMA - Equipe Sons, Marseille, France, `http://www.lma.cnrs-mrs.fr/spip/?lang=fr`.

14. Martin Geijer, Director and Founder at Improvisationsteater Svea AB, Stockholm, Sweden, `http://improvisationsteater.se`.

15. Richard Dubelski, Performer, Singer, Musician, Paris, France.

16. Sylvie-Bobette Levesque, Performer, Singer, Paris, France.

17. Luigi Maffei, Full Professor at Department of Architecture and Industrial Design, Seconda Università di Napoli, Napoli, Italy, http://www.architettura.unina2.it/docenti.asp?ID=67.

18. T. Metin Sezgin, Assistant Professor at Department of Computer Engineering, Koc University College of Engineering, Istanbul, Turkey, http://iui.ku.edu.tr/.

19. Martin Luccarelli, Assistant Professor at Department of Design and Art, Libera Università di Bolzano, Bolzano, Italy http://www.unibz.it/en/design-art/welcome/default.html.

20. Mark Cartwright, PhD Candidate in Electrical Engineering and Computer Science at Northwestern University - Interactive Audio Lab, Evanston, Illinois, http://music.cs.northwestern.edu/.

21. Carlo Drioli, Assistant Professor at Department of Mathematics and Computer Science, Università degli Studi di Udine, Italy http://people.uniud.it/page/carlo.drioli

Dissemination Activities:

1. The authors of "Combining gestures and vocalizations during sound imitation", being presented at the 170th ASA Meeting in Jacksonville (USA) have been invited to submit a lay-language version of the paper, which will be posted on the online press room of the Acoustical Society of America, http://acoustics.org/world-wide-press-room/.

2. "Imagining, sketching and prototyping sound", invited talk at the Sound and Music Computing Colloquium, Aalborg University Copenhagen, 26 November, 2015.

3. SkAT-VG at ICT2015, Lisbon, Innovate Area, Booth number: i09, 20-22 October, 2015, http://ec.europa.eu/digital-agenda/events/cf/ict2015/item-display.cfm?id=14872.

4. "Sketching Audio Technologies using Vocalization and Gestures: Le projet SkAT-VG, Projet Européen FP7-ICT FET-Open: challenging current thinking". Séminaire Recherche et Technologie, Ircam, Paris, France, December 2014.

5. Ca' Foscari University Opening of the Doctoral Year. 14 November, 2014.

6. "Sketching Audio Technologies using Vocalizations and Gestures". Workshop and demo. Ircam open days, 11 June, 2014, Paris.

7. Research seminar at KTH, the Marcus Wallenberg Laboratories for Sound and Vibration, Pétur Helgason and Sten Ternström presented SkAT-VG and EUNISON projects, 4 April, 2014.

8. World Voice Day. 16 April, 2014. Open seminar "Music and Speech Sounds" at KTH, Stockholm, with several high-profile composers and litterati. SkAT-VG was presented by Sten Ternström and Pétur Helgason.

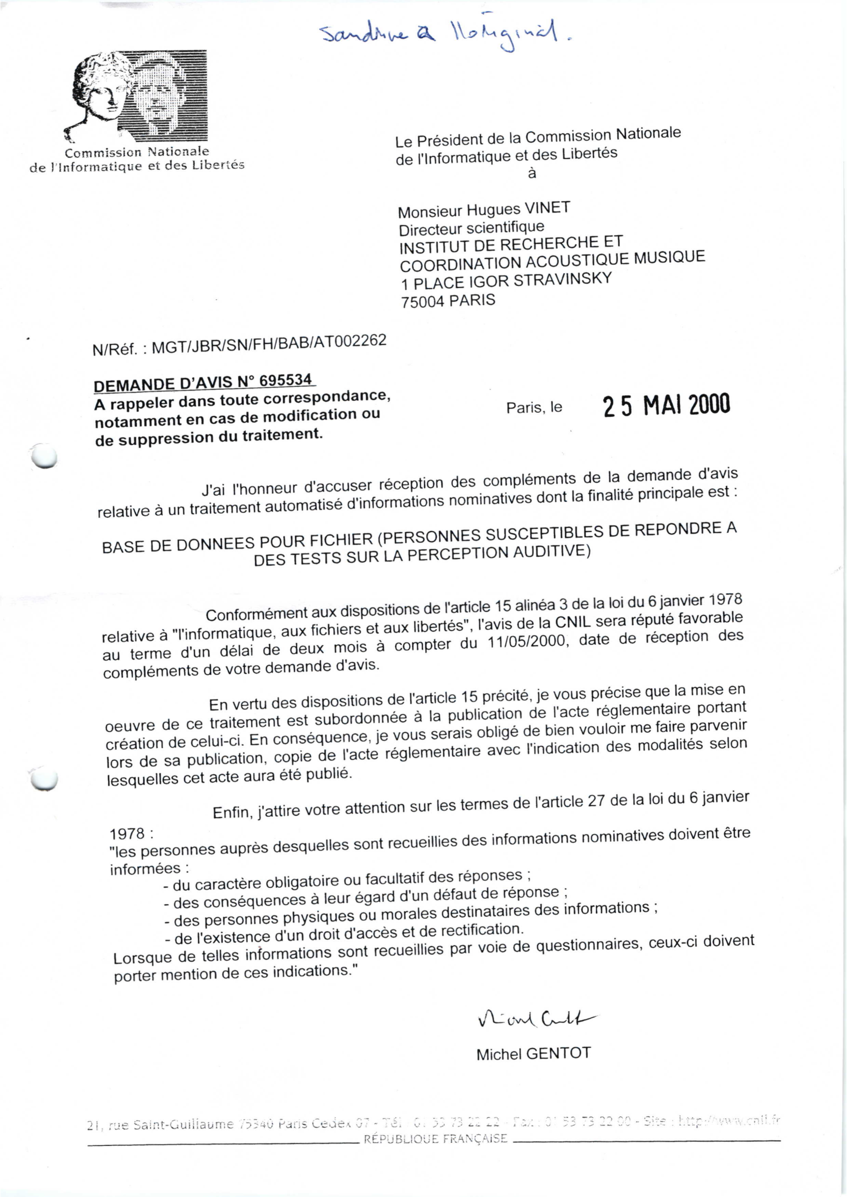
9. World Voice Day. 16 April, 2014. Performance "Textures from an Exhibition" and SkAT-VG presentation.

6 Ethical issues

SkAT-VG is performing psychological experiments, which are taking place in Paris, within the scope of WP4. A list of subjects for experiments is composed of adult healthy volunteers taken from a database that respects French legislation, and already registered at the CNIL (Commision Nationale de l'Informatique et des Libertés, see Figure 8). This database contains personal information about people (address, age, sex, musical practice) who register themselves to participate in psychological experiments. Subjects have the possibility to leave the experiment any time they want. Subjects are reimbursed for their participation, even if they decide to leave the experiment before the end. A new registration for the database of video recordings has been accepted by the CNIL in 2015 (see Figure 9). The experiments performed at IRCAM have been approved by the ethical committee of the French National Institute for Medical Research (CEEI IRB of Inserm) under the number IRB00003888 (pending minor modifications). This committee is a registered Institutional Review Board (IRB) that meets the international ethic standards. A practical consequence is that the results of such experiments can be published in journals such as Plos One, for which an official IRB number is mandatory.

This database is used only within the framework of auditory experiments at IRCAM and will never be shared. The consent form is reproduced in Figure 10. Each subject is identified by a code, and each subject's data is only labeled with this code. The correspondence between subject's code and identity is stored in a separate place.

Within the scope of WP2 and WP3, SkAT-VG has also performed observational studies of performers/imitators. The studies included audio, video and EGG recordings. These experiments have been carried out at KTH, and the participants are improvisational actors recruited through an agency and paid for their participation. Normally, when performing experiments with lay subjects KTH applies for ethical approval from Regionala Etikprövningsnämnden i Stockholm. In this case, these professional stakeholders were requested to sign the consent form reproduced in Figure 11. Ethical approval for making recordings of lay subjects was sought from the regional ethical vetting committee in Stockholm (EPN). In its reply of 2015-09-16, and reported in figure 12, the committee stated that the proposed procedure does not present any ethical objections, and therefore does not need to be considered for approval.



Sandrine a l'original.

Commission Nationale
de l'Informatique et des Libertés

Le Président de la Commission Nationale
de l'Informatique et des Libertés
à

Monsieur Hugues VINET
Directeur scientifique
INSTITUT DE RECHERCHE ET
COORDINATION ACOUSTIQUE MUSIQUE
1 PLACE IGOR STRAVINSKY
75004 PARIS

N/Réf. : MGT/JBR/SN/FH/BAB/AT002262

DEMANDE D'AVIS N° 695534
A rappeler dans toute correspondance, notamment en cas de modification ou de suppression du traitement.

Paris, le 25 MAI 2000

J'ai l'honneur d'accuser réception des compléments de la demande d'avis relative à un traitement automatisé d'informations nominatives dont la finalité principale est :

BASE DE DONNEES POUR FICHIER (PERSONNES SUSCEPTIBLES DE REPONDRE A DES TESTS SUR LA PERCEPTION AUDITIVE)

Conformément aux dispositions de l'article 15 alinéa 3 de la loi du 6 janvier 1978 relative à "l'informatique, aux fichiers et aux libertés", l'avis de la CNIL sera réputé favorable au terme d'un délai de deux mois à compter du 11/05/2000, date de réception des compléments de votre demande d'avis.

En vertu des dispositions de l'article 15 précité, je vous précise que la mise en oeuvre de ce traitement est subordonnée à la publication de l'acte réglementaire portant création de celui-ci. En conséquence, je vous serais obligé de bien vouloir me faire parvenir lors de sa publication, copie de l'acte réglementaire avec l'indication des modalités selon lesquelles cet acte aura été publié.

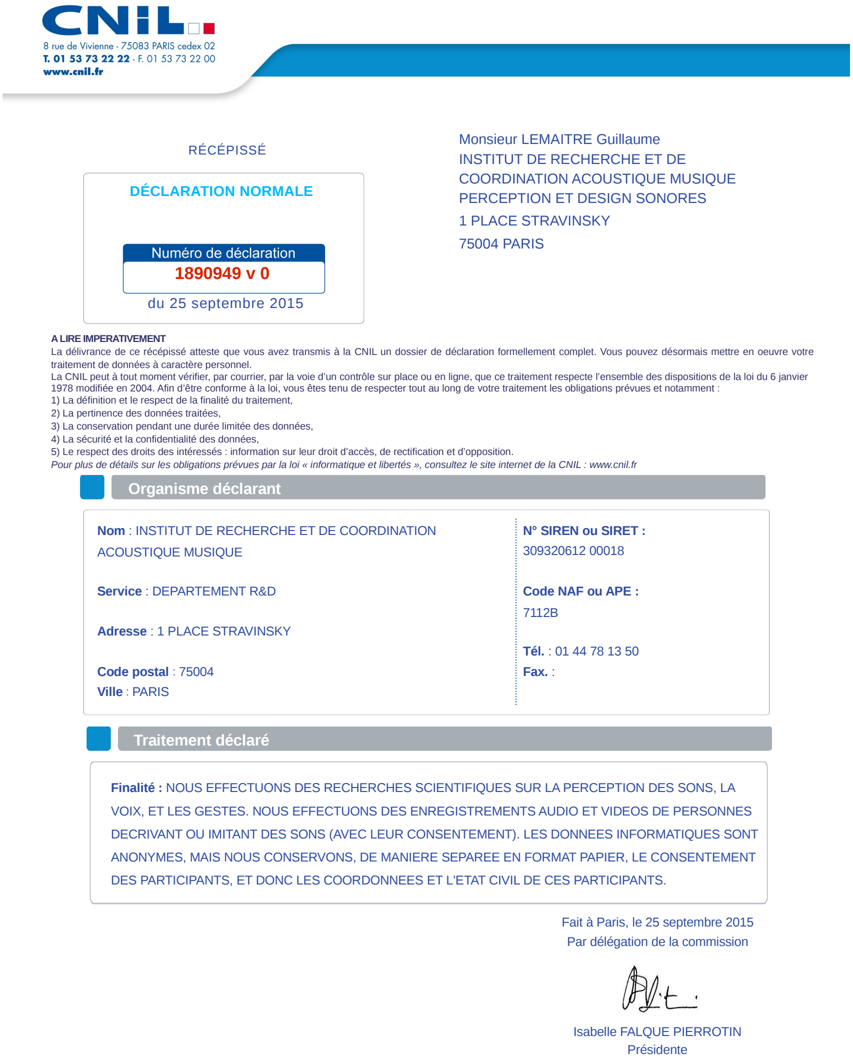
Enfin, j'attire votre attention sur les termes de l'article 27 de la loi du 6 janvier 1978 :
"les personnes auprès desquelles sont recueillies des informations nominatives doivent être informées :
- du caractère obligatoire ou facultatif des réponses ;
- des conséquences à leur égard d'un défaut de réponse ;
- des personnes physiques ou morales destinataires des informations ;
- de l'existence d'un droit d'accès et de rectification.

Lorsque de telles informations sont recueillies par voie de questionnaires, ceux-ci doivent porter mention de ces indications."

Michel GENTOT

21, rue Saint-Guillaume 75340 Paris Cedex 07 - Tél : 01 53 73 22 22 - Fax : 01 53 73 22 00 - Site : http://www.cnil.fr
RÉPUBLIQUE FRANÇAISE

Figure 8: IRCAM Acknowledge of receipt of the IRCAM database by the Commission Nationale de l'Informatique et des Libertés (CNIL). According to this letter, no answer within two months indicates that the database is accepted.



CNIL
8 rue de Vivienne - 75083 PARIS cedex 02
T. 01 53 73 22 22 - F. 01 53 73 22 00
www.cnil.fr

RÉCÉPISSÉ

DÉCLARATION NORMALE

Numéro de déclaration
1890949 v 0
du 25 septembre 2015

Monsieur LEMAITRE Guillaume
INSTITUT DE RECHERCHE ET DE
COORDINATION ACOUSTIQUE MUSIQUE
PERCEPTION ET DESIGN SONORES
1 PLACE STRAVINSKY
75004 PARIS

A LIRE IMPERATIVEMENT

La délivrance de ce récépissé atteste que vous avez transmis à la CNIL un dossier de déclaration formellement complet. Vous pouvez désormais mettre en oeuvre votre traitement de données à caractère personnel.

La CNIL peut à tout moment vérifier, par courrier, par la voie d'un contrôle sur place ou en ligne, que ce traitement respecte l'ensemble des dispositions de la loi du 6 janvier 1978 modifiée en 2004. Afin d'être conforme à la loi, vous êtes tenu de respecter tout au long de votre traitement les obligations prévues et notamment :

1) La définition et le respect de la finalité du traitement,
2) La pertinence des données traitées,
3) La conservation pendant une durée limitée des données,
4) La sécurité et la confidentialité des données,
5) Le respect des droits des intéressés : information sur leur droit d'accès, de rectification et d'opposition.

Pour plus de détails sur les obligations prévues par la loi « informatique et libertés », consultez le site internet de la CNIL : www.cnil.fr

Organisme déclarant

Nom : INSTITUT DE RECHERCHE ET DE COORDINATION ACOUSTIQUE MUSIQUE

Service : DEPARTEMENT R&D

Adresse : 1 PLACE STRAVINSKY

Code postal : 75004
Ville : PARIS

N° SIREN ou SIRET :
309320612 00018

Code NAF ou APE :
7112B

Tél. : 01 44 78 13 50
Fax. :

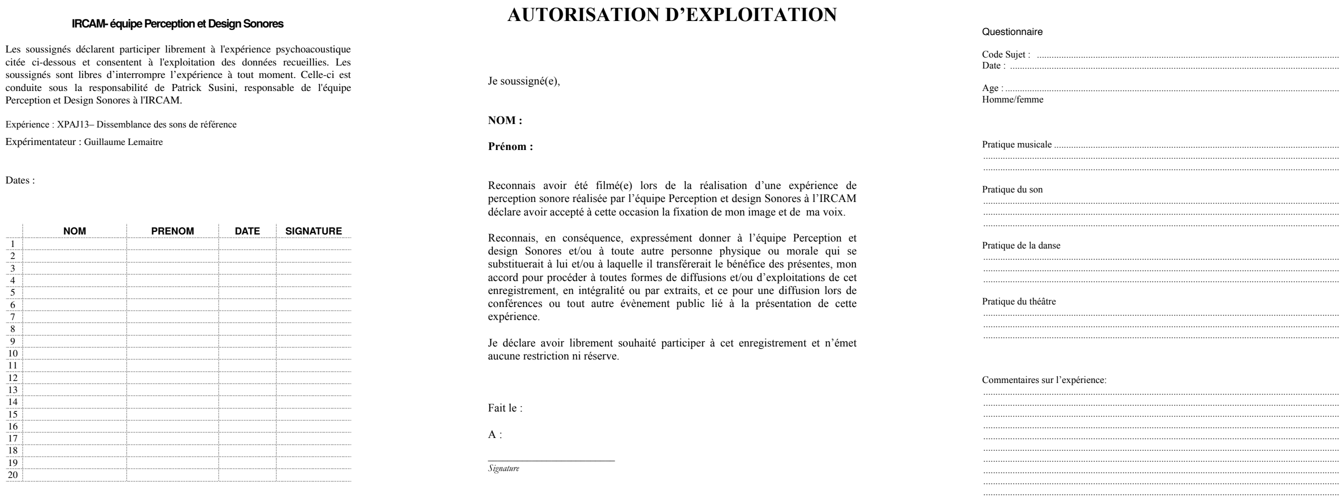
Traitement déclaré

Finalité : NOUS EFFECTUONS DES RECHERCHES SCIENTIFIQUES SUR LA PERCEPTION DES SONS, LA VOIX, ET LES GESTES. NOUS EFFECTUONS DES ENREGISTREMENTS AUDIO ET VIDEOS DE PERSONNES DECRIVANT OU IMITANT DES SONS (AVEC LEUR CONSENTEMENT). LES DONNEES INFORMATIQUES SONT ANONYMES, MAIS NOUS CONSERVONS, DE MANIERE SEPAREE EN FORMAT PAPIER, LE CONSENTEMENT DES PARTICIPANTS, ET DONC LES COORDONNEES ET L'ETAT CIVIL DE CES PARTICIPANTS.

Fait à Paris, le 25 septembre 2015
Par délégation de la commission

Isabelle FALQUE PIERROTIN
Présidente

Figure 9: IRCAM Acknowledge of receipt of the IRCAM database of video recordings (2015) by the Commission Nationale de l'Informatique et des Libertés (CNIL)



IRCAM- équipe Perception et Design Sonores

Les soussignés déclarent participer librement à l'expérience psychoacoustique citée ci-dessous et consentent à l'exploitation des données recueillies. Les soussignés sont libres d'interrompre l'expérience à tout moment. Celle-ci est conduite sous la responsabilité de Patrick Susini, responsable de l'équipe Perception et Design Sonores à l'IRCAM.

Expérience : XPAJ13– Dissemblance des sons de référence

Expérimentateur : Guillaume Lemaitre

Dates :

	NOM	PRENOM	DATE	SIGNATURE
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

AUTORISATION D'EXPLOITATION

Je soussigné(e),

NOM :

Prénom :

Reconnais avoir été filmé(e) lors de la réalisation d'une expérience de perception sonore réalisée par l'équipe Perception et design Sonores à l'IRCAM déclare avoir accepté à cette occasion la fixation de mon image et de ma voix.

Reconnais, en conséquence, expressément donner à l'équipe Perception et design Sonores et/ou à toute autre personne physique ou morale qui se substituerait à lui et/ou à laquelle il transférerait le bénéfice des présentes, mon accord pour procéder à toutes formes de diffusions et/ou d'exploitations de cet enregistrement, en intégralité ou par extraits, et ce pour une diffusion lors de conférences ou tout autre évènement public lié à la présentation de cette expérience.

Je déclare avoir librement souhaité participer à cet enregistrement et n'émet aucune restriction ni réserve.

Fait le :

A :

____

Signature

Questionnaire

Code Sujet : ....
Date : ....

Age : ....
Homme/femme

Pratique musicale ....
....
....

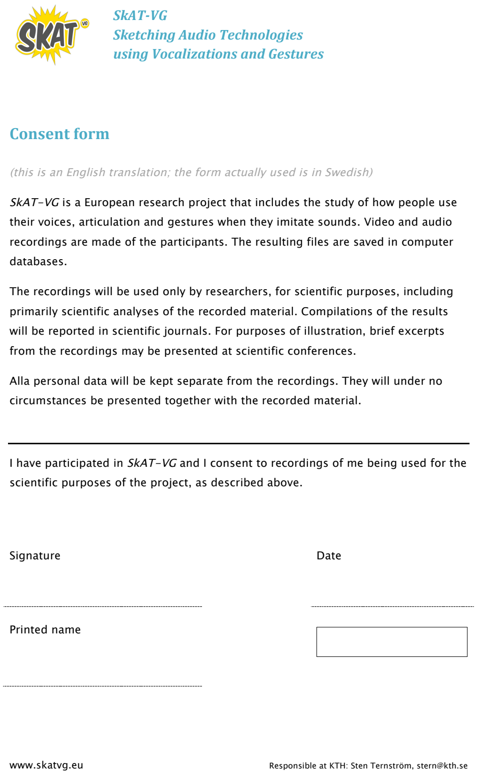
Pratique du son
....
....
....

Pratique de la danse
....
....
....

Pratique du théâtre
....
....
....

Commentaires sur l'expérience:
....
....
....
....
....
....
....
....
....
....

Figure 10: IRCAM Consent forms and questionnaire for subjects.



SkAT-VG
Sketching Audio Technologies
using Vocalizations and Gestures

Consent form

(this is an English translation; the form actually used is in Swedish)

SkAT-VG is a European research project that includes the study of how people use their voices, articulation and gestures when they imitate sounds. Video and audio recordings are made of the participants. The resulting files are saved in computer databases.

The recordings will be used only by researchers, for scientific purposes, including primarily scientific analyses of the recorded material. Compilations of the results will be reported in scientific journals. For purposes of illustration, brief excerpts from the recordings may be presented at scientific conferences.

Alla personal data will be kept separate from the recordings. They will under no circumstances be presented together with the recorded material.

I have participated in *SkAT-VG* and I consent to recordings of me being used for the scientific purposes of the project, as described above.

Signature ....

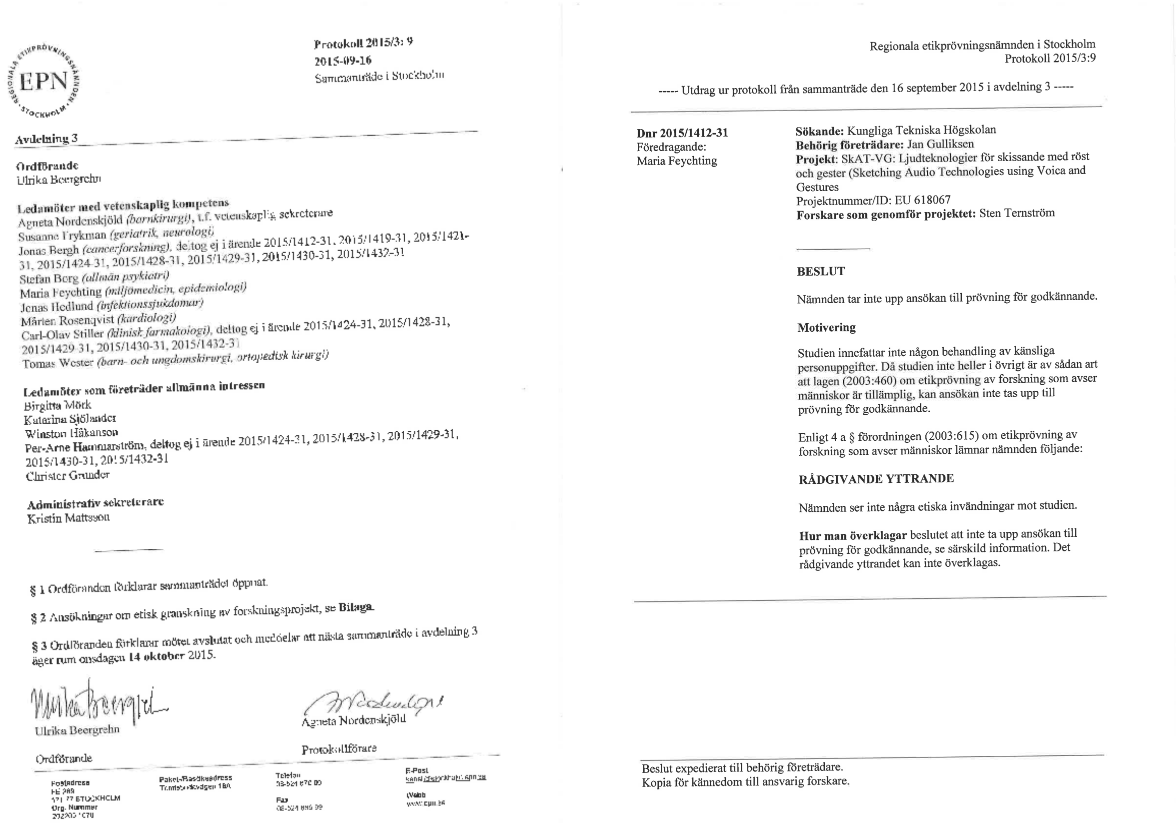
Date ....

Printed name ....

www.skatvg.eu

Responsible at KTH: Sten Ternström, stern@kth.se

Figure 11: KTH consent form.



EPN
REGIONALA ETIKPRÖVNINGSNÄMNDEN · STOCKHOLM ·

Protokoll 2015/3: 9
2015-09-16
Sammanträde i Stockholm

Avdelning 3

Ordförande
Ulrika Beergrehn

Ledamöter med vetenskaplig kompetens
Agneta Nordenskjöld *(barnkirurgi)*, t.f. vetenskaplig sekreterare
Susanne Frykman *(geriatrik, neurologi)*
Jonas Bergh *(cancerforskning)*, deltog ej i ärende 2015/1412-31, 2015/1419-31, 2015/1421-31, 2015/1424-31, 2015/1428-31, 2015/1429-31, 2015/1430-31, 2015/1432-31
Stefan Berg *(allmän psykiatri)*
Maria Feychting *(miljömedicin, epidemiologi)*
Jonas Hedlund *(infektionssjukdomar)*
Mårten Rosenqvist *(kardiologi)*
Carl-Olav Stiller *(klinisk farmakologi)*, deltog ej i ärende 2015/1424-31, 2015/1428-31, 2015/1429-31, 2015/1430-31, 2015/1432-31
Tomas Wester *(barn- och ungdomskirurgi, ortopedisk kirurgi)*

Ledamöter som företräder allmänna intressen
Birgitta Mörk
Katarina Sjölander
Winston Håkanson
Per-Arne Hammarström, deltog ej i ärende 2015/1424-31, 2015/1428-31, 2015/1429-31, 2015/1430-31, 2015/1432-31
Christer Grunder

Administrativ sekreterare
Kristin Mattsson

§ 1 Ordföranden förklarar sammanträdet öppnat.

§ 2 Ansökningar om etisk granskning av forskningsprojekt, se **Bilaga**.

§ 3 Ordföranden förklarar mötet avslutat och meddelar att nästa sammanträde i avdelning 3 äger rum onsdagen **14 oktober** 2015.

Ulrika Beergrehn
Ordförande

Agneta Nordenskjöld
Protokollförare

Postadress
FE 289
171 77 STOCKHOLM
Org. Nummer
[illegible]

Paket-/Besöksadress
Tomtebodavägen 18A

Telefon
08-524 870 00
Fax
08-524 886 99

E-Post
kansli@stockholm.epn.se
Webb
www.epn.se

Regionala etikprövningsnämnden i Stockholm
Protokoll 2015/3:9

----- Utdrag ur protokoll från sammanträde den 16 september 2015 i avdelning 3 -----

Dnr 2015/1412-31
Föredragande:
Maria Feychting

Sökande: Kungliga Tekniska Högskolan
Behörig företrädare: Jan Gulliksen
Projekt: SkAT-VG: Ljudteknologier för skissande med röst och gester (Sketching Audio Technologies using Voica and Gestures
Projektnummer/ID: EU 618067
Forskare som genomför projektet: Sten Ternström

BESLUT

Nämnden tar inte upp ansökan till prövning för godkännande.

Motivering

Studien innefattar inte någon behandling av känsliga personuppgifter. Då studien inte heller i övrigt är av sådan art att lagen (2003:460) om etikprövning av forskning som avser människor är tillämplig, kan ansökan inte tas upp till prövning för godkännande.

Enligt 4 a § förordningen (2003:615) om etikprövning av forskning som avser människor lämnar nämnden följande:

RÅDGIVANDE YTTRANDE

Nämnden ser inte några etiska invändningar mot studien.

Hur man överklagar beslutet att inte ta upp ansökan till prövning för godkännande, se särskild information. Det rådgivande yttrandet kan inte överklagas.

Beslut expedierat till behörig företrädare.
Kopia för kännedom till ansvarig forskare.

Figure 12: Reply from EPN-Stockholm

7 Relations with other projects

All the reported work in voice production, perception, and machine learning has been done in SkAT-VG, with no interaction with other sources of funding. The work on gestures has benefited from previous and ongoing work at IRCAM, such as the use of motion sensor to control sound. Nevertheless, the studies on vocal and gesture imitation and the analysis of the database have been completely done within SkAT-VG. In particular, a totally new approach was developed to analyse gesture in real-time based on wavelet. The analysis of the different multimodal strategies the listeners use to describe sound was performed entirely within SkAT-VG. The classification work of WP5 is being largely based on the findings of WP3 and WP4, well within the boundary of SkAT-VG, and brand new features have been derived for the project purposes. The classification task in SkAT-VG has led to the extension of an existing framework. Thanks to SkAT-VG funds the automatic classifiers now model the evolution in time of sounds, thus tackling a different, and open, scientific problem. The automatic prediction of articulatory classes, as developed at KTH, is based on the auditory receptive fields toolbox developed in a previous project. All articulatory models of the phonetic, myoelastic, and turbulent components including all used audio features are new and are being developed in Matlab in SkAT-VG. The work on sound models at IUAV is the continuation of over a decade of studies and developments on sound synthesis by physical modeling, partly funded by previous FET and NEST projects (SOb and CLOSED). The Sound Design Toolkit, as it is being distributed by SkAT-VG in fall 2015, contains some new sound models that are necessary to represent the sound categories emerging from WP3 and WP4: gases flowing against solid surfaces, through hollow cavities and across thin objects, liquid sounds by means of a stochastic population of bubbles, supersonic blasts and explosions, electric motors and combustion engines. The inherited sound models (mechanical interactions between solid objects) have been completely rewritten to comply with a new software architecture. The Sound Design Toolkit distributed by SkAT-VG is composed of a core framework entirely developed in ANSI C and a collection of wrappers for Max and PureData. The code is designed to be portable across different operating systems (Windows, Mac OS X, Linux), and the APIs exposed by the core framework allow the reuse of the synthesis algorithms in a wide variety of developing environments other than Max and Pd. The software SkAT-Studio is being totally developed with SkAT-VG, as well as the sound designers interviews.

SkAT-VG is developing knowledge, methods and tools that are partly derived from previous researches carried out by the Consortium. In particular, Table 10 summarizes the exploitation of technologies and data previously developed or collected in other projects. All the products, tools, and datasets used in SkAT-VG, mentioned in this Periodic Report and not mentioned in Table 10, have been entirely developed or collected in SkAT-VG.

As a side note, a detailed list of Background Included is provided in the Consortium Agreement, to provide access rights to Background made available to the Parties. It also provides factual information about the state of the art upon which the Consortium is building its research.

Partner	Object:	Related Project:
IUAV	Sound Design Toolkit	Started being developed in project IST-2000-25287 (The Sounding Object). Further developed in project FP6-NEST-PATH-29085 (CLOSED). SkAT-VG is providing an extension of the palette of models and a re-writing of some models: at least 30% extension of prior work.
IRCAM	Collection of everyday sound categories	Defined in projects FP6-NEST-PATH-29085 (CLOSED) and Sample Orchestrator (ANR France). Exploited in SkAT-VG as follows: Selecting a subset of the categories, populating these categories with exemplars, conducting identification experiment to select only the categories that are not confused and the best exemplars within each category. About 90% of the work is new in SkAT-VG.
IRCAM	Physical objects used in MIMES	The physical objects for the MIMES installation/demonstration were designed during a previous project (ANR Legos) to study sensori-motor learning. These interactive objects have been further developed, adding the voice component, and are being used in a completely different research context, with very different goals.
IRCAM	MuBu	The MuBu multi-buffer is a container for sound and motion data. It provides a structured memory for the real-time processing of recorded sound and action together with interfaces and operators as a set of complementary Max externals. The ensemble of MuBu externals for Max allows for sound synthesis such as granular, concatenative and additive synthesis and interactive machine learning using algorithms such as Knn, Gaussian Mixture Models and Hidden Markov Models. In SkAT-VG, some externals were added to perform real-time wavelet analysis of gestures, and improved interactive machine learning.
KTH	ELAN annotation procedures	ELAN is provided by The Language Archive project at the Max Planck Institute for Psycholinguistics, Nijmegen, NL. The annotation procedures have been developed in SkAT-VG.
Genesis	XTract	Largely developed within the projects FET-Open-255931 (UNLocX) and EU-FP7-233980 (BESST). SkAT-VG provided a 10% extension and allowed industrialization of the product.
Genesis	Active Sound Design	Largely developed on Genesis own funds. SkAT-VG allowed to improve the overall measure process.

Table 10: Relations with other projects.

References

[BLDB15] Stefano Baldan, Hélène Lachambre, Stefano Delle Monache, and Patrick Boussard. Physically informed car engine sound synthesis for virtual and augmented environments. In *Proceedings of the 2nd Workshop on Sonic Interactions in Virtual Environments - IEEE VR 2015*, Arles, France, 2015.

[CP15] Mark Cartwright and Bryan Pardo. Vocalsketch: vocally imitating audio concepts. In *Proceedings of CHI 2015*, Seoul, Republic of Korea, 2015.

[DBMR14] Stefano Delle Monache, Stefano Baldan, Davide Andrea Mauro, and Davide Rocchesso. A design exploration on the effectiveness of vocal imitations. In *Proc. of the Sound and Music Computing Conference*, Athens, Greece, September 2014.

[DPR10] Stefano Delle Monache, Pietro Polotti, and Davide 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, pages 1:1–1:7, New York, NY, USA, 2010. ACM.

[DRBM15] Stefano Delle Monache, Davide Rocchesso, Stefano Baldan, and Davide Andrea Mauro. Growing the practice of vocal sketching. In *Proceedings of the 21st International Conference on Auditory Display (ICAD 2015)*, pages 58 – 65, Graz, Austria, 2015.

[DRP14] Stefano Delle Monache, Davide Rocchesso, and Stefano Papetti. Sketch a scratch. In *Eight International Conference on Tangible, Embedded and Embodied Interaction*, TEI '14, 2014.

[Hel14] Pétur Helgason. Sound initiation and source types in human imitations of sounds. In *Proceedings FONETIK*, pages pp 83–89, Stockholm (Sweden), 09-11 June 2014.

[LEP+82] Norman J. Lass, Sandra K. Eastham, William C. Parrish, Kathleen A. Sherbick, and Dawn M. Ralph. Listener's identification of environnmental sounds. *Perceptual and Motor Skills*, 55:75–78, 1982.

[LEW+83] Norman J. Lass, Sandra K. Eastham, Tammie L. Wright, Audrey HR. Hinzman, Karen J. Mills, and Amy L. Hefferin. Listener's identification of human-imitated sounds. *Perceptual and Motor Skills*, 57:995–998, 1983.

[LF15a] Tony Lindeberg and Anders Friberg. Idealized computational models for auditory receptive fields. *PloS one*, 10(3), 2015.

[LF15b] Tony Lindeberg and Anders Friberg. Scale-space theory for auditory signals. In *Scale Space and Variational Methods in Computer Vision*, pages 3–15. Springer, 2015.

[LHE$^+$84] Norman J. Lass, Audrey HR. Hinzman, Sandra K. Eastham, Tammie L. Wright, Karen J. Mills, Bonita S. Bartlett, and Pamela A. Summers. Listener's discrimination of real and human-imitated sounds. *Perceptual and Motor Skills*, 58:453–454, 1984.

[LJH$^+$15] Guillaume Lemaitre, Ali Jabbari, Olivier Houix, Nicolas Misdariis, and Patrick Susini. Vocal imitations of basic auditory features. *The Journal of the Acoustical Society of America*, 137(4):2268–2268, 2015.

[LR14] Guillaume Lemaitre and Davide Rocchesso. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135(2):862–873, 2014.

[MP15] Enrico Marchetto and Geoffroy Peeters. A set of audio features for the morphological description of vocal imitations. In *Proceedings of the International Conference on Digital Audio Effects*, Trondheim, Norway, 2015.

[MR15] Davide Andrea Mauro and Davide Rocchesso. Analyzing and organizing the sonic space of vocal imitations. In *Proceedings of Audio Mostly*, Thessaloniki, Greece, october 2015. ACM.

[New04] Fred Newman. *MouthSounds: How to Whistle, Pop, Boing, and Honk... for All Occasions and Then Some*. Workman Publishing, 2004.

[RBD15] Davide Rocchesso, Stefano Baldan, and Stefano Delle Monache. Reverberation still in business: Thickening and propagating micro-textures in physics-based sound modeling. In *Proceedings of the International Conference on Digital Audio Effects*, Trondheim, Norway, 2015.

[RDA16] Davide Rocchesso, Stefano Delle Monache, and Mauro Davide A. miMic: The microphone as a pencil. In *Tenth International Conference on Tangible, Embedded and Embodied Interaction*, TEI '16, 2016.

[RDP15] Davide Rocchesso, Stefano Delle Monache, and Stefano Papetti. Multisensory texture exploration at the tip of the pen. *International Journal of Human-Computer Studies*, pages –, 2015.

[RM14] Davide Rocchesso and Davide Andrea Mauro. Self-organizing the space of vocal imitations. In *Proceedings of the XX CIM*, Rome, Italy, october 2014. AIMI.

[SDPD13] Clara Suied, Angélique Drémeau, Daniel Pressnitzer, and Laurent Daudet. Auditory sketches: sparse representations of sounds based on perceptual models. In Mitsuko Aramaki, Mathieu Barthet, Richard Kronland-Martinet, and Sølvi Ystad, editors, *From Sounds to Music and Emotions, 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers*, volume 7900 of *Lecture Notes in Computer Science*, pages 154–170. Springer, Berlin/Heidelberg, Germany, 2013.

[Zoe02] Udo Zoelzer, editor. *Dafx: Digital Audio Effects*. John Wiley & Sons, Inc., New York, NY, USA, 2002.