SEVENTH FRAMEWORK PROGRAMME

Grant agreement for: Collaborative project[*]

Annex I - "Description of Work"

Project acronym: SkAT-VG

Project full title: " Sketching Audio Technologies using Vocalizations and Gestures "

Grant agreement no: 618067

Version date: 2013-08-30

Table of Contents

Part A

Workplan Tables

A1:
Project summary

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per project

General information

Project title [3]	Sketching Audio Technologies using Vocalizations and Gestures		
Starting date [4]	01/01/2014		
Duration in months [5]	36		
Call (part) identifier [6]	FP7-ICT-2013-C		
Activity code(s) most relevant to your topic [7]	:		
Free keywords [8]		voice, sound analysis and classification, gesture analysis, sound synthesis, sonic interaction design	

Abstract [9]

Sketching is at the root of any design activity. In visual design, hand and pencil are still the primary tools used to produce a large variety of initial concepts in a very short time. However, in product and media design the sonic behavior of objects is also of primary importance, as sounds may afford seamless and aesthetically pleasing interactions. But how might one sketch the auditory aspects and sonic behavior of objects, in the early stages of the design process? Non-verbal sounds, more than speech, are naturally and spontaneously used in everyday life to describe and imitate sonic events, often accompanied by manual expressive gestures that complement, qualify, or emphasize them. The SkAT-VG project aims at enabling designers to use their voice and hands, directly, to sketch the auditory aspects of an object, thereby making it easier to exploit the functional and aesthetic possibilities of sound. The core of this framework is a system able to interpret users' intentions trough gestures and vocalizations, to select appropriate sound synthesis modules, and to enable iterative refinement and sharing, as it is commonly done with drawn sketches in the early stages of the design process. To reach its goal, the SkAT-VG project is based on an original mixture of complementary expertise: voice production, gesture analysis, cognitive psychology, machine learning, interaction design, and audio application development. The project tasks include case studies of how people naturally use vocalizations and gestures to communicate sounds, evaluation of current practices of sound designers, basic studies of sound identification trough vocalizations and gestural production, gesture analysis and machine learning, and development of the sketching tools.

A2:
List of Beneficiaries

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

List of Beneficiaries

No	Name	Short name	Country	Project entry month[10]	Project exit month
1	UNIVERSITA IUAV DI VENEZIA	IUAV	Italy	1	36
2	INSTITUT DE RECHERCHE ET DE COORDINATION ACOUSTIQUE MUSIQUE - IRCAM	IRCAM	France	1	36
3	KUNGLIGA TEKNISKA HOEGSKOLAN	KTH	Sweden	1	36
4	GENESIS S. A.	GENESIS	France	1	36

A3:
Budget Breakdown

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One Form per Project

Participant number in this project [11]	Participant short name	Fund. %[12]	Ind. costs[13]	Estimated eligible costs (whole duration of the project)					Requested EU contribution
				RTD / Innovation (A)	Demonstration (B)	Management (C)	Other (D)	Total A+B+C+D	
1	IUAV	75.0	T	712,640.00	0.00	103,240.00	0.00	815,880.00	637,720.00
2	IRCAM	75.0	A	998,802.00	0.00	40,113.00	0.00	1,038,915.00	789,214.00
3	KTH	75.0	T	785,920.00	0.00	52,040.00	0.00	837,960.00	641,480.00
4	GENESIS	75.0	T	433,868.00	0.00	41,676.00	0.00	475,544.00	367,077.00
Total				2,931,230.00	0.00	237,069.00	0.00	3,168,299.00	2,435,491.00

Note that the budget mentioned in this table is the total budget requested by the Beneficiary and associated Third Parties.

*** The following funding schemes are distinguished**

Collaborative Project (if a distinction is made in the call please state which type of Collaborative project is referred to: (i) Small of medium-scale focused research project, (ii) Large-scale integrating project, (iii) Project targeted to special groups such as SMEs and other smaller actors), Network of Excellence, Coordination Action, Support Action.

1. Project number

The project number has been assigned by the Commission as the unique identifier for your project, and it cannot be changed. The project number **should appear on each page of the grant agreement preparation documents** to prevent errors during its handling.

2. Project acronym

Use the project acronym as indicated in the submitted proposal. It cannot be changed, unless agreed during the negotiations. The same acronym **should appear on each page of the grant agreement preparation documents** to prevent errors during its handling.

3. Project title

Use the title (preferably no longer than 200 characters) as indicated in the submitted proposal. Minor corrections are possible if agreed during the preparation of the grant agreement.

4. Starting date

Unless a specific (fixed) starting date is duly justified and agreed upon during the preparation of the Grant Agreement, the project will start on the first day of the month following the entry info force of the Grant Agreement (NB : entry into force = signature by the Commission). Please note that if a fixed starting date is used, you will be required to provide a detailed justification on a separate note.

5. Duration

Insert the duration of the project in full months.

6. Call (part) identifier

The Call (part) identifier is the reference number given in the call or part of the call you were addressing, as indicated in the publication of the call in the Official Journal of the European Union. You have to use the identifier given by the Commission in the letter inviting to prepare the grant agreement.

7. Activity code

Select the activity code from the drop-down menu.

8. Free keywords

Use the free keywords from your original proposal; changes and additions are possible.

9. Abstract

10. The month at which the participant joined the consortium, month 1 marking the start date of the project, and all other start dates being relative to this start date.

11. The number allocated by the Consortium to the participant for this project.

12. Include the funding % for RTD/Innovation – either 50% or 75%

13. Indirect cost model
A: Actual Costs
S: Actual Costs Simplified Method
T: Transitional Flat rate
F :Flat Rate

Workplan
Tables

Project number

618067

Project title

SkAT-VG—Sketching Audio Technologies using Vocalizations and Gestures

Call (part) identifier

FP7-ICT-2013-C

Funding scheme

Collaborative project

WT1
List of work packages

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

LIST OF WORK PACKAGES (WP)

WP Number [53]	WP Title	Type of activity [54]	Lead beneficiary number [55]	Person-months [56]	Start month [57]	End month [58]
WP 1	Project coordination	MGT	1	22.00	1	36
WP 2	Case studies	RTD	3	28.00	1	12
WP 3	Phonetic listening and world-event representation	RTD	3	39.00	1	36
WP 4	Perception and cognition of vocalizations and expressive gestures	RTD	2	59.00	1	36
WP 5	Automatic imitation recognition	RTD	2	44.00	12	35
WP 6	Imitation-driven sound synthesis	RTD	1	45.00	15	31
WP 7	Sonic interaction design	RTD	4	43.00	3	36
			Total	280.00		

WT2:
List of Deliverables

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

List of Deliverables - to be submitted for review to EC

Deliverable Number [61]	Deliverable Title	WP number [53]	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D1.1.1	First Periodic Report	1	1	5.00	R	PP	12
D1.1.2	Second Periodic Report	1	1	8.00	R	PP	36
D1.1.3	Final Report	1	1	9.00	R	PU	36
D2.2.1	Explorative collection of imitated sounds	2	3	10.00	R	PU	4
D2.2.2	Extensive set of recorded imitations	2	3	18.00	R	PU	12
D3.3.1	Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives	3	3	10.00	R	PU	12
D3.3.2	Final comprehensive annotation of the database of imitations	3	3	20.00	R	PU	24
D3.3.3	Report on how non-vocal world events are represented phonetically in SkAT-VG	3	3	9.00	R	PU	36
D4.4.1	A large set of vocal and gestural imitations	4	2	30.00	R	PU	21
D4.4.2	An analysis of how vocal and gesture primitives are sequenced	4	2	29.00	R	PU	26
D5.5.1	Blind classifiers of imitations	5	2	12.00	R	PU	23
D5.5.2	Informed classifiers of imitations	5	2	12.00	R	PU	28

Deliverable Number [61]	Deliverable Title	WP number [53]	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D5.5.3	Integrated system that predicts the category of imitated sound sources	5	2	20.00	P	PU	31
D6.6.1	Automatic system for the generation of sound sketches	6	1	20.00	P	PU	24
D6.6.2	Front-end application for interactive sound prototyping	6	1	25.00	R	PU	30
D7.7.1	Interactive prototypes realized with the SkAT-VG tool	7	4	31.00	P	PU	36
D7.7.2	Applications of vocal sketching	7	4	12.00	R	PU	36
			Total	280.00			

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP1	Type of activity [54]	MGT
Work package title	Project coordination		
Start month	1		
End month	36		
Lead beneficiary number [55]	1		

Objectives

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.

Description of work and role of partners

WP1 [IUAV 12PM; IRCAM 3PM; KTH 3PM; GENESIS 4PM] 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.

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	12.00
2	IRCAM	3.00
3	KTH	3.00
4	GENESIS	4.00
	Total	22.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D1.1.1	First Periodic Report	1	5.00	R	PP	12
D1.1.2	Second Periodic Report	1	8.00	R	PP	36
D1.1.3	Final Report	1	9.00	R	PU	36
		Total	22.00			

Description of deliverables

D1.1.1) First Periodic Report: [month 12]

D1.1.2) Second Periodic Report: [month 36]

D1.1.3) Final Report: [month 36]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP2	Type of activity [54]	RTD
Work package title	Case studies		
Start month	1		
End month	12		
Lead beneficiary number [55]	3		

Objectives

To study existing instances of effective vocal and gestural imitations, and their execution. To define from these examples a list of action/sound primitives, and record them.

Description of work and role of partners

Just as there are bad, good, and virtuoso freehand sketchers, we expect to find similar degrees of abilities in vocal sketching. Examining the cases of virtuosos (e.g. Cathy Berberian, David Moss, or beatboxers in hip-hop performances) will be the starting point of WP2.

Task 2.1 [IUAV 1PM; IRCAM 1PM; KTH 6PM; GENESIS 2PM]: Case studies. 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 Task 2.1.

Task 2.2 [IUAV 3PM; IRCAM 1PM; KTH 10PM; GENESIS 4PM]: Set of imitations. 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.

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	4.00
2	IRCAM	2.00
3	KTH	16.00
4	GENESIS	6.00
	Total	28.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D2.2.1	Explorative collection of imitated sounds	3	10.00	R	PU	4
D2.2.2	Extensive set of recorded imitations	3	18.00	R	PU	12
		Total	28.00			

Description of deliverables

D2.2.1) Explorative collection of imitated sounds: Commercially available sound examples plus an initial set of newly recorded imitations, both categorised into classes of action primitives [month 4]

D2.2.2) Extensive set of recorded imitations: Newly recorded imitations of action primitives, covering all types of sounds that the SkAT-VG system will be able to handle. [month 12]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS1	Accumulation of a large enough database of recorded, sorted, and labeled imitations	3	12	Verified by documentation and scientific publications. It will allow the effective start of WP5.

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP3	Type of activity [54]	RTD
Work package title	Phonetic listening and world-event representation		
Start month	1		
End month	36		
Lead beneficiary number [55]	3		

Objectives

To annotate the database of imitations of interaction primitives with a time-continuous representation of phonatory and articulatory actions (vocal primitives). The database will form the input to the informed classification process to be developed in WP5. The vocal primitives will eventually by used to analyze the results of WP4.

Description of work and role of partners

Expertise in phonetics of spoken and sung voice, as expressed by the KTH team, will chart how the physical dynamics of the primitives obtained in WP2 are mimicked by using the human voice. This will be done in the paradigm of articulatory phonology. Manual gestures will be analyzed at IRCAM. The resulting estimations of articulatory data will be used to train the front-end application developed in WP5, and integrated with the results of WP4. If successful, the results of WP5 can be fed back to WP3 to speed up further manual classification work.

Task 3.1 [IUAV 1PM; IRCAM 1PM; KTH 6PM; GENESIS 0PM]: Editor. 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.

Task 3.2 [IUAV 1PM; IRCAM 1PM; KTH 20PM; GENESIS 0PM]: Primitives. 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. 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.

Task 3.3 [IUAV 1PM; IRCAM 2PM; KTH 6PM; GENESIS 0PM]: Linguistic aliases for non-vocal sounds. 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.

WT3:
Work package description

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	3.00
2	IRCAM	4.00
3	KTH	32.00
	Total	39.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D3.3.1	Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives	3	10.00	R	PU	12
D3.3.2	Final comprehensive annotation of the database of imitations	3	20.00	R	PU	24
D3.3.3	Report on how non-vocal world events are represented phonetically in SkAT-VG	3	9.00	R	PU	36
		Total	39.00			

Description of deliverables

D3.3.1) Preliminary annotation of the database of imitations of action primitives in terms of vocal primitives: [month 12]

D3.3.2) Final comprehensive annotation of the database of imitations: [month 24]

D3.3.3) Report on how non-vocal world events are represented phonetically in SkAT-VG: [month 36]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS1	Accumulation of a large enough database of recorded, sorted, and labeled imitations	3	12	Verified by documentation and scientific publications. It will allow the effective start of WP5.
MS2	Automatic classifiers of vocal and gestural imitations into cagetories of imitated sounds	2	24	First implementation, verified by prototype classifiers and scientific publications
MS3	Integrated sketching tools	1	36	Verified by prototype sketching tools,

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
				documentation, and scientific publications

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP4	Type of activity [54]	RTD
Work package title	Perception and cognition of vocalizations and expressive gestures		
Start month	1		
End month	36		
Lead beneficiary number [55]	2		

Objectives

To record and select a large set of successful imitations, and to analyze how vocal and gestural primitives are combined to create them. WP4 will proceed from the perspective of perception of sounds and imitations. Results will be integrated with those of WP3, and the other WPs will be given a large set of successful examples, and a description of the imitation mechanisms.

Description of work and role of partners

Task 4.1 [IUAV 1PM; IRCAM 12PM; KTH 0PM; GENESIS 2PM]: Perceptually relevant categories of sounds. 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 classes of sounds in terms of interaction, temporal and timbral properties. A strong interaction with IUAV will 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.

Task 4.2 [IUAV 1PM; IRCAM 20PM; KTH 1PM; GENESIS 2PM]: 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 IUAV in 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.

Task 4.3 [IUAV 1PM; IRCAM 16PM; KTH 1PM; GENESIS 2PM]: Analysis of the successful imitations. 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 Tasks 3.3 and 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 IUAV in Task 6.2 about the aspects that are important for the fine tuning of the timbral families.

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	3.00
2	IRCAM	48.00

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
3	KTH	2.00
4	GENESIS	6.00
	Total	59.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D4.4.1	A large set of vocal and gestural imitations	2	30.00	R	PU	21
D4.4.2	An analysis of how vocal and gesture primitives are sequenced	2	29.00	R	PU	26
		Total	59.00			

Description of deliverables

D4.4.1) A large set of vocal and gestural imitations: This set of imitations correspond to psychologically relevant categories of sounds [month 21]

D4.4.2) An analysis of how vocal and gesture primitives are sequenced: To understand how primitives are combined to create successful imitations. [month 26]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS1	Accumulation of a large enough database of recorded, sorted, and labeled imitations	3	12	Verified by documentation and scientific publications. It will allow the effective start of WP5.
MS2	Automatic classifiers of vocal and gestural imitations into cagetories of imitated sounds	2	24	First implementation, verified by prototype classifiers and scientific publications
MS3	Integrated sketching tools	1	36	Verified by prototype sketching tools, documentation, and scientific publications

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP5	Type of activity [54]	RTD
Work package title	Automatic imitation recognition		
Start month	12		
End month	35		
Lead beneficiary number [55]	2		

Objectives

To technically integrate the developments of WP3 and the results of WP4 into a system that will feed WP6. Transforming vocal and gestural imitations into synthetic sounds requires a system to automatically identify the category of imitated sound source. Hence such a system should fulfill three functions: estimating the acoustical and gestural features of the imitation and type of articulatory mechanisms, segmenting the signal into sequences of meaningful elements, and predicting the category of the imitated sound source.

Description of work and role of partners

Task 5.1 [IUAV 1PM; IRCAM 4PM; KTH 4PM; GENESIS 2PM]: Representation and Segmentation. 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.

Task 5.2 [IUAV 0PM; IRCAM 6PM; KTH 0PM; GENESIS 1PM]: Blind classifiers that predict the categories of imitated sound sources (from WP4) directly from low-level multimodal (sound and gesture) representations.

Task 5.3 [IUAV 0PM; IRCAM 6PM; KTH 2PM; GENESIS 1PM]: Informed 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.

Task 5.4 [IUAV 1PM; IRCAM 12PM; KTH 2PM; GENESIS 2PM]: 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).

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	2.00
2	IRCAM	28.00
3	KTH	8.00
4	GENESIS	6.00
	Total	44.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D5.5.1	Blind classifiers of imitations	2	12.00	R	PU	23
D5.5.2	Informed classifiers of imitations	2	12.00	R	PU	28
D5.5.3	Integrated system that predicts the category of imitated sound sources	2	20.00	P	PU	31
		Total	44.00			

Description of deliverables

D5.5.1) Blind classifiers of imitations: [month 23]

D5.5.2) Informed classifiers of imitations: [month 28]

D5.5.3) Integrated system that predicts the category of imitated sound sources: The system operates on vocal and gestural signals. [month 31]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS2	Automatic classifiers of vocal and gestural imitations into cagetories of imitated sounds	2	24	First implementation, verified by prototype classifiers and scientific publications
MS3	Integrated sketching tools	1	36	Verified by prototype sketching tools, documentation, and scientific publications

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP6	Type of activity [54]	RTD
Work package title	Imitation-driven sound synthesis		
Start month	15		
End month	31		
Lead beneficiary number [55]	1		

Objectives

To implement the sound synthesis tools able to simulate the sound sources represented by imitations. These tools will form the generative component of the SkAT-VG system, ensuing from the automatic classifier developed in WP5. Once selected by the classifier, the output synthetic sounds will be adaptively refinable as in a usual iterative design process. The classifier, the synthesis engine and the post processing section will be integrated in order to constitute altogether the complete SkAT-VG system.

Description of work and role of partners

Task 6.1 [IUAV 8PM; IRCAM 1PM; KTH 1PM; GENESIS 2PM]: Definition of the timbral families. 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.

Task 6.2 [IUAV 10PM; IRCAM 1PM; KTH 1PM; GENESIS 2PM]: Temporal behaviour and sound post processing. 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.

Task 6.3 [IUAV 12PM; IRCAM 2PM; KTH 1PM; GENESIS 4PM]: Fine parameterization of timbral subclasses, 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.

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	30.00
2	IRCAM	4.00
3	KTH	3.00
4	GENESIS	8.00
	Total	45.00

WT3:
Work package description

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D6.6.1	Automatic system for the generation of sound sketches	1	20.00	P	PU	24
D6.6.2	Front-end application for interactive sound prototyping	1	25.00	R	PU	30
		Total	45.00			

Description of deliverables

D6.6.1) Automatic system for the generation of sound sketches: The system is based on imitation categories [month 24]

D6.6.2) Front-end application for interactive sound prototyping: [month 30]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS2	Automatic classifiers of vocal and gestural imitations into cagetories of imitated sounds	2	24	First implementation, verified by prototype classifiers and scientific publications
MS3	Integrated sketching tools	1	36	Verified by prototype sketching tools, documentation, and scientific publications

WT3:
Work package description

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

One form per Work Package

Work package number [53]	WP7	Type of activity [54]	RTD
Work package title	Sonic interaction design		
Start month	3		
End month	36		
Lead beneficiary number [55]	4		

Objectives

To provide a reference framework based on sonic interaction design cases that will serve both as a guide and inspiration source for the basic research work packages, and as an assessment bench for all of the results of the project. For the latter point, the implementation of sonic interaction design mock-ups is envisaged as a necessary step, through which to test and assess the whole sketching strategy.

Description of work and role of partners

Task 7.1 [IUAV 6PM; IRCAM 1PM; KTH 1PM; GENESIS 6PM]: Definition of sonic interactive scenarios and applications. 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.

Task 7.2 [IUAV 7PM; IRCAM 1PM; KTH 1PM; GENESIS 8PM]: Realization of sonic interactive experiences, using the SkAT-VG tool. 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.

Task 7.3 [IUAV 2PM; IRCAM 2PM; KTH 2PM; GENESIS 6PM]: Side applications of vocal and gestural sketching. 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.

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
1	IUAV	15.00
2	IRCAM	4.00
3	KTH	4.00
4	GENESIS	20.00

Person-Months per Participant

Participant number [10]	Participant short name [11]	Person-months per participant
	Total	43.00

List of deliverables

Deliverable Number [61]	Deliverable Title	Lead beneficiary number	Estimated indicative person-months	Nature [62]	Dissemination level [63]	Delivery date [64]
D7.7.1	Interactive prototypes realized with the SkAT-VG tool	4	31.00	P	PU	36
D7.7.2	Applications of vocal sketching	4	12.00	R	PU	36
		Total	43.00			

Description of deliverables

D7.7.1) Interactive prototypes realized with the SkAT-VG tool: [month 36]

D7.7.2) Applications of vocal sketching: [month 36]

Schedule of relevant Milestones

Milestone number [59]	Milestone name	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS3	Integrated sketching tools	1	36	Verified by prototype sketching tools, documentation, and scientific publications

WT4:
List of Milestones

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

List and Schedule of Milestones

Milestone number [59]	Milestone name	WP number [53]	Lead beneficiary number	Delivery date from Annex I [60]	Comments
MS1	Accumulation of a large enough database of recorded, sorted, and labeled imitations	WP2, WP3, WP4	3	12	Verified by documentation and scientific publications. It will allow the effective start of WP5.
MS2	Automatic classifiers of vocal and gestural imitations into cagetories of imitated sounds	WP3, WP4, WP5, WP6	2	24	First implementation, verified by prototype classifiers and scientific publications
MS3	Integrated sketching tools	WP3, WP4, WP5, WP6, WP7	1	36	Verified by prototype sketching tools, documentation, and scientific publications

WT5:
Tentative schedule of Project Reviews

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

Tentative schedule of Project Reviews

Review number [65]	Tentative timing	Planned venue of review	Comments, if any
RV 1	13	Paris	
RV 2	36	Venice	

WT6:
Project Effort by Beneficiary and Work Package

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

Indicative efforts (man-months) per Beneficiary per Work Package

Beneficiary number and short-name	WP 1	WP 2	WP 3	WP 4	WP 5	WP 6	WP 7	Total per Beneficiary
1 - IUAV	12.00	4.00	3.00	3.00	2.00	30.00	15.00	69.00
2 - IRCAM	3.00	2.00	4.00	48.00	28.00	4.00	4.00	93.00
3 - KTH	3.00	16.00	32.00	2.00	8.00	3.00	4.00	68.00
4 - GENESIS	4.00	6.00	0.00	6.00	6.00	8.00	20.00	50.00
Total	22.00	28.00	39.00	59.00	44.00	45.00	43.00	280.00

WT7:
Project Effort by Activity type per Beneficiary

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

Indicative efforts per Activity Type per Beneficiary

Activity type	Part. 1 IUAV	Part. 2 IRCAM	Part. 3 KTH	Part. 4 GENESIS	Total
1. RTD/Innovation activities					
WP 2	4.00	2.00	16.00	6.00	28.00
WP 3	3.00	4.00	32.00	0.00	39.00
WP 4	3.00	48.00	2.00	6.00	59.00
WP 5	2.00	28.00	8.00	6.00	44.00
WP 6	30.00	4.00	3.00	8.00	45.00
WP 7	15.00	4.00	4.00	20.00	43.00
Total Research	57.00	90.00	65.00	46.00	258.00
2. Demonstration activities					
Total Demo	0.00	0.00	0.00	0.00	0.00
3. Consortium Management activities					
WP 1	12.00	3.00	3.00	4.00	22.00
Total Management	12.00	3.00	3.00	4.00	22.00
4. Other activities					
Total other	0.00	0.00	0.00	0.00	0.00
Total	69.00	93.00	68.00	50.00	280.00

WT8:
Project Effort and costs

Project Number [1]	618067	Project Acronym [2]	SkAT-VG

Project efforts and costs

Beneficiary number	Beneficiary short name	Estimated eligible costs (whole duration of the project)						Requested EU contribution (€)
		Effort (PM)	Personnel costs (€)	Subcontracting (€)	Other Direct costs (€)	Indirect costs OR lump sum, flat-rate or scale-of-unit (€)	Total costs	
1	IUAV	69.00	394,800.00	5,000.00	112,000.00	304,080.00	815,880.00	637,720.00
2	IRCAM	93.00	487,753.00	4,000.00	87,200.00	459,962.00	1,038,915.00	789,214.00
3	KTH	68.00	445,600.00	5,000.00	75,000.00	312,360.00	837,960.00	641,480.00
4	GENESIS	50.00	264,550.00	0.00	32,666.00	178,328.00	475,544.00	367,077.00
Total		280.00	1,592,703.00	14,000.00	306,866.00	1,254,730.00	3,168,299.00	2,435,491.00

1. Project number

The project number has been assigned by the Commission as the unique identifier for your project. It cannot be changed. The project number **should appear on each page of the grant agreement preparation documents (part A and part B)** to prevent errors during its handling.

2. Project acronym

Use the project acronym as given in the submitted proposal. It cannot be changed unless agreed so during the negotiations. The same acronym **should appear on each page of the grant agreement preparation documents (part A and part B)** to prevent errors during its handling.

53. Work Package number

Work package number: WP1, WP2, WP3, ..., WPn

54. Type of activity

For all FP7 projects each work package must relate to one (and only one) of the following possible types of activity (only if applicable for the chosen funding scheme – must correspond to the GPF Form Ax.v):

• **RTD/INNO =** Research and technological development including scientific coordination - applicable for Collaborative Projects and Networks of Excellence

• **DEM =** Demonstration - applicable for collaborative projects and Research for the Benefit of Specific Groups

• **MGT =** Management of the consortium - applicable for all funding schemes

• **OTHER =** Other specific activities, applicable for all funding schemes

• **COORD =** Coordination activities – applicable only for CAs

• **SUPP =** Support activities – applicable only for SAs

55. Lead beneficiary number

Number of the beneficiary leading the work in this work package.

56. Person-months per work package

The total number of person-months allocated to each work package.

57. Start month

Relative start date for the work in the specific work packages, month 1 marking the start date of the project, and all other start dates being relative to this start date.

58. End month

Relative end date, month 1 marking the start date of the project, and all end dates being relative to this start date.

59. Milestone number

Milestone number:MS1, MS2, …, MSn

60. Delivery date for Milestone

Month in which the milestone will be achieved. Month 1 marking the start date of the project, and all delivery dates being relative to this start date.

61. Deliverable number

Deliverable numbers in order of delivery dates: D1 – Dn

62. Nature

Please indicate the nature of the deliverable using one of the following codes

R = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

63. Dissemination level

Please indicate the dissemination level using one of the following codes:

• **PU =** Public

• **PP =** Restricted to other programme participants (including the Commission Services)

• **RE =** Restricted to a group specified by the consortium (including the Commission Services)

• **CO =** Confidential, only for members of the consortium (including the Commission Services)

- **Restreint UE =** Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments

- **Confidentiel UE =** Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments

- **Secret UE =** Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

64. Delivery date for Deliverable

Month in which the deliverables will be available. Month 1 marking the start date of the project, and all delivery dates being relative to this start date

65. Review number

Review number: RV1, RV2, ..., RVn

66. Tentative timing of reviews

Month after which the review will take place. Month 1 marking the start date of the project, and all delivery dates being relative to this start date.

67. Person-months per Deliverable

The total number of person-month allocated to each deliverable.

PART B
COLLABORATIVE PROJECT

Contents

1 Concept and objectives, progress beyond state-of-the-art, S/T methodology and work plan

1.1 Concept and project objectives

Humans live immersed in a sonic world. Most of the objects or other animals we interact with have sonic manifestations that we naturally exploit for monitoring purposes, or to guide our actions. Even though acoustic pollution is a problem in modern societies, we can not reasonably think of living in a silent world. Conversely, we should aim at **sounds of objects** that are functional and aesthetically pleasant, and we should give designers the tools and knowledge they need to reach this goal. For a long time, sound in industrial design has been mainly seen as unwanted noises, a sort of side effect of the design process. This is not true anymore. For instance, the automotive industry is already investing a lot of energy to carefully design every sound produced by a car (engine noise, alarms, etc.). But the tools for fast sound prototyping are still missing and industrial sound design is still a daunting process. The SkAT-VG project aims at facilitating the constructive use of sound by developing **sketching tools** that can be used to include sound even in the early stage of the design process, when ideas are rapidly produced in the form of numerous sketches that are used to spur design thinking and discussions with stakeholders. If visual sketching is mainly done by hand and pencil, sonic sketching is naturally done by **voice and gesture**. However, the ephemerality of utterances and gestures makes it challenging to transform them into manipulable **sound models**, and only a collaborative effort between phoneticians, psychoacousticians, computer scientists, and designers can address this goal.

1.1.1 Long-term vision

Sketching is at the roots of any design activity. Sketches are the materials that are passed through the early loops of the design process, as through them the designer produces, verifies, selects, communicates, and refines ideas. In this context, hand and pencil are still the primary tools used by designers to produce a large variety of initial concepts in a very short time. The community of graphic designers has been paying attention to the cognitive and creative processes involved in sketching activities, and has developed tools to facilitate them.

Design is however embracing new areas where non-graphical aspects become important. Specifically, within the wide area of interaction design there is a growing interest in interaction mediated by sound. Examples include mobile devices, video-games, social networks, etc. Furthermore, technological developments are rapidly populating our everyday environments with networks of interconnected objects[1] dressed with technological capabilities (computing, sensing, and communicating), which seriously raises the issue of communication between users and ubiquitous “machines” devoid of traditional interfaces (displays and keyboards). In product design, designers and theorists are becoming aware that the sonic manifestations of objects can afford natural, powerful, and useful interactions, and participate to the aesthetic appraisal of a product. The transportation industry, in particular, is deeply concerned with the sound quality of the vehicles. The most notable example is probably that of the sounds of quiet vehicles (electric and hybrid), which addresses the issues of pedestrian safety, product aesthetics, and brand image[2].

In comparison to graphic design, however, the problem of sketching, and more generally of communicating the auditory aspects and sonic behavior of objects, has not been seriously addressed yet. Tools that enable intuition, expressiveness, and cooperation are still missing in sound design. In particular, there is no single analogue to hand and pencil, but rather a plethora of tools and methods that can be used to rapidly

[1] This potentially concerns 50-70 billion ‘machines’
http://ec.europa.eu/information_society/policy/rfid/documents/commiot2009.pdf

[2] The forecast of worlwide sales of electric vehicles is 3.8 million annually by 2020
http://www.pikeresearch.com/newsroom/worldwide-electric-vehicle-sales-to-reach-3-8-million-annually-by-2020

obtain a variety of pre-prototypes that are expressive enough to allow communication, testing, and development of ideas. Notably, sketching in interaction design is an open problem as these tools and methods are not nearly as effective as hand and pencil. As regards the sounds of industrial products, the common practice of forgetting sound until the final stage of the design process, where sound designers or acousticians are often asked to fix acoustic misbehaviors or to dress up prototypes with sonic widgets, is deleterious for the quality of objects, and contributes to the increase of the sonic clutter of our everyday environments.

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. Non-verbal vocalizations and manual gestures are easily the most natural analogues to hand and pencil, having the innate capacity of conveying straightforward information on several attributes of a given sound. Intermixed with ordinary verbal items, vocalizations are used to communicate emotional states, to integrate the rhetoric of a sentence, or to imitate non-human sounds. Expressive gestures are very often used to complement these vocalizations, qualify or emphasize certain aspects. Sound representation by vocal and gestural imitation is particularly close to the idea of sketching. For instance, if a person were asked to describe how her house looks like, she would probably sketch it on paper. Similarly, if a motor biker were asked to describe how his bike behaves, he would probably use a number of vocal imitations and expressive gestures, where hands would be used to emphasize, supplement, or substitute part of the information conveyed by the voice.

The goal of SkAT-VG is to develop technologies that use a designer's vocal or/and manual gestures to create a synthetic sound that can actually be reproduced, exchanged between collaborators, edited and refined. The vision is that technologies that enable fast and intuive 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.

1.1.2 Breakthroughs

Incorporating vocal and gestural sketching in the design process will allow the designers to exploit the functional and aesthetic possibilities of sound. Hence, designers will be able to use their voice and hands to sketch sonic behaviors, with just the same naturalness, efficiency, and expressiveness they use while drawing on paper.

Three breakthroughs are necessary to achieve this goal. First, the SkAT-VG project will need to understand how non-verbal vocalizations and manual gestures are related to sound perception. SkAT-VG starts from several hypotheses that will be assessed during the project. These hypotheses are that listeners identify the mechanical events at the source of everyday sounds; That these different sound events are communicated through articulatory gestures recognized by the listeners; And that the fine temporal evolution of the sounds, as well as expressive intentions of the producer, are highlighted by manual gestures. Therefore, major new insights as regards sound perception and identification are expected. Studying vocal imitations and imitative gestures of sounds is expected to increase our understanding of how the human cognition deals with identification of the auditory environment, now considered as one of the most important functions of audition, because they allow identification through simplification and exaggeration of features. Studying the articulatory phonology of non-verbal vocalizations is expected to produce important new results in phonology, since this type of signal has almost never been studied. The original approach enabled by the SkAT-VG consortium, consisting in two perspectives - phonetic production and sound perception - will facilitate this objective.

The second breakthrough will result from the development of tools that convert the designer's vocalisations and gestures into usable sound design materials. For example, if a designer says "I want this electric car to make a wooeeeeesh sound when I turn it on" and rapidly raises up her hands, it should be possible to convert the "wooooeeeesh" into a synthetic sound with a rapid pitch sweep that can be actually reproduced, modified, and refined, until it is actually ready to be included into a prototype. This task is only apparently simple, as it implies the understanding of how humans produce, perceive and make sense of non-verbal vocalizations and manual gestures. In fact, the scientific advances introduces by the SkAT-VG project will also have an impact on the field of machine learning.

The final breakthrough will come from the development of intuitive tools, designed to enlarge the range of action of sound designers. These sets of tools will allow designers to use their voice and hands to create and iteratively refine sound ideas. Such a goal requires several challenging developments. For instance, how to turn a well-defined sound image into a flexible synthesis model is an open problem that, if solved, will pave the road to many applications in a variety of fields. Another challenging example is the development of intuitive manipulation tools (e.g. using continuous gestures to control the temporal evolution of synthesis parameters).

In sum, the major breakthrough that SkAT-VG is targeting to achieve is to give the human expressiveness the power to affect the world: Voice and gestures not only as a communication means but actually as the most powerful way to determine and shape the sonic manifestations of future interactive objects.

1.1.3 Concrete Objectives

The achievement of the breakthroughs will be measured by the success of three concrete objectives:

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.

Add timing: by which date/milestone the objectives will be achieved

Objective 1 – [Milestone 1, month 12] This is a prerequisite for the other objectives, as current knowledge in perception and production of vocal imitations and expressive gestures is far from complete. The original approach of SkAT-VG will address this issue from two parallel perspectives: by studying what humans can perceive from the imitations, and how humans actually produce the imitations. The first perspective will highlight categories of sounding events that are psychologically relevant for a listener. The second perspective will reveal the vocal and gestural features that allow a listener to recover what the imitation intends to convey. The results of these two perspectives will be integrated to identify the correspondences between produced vocal-gestural features and perceived sound sources. Moving the focus from signals to sources will give more compact representations of segments of imitations, and will ease the task of organizing a re-synthesis using sound models.

Objective 2 – [Milestone 2, month 24] Once the main categories of perceived/produced imitations and their relationships to the perceived categories of sound sources are known from objective 1, machine learning techniques can be used to recover the categories of perceived/produced imitations from the acoustic (for vocalizations) and kinetic signals (for gestures). However, such "blind" approach can be extended and improved by exploiting another outcome of objective 1, i.e., the characterization of vocal productions based on the study of imitations. The result will be an "informed" classifier.

Objective 3 – [Milestone 3, month 36] In a sonic interaction design context, the designer will develop vocal and gestural sketching skills through practice, similarly to how manual dexterity for drawing naturally develops. In a design scenario, the designer imitates a certain sound-producing phenomenon, the system classifies the imitation and returns a plausible sound model that can be used to resynthesize such imitation. The parameters of the sound model can then be vocally or manually tweaked until a satisfactory sonic behavior of the designed object is obtained. The third objective is therefore to use the results of the two previous objectives to develop the design tools that enable sound creation and sound manipulation with the voice and the hands. An evaluation of how these tools improve and facilitate sound creation and manipulation will be conducted.

These three objectives define the three Milestones that will help monitoring the progress of the project. The details about the implementation of the strategy of the SkAT-VG project, and about how the planned tasks will contribute to reach the Milestones, are given in Section 1.3.

1.2 Progress beyond the state of the art

The SkAT-VG project aims at exploiting, in the context of interaction and sound design, the human capabilities of producing a wide range of utterances, with special attention to imitations of non-animal sounds. Hand gestures will be also used as an additional input channel. Looking at the portions of the human brain responsible for sensory-motor information of the rest of the body it is evident that the vocal production apparatus is at least as important as the hand. Moreover, anatomical connectivity at the level of brain cortex shows that auditory perception and vocal production are tightly linked [55]. Recent neuropsychological results have highlighted that listening to speech sounds activates somatotopic representations (activations of the speech motor area) [68, 21, 16], providing support for the motor theory of speech perception [46], and even for the somewhat controversial idea that speech has originated from the imitation of facial and manual gestures [14].

SkAT-VG has the ambitious goal of giving the designer a flexible tool that exploits human vocal production and manual gestures for sketching the sonic behavior of objects, with the same naturalness and expressiveness as a hand-driven pencil on paper.

So far, no European funded initiatives have been dedicated to vocal imitations, although some projects did focus on non-verbal aspects of human speech. For example, the FP6-IST ACORNS project aimed at developing a conversational agent that can learn emergent patterns of speech and non-speech sounds. Some efforts were devoted to the study of human vocal production, especially for the purpose of speech recognition. The ESPRIT projects ACCOR and ACCOR2 were motivated by the observation that a major barrier to robust speaker-independent systems is the great variability in the relationship between the acoustic level of representation and the phonological structure of a given utterance. The projects exploited this variability as a productive source of information by modeling the underlying physiological and linguistic constraints (coarticulation). The ESPRIT 3 project SPEECH MAPS was concerned with sound-to-gesture inversion in speech. A robotics-based approach to action and perception was used. The recently-launched FET-Open project EUNISON aims at building a complete voice simulator that is based on physical first principles. This will be important for understanding how different kinds of utterances may be actually produced.

In an artistic context, the use of the voice to control sound synthesis in networked music performances was pioneered by Max Neuhaus and his Auracle project (http://www.auracle.org). In Paul Elliman's work the voice has been used to imitate the sounds of the city, including emergency vehicle sirens, radio transmissions and the acoustics of architectural space.

Among the initiatives on recognition of acoustic sources, we mention the MAT project MADRAS, that proposed a new generation of instruments that can automatically identify and quantify, in real time, the various acoustic sources that make up a given acoustic environment. Other projects may be found that insert everyday sound recognition as a component of a surveillance system.

European projects about sketching, even in the visual domain, have been rare. IST SMARTSKETCHES proposed computer-based tools in the initial stages of product design and development. It stemmed from the observation that CAD systems enable us to manufacture very complicated models, but at the early stages of product design pencil and paper still reign. IST projects T'ND and SATIN introduced haptic interfaces for direct manipulation of early prototypes, and these were extended with auditory feedback.

The use of sound as a mediator of meaningful and expressive interactions with everyday objects was explored by the FP5 Sounding Object (SOb) project, as part of the Disappearing Computer FET Proactive Initiative. In SOb, a library of physics-based sound models was developed and demonstrated in tasks of human-object continuous interaction. That project highlighted the need of a structured design approach to product sounds, where design iterations are guided by stages of functional- aesthetic evaluation. This necessity led to the project CLOSED (Closing the Loop of Sound Evaluation and Design) which was part of the FP6-NEST Measuring-the-Impossible initiative.

Sound design for interactive products was indicated in the Roadmap for Sound and Music Computing [67] - delivered by the Coordination Action $S2S^2$ - as one of the key directions where research in sound and music computing should be aiming. An international effort to coordinate research in this area, now called Sonic Interaction Design (SID), was the COST IC-0601 SID Action, chaired at IUAV. As part of SID, the first workshop on Sketching Sonic Interaction Design, especially focused on vocal sketching, was held in 2009 [20]. Before SID, the COST ICT-287 ConGAS Action dealt with the role of gestures in audio systems, including sound synthesis and processing, and helped establishing a European leadership in this research area.

1.2.1 State of the Art

Vocal imitations – Vocal imitations of sound sources occur naturally and spontaneously in conversations. Onomatopoeias can be in some cases better identified than the sounds they imitate [37], but speakers often resort to "wild" non-verbal *vocal imitations* when no standardised word is available. These imitations are efficient, for they allow the interlocutor to identify what has been heard. Several studies have stressed the existence of relationships between the phonetic properties of the onomatopoeias and the acoustical properties of the imitated sounds [75, 66], or between the acoustical properties of the imitations and the acoustical properties of the imitated sounds [54, 59]. In Japan, Ishihara and coworkers made a system to reduce everyday or natural sounds to phonemes and, in turn, to sound-imitation words [31], thus going into the opposite direction as to the objectives of the SkAT-VG project. Very little is known about the processes that allow an imitation to trigger the identification of a sound source. There is also a growing body of research in "mirror-neuron mechanisms" in relation to how we interpret speech sounds. This suggests that one interprets vocal sounds in terms of how one might have produced them oneself [71]. The novel perspective of SkAT-VG is to exploit this hard-wired mechanism across to non-vocal sounds.

Perception and identification of sound sources – The SkAT-VG project relies on a fundamental idea: that vocal and gestural imitations of sounds convey relevant information about the sounds that users have in mind. Many publications have reported the listeners' ability to recover auditory properties of events causing sound. Some were related to the objects causing the sound: dimensions [12, 36, 30], shape [34, 27], materials [33, 24]. Others were related to the actions, like bouncing, breaking [70], or pouring [9]. One important question raised by these results is thus to identify the acoustic information used (or needed) by the listeners to recover these properties [51, 49, 48]. However, the relationships between signal properties and information about the source properties appear sometimes to be complex [23], involving the context and previous knowledge of the listeners, and stereotypical (arbitrary) associations [58, 45]. Several studies have also suggested that the processing of meaningful sounds and language share common neural networks, using cross-form semantic priming [69, 53, 17, 1], brain imagery [38, 15] and clinical observations [62].

Inferring the voice production from the acoustic waveform – This is a classical problem called "acoustic-to-articulatory inversion". Although this can be done for a limited number of speech sounds, the general problem is known to be ill-posed. Machine learning has been used in the audio field for several decades essentially in the speech domain for the development of speech-to-text or speaker-identification applications. Its use in other audio domains started in the 1990s, with studies on instrument class recognition [50, 29], audio stream segmentation [63], sound-effect recognition and more recently for tasks related to music description (automatic genre, mood recognition) [74, 57]. Researches in this field focus on the two complementary parts of every audio indexing system: (i) improving the extraction of pertinent information from the audio signal (the audio features) [56] and (ii) improving the modelling of the class (machine learning). Many machine learning methods have been used in the audio field: Decision Trees, Gaussian Mixture Models, Support Vector Machines, Hidden Markov Models, etc. The choice of a specific method depends on the dimensionality of the feature space, the size of the training set, the static/dynamic characteristics of the data, the required performances, the available training time and the required complexity of the models. When various classifiers are assigned to various parts of the problem, or are in competition to solve the same problem (Ada-Boost), late-integration and fusion techniques must be used.

Gesture analysis and following – In the context of embodied music cognition [44], the general belief is that there is an intimate relationship between sound perception and human motion: Sound-accompanying movements partially parse temporal sonorous features. For example, Godøy et al. [25] studied the "traces" that participants performed on a two-dimensional surface in response to sound stimuli, and Nymoen et al. [52] investigated the relationships between sounds and hand gestures. Most of the techniques to recognize gestures have been designed for simple triggering tasks in human-computer interaction. The most widely used technique is Dynamic Time Warping (DTW), that requires the storage of the whole gesture temporal structure [47]. In general these techniques do not provide an explicit noise model, and do not prevent from errors due to unexpected or lost observations in the incoming sequence. Statistical methods, such as Hidden Markov Models (HMM) are generally robust since they rely on learning procedures based on large databases, modeling the variations occurring within a gesture class [7]. Several extensions of HMM have been proposed. Wilson and Bobick [72] proposed a model that takes into account parametric changes in execution. Even though they described an online learning method that can be applied to each different user [73], this method does not allow for the continuous adaptation of gesture classes.

Sketching for sonic interaction design – In design and architecture, the inherent complexity and ambiguity of freehand sketches is widely appreciated in the creative, explorative stage of problem solving [26]. They work both as communication devices, triggers of new ideas, and facilitators of iterative transformations [32, 65]. The main advocate of sketches as crucial in-embryo manifestations of interactive products is certainly W. Buxton [8]. In interaction design, several researchers have recently proposed methods and tools to facilitate the early stage of the process of designing interactive artefacts. However, none of these tools is used with the same immediacy and naturalness as the pencil is used in drawing by hand. Despite evident progress in the area, interaction remains difficult to sketch. Moreover, most researchers overlooked the fact that many artefacts have auditory manifestations, that these manifestations are crucial to many interactions, and that humans are naturally equipped with an acoustic sketching apparatus, namely the articulatory phonetic system, or vocal tract.

1.2.2 Baseline of the project

The SkAT-VG project starts from recent work by partners in the different subfields involved. This consortium "baseline" is here outlined:

Vocal imitations – When no proper vocabulary is available to describe the sounds, vocal imitations facilitate the communication of an acoustic experience [19]. It has been recently shown that, even when there are appropriate words to describe a sound, vocal imitations are at least as effective as verbalizations [43]. Vocal imitations can therefore be thought of as a convenient device used by a speaker to highlight the properties of a sound, thus allowing identification of the source. Vocal imitations are hybrid linguistic devices, arbitrary and dependent on a linguistic system, yet related to the perceived acoustic properties of sounds. Some recent experiments [2, 19] suggest that the phonetic properties of the imitations of identifiable everyday sounds are more related to the mechanism causing the sounds than to the sound signal properties. Both in speech and in other vocalizations, there is a wide variability in the waveforms produced by persons who differ in size, age, or gender, even if the message is consistently received by all listeners. So, some form of normalization across the internal characteristics of the voice production system (e.g., vocal-tract length, glottal pulse rate, etc.), must occur in the auditory system [55]. This is consistent with both the motor theory of speech perception [46], and with recent insights into the causal perception of sound sources [42].

Perception and identification of sound sources – Recent results have shown that listeners of vocal imitations of sound sources can recover the type of sound source (mostly the action) that is being imitated [39]. They actually outperformed listeners listening to verbal descriptions of a large variety of sounds when required to recognize the sounds [43]. It was also shown that different speakers use similar articulatory mechanisms when required to imitate sources of the same kind and are also able to closely mimic some acoustical properties (fundamental frequency, spectral centroid) [3, 19]. In neuroscience, recent results showed that hearing a sound made by a manual action can actually prime the execution of that action [13, 64] and suggested that auditory perception may be more effective to recover actions' than objects' properties [40] and that the most accessible information is the gesture that created the sound [41].

Inferring the voice production from the acoustic waveform – A detailed articulatory synthesis model could be used to back-track to the target sound, but to cover the practically unconstrained range of all possible vocal imitations would require a model of higher sophistication than any that is currently available. The realization of a model that begins to approach such a level of complexity is a main goal of the currently-running FET-Open EUNISON project, coordinated by SkAT-VG partner KTH.

Gesture analysis and following – Studies of gestural embodiment of environmental sounds [11] showed that listeners choose different strategies to mimick a sound, depending whether they can recognize the sound source or not. A recent method proposed by Bevilacqua et al. [5, 6] at IRCAM, called *Gesture Follower* makes use of the HMM statistical framework, but with an approach that differs from standard implementations. With the aim of estimating the progression of a gesture in real time [4] the system uses a concurrent forward procedure on several template gestures, allowing for the estimation of the likelihood related to each template. The effectiveness of adaptive recognition for the continuous manipulation of media has been recently demonstrated [10], where the gesture is used to select an effect, and variations are used to modulate the effect.

Sketching for sonic interaction design – The repertoire of audio sensors, actuators, processing and synthesis techniques that are widely accessible nowadays make sonic interaction design the preferred framework for experimental design of continuous interaction in embodied interfaces, as demonstrated in recent activities coordinated at IUAV [61, 60, 22].

1.2.3 Advances and Performance/research indicators

The SkAT-VG project proposes to expand the baseline work on vocal imitations and perception/identification of sound sources via an original mixture of experimental psychology and phonetics.Although the SkAT-VG project will take advantage of the EUNISON voice simulator when it will become available, its main methods will be based on a combination of human phonetic expertise and machine-learning techniques.

In the framework of the SkAT-VG project we propose to compare systems based on low-level and on high-level features (including new phonetic representations) and to study fusion techniques of both systems. Instead of directly trying to solve "the inversion problem" algorithmically, the SkAT-VG strategy for describing the vocal production of imitations is inspired by articulatory phonology [35]. Here, the production mechanism for any given utterance is described in terms of a time-varying linear combination of phonetic gestures, combining (a) the source production mechanisms (e.g., phonation, frication, tremulants and plosives); and (b) the filter properties (i.e., location(s) of the source(s) and vocal tract shape, vowel). Unlike standard phoneme-discretised phonetics, there is a continuous, if approximate, quantification of the time-varying degree of activation of each attribute. This is needed in SkAT-VG to handle gradual transitions between sounds. Some imitations resort to mechanisms that are not commonly used in Western languages, such as grating and plopping sounds. SkAT-VG would handle these as special phonemes, introduced as needed. This articulatory phonology approach naturally exploits our familiarity with phonemes; it allows SkAT-VG to accept wild and semi-wild onomatopoetic words with little extra effort; and in some sense, it also implicitly deals with scaling between different speakers. A disadvantage is that this approach will require a lot of manual phonetic transcription work, but this is building knowledge.

In SkAT-VG, gesture will be analyzed to select and manipulate expressive temporal evolution of sound properties.

The SkAT-VG project focuses on how a designer may sketch interactions mediated by sound, especially where the user is continuously involved in object manipulation, and human-object interaction may be altered by specific design choices in continuous feedback. As observed by Hallnäs and Redström [28], "a proper aesthetics of computational things concerns time gestalt", and continuous sonic feedback is one of the crucial design dimensions in this respect.

The first milestones of the project consists of defining categories of sounds that are meaningful to the listeners, and accumulate examples of vocalizations and gestures that adequately describe the sounds (IRCAM). On the other hand, phonetic expertise by KTH and gesture analysis experience at IRCAM will provide analysis tools specific to these productions. A first indicator of success for the project will assess how well these two approaches converge. Suitability of the analysis tools to analyses the database of vocal and gestural imitations will be evaluated before feeding the machine learning work packages.

The success of the machine learning/gesture following techniques should also be measured with a proper indicator.

The consortium will also use several indicators to measure the progress in developing the final sketching tools. First, interactions between GENESIS and IUAV (workshops) should result in precise specifications for the final tools. The variety and quality of these specifications and ideas will be a first indicator of quality of this interaction. The last part of the project will consist of creating the actual prototypes of the sketching tools. In the meantime, the interaction between GENESIS and IUAV will have defined a set of scenarios to test the effectiveness of these tools. We can imagine, for instance, objects or products that designers will have to augment with sound, with the newly-developed sketching tools or with more traditional tools. The originality, adequacy, functionality, aesthetics of the resulting designs will indicate how well the project will have matched its objectives. Evaluation of the quality of the designs will use the methodologies developed during the CLOSED project.

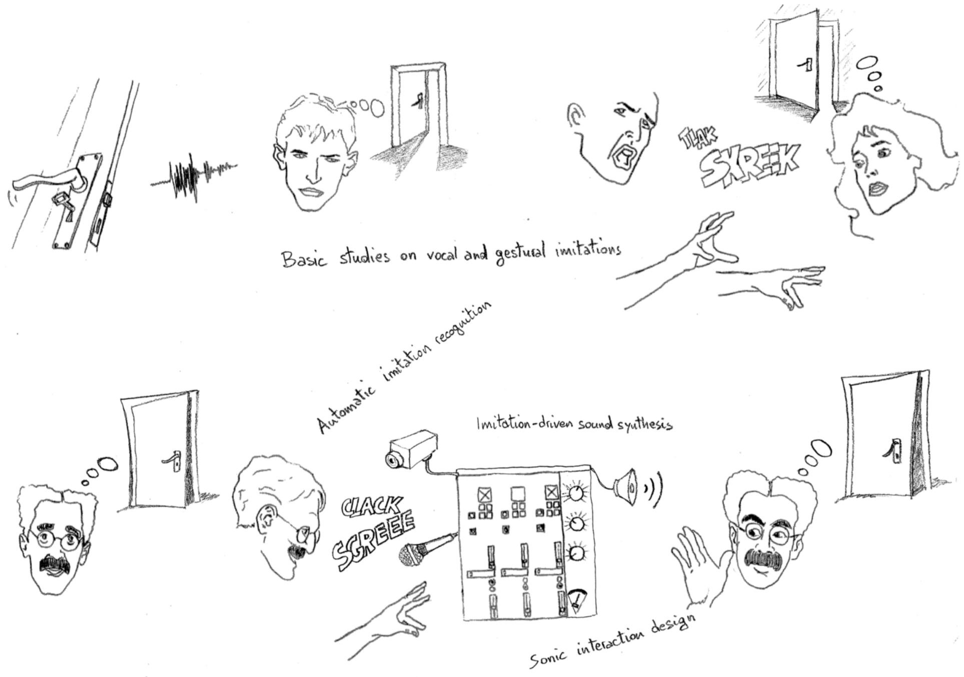
TLAK
SKREEK
Basic studies on vocal and gestural imitations
Automatic imitation recognition
Imitation-driven sound synthesis
CLACK
SCREEE
Sonic interaction design

Figure 1: Foundational sketch of SkAT-VG.

1.2.4 Vision

Eventually, designers should be able to sketch the sonic behavior of objects with the same ease as they render their visual appearance by paper and pencil. As in everyday conversations, the designer will evoke or suggest a sound by vocal production and hand gesturing. An automatic system will recognize the kind of physical phenomenon that may produce such sound and select a sound synthesis algorithm that is capable of generating it. Then, the algorithm may be parametrically controlled to fine tune the sound, thus enabling a sketch-to-prototype conversion.

A basic sketch of the SkAT-VG project is portrayed in figure 1. Basic studies (WP2, WP3, WP4) are depicted at the top of the figure, showing investigations of the SkAT-VG project on the perception of everyday sound, on the one hand, and on the production and perception of their vocal and gestural imitations, on the other hand. Then, based on the description of imitations obtained at this stage, an automatic imitation recognition system (WP5) will predict the categories of the imitated sound related to relevant categories of everyday sounds. The results will be implemented in a sound synthesis tool able to simulate the sounds represented by vocal imitations and hand gestures (WP6). These tools will eventually be applied in sonic interaction design frameworks, as an assessment bench for all of the results coming out from the project (WP7).

1.2.5 Novelty of the proposal

In order to achieve its vision, the SkAT-VG project has to bring about a range of advances beyond the state of the art. The project is **challenging current thinking** along several directions.

- **How listeners identify a sound source**, either directly or through vocal and gestural imitations, is a fundamental question in auditory cognition, but is however a little-understood issue. With its original mixture of psychology, gesture analysis, and articulatory phonology, the SkAT-VG project provides a new approach to this issue;

- Vocal and gestural inputs are already used in a number of applications, such as querying multimedia databases, triggering events, or controlling sound parameters. The **extraction and parameterization of a sound model** from vocal-gestural imitations is a challenge that has never been attempted before;
- The realization of **a sketching tool for sound design** is an endeavor that challenges the ephemeral nature of sound. A framework for sonic interaction design will enable rapid and creative exploration of the sonic behavior of objects with the same ease as their visual appearance is rendered by paper and pencil.

The basic research conducted in the project will result in the development of a computational system mimicking what humans do when interpreting vocal and gestural imitations of sounds. The system will extract important features of vocalizations and manual gestures and map them to psychologically relevant sound categories, using machine-learning and perceptually-based signal processing. The system will eventually connect vocal and manual inputs to appropriate sound synthesis modules and allow the user to edit and fine-tune the resulting sounds. Even in case of failure of the development of such system, the SkAT-VG project will develop an extensive **scientific foundation** on perception and identification of imitated sound sources and will contribute to the understanding of how humans effectively use their voice and hands to communicate sounds. That these innate abilities can actually be exploited for sketching sonic behaviors is a property that may emerge from a **new combination of insights** from different scientific areas as well as from research through design.

1.3 S/T methodology and associated work plan

1.3.1 Overall strategy and general description

The SkAT-VG project is multidisciplinary: **Production and perception** (KTH, IRCAM); **Automatic classification** (IRCAM); **Sonic Interaction Design** (IUAV, GENESIS).

Production and perception of vocalizations and expressive gestures. – A person imitates a sounding object to let an interlocutor identify what she has in mind. Through an original mixture of psychology, phonetics, and gesture analysis SkAT-VG studies two hypotheses: first, that the articulatory mechanisms used to imitate sounds are related to the mechanical characteristics and behavior of the source; second, that expressive gestures communicate the temporal evolution of fine timbral properties.

Production and perception of vocal imitations are inextricably connected, as humans often tune their listening to the constraints of human vocal production (phonetic listening). Expertise in phonetics, as expressed by KTH, helps understanding how the physical dynamics of an event is mimicked by the voice. By adopting an analytical approach to the craft of vocal imitation, SkAT-VG aims at clearly characterizing how humans vocally imitate sounds. From the cognition side sound source identification is still an open issue. IRCAM focuses on relevant elementary auditory phenomena to understand which sound features allow identification. Similarly, studying expressive gestures helps understanding when and how humans mimic the causes or the effects of sounding actions and their temporal evolution.

Automatic identification of vocalizations and expressive gestures. – Transforming imitations into synthetic sounds has two parts: automatic recognition of the sound source and estimation of the sound features. Automatic recognition requires three functions: i) providing a relevant representation of the signals (acoustical features, articulatory mechanisms, gestural features) ii) segmenting signals into meaningful elements, iii) Predicting the category of the imitated sound. IRCAM will embody the results of the basic studies into state-of-the-art machine-learning techniques (classifiers), different from conventional speech recognition in that there is here no linguistic context. As regards estimation of the sound features novel techniques of adaptation and estimation of gesture characteristics allow to exploit the expressiveness of vocal and manual gestures for continuous interaction. In this context, this means that the recognizer is able to adapt to user-controlled variations, in such a way that continuous classification and early estimation of variations will be possible while performing the recognition task.

Sonic interaction design. – From the beginning of the project user studies precisely specify the resulting sketching tools. Technically, these tools process vocalizations and gestures and transform them into synthetic sounds, further molded and included in actual prototypes. Various sound models are already available that allow parametric exploration of a wide sound space. IUAV has been experimenting with fundamental interaction primitives mediated by continuous sound feedback. This is extended in SkAT-VG by inclusion of vocal and gestural sketching in the design process, thus allowing the designer to rapidly specify a sonic behavior by directly acting on an object mockup. Basic articulatory mechanisms recognized from the vocalizations are therefore used to select appropriate synthesis methods. The selected synthesis models are driven so as to adaptively fine tune their parameters to match the target sound and the evolution of the expressive gestures as closely as they can. Manual gestures, already exploited as a source of expressive information at the imitation stage, are also used for real-time continuous control of sound synthesis.

The work plan is broken down 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, as described in section 3.2. A work package list is reported in table WT1. Work packages are further decomposed into tasks, whose verifiable outcomes are called deliverables. A list of deliverables is reported in table WT2.

The development of the project includes three Milestones, as shown in Table WT4. The first period is 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 will lead to Milestone 1 at month 12. The remaining Milestones are all found in the second period. Even though scientific advances will be made in the first reporting period of the project as well, the major decisions on how to turn technologies into tools and applications will be made in the second reporting period.

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 human effort required to achieve the objectives, subdivided among the project partners and assigned to each work package, is summarized in table WT6.

The idea of Sketching Audio Technologies using Vocalizations and Gestures, albeit being straightforward, requires some preliminary achievements that have a degree of uncertainty. **First**, the assumption that imitations are based on the dynamics of sound production needs to be verified experimentally. If this is reasonably true when a human imitates another human, such an assumption is far less obvious when humans imitate inanimate sources. In fact, preliminary research indicates that a mixture of physical and acoustic cues seem to be used by humans to produce and perceive vocal or gestural imitations. **Second**, it is not guaranteed that automatic classification of vocalizations and gesturalizations can be implemented robustly for a useful range of sounds. One reason for this is that mainstream recognition approaches rely on data-driven machine-learning and typically require a large corpus of sounds for training, but sketching is in some ways original by definition and so a large corpus with many instances of similar tokens could be difficult to build. Ideally, classification would be easier if the inverse problem of deducing articulatory parameters from the acoustic signal would admit a usable solution. Unfortunately, such solution is still far from being available, both because of the unavailability of complete and accurate articulatory models and because the inverse problem is intrinsically ill-posed. However, interaction with the recently-started FET project EUNISON may give pleasant surprises on the modelling side. To handle the variability of sketching gestures, recent advances in adaptive machine learning will be exploited. **Third**, the automatic translation of vocal imitations into specifications of sound models is not free of difficulties, as it depends on the range, the nature, and the characterization of the models that are available for sound synthesis. Eventually, the success of the project will be measurable in terms of versatility of vocal and gestural sketching as referred to the range of sounds that can be produced and to the variety of application contexts.

The simple temporal organization of activities will help reduce risks. At any given stage, the development of a task will not depend on more than two prior tasks, whose accomplishment is easily verified.

The four partners complement each other in the different areas of expertise that are required by the project. Overlap is reduced at a minimum, so that research progress can be achieved independently and concurrently at the different sites. This may imply a risk of the research units drifting apart. However, each partner has devoted some person-months to research visits or collaborations with other partners, so that the activities can be mutually monitored.

The SkAT-VG core ideas are quite innovative, and it is likely that some groundbreaking research will have to be performed in order to realize them. However, the partners do not start from scratch, as a significant amount of established research results, experiments, and implementations are available at the participants' sites. For example, some experiments on perception of imitations have already been made [43], online and adaptive classifiers have already demonstrated their effectiveness to handle variations in gestural interactions [5, 10], and a Sound Design Toolkit [18] is already at an advanced stage of development. The sharing of earlier or new results among the SkAT-VG partners will be relatively smooth, as pairs of partners have been collaborating at previous projects and research initiatives.

Specific risks are enumerated as follows, together with the measures that will be taken to overcome them:

- Effective vocal imitations often combine several sound generation mechanisms, such as "zzz" with frication and periodic modulation at the same time. This means that the automatic classifier should be ideally able to process compound audio signals. The supporting articulatory information will indicate when multiple sources are believed to be present, but there will be complex cases which will challenge automatic classification (such as phase-locking of vibrations in the vocal folds and the lips, which is one imitation technique). A risk is that the system will have problems with compound sounds in imitations, but these problems will be treated on a case-by-case basis as they emerge.

- Ideally, SkAT-VG aims at developing a user-independent system. The role of user specificities, however, may be significant in determining the potential and usability of the system. Just as there are good and poor drawers, there are also good and poor vocal and gestural imitators. The approach of looking at the vocal production mechanisms reduces the variability that would be associated with an approach purely based on acoustic and video signals. Still, user-dependent variations can be handled by adaptive classifiers.

  There are many possible sources of variation. They include:

  - speaker physiology,
  - imitation strategies (probably depending on the type of sounds),
  - native language.

  The first point is addressed by referring, at the beginning of WP2 and WP3, to a small set of selected skilled imitators producing a reasonably rich set of vocalizations. The second point is addressed by WP4, which will specifically highlight the different categories of sounds (and sound sources) leading to different types of imitations. The third point will not be problematic until the project addresses 'tamed' linguistic imitations, i.e., sound-words.

- Even if the main objective of SkAT-VG deals only with non-linguistic imitations, there is a risk that limiting the system to only such inputs will practically limit its flexibility. In fact, we can forecast that some sounds will be intrinsically difficult to imitate because of physiological "limitations". To overcome such limitations, the SkAT-VG approach leaves the door open to another kind of input: sound symbolism or caricature words (in Tasks 3.3, 4.3 and 5.4). For instance, there are phonemes that are conventionally used to represent certain features, such as /k/ for a rapid attack and /ng/ for a damped oscillation; /i/ for a tone with a bright spectrum and /u/ or /m/ for tones with a more dominating fundamental partial. However, rather than implementing a general-purpose word recognition module (which would be completely off topic), SkAT-VG will impose a predefined (and probably simplified) set of caricature words to the user. This will allow us to test the efficiency of such an approach, to be possibly developed more accurately in the future.

- Physical models will be the preferred synthesis paradigm, as these are consistent with an ecological view of perception and action. However, for some imitations it will be difficult to find the appropriate physical model of the imitated source. And when designing new sounds, the designer will often be searching for unconventional, previously unheard, yet still appropriate sounds. In these cases, the repertory of sound synthesis models will include classic signal-based techniques and modulations. The mapping between imitation and synthesis model will be treated on a case-by-case basis as soon as such a need for extended synthesis techniques will emerge from the practices of WP7.

- In a first approach to the SkAT-VG project, a sequential timing of the objectives (understanding the phenomena of vocal and gestural imitations, building a system that automatically identifies the imitated sounds, and integrating it in sketching tools) might seem natural. However such a timing, practically implementing one-way dependencies between the work packages would be very risky. To overcome this problem, the overall strategy of the SkAT-VG project is based on interacting and, when possible, parallel threads of research. The idea is that the challenges of the SkAT-VG project will be best approached from multiple perspectives (see section 1.3.2). Technically, interactions will be favored by research missions and workshops.

Each Work Package can be associated with its potential risks. These are listed in table 3.a together with mitigation or treatment activities. These may be preventive, when actions are conceived at the project design stage to reduce such risks. Or there may be contingency plans to be activated when a risk actually occurs.

Work-package	Potential Risk	Prevention / Contingency Plan
WP1	Failure to achieve effective project management	The consortium has significant collective experience of management of collaborative projects.
WP2	Failure to collect and analyze cases that are significant for the project	Significance of imitations will be assessed based on scenarios and design expertise collected in WP7. / Additional cases can be collected throughout the whole project duration.
WP3	Failure to derive accurate descriptions of phonatory and articulatory actions	The methods of articulatory phonology are robust and well established. / Analysis-by-Synthesis using an articulatory synthesizer will also be performed.
WP4	Failure to reveal how the phonetic mechanisms are used to convey an idea of the imitated sounds	Reduction of risks for WP3 should also prevent this risk. / Even without thorough understanding of the vocal production of imitations it will be possible to use in WP5 the results of classification experiments.
WP5	Failure to build a classifier for vocal and gestural imitations	The telescopic organization of WP5 ensures that at least a simple classifier based on low-level representations will be achievable. / As an extreme contingency measure, for the class of imitations based on onomatopoeias, an existing speech recognizer may be custom modified.
WP6	Failure to provide a re-synthesis of imitated sounds	Sound synthesis models will be prepared and organized according to the results of WP4. / Although physical models are the preferred synthesis framework, other synthesis methods are readily available for those sounds that are not achievable within such framework.
WP7	Failure to provide “killer applications” of vocal and gestural sketching	The involvement of the design community at large will let significant application scenarios emerge. / Although SkAT-VG is mainly directed towards interaction design, there are significant areas of application in the film and game industries. Beyond the area of product sounds, applications in these domains will also be demonstrated.

Table 3.a: Risks identified in the project WPs

1.3.2 Timing of work packages and their components

Figure 2 shows the timing of the different work packages, and their interactions and dependencies (see the details of each work package in section WT3).

Tasks 2.1 and 2.2 (recording a set of selected imitations from skilled imitators) aim at directly feeding WP3 (phonetic listening and world-event representations). Tasks 2.1 and 2.2, and 4.1 and 4.2, which are similar in nature but different in scale and objectives (the latter aiming at a large database of imitations from lay persons), run in parallel, and feed Task 4.3 (analysis of successful imitations) and Task 5.2 (blind classifiers). Task 4.2 and Task 2.2 will use the same recording setup, to acquire both audio and frontal/profile video of mouth and hands. Then some of the materials acquired in Task 4.2 will be passed to Task 3.2 (tracks of articulatory parameters) for supplementary articulatory analysis. Task 5.2 (blind classifiers) will start as soon as a critical mass of recordings of imitations are accumulated: this is **Milestone 1**.

The results of Task 4.3 (analysis of successful imitations) also feed Tasks 5.3 (informed classifiers), but for a successful achievement Tasks 4.3 and 5.3 run in parallel and benefit from mutual interactions. Task 5.3 will have to wait for a sufficiently large database from Task 3.2 (namely, the articulatory analyses), but only for each class of imitated sound. The ongoing results of Task 3.2 also inform Task 5.1 (segmentation) and Task 5.2 (inferring vocal and gestural production). This latter result derives from a tight collaboration between KTH and IRCAM. Task 5.1 (segmentation and computation of the signal representation) could almost run independently, but requires also at some stage the results of Task 4.3. The additional module of simple linguistic aliases (for the case of sounds impossible to vocally convey) will be developed in Task 5.4, based on the results of Tasks 4.3 and 3.3.

WP5 (automatic imitation recognition) collaborates continuously with WP6 (imitation-driven sound synthesis) to implement the sketching tools. **Milestone 2** occurs when robust enough classifiers are achieved, then allowing the first tests of the tools in WP6. The confrontation of Tasks 5.2 and 5.3 results in Task 5.4 (an integrated classifier), and these three tasks feed (and therefore require an interaction with) Tasks 6.1 and 6.2 (parameterization of the timbral families), which result eventually in the developments carried in Task 6.3 (the development of a sketching tool). Furthermore, the results of Task 5.4 will eventually complement and validate the studies in WP3 and WP4 (confirmation of the power of the models). Tasks 7.1 and 7.2 (design scenarios and prototypes), which together span the whole project, will be continously fed by the progresses of the other tasks. Once we will have robust enough results, side applications will be considered in Task 7.3. The results of WP6 and WP7 will be evaluated with user tests and result in **Milestone 3**, the final achievement of the SkAT-VG project.

This timing is conceived in such a way to maximize overlaps and, thus, collaborations between the work packages. These collaborations will be technically implemented in research missions and workshops (budgeted in Section 2.4 and described in table 1.b).

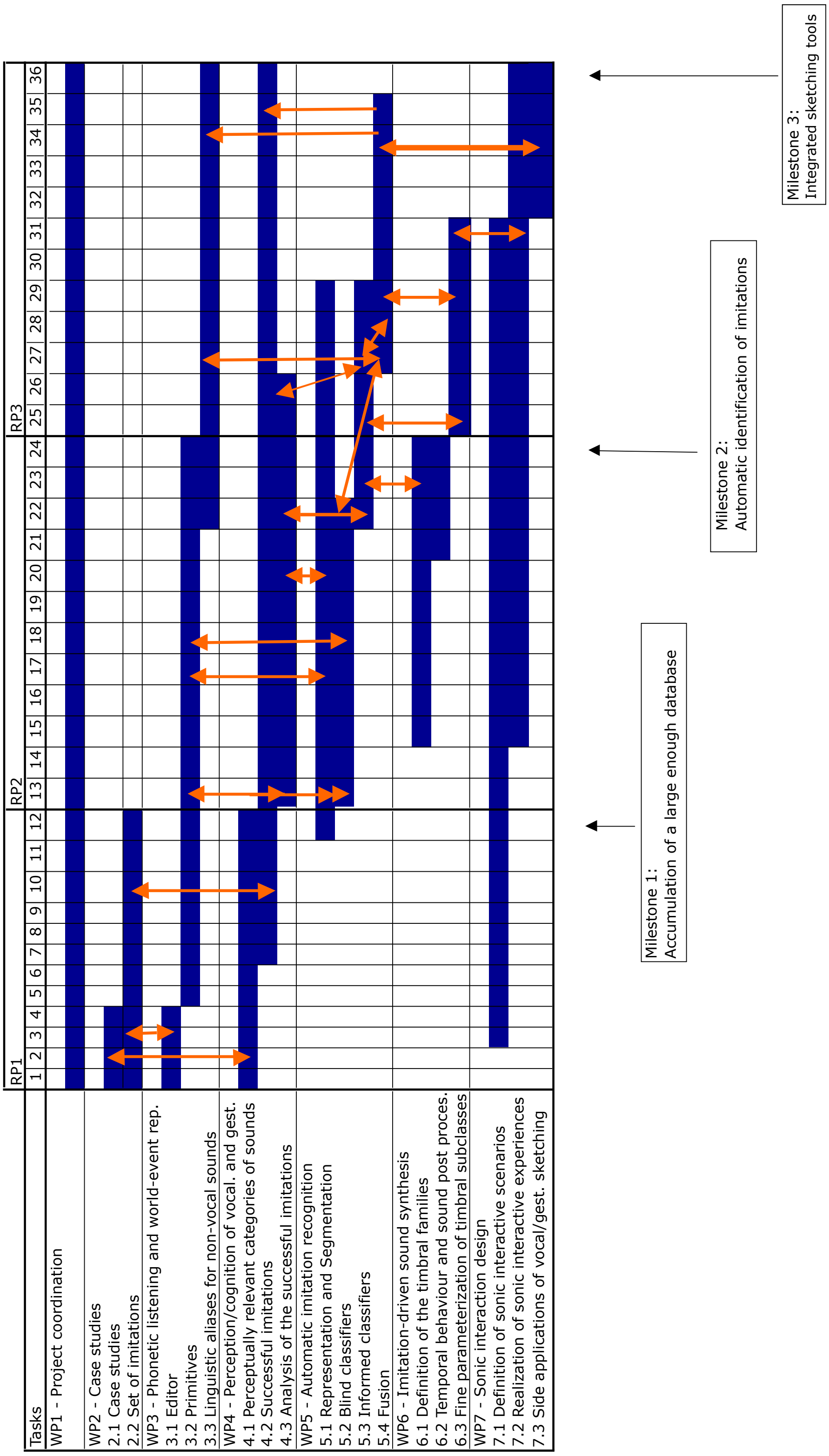

Figure 2: Timing of the different work packages. Orange arrows indicate interaction between the work packages.

The temporal organization of activities, with their inter-dependencies, is represented in the PERT chart of Figure 3, where activities are assigned to branches, and the nodes represent significant stages of the project, including milestones. This representation, as compared to the Gantt chart of Figure 2, provides a more 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.

It is clear that research in work packages 2, 3, 4, and 7 get started since the early months of the project, and all participants are involved. The tasks leading to the construction of automatic classifiers of imitations and to imitation-driven synthesizers should be completed when reaching the last semester, so that the final months of the project can be devoted to integration, prototyping, and refinement.

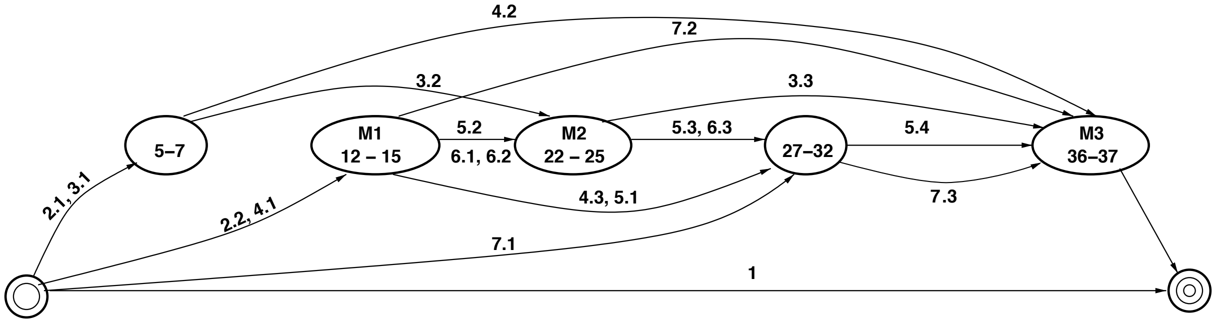

Figure 3: Interaction between the work packages

2 Implementation

2.1 Management structure and procedures

The SkAT-VG project involves four partners, each being specialized in an area of activity. Due to the size of consortium and the need of fast dispatching of research directives, a centralized and light management structure will be adopted, consisting in a Project Manager (Davide Rocchesso) at the coordinator IUAV, and work package leaders. The project management of the SkAT-VG project will be orchestrated and conducted by the coordinator and will consist of administrative and scientific components. The main goal of the project management will be to provide an appropriate, clear, consistent and efficient management of internal affairs as well as relations with the Commission and third entities. A centralized internet-based repository will be installed and maintained to make such management structure efficient. Tools for concurrent development, such as `svn` and `redmine`, will be used for such purpose. IUAV has experience in managing European and national research projects. The Project Manager acted as coordinator of the IST-FET project SOb and was Chair of the COST IC-0601 Action on Sonic Interaction Design.

2.1.1 Structure of the management

The structure of the management and the relations between the key roles are represented in figure 4.

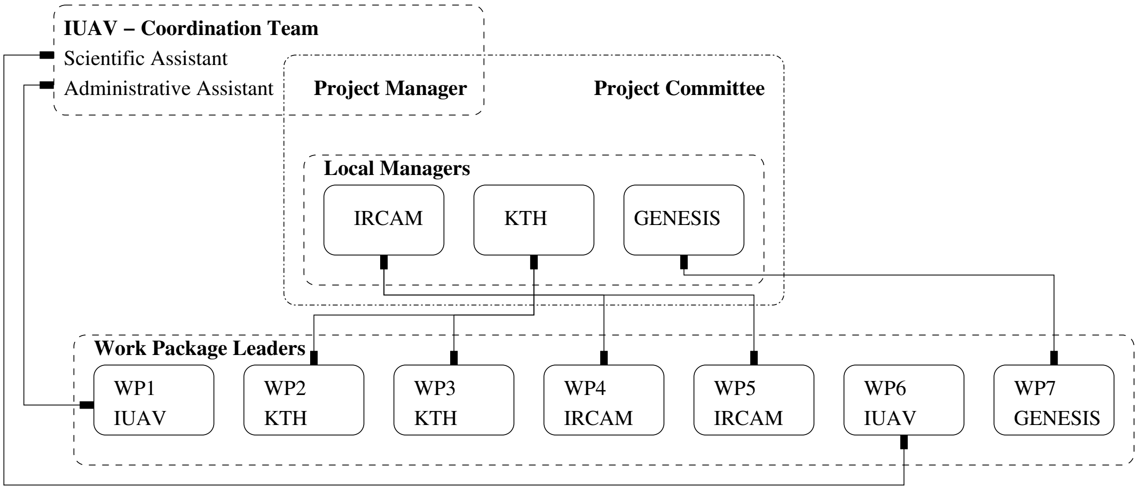

Figure 4: Management structure

Coordination Team and Project Manager: The position of coordinator is entrusted to IUAV, which nominates a Coordination Team. The Coordination Team is responsible for the overall project management. The Coordination Team is headed by the Project Manager, who acts as a focal point for all activities within the consortium and for all relations between the consortium and the Commission. The Project Manager is in charge of the consortium management and satisfies all of the coordinator obligations, including coordination of work packages, deadline monitoring, and exchange of information with the outside world. In the Coordination Team, the Project Manager is supported by a Scientific Assistant and by an Administrative Assistant.

Local Managers: Each non-coordinating participant institution nominates a Local Manager, who is responsible for the work assigned to the institution within each Work Package. The Local Manager (i) organizes and coordinates the execution of work within the participant institution, (ii) is responsible for the timely delivery of expected results before expiring of deadlines, and handles all relations with the Project Manager and with the other partners. The Local Manager also refers to the Work Package Leader about any issue arising during the work assigned to his institution within a Work Package.

Project Committee: The Project Committee groups together the Project Manager and the Local Managers. It is the ultimate authority for all decisions and is responsible for the overall direction of the project. The Project Committee will meet twice a year, in coincidence with general project meetings, in order to assess the progress towards the objectives and to review the project strategies. Conflicts are resolved in the Project Committee by vote: each partner has a vote and the Project Manager has a casting vote.

Work Package Leaders: For each Work Package of the project Workplan, the responsible partner nominates a Work Package Leader, who is in charge of coordinating the work of all partners involved in the Work Package and in the production of resulting deliverables. The Work Package Leader (i) presents a detailed preliminary plan to the partners involved in the Work Package, taking into account the roles and resources of each partner, as well as the subdivision into Tasks, (ii) coordinates the execution of the tasks, (iii) reports to the Project Manager on the work progress, (iv) is responsible for preparing and submitting the Work Package deliverables, (v) tries to solve any issues or conflicts that may arise within the Work Package.

2.1.2 Management procedures

Administrative and financial co-ordination: A periodicity for compiling and sending reports to the EC is established as follows: two reporting periods for months 1-12, and 13-36. Management and Activity Reports will be delivered to the Commission after each reporting period (also see section 2.1.3). A schedule will be prepared for the distribution of the funding. Advance payment will be designated with the agreement of the partners. The coordinator IUAV will be in charge of the distribution of the payments to the partners of the project. A safe and quick accounting will be supplied and will be held open to inspections to provide clarity. Periodic financial reports will be prepared and submitted to the Commission after each reporting period. Audits will be led to certify the financial reports.

Scientific coordination: This has two main goals: the coordination of different work groups and the transfer of the deliverables to the Commission. Regular meetings every six months will be the main opportunity for coordination and for keeping the work on schedule. Reports of these meetings will be delivered to the Commission by the project coordinator IUAV. Work Package Leaders will have to deliver the scientific deliverables to IUAV, who will be in charge of delivering them to the Commission. Short-term scientific visits across partners will allow effective integration of work and knowledge, as well as scientific progress in specific areas. A summary of meetings and visits, using a mixture of face-to-face and video-conference contacts, is reported in table 1.b.

Organization and Conflict Resolution: For this project, the management structure is largely centralized, being dominated by one central figure, the Project Manager, who will be responsible for the coordination of the whole project. The Project Manager will interact with the Work Package Leaders, who will be responsible for the coordination of each work package, and with the Local Managers, responsible for work carried out by each partner. A Project Committee grouping together the Project Manager and the Local Managers is the ultimate authority for all decisions and is responsible for the overall direction of the project. Conflicts are resolved in the Project Committee by vote: each partner has a vote and the Project Manager has a casting vote.

Meeting	When	Who	What	How
Project kick-off	Project start-up, first month	All participants	Work plan, project organization, circulation of information; Web infrastructure.	Face-to-face
General project meeting	Every six months	All participants	Overall progress; Information exchange between WPs; Refinements and alterations.	Face-to-face
Project committee meeting	Every six months, at general project meetings	Project committee members	Monitoring progress towards objectives; Costs; Deadlines; Contingency plans.	Face-to-face
WP meeting	Quarterly	WP members	Monitoring work in the WP; Reporting to Project Manager; Solving issues	Video conference
Visit	When needed	Single researcher	Short-term scientific mission at another partner	Face-to-face
Research workshop	At some general project meetings	WP members	Hands-on work; exchange of materials and experiences	Face-to-face

Table 1.b: Description of meetings and visits

2.1.3 Access to documents

A web-based infrastructure for document exchange between participants will be established, and joint document and code development will be facilitated by tools such as `svn` and `redmine`.

All the reports associated with deliverables (see table WT2) will be publicly available through the project web site. The recordings, including annotations, and the developed software modules will also be made available with Creative Commons and GPL licenses, respectively.

Progress and meeting reports will be kept internal to the project, and available through the web infrastructure under authenticated access.

2.2 Beneficiaries

2.2.1 Participant 1. IUAV

The main educational and research activities of IUAV are focused on the construction and continual transformation of the physical environment: from cities and territories, to buildings and objects. Research and teaching intermix in an environment where Master students are involved in prototyping activities, and Doctoral students push the hottest theoretical issues. **The Interaction unit**, hosted by the Dipartimento di Culture del Progetto and headed by D. Rocchesso, is one of the most active research groups at IUAV. In the area of Sonic Interaction Design, the unit has coordinated a large international research network, the COST Action IC0601 on Sonic Interaction Design. The local coordinator and key researchers have been involved in previous EU projects. D. Rocchesso coordinated the project SOb (the Sounding Object), that developed versatile and efficient physics-based sound models, that could be easily integrated within artefacts and appliances. In the CLOSED (Closing the Loop Of Sound Evaluation and Design) project, the key researchers provided sound synthesis modules linked to the perceived characteristics of basic sound events. Within SkAT-VG, IUAV is mainly active in research on sound synthesis and sonic interaction design. The Interaction unit takes the responsibility of project coordination (WP1) and leads WP6 on sound synthesis.

Key Staff *Davide Rocchesso* (Project Manager) is Associate Professor at IUAV where he teaches media processing and interaction design. He obtained his Ph.D. from the University of Padova in the field of sound processing, with special focus on physical modelling and spatial processing. Until 2006, Davide Rocchesso led the sound processing group at the University of Verona, where he coordinated the IST-FET project SOb, and he was local coordinator in the IST-FET Coordination Action $S2S^2$ and in the NEST project CLOSED. He was Chair of the COST Action IC0601 on Sonic Interaction Design.

Gillian Crampton Smith was the Head of the Department of Computer Related Design at the Royal College of Art, London, and the Director of Interaction Design Institute Ivrea, Italy. She has been directing the Interaction Design programme at IUAVsince 2006.

Pietro Polotti is professor of Electronic Music at the Conservatory of Trieste in Italy. He also collaborates with IUAV within the Interaction unit. He collaborated with the University of Verona within the CLOSED project. Previously, he was employed by the Politecnico di Milano as part of the IST TAI-CHI European project. He obtained a Ph.D. in Communication Systems from EPFL, Lausanne in 2002 with a thesis on fractal additive sound synthesis. In recent years, his interests moved towards Sonic Interaction Design (SID).

Stefano Delle Monache graduated in law at the Università Cattolica di Milano, and in electronic music and technologies of sound at Conservatorio di Como. In 2007/2008, he received a research grant at University of Verona in the scope of the CLOSED project. He obtained a PhD in design sciences at IUAV, where his research on product sounds is mainly related to the application of design methods and practices, including sketching, to a context of interaction design where auditory feedback is of central importance.

2.2.2 Participant 2. IRCAM

The Sound Perception and Design group (PDS) at IRCAM is Work Package leader for WP4 (perception and cognition of imitations) and WP5 (automatic recognition of imitations). Members of two other groups at IRCAM, the Real-Time Musical Interaction group (IMTR) and the Analysis/Synthesis group (A/S), will also be involved respectively in WP4 and WP5.

The Institut de Recherche et de Coordination Acoustique/Musique (IRCAM) is a non-profit research organisation founded in 1976. Its main activities include contemporary music production, R&D of music technologies, and music-related basic scientific research. The research at IRCAM covers the domains of sound analysis and synthesis, acoustics of musical instruments and concert halls, sound perception and design, music cognition, computer science. IRCAM has a research laboratory association with the French Centre National de la Recherche Scientifique and the University Pierre-et-Marie-Curie (STMS-IRCAM-CNRS-UPMC). IRCAM has participated to several European funded projects: Esprit (FP3), CUIDAD (FP4), WedelMusic, Carrouso, Listen, Agnula, Music Network, Listen (FP5), Semantic Hi-Fi , CROSS-MOD, i-Maestro (FP6), CLOSED, MINET, SAME, the SID COST action, Quaero, MIReS, Verve, 3DTVS and HC2(FP7). IRCAM is also participating to several national projects funded by the French National Research Agency (ANR).

The Sound Perception and Design group: Basic research activities of the group include: loudness of non stationary sounds, everyday sound perception and recognition, sound signalling, and sonic interaction design. The group participates to several applied projects in sound quality. The group has coordinated the CLOSED European project (FP7) and is participating to several national projects (ANR).

The Real-Time Musical Interaction group: The IMTR group conducts research and development on interactive music systems, gesture and sound modeling, interactive music synthesis, gesture capture systems and interfaces. The targeted applications concern primarily music performance and the performing arts, but the team also collaborates regularly on industrial projects for the development of audio software, sound simulation and gaming.

The Analysis and Synthesis group: The A/S group carries out research and development activity into sound analysis, transformation and synthesis. The activities of the group include additive analysis/synthesis, automatic music indexing, control of sound synthesis for musical composition, orchestration, pitch recognition in a polyphonic context, processing by phase vocoder, and score following and alignment.

Key Staff *Patrick Susini* (Local Manager) received a Ph.D degree in Acoustics in 1999, and a Habilitation in 2011. He is the head of the Sound Perception and Design group. His research activities include everyday sound perception, loudness and sound quality. He organised the 1st and 2nd international symposium in sound design in 2002 and 2004.

Nicolas Misdariis graduated from a college of university level specialized in professional training on mechanics (CESTI-SupMeca), completed by a specialization on acoustics (LAUM, Le Mans), in 1995. His current research topics are related to environmental sound perception, auditory display and interactive sonification.

Olivier Houix received the PhD degree in acoustics in 2003 from the Université du Maine, Le Mans. His research interests concern the perception of environmental sounds and the gesture-sound relationship in sound design. He teaches audio engineering. He has been involved in national and European projects such as CLOSED.

Geoffroy Peeters received his Ph.D. degree in computer science from the Université Paris VI in 2001. His current research interests are in signal processing and pattern matching applied to audio and music indexing: timbre description, sound classification, music classification, audio identification, rhythm description, music structure discovery.

Frédéric Bevilacqua is the head of the Real Time Musical Interactions team (IMTR). He holds a PhD from the Swiss Federal Institute of Technolgy in Lausanne (EPFL) in biomedical engineering. Since 2003, he has been conducting research at IRCAM on gesture analysis and on gesture-based musical interaction systems.

Norbert Schnell is a researcher and developer in the IMTR group focusing on real-time digital audio processing techniques for interactive music applications. He studied Telecommunications and Music in Graz/Austria and worked as programmer and sound designer with the Musiklabor Wien. He chaired the 6th International Conference NIME in 2006.

2.2.3 Participant 3. KTH

Kungliga Tekniska Högskolan (KTH) in Stockholm is Sweden's largest technical university. The Sound and Music Computing Group (SMC, formerly Music Acoustics) of the Department of Speech, Music and Hearing at KTH is the leader for WP2 (Case studies) and WP3 (Phonetic listening and world-event representation). In the KTH 2012 International Research Assessment Exercise, the Department was evaluated as part of Applied Computer Science, as, quote "Research output is internationally excellent in all fields, with a substantial number of units reaching the level of world-leading quality."

The SMC group has considerable experience of European projects and actions, such as MOSART, SoundingObject, Agnula, Marie Curie Training Site, $S2S^2$, IMUTUS, VEMUS, and BrainTuning; as well as COST Actions IC0601 Sonic Interaction Design and 2103 Advanced Voice Function Assessment. One currently EU-funded project is EUNISON, which in its topic is complementary to the proposed SkAT-VG project. In addition, the SMC Group is in the same department as the well-known Speech Communication and Technology Group, with many points of convergence with the SkAT-VG project, and with very extensive experience of EU-funded projects.

Key Staff *Sten Ternström* (Local Manager) received his Ph.D. in Music Acoustics at KTH in 1989, and heads the KTH SMC group as its professor since 2003. His research interests center on the human voice, but extend also across music, electronics, audio technology, computing and signal processing. He has been principal investigator for numerous voice-related research projects, funded mostly by Swedish research councils and foundations. Ternström is the project coordinator of the new FET-Open EUNISON project. Ternström also co-founded and worked in a signal processing company 1995-2000 and has written commercial voice analysis software, in addition to some 40 scientific papers and book chapters. He will be work package Leader and Local Coordinator for WP2 and WP3, and will also work on the phonetic analysis.

Olov Engwall is professor of speech communication in the Speech group, who works on articulatory modelling of speech. He will supervise PhD student work on the phonetic analysis and also on the inversion problem for identifying articulation automatically.

Roberto Bresin is associate professor with research interests in Sound Design. He was Swedish delegate to the COST IC-0601 action on Sonic Interaction Design. He will liaise for KTH especially with WP7 in years 2 and 3 of SkAT-VG.

2.2.4 Participant 4. GENESIS

GENESIS ACOUSTICS is a high-tech company founded in 1999 (SME), spin-off from CNRS-LMA. Its main activities are Sound Simulation, Sound Perception, and Sound Design and its main markets are the automotive (Renault, PSA, Volkswagen, Hyundai, Nissan, etc.) and the aeronautic industries (Airbus, Eurocopter, Dassault Aviation, Snecma, etc.). GENESIS spends more than 30 % of its activities in Research and Development projects (FUI, EC FP7 projects). GENESIS received an Award from the SFA (Société Française d'Acoustique) in 2005 for its technologies and involvement in research projects. The key know-hows of GENESIS are:

- Signal Processing, software development, real-time sound simulators
- 3D sound and virtual reality
- Psychoacoustics
- Speech intelligibility

Among other realizations, GENESIS has developed LEA, a tool for sound quality assessment and sound design used by Dassault, Airbus, Snecma, Eurocopter. GENESIS was also expert on EC FP7 project, formerly SEFA and now COSMA (Patrick Boussard), about the sound simulation of aircraft and sound perception for community around airports.

Key staff *Patrick Boussard* (Local Manager) is the executive director and founder of GENESIS. He received master's degree in engineering from l'École Nationale des Arts and a master's degree in acoustics from the University of Aix-Marseille (1991). He has worked with the Laboratoire de Mécanique et d'Acoustique (LMA) du CNRS and for real-time sound simulation.

Christophe Lambourg is the technical director of GENESIS. He received a PhD degree in acoustics from the Université du Maine (1997, physical modeling of boards). He has worked for several acoustical engineering agencies and joined GENESIS in 2008. His specialties include signal processing, sound perception, microphone array design, speech intellegibillity.

Guillaume Lemaitre received a Ph.D in Acoustics from the Université du Maine (Le Mans) in 2004. Since 2000, he has been involved in several projects with IRCAM. He has worked with the department of Psychology at Carnegie Mellon University (Pittsburgh, PA) and with Davide Rocchesso at IUAV. His research activities include psychoacoustics, perception and cognition of sound sources, and vocal imitations of sounds.

2.3 Consortium as a whole

The SkAT-VG consortium is at the same time small enough to allow for flexibility and adaptation, and based on a solid experience that covers the scientific domains necessary for the success of the project. It includes four participants from different disciplines all having an important experience in sound and interaction design, by working in close relations with industry professionals and sound designers, providing tools and methodologies for designers, and being involved in several projects related to sound and interaction design. This expertise ensures the relevance of the SkAT-VG objectives for the community of designers. Particularly, the reference framework provided by WP7, and the imitation-driven sound synthesis tools provided by WP6 will benefit from such kind of expertise. Reaching the objectives of SkAT-VG also requires a strong scientific background, provided by the partners. One the one hand, IUAV has a strong experience in sound synthesis, and particularly in the conception and development of algorithms based on the physics of mechanical sources. On the other hand, the collaboration between KTH and IRCAM will allow to study the phenomenon of imitations from two perspectives: from the perspective of the phonetic production of imitations (KTH), and from the perspective of the perception of imitations (IRCAM). This collaboration between IRCAM and KTH will therefore provide a unique approach to this phenomenon in WP3 and WP4. KTH has additional experience in gesture analysis and is currently building a Performance and Multimodal Interaction Lab which will include a motion-capture system that will be operational for the SkAT-VG project. This facility will strengthen collaboration between KTH and IRCAM, especially through short-term scientific missions. The link between the basic study of imitations of sounds, and the final development of the sketching tools will be made in WP5: The results of WP3 and WP4 will feed the system of automatic recognition of imitations developed in WP5. The expertise of the IMTR and the A/S group at IRCAM in machine learning and automatic classification will ensure the success of this work package. The system will be used for the tools developed in WP6 and WP7, where the role of industrial partner GENESIS (SME) will be exploited.

The relationship among the SkAT-VG participants and their best cooperation for the internal operation of the consortium will be regulated by a Consortium Agreement, whose terms shall complement but not conflict with the requirements, terms and conditions provided under this Description of Work. The Consortium agreement contains technical, organizational, financial, and legal provisions, as well as rules for dissemination and use, including protection of intellectual property rights.

2.4 Resources to be committed

The costs indicated on form A3 are further detailed in this section. In the following tables, the gross figures include indirect costs, and the third column reports the requested EC contributions. Most of the budget is dedicated to human resources, with research positions and management positions distributed according to the detailed work description of section 1.3. All partners have computing and prototyping facilities that will be exploited by SkAT-VG and further expanded with project-specific purchases of hardware and software equipment.

All partners have software frameworks, largely developed in previous projects, that will be made available to the consortium. These include software and databases for sound and gesture analysis, sound synthesis, and vocal tract morphology.

Mobility of researchers is considered to be important for effective intra-project communication and for effective out-of-lab assessment of research results.

Description	Resources	Cost	Gross	Request
	IUAV			
Personnel	57 person-months for research position	338400	541440	406080
	12 person-months for management position	56400	90240	90240
Equipment	Computers	13000	20800	15600
	Software licenses	5000	8000	6000
	Audiovisual equipment	13000	20800	15600
	Fabrication costs for prototypes	4000	6400	4800
Travel	6 conference trips (3 persons, 2 per year)	27000	43200	32400
	6 project meetings (3 persons, 2 per year)	27000	43200	32400
	Organization of research workshop	10000	16000	12000
	Research missions (4 PMs across partners)	8000	12800	9600
Dissemination	Publications and promotion	5000	8000	8000
Subcontr.	Audits	5000	5000	5000
Total		511800	815880	637720

Description	Resources	Cost	Gross	Request
		IRCAM		
Personnel	90 person-months for research position	451890	813402	610052
	3 person-months for management position	15063	27113	27113
	4 master students	10800	19440	14580
	800 subjects for perception tests	10000	18000	13500
Equipment	Computers	10400	18720	14040
	Software licenses	8000	14400	10800
	Audio equipment	800	1440	1080
Travel	6 conference trips (4 persons, 2 per year)	36000	64800	48600
	6 project meetings (2 persons, 2 per year)	18000	32400	24300
	Organization of research workshop	5000	9000	6750
	Research missions (visits/invitations)	4000	7200	5400
Dissemination	Publications and promotion	5000	9000	9000
Subcontr.	Audits	4000	4000	4000
Total		578953	1038915	789215
		KTH		
Personnel	65 person-months for research position	417200	667520	500640
	3 person-months for management	24400	39040	39040
	20 expert subjects for production tests	4000	6400	4800
Equipment	Computers	5000	8000	6000
	Audiovisual equipment	9000	14400	10800
Travel	6 conference trips (2 persons, 2 per year)	18000	28800	21600
	6 project meetings (2 persons, 2 per year)	27000	43200	32400
	Organization of research workshop	6000	9600	7200
	Research missions (3 PMs across partners)	5000	8000	6000
Dissemination	Publications and promotion	5000	8000	8000
Subcontracting	Audits	5000	5000	5000
Total		525600	837960	641480
		GENESIS		
Personnel	46 person-months for research position	238502	381603	286202
	4 person-months for management	26048	41676	41676
Equipment	Audiovisual equipment	15000	24000	18000
Travel	6 project meetings (1 person, 2 per year)	9000	14400	10800
	3 conference meetings (1 person, 2 per year)	8666	13865	10399
Subcontracting	Audits	0	00	0
Total		297216	475544	367077

3 Impact

3.1 Strategic impact

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).

3.1.1 Science

Functional to the success of SkAT-VG will be a significant advance in fundamental knowledge in voice, hearing, and motor behavior, including sound perception and cognition. The challenge is to shed some light on the processes that allow the transformation of the signals coming from an imitation into a mental model of a sound source. As compared to the vast literature on speech production and perception, research on vocal and gestural imitations has been quite sparse. However, imitations and non-linguistic utterances are at the roots of highly-evolved forms of communication, both in phylogenetic and in ontogenetic sense. Beside contributing to evolutionary developmental psychology, studying these processes should provide references to scientists studying environmentally-situated acoustically-based interactions between intelligent agents. As humans, we exploit the subtleties of sounds to communicate with our similars and to connect to the world we live in. Even though many of the everyday sounds we hear today did not exist a century ago, we have been able to develop imitative capabilities by exploiting the flexibility of voice and gesture, and the analytic features of hearing. By basing the investigations on both production and reception, SkAT-VG will help understanding how the structural and transformational characteristics of sounding objects are translated into characteristics of human vocal and gestural production and, in turn, into an organisation of cognitive representations.

The SkAT-VG studies in production and perception of vocal imitations are complementary to those of the FET-Open project EUNISON. Eventually, the two projects may set the basis for a system that could “listen” to sounds and learn to direct a model of the human voice to imitate those sounds.

3.1.2 Technology

Describing a sound by imitation is very natural for humans. It is at least as natural as drawing the representation of an object on paper. However, as opposed to pencil and paper, whose role in communication and design has a very long history, vocal and gestural imitations could be used in the past only for face-to-face, instantaneous communication. Before recording and broadcasting devices became commonplace, there wasn’t much support for any other use of utterances. With the availability of recorders, and the possibility to manipulate recordings, the human voice started to become usable as a design tool and as a design material at the same time, thus acquiring a level of flexibility similar to that of hand drawing. Nowadays, drawing and gesturing capabilities of humans are being increasingly exploited in research and in actual human-machine interfaces. This requires some understanding of how humans interpret and represent the world through drawn lines or traces in the air. A similar research line is only at its beginnings in the auditory domain, but it is likely to have an impact just as strong. If the SkAT-VG project will be able to provide tools for automatic classification of vocal and gestural imitations, it will become possible to use voice and hands as fast specifiers of sounding objects and processes. More generally, the tools developed during the SkAT-VG project will allow a variety of applications (e.g. retrieval in large databases of sounds, human-computer communication interfaces) to use non-speeech sounds as an intuitive input. As a further step, the automatic translation of an imitation into an instance of a sound model will make it possible to move from vocal sketching to vocal and gestural prototyping of sonic interactive devices. This means that, beyond being able to specify the sonic behavior of an object, the designer will also be able to rapidly obtain a prototype that implements such sonic behavior and that allows fine adjustment through the manipulation of model parameters. Obvious applications of such breakthroughs are found in sound effects for films and multimedia shows, but it is in interactive contexts, such as that of games, that the availability of controllable sound models will have a strong impact, as the sense of agency is largely mediated by sound. The availability of low-cost processors, sensors and actuators, enables the embedding of sound models into everyday objects, whose dynamic sonic character can be explicitly designed. Sounds will not be only side effects of actions anymore, but rather designed components of the whole aesthetics of an object in a given context. As a result, a larger part of our surrounding landscape will result from searching for an informative equilibrium among multitudes of sounding objects.

3.1.3 Society

Technological noises, from motors to ringtones, pervade our everyday environments and are usually perceived as principal sources of acoustic pollution. Therefore, any attempt to talk about how to augment objects with sounds is often accompanied by questions about the acoustic sustainability of such sounding objects. Auditory clutter is perceived as more problematic than visual clutter, probably because humans know they have no earlids, and that turning their head is no help to get rid of a noise. However, even if it were attainable, a silent world wouldn't be better for human life. Especially because of these concerns and reasons, it is important to take a design attitude towards the acoustic manifestations of objects, and to think about the sounding objects as situated in a context. In natural ecosystems, phylogenetical processes govern the evolution of species in such a way that, for example, certain birds occupy specific time-frequency windows of the acoustic space and have their voices communicated with minimal effort from the emitter to the listener. These processes, naturally reducing acoustic clutter and leading to pleasant soundscapes, should be taken as a reference when designing auditory displays of any kind. The fact of starting the design process by using the most natural of the means – the human voice and hands – is crucial, as it avoids the emergence of those beeps and sweeps that are just trivial byproducts of technological components. In using their voice, humans naturally adapt to different contexts in order to get effective communication with minimal disturbance. There are contexts that require whispering and contexts that require shouting, and changing from one to the other is not just a matter of decibels. A shouting voice will sound shouting even at a very low volume, and its appropriateness for a given context wouldn't change much with a change in volume. If a designer wants to impress a certain sound character to an object for a certain context, vocal sketching is the most natural mean to rapidly convey the acoustic idea.

The technological and social impacts will largely depend on the degree of adoption of SkAT-VG technologies in the world of interaction design. In order to foster the interest of professionals, it is important that the tools are made available in stable and usable form, and that a number of practitioners are involved in dissemination activities, both during and after the project. All the partners are equally committed to such program, and are sufficiently scattered over Europe to promote vocal sketching in sonic interaction design among a large community of designers. After the recent conclusion of the COST IC-0601 Action SID, a network of researchers and practitioners interested in use and demonstration of SkAT-VG technologies is well established all over Europe.

3.1.4 Contribution to the European level towards the expected impacts listed in the work programme

Often the birth of design as a discipline is attributed to the Bauhaus school, founded in 1919 by Walter Gropius in Weimar. Initially design was defined as an original combination of crafts and fine arts, but in later developments, both in and outside Europe (Ulm, Chicago, Yale, etc.), technological innovation and scientific knowledge became increasingly important components of design practices. Nowadays, the importance of technological tools to assist and feed human creativity is obvious, especially in all those cases where the objective of design is itself a technological product. Even ancestral gestures such as those of free-hand drawing need to be considered as part of a complex design process, that should go smoothly from sketching to modeling to prototyping. All of these stages are increasingly being supported by information and computing technologies, but the importance of human agency is also being increasingly felt.

As human-object interactions overcome the limits of visual displays to embrace all human senses, the role of sound for function and aesthetics become crucial. Because of its century-long tradition in design and its prominent role in advancing the sciences of sound and voice, Europe can take the lead in promoting new techniques, tools and practices for sonic interaction design. Besides specific projects on sound and design, such as the FET project SOb (The Sounding Object) or the NEST project CLOSED (Closing the Loop of Sound Evaluation and Design), Europe has been coordinating major networking efforts on sound-based research. A Roadmap on research in sound and music computing [67] was prepared in 2007 by the FET Coordination Action $S2S^2$ (Sound to Sense, Sense to Sound). One of the five challenges launched in the Roadmap is “to design better sound objects and environments”. One of the six strategies that are proposed to address such challenge is to “improve technologies for pervasively producing, transforming and delivering sounds”. The tools provided by the SkAT-VG project will fit this strategy, and have their core objective in the transformation of vocal and gestural imitations into object representations and synthetic sound models. Another strategy is to “promote research in fields involved in the shaping of natural, artificial and cultural acoustic ecosystems”. Design at large is clearly involved in such process of shaping, and designers are the target users of SkAT-VG innovations. The discipline of Sonic Interaction Design (SID) has emerged from the activities of another Europe-centered networking initiative, that was the COST IC0601 SID Action[3]. SID was the real incubator that has been nurturing the SkAT-VG research project. In that network the topic of vocal sketching for sonic interaction design was discussed in various meetings and was the topic of an interdisciplinary workshop (Holon, Israel, november 2009). Before SID, the COST Action ConGAS had the important function of boosting European research in gesture-controlled audio systems. Most of the leading research groups in the field of sound and gesture are now based in Europe, and some are represented in the SkAT-VG consortium.

It can be said that in Europe the ground has been prepared and it is ready to absorb and process the seeds of innovation that the SkAT-VG project will introduce and disseminate. Although the range of disciplines involved in SkAT-VG is not tremendously large, the project is still multidisciplinary in nature. It is difficult to coordinate, in a single country, research efforts in perception and cognition, design, phonetics, and computing, especially because it is impossible to find research groups in each of these disciplines, that are aware of what the others do and that are all committed to a common objective. These groups have been put together at the European scale, under the facilitatory role of prior research and networking initiatives.

3.1.5 Contribution to portfolio and concertation activities at FET-Open level

In order to support scientific cooperation at the FET-Open level and broad public awareness of project achievements, consortium members will ensure within the areas of interest of the project:

- Project results shall be published throughout the duration of the project in widely accessible science and technology journals, as well as through conferences and through other channels, including the Web, reaching audiences beyond the academic community.
- Beneficiaries shall deposit an electronic copy of the published version or the final manuscript accepted for publication of a scientific publication relating to foreground published before or after the final report in an institutional or subject-based repository at the moment of publication.

[3] http://sid.soundobject.org/

- Beneficiaries are required to make their best efforts to ensure that this electronic copy becomes freely and electronically available to anyone through this repository:
  - immediately if the scientific publication is published “open access”, i.e. if an electronic version is also available free of charge via the publisher, or
  - within 6 months of publication.
- Periodic press releases shall be issued, and other means of disseminating project progress to a wider audience e.g. via video.
- Participation in FET-organised events, for example conferences, dedicated workshops and working groups, consultation meetings, summer schools, online fora, etc.
- International Co-operation - contribution to relevant national and international activities (e.g., Joint workshops, calls, etc., for example with NSF).

The above activities will be reported in the project’s Dissemination Plan and in periodic progress reports. In addition, the consortium agrees to include the following reference in all project-related publications, activities and events:

“The project SkAT-VG acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: XXXXXXX”.

3.2 Plan for the use and dissemination of foreground

Even though no specific Work Package has been exclusively designed for it, dissemination of project findings and results is crucial to success of the SkAT-VG project. Every partner has a portion of the budget dedicated to cover the expenses for publications and promotion.

The general project strategy for dissemination is based on openness: project reports, experimental data, models, and tools will be made available to all through the project website as soon as they are sufficiently solid. This approach will enable independent evaluation and cross-validation of results. Moreover, the availability of research tools may convince some practitioners to embark on new approaches to sonic interaction design based on vocal sketching and gestural interaction. The SkAT-VG web site will be the main source of dissemination of research results, demonstrations, and software tools.

Rules for dissemination and use, such as confidentiality, mandatory communications, and ownership of results are defined in the Consortium Agreement.

3.2.1 An enlarged SkAT-VG community

Since the early stages there will be an extensive involvement of external stakeholders, such as expert sound designers, whose advice is needed to better tailor the design of sketching tools. Another important class of stakeholders that will be involved since the beginning of the project is that of expert vocal imitators and professionals who make extensive use of vocal abilities. They will help identifying the most relevant vocal primitives. Interaction designers, both students and professionals, will be involved in workshops where vocal and manual sketching will be at the center of design exercises. In the framework of Work Package 7, the exercises will be directed toward the definition of scenarios and applications. Then, actual designs of sonic interactive objects will be performed and documented, with involvement of the Sonic Interaction Design community.

3.2.2 Publications

Since basic scientific research is an important component of the project, it is expected that the results of experimental and modeling activities will be disseminated by means of contributions to key scientific conferences and journals. A selection of targeted journals is: Speech Communication; Journal of the Acoustical Society of America; Acustica / Acta Acustica; IEEE Trans. on Speech, Audio and Language Processing; Journal of Experimental Psychology; Perception and Psychophysics; ACM Transactions on Computer-Human Interaction; ACM Transactions on Applied Perception; Journal of Human-Computer Studies; International Journal of Design. A set of targeted conferences will be decided at each semestral meeting according to the results that are being achieved. Certainly, the project will organize at least one demonstration in one of the leading conferences that deal with interactive technologies, such as ACM SIGGRAPH or CHI.

The writing of articles on vocal sketching, gestures, and sonic interaction design in the Free Encyclopedia Wikipedia will be a means to convey the project ideas to the largest audience. For the same purpose non-scholarly journals (e.g., ACM Interactions) will be considered for dissemination of project results.

3.2.3 Intellectual property

The sound synthesis tools will be based on refinement and extension of existing tools, such as the ones developed in the CLOSED project. These tools will continue to be distributed as free software under GPL license, thus encouraging experimentation in a large community of practitioners, whose contributions and feedback will be considered for maintaining and improving the software. Industries may benefit from such tools as long as they keep their free and open form. The automatic classifier of vocal and gestural imitations will be originally developed in SkAT-VG, and a license for such project will be established at the project kickoff meeting. The general policy is to make sure that such tool is widely accessible and usable in a variety of contexts. The sharing of rights that may arise from products or applications deriving from the results of the project will be agreed among the project participants, based on the proportion of contribution of each partner. The participants will not claim *a posteriori* any intellectual property of processes and products derived from the results of the project, unless they were formally involved in such developments.

At conclusion of the project, an Exploitation Plan will be prepared to consolidate the policies towards the usage of the results in prospective applications or commercial products. This document will provide a strategy to seek engagements with market players.

4 Ethical issues

The SkAT-VG project involves psychological experiments, within the scope of WP4. These experiments will take place in Paris. A list of subjects for experiments will be 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). This database contains personal information about people (address, age, sex, musical practice) who register themselves to participate in psychological experiments. This database is used only within the framework of auditory experiments at IRCAM and will never be shared. Subjects will be reimbursed for their participation.

Within the scope of WP3, the SkAT-VG project involves also observational studies of performers/imitators. The studies include audio and video recordings and possibly also other noninvasive physiological measurements such as electroglottography. These experiments will be carried out in Sweden. Ethical approval will be sought from Regionala Etikprövningsnämnden i Stockholm.

5 Gender aspects

In recruiting personnel for the project, the SkAT-VG consortium will adopt an equal-opportunity policy. Within the enlarged SkAT-VG community, the consortium will stimulate the presence of women in the areas of research, dissemination, workshops, and administration. Gender balance will be sought when composing the pools of subjects or experts for experiments or tests. Ensuring gender equality means giving equal consideration to the life patterns, needs and interests of both women and men. Gender mainstreaming thus includes also changing the working culture.

References

[1] M. Aramaki, C. Marie, R. Kronland-Martinet, S. lvi Ystad, and M. Besson. Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. *Journal of cognitive neuroscience*, 22(11):2555–2569, 2010.

[2] K. Aura, G. Lemaitre, and P. Susini. Verbal imitations of sound events enable recognition of the imitated sound events. In *Acoustics 08*, Paris, France, Juillet 2008.

[3] K. Aura, G. Lemaitre, and P. Susini. Verbal imitations of sound events enable recognition of the imitated sounds. In *Proceedings of the EAA/ASA joint meeting Acoustics'08*, Paris, France, 2008. Also in *Journal of the Acoustical Society of America*, vol. 123 p. 3414.

[4] F. Bevilacqua, F. Baschet, and S. Lemouton. The augmented string quartet: experiments and gesture following. *Journal of New Music Research (Accepted)*, 2012.

[5] F. Bevilacqua, N. Schnell, N. Rasamimanana, B. Zamborlin, and F. Guédy. Online gesture analysis and control of audio processing. In J. Solis and K. Ng, editors, *Musical Robots and Interactive Multimodal Systems*, volume 74 of *Springer Tracts in Advanced Robotics*, pages 127–142. Springer Berlin Heidelberg, 2011.

[6] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, pages 73—84. Springer Berlin / Heidelberg, 2010.

[7] J. Bilmes. What HMMs can do. Technical report, University of Washington, Department of EE, Seattle WA, 98195-2500, January 2002.

[8] W. Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann, 2007.

[9] P. A. Cabe and J. B. Pittenger. Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology: human perception and performance*, 26(1):313–324, 2000.

[10] B. Caramiaux, F. Bevilacqua, and A. Tanaka. Beyond recognition: Using gesture variation for continuous interaction. In *CHI '13 extended abstracts on Human factors in computing systems*, New York, NY, USA, 2013. ACM.

[11] B. Caramiaux, P. Susini, T. Bianco, F. Bevilacqua, O. Houix, N. Schnell, and N. Misdariis. Gestural embodiment of environmental sounds: An experimental study. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 144–148, Oslo, Norway, 2011.

[12] C. Carello, K. L. Anderson, and A. J. Kunkler-Peck. Perception of object length by sound. *Psychological science*, 9(3):211–214, May 1998.

[13] U. Castiello, B. L. Giordano, C. Begliomini, C. Ansuini, and M. Grassi. When ears drive hands: the influence of contact sound to reaching grasp. *PLoS ONE*, 5(8):e12240, 2010.

[14] M. C. Corballis. *From hand to mouth: the origins of language*. Princeton University Press, Princeton, NJ, 2002.

[15] A. Cummings, R. Čeponienė, A. Katoma, A. P. S. n, J. Townsend, and F. Dick. Auditory semantic networks for words and natural sounds. *Brain research*, 115:92–107, 2006.

[16] A. D'Ausilio, F. Pulvermüller, P. Salmas, I. Bufalari, C. Begliomini, and L. Fadiga. The motor somatotopy of speech perception. *Current Biology*, 19:1–5, 2009.

[17] M. De Lucia, L. Cocchi, R. Martuzzi, R. A. Meuli, S. Clarke, and M. M. Murray. Perceptual and semantic contributions to repetition priming of environmental sounds. *Cerebral cortex*, 20(7):1676–1684, 2009.

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

[19] A. Dessein and G. Lemaitre. Free classification of vocal imitations of everyday sounds. In *Proc. of Sound and Music Computing conference*, pages 213–218, 2009.

[20] I. Ekman and M. Rinott. Using vocal sketching for designing sonic interactions. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, pages 123–131, New York, NY, USA, 2010. ACM.

[21] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolati. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, 15:399–402, 2002.

[22] K. Franinović and S. Serafin, editors. *Sonic Interaction Design*. The MIT Press, Cambridge, Massassuchetts, 2013.

[23] D. J. Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America*, 87(1):311–322, January 1990.

[24] B. L. Giordano and S. McAdams. Material identification of real impact sounds: effect of size variation in steel, glass, wood and plexiglass plates. *Journal of the Acoustical Society of America*, 119(2):1171–1881, February 2006.

[25] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing – a preliminary study. In *2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*, 2006.

[26] V. Goel. *Sketches of Thought*. MIT Press, Cambridge, MA, 1995.

[27] M. Grassi. Do we hear size or sound? Balls dropped on plates. *Perception and Psychophysics*, 67(2):274–284, 2005.

[28] L. Hallnäs and J. Redström. From use to presence: on the expressions and aesthetics of everyday computational things. *ACM Trans. Comput.-Hum. Interact.*, 9(2):106–124, 2002.

[29] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of new music research*, 32(1):3–21, 2003.

[30] M. M. J. Houben, A. Kohlrausch, and D. J. Hermes. Perception of the size and speed of rolling balls by sound. *Speech communication*, 43:331–345, 2004.

[31] K. Ishihara, T. Nakatani, T. Ogata, and H. G. Okuno. Automatic sound-imitation word recognition from environmental sounds focusing on ambiguity problem in determining phonemes. In *Proc. of PRICAI-2004 Topics in Artificial Intelligence (Eighth Pacific Rim International Conference on Artificial Intelligence)*, volume LNAI 3157, pages 909–918. Springer Verlag, 2004.

[32] M. Kavakli and J. Gero. Sketching as mental imagery processing. *Design Studies*, 22(4):347–364, 1997.

[33] R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence*, 9(4):399–410, 2000.

[34] A. J. Kunkler-Peck and M. T. Turvey. Hearing shape. *Journal of Experimental psychology: human perception and performance*, 26(1):279–294, 2000.

[35] H. Laboratories. Introduction to articulatory phonology and the gestural computational model. URL: http://www.haskins.yale.edu/research/gestural.html.

[36] S. Lakatos, S. McAdams, and R. Caussé. The representation of auditory source characteristics: simple geometric sources. *Perception and psychophysics*, 59(8):1180–1190, 1997.

[37] N. J. Lass, S. K. Eastham, T. L. Wright, A. H. Hinzman, K. J. Mills, and A. L. Hefferin. Listener's identification of human-imitated sounds. *Perceptual and Motor Skills*, 57:995–998, 1983.

[38] N. Lebrun, P. Clochon, P. Etévenon, J. Lambert, J. Baron, and F. Eustache. An ERD mapping study of the neurocognitive processes involved in the perceptual and semantic analysis of environmental sounds and words. *Cognitive brain research*, 11:235–248, 2001.

[39] G. Lemaitre, A. Dessein, P. Susini, and K. Aura. Vocal imitations and the identification of sound events. *Ecological Psychology*, 23:267–307, 2011.

[40] G. Lemaitre and L. M. Heller. Auditory perception of material is fragile, while action is strikingly robust. *Journal of the Acoustical Society of America*, 131(2):1337–1348, 2012.

[41] G. Lemaitre and L. M. Heller. Probing a taxonomy of simple sound events via identification time and semantic priming. *Experimental Brain Research*, 2013. In press.

[42] G. Lemaitre, O. Houix, N. Misdariis, and P. Susini. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16–32, 2010.

[43] G. Lemaitre and D. Rocchesso. Vocal imitations communicate sounds more effectively than verbalizations. 2013. Under review.

[44] M. Leman. *Embodied music cognition and mediation technology*. The MIT Press, Cambridge, Massassuchetts, 2007.

[45] X. Li, R. J. Logan, and R. E. Pastore. Perception of acoustic source characteristics: walking sounds. *Journal of the Acoustical Society of America*, 90(6):3036–3049, December 1991.

[46] A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, October 1985.

[47] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657 – 675, 2009. <ce:title>PerCom 2009</ce:title>.

[48] R. A. Lutfi. Auditory detection of hollowness. *Journal of the Acoustical Society of America*, 110(2):1010–1019, August 2001.

[49] R. A. Lutfi and E. L. Oh. Auditory discrimination of material changes in a struck-clamped bar. *Journal of the Acoustical Society of America*, 102(6):3647–3656, December 1997.

[50] K. Martin and Y. Kim. Instrument identification: a pattern-recognition approach. In *Proceedings of 136th Meeting of the Acoustical Society of America*, 1998.

[51] S. McAdams, A. Chaigne, and V. Roussarie. The psychomechanics of simulated sound sources: material properties of impacted bars. *Journal of the Acoustical Society of America*, 115(3):1306–1320, March 2004.

[52] K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen. Analyzing sound tracings: a multimodal approach to music information retrieval. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, MIRUM '11, pages 39–44, New York, NY, USA, 2011. ACM.

[53] G. Orgs, K. Lange, J.-H. Dombrowski, and M. Heil. Conceptual priming for environmental sounds: an ERP study. *Brain and cognition*, 62:267–272, 2006.

[54] A. Patel and J. Iversen. Acoustical and perceptual comparison of speech and drum sounds in the North India tabla tradition: an empirical study of sound symbolism. In *Proceedings of the 15^{th} International Congress of Phonetic Sciences*, Barcelona, Spain, 2003.

[55] R. Patterson and I. Johnsrude. Functional imaging of the auditory processing applied to speech sounds. *Philosophical Transactions B*, 363(1493):1023, 2008.

[56] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Cuidado project report, Institut de Recherche et de Coordination Acoustique Musique (IRCAM), 2004.

[57] G. Peeters. A generic system for audio indexing: application to speech/ music segmentation and music genre. In *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, Bordeaux, France, 2007.

[58] B. H. Repp. The sound of two hands clapping: an exploratory study. *Journal of the Acoustical Society of America*, 81(4):1100–1109, April 1987.

[59] V. Rioux. *Sound quality of flue organ pipes. An interdisciplinary study on the art of voicing*. PhD thesis, Chalmers University of Technology, Göteborg, Sweden, 2001.

[60] D. Rocchesso. *Explorations in Sonic Interaction Design*. Logos Verlag, Berlin, Germany, 2011.

[61] D. Rocchesso, P. Polotti, and S. Delle Monache. Designing continuous sonic interaction. *International Journal of Design*, 3(3):13–25, 2009.

[62] A. P. Saygın, F. Dick, S. M. Wilson, N. F. Dronkers, and E. Bates. Neural resources for processing language and environmental sounds. evidence from aphasia. *Brain*, 126:928–945, 2003.

[63] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE International Conference and Acoustics, Sound And Signal Processing (ICASSP)*, Munich, Germany, 1997.

[64] A. Sedda, S. Monaco, G. Bottini, and M. A. Goodale. Integration of visual and auditory information for hand actions: preliminary evidence for the contribution of natural sounds to grasping. *Experimental Brain Research*, 209:365–374, 2011.

[65] M. Suwa and B. Tversky. What do architects and students perceive in their design sketches? A protocol analysis. *Design Studies*, 18(4):385 – 403, 1997.

[66] M. Takada, K. Tanaka, and S.-I. Iwamiya. Relationships between auditory impressions and onomatopoeic features for environmental sounds. *Acoustical science and technology*, 27(2):67–79, 2006.

[67] The S2S^2 Consortium. *A Roadmap for Sound and Music Computing, Version 1.0.* 2007.

[68] H. Théoret and A. Pascual-Leone. Language acquisition: do as you hear. *Current Biology*, 12:R736–R737, 2002.

[69] C. van Petten and H. Rheinfelder. Conceptual relationships between spoken words and environmental sounds: event related brain potential measures. *Neuropsychologia*, 33(4):485–508, 1995.

[70] W. H. Warren and R. R. Verbrugge. Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance*, 10(5):704–712, 1984.

[71] K. Watkins, A. Strafella, and T. Paus. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41:989–994(6), 2003.

[72] A. Wilson and A. Bobick. Parametric Hidden Markov Models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9):884–900, 1999.

[73] A. Wilson and A. Bobick. Realtime online adaptive gesture recognition. In *Pattern Recognition, Proc. 15th International Conference on*, volume 1, pages 270–275 vol.1, 2000.

[74] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, September 1996.

[75] R. Zuchowski. Stops and other sound-symbolic devices expressing the relative length of referent sounds in onomatopoeia. *Studia Anglica Posnaniensia*, 33:475–485, 1998.