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

SkAT-VG: Sketching Audio Technologies using Vocalizations and Gestures

D4.4.1
A large set of vocal and gestural imitations

First Author	Guillaume Lemaitre
Responsible Partner	IUAV
Status-Version:	Final-1.1
Date:	November 4, 2015
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	A large set of vocal and gestural imitations
Date of delivery to the EC:	30/09/2015

Workpackage responsible for the Deliverable	WP4
Editor(s):	-
Contributor(s):	Guillaume Lemaitre, Frédéric Voisin, Hugo Scurto, Olivier Houix, Patrick Susini, Nicolas Misdariis, Frédéric Bevilacqua
Reviewer(s):	-
Approved by:	All Partners

Abstract	This deliverable reports on the construction of the database of vocal and gestural imitations, and the database of referent sounds used to elicit the imitations
Keyword List:	Periodic report

The SkAT-VG Consortium consists of the following entities:

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

Document Revision History

Deliverable D4.4.1

Version	Date	Description	Author
First draft	2015/04/09	Import Word documents	GL
v0.2	2015/06/29	Update with new sections	GL
v0.3	2015/07/15	Update: only work related to the database	GL
v0.4	2015/09/10	New parts: statistical analyses and identification experiment	GL
v0.5	2015/09/18	Changed the methodology for identification experiment	GL
v0.6	2015/09/22	Statistical analyses	OH
v0.7	2015/09/28	Update of statistical analyses	OH/GL/PS
v0.8	2015/09/29	Patrick's take on it	GL/PS
v0.9	2015/09/30	Including Nicolas's comments	GL/NM
v1.0	2015/09/30	Final tweaking and delivery	GL/PS
v1.1	2015/10/01	Include Biblio in the file	GL

Table of Contents

Index of Figures

List of Acronyms and Abbreviations

DoW Description of Work

EC European Commission

PM Person Month

WP Work Package

GA Grant Agreement

CA Consortium Agreement

M Milestone

Mo Month

Q Quarter

1 Executive summary

1.1 Introduction: content of this deliverable

Work package 4 (WP4) of the SkAT-VG project studies how people produce and perceive vocal and gestural imitations when they communicate about sounds. In terms of terminology, we distinguish between: the *referent sounds* (the sounds that are imitated), the *imitator* (the person that produces the imitations), the *imitations* (vocal or gestural), and the *receiver* (the person that perceives and makes sense of the imitations, see Figure 1).

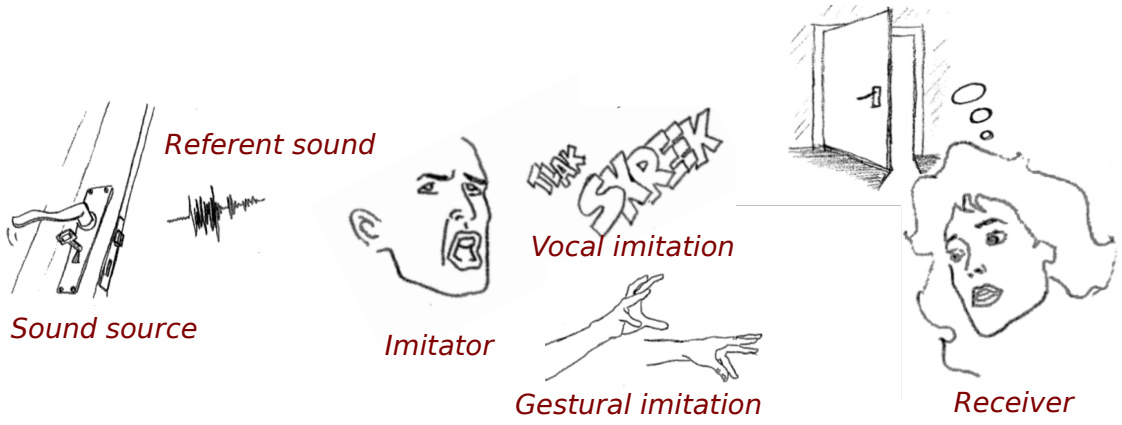

Figure 1: An imitator produces vocal and gestural imitations of a referent sounds, perceived by a receiver.

We use in this document the terms "vocal and gestural imitations" for the sake of simplification and to insist on the multimodal aspects. Nevertheless, one should note that the gestural task is not really an imitation in the same sense as the voice can imitate a sound. It corresponds to a gestural representation of a sound and/or gestural mimicry of sound.

Overall, WP4 has three main objectives. First, WP4 aims at studying how people produce and perceive vocal and gestural imitations with *experimental studies*. The second objective is to provide the project (WP5, WP6, and WP7 in particular) with *datasets and new insights* on how vocal and gestural imitations can be practically used in the context of sound design. The third objective is to use vocal and gestural imitations as new tools to *investigate sound perception and cognition in general*. Figure 2 summarizes these three objectives.

To reach these goals, WP4 is divided in three Tasks. Task 4.1 and 4.2 provide the other parts of the project with *databases*: a database of referent sounds (Task 4.1), sampling the potential applications of sound design in SkAT-VG; and a large database of vocal and gestural imitations of these referent sounds (Task 4.2). These databases are feeding WP5 (machine learning) and WP3 (study of vocal production of imitations). The other tasks of the project consist in assessing the *perception* of the imitations (and in particular the identification of the referent sounds via the imitations) in Task 4.2, and analyzing the *production* of imitations in Task 4.3 to understand what makes an imitation successful or not. **The core of this first deliverable reports on the elaboration of the multimedia databases of referent sounds and imitations: the rationale behind the elaboration, the description of the setups,**

the definition of the procedures, the description of the content of the databases (Sections 2 and 3), and the statistical analyses of the imitations (Section 4).

With the completion of the tasks reported in this document (databases of referent sounds and imitations), the project has reached its first Milestone: "MS1: An accumulation of a large enough database of recorded, sorted, and labeled imitations". It is an exceptional material to study what people perceive from these imitations, and, as consequence, use imitations to study sound perception and cognition. For instance, the datasets of imitations are currently used to investigate the semantic content of imitations: what do listeners perceive and recognize from imitations when they are not provided with the referent sounds? How do human-made imitations compare with "auditory sketches" created by automatically sparsifying mathematical representations of the referent sounds? This study is described in the "Work in progress" section (Section 5.3).

Finally, WP4 studies the intriguing question of the production of imitations in Task 4.3, in collaboration with KTH: how do speakers produce imitations that are understood by receivers? What features do they imitate, given the constraints of vocal production? What is the role of gestures, what pieces of information do they convey? This del¡iverable also touches upon these questions by reporting statistical analyses of the database of vocal imitations, and qualitative analyses of the database of vocal and gestural imitations (Section 4). This latter analysis is currently used by WP5 to derive new descriptors of gesture primitives (see WP5 deliverables). Another example of such study is currently revised for a publication in the Journal of the Acoustical of America: we studied how expert singers and lay participants vocalize a set of basic auditory features.

More precisely, the deliverable details four main sections:

- The creation of the **database of referent sounds** that samples across potential applications of sound design, perceptually relevant categories, and articulatory mechanisms (Task 4.1, Section 2),
- The creation of the **database of imitations**. This huge database contains about 8000 imitations made by 50 imitators, consisting of carefully controlled audio, video, and motion capture data (Task 4.2, Section 3),
- **Statistical and qualitative analyses** of the database of imitations, identifying the vocal and gestural imitations used across speakers, and looking for idiosyncratic behaviors (Task 4.3, Section 4),
- An overview of completed work and **work in progress** to appear in the next deliverable: a study of what features can the voice reproduce, a study of the respective roles of voice and gesture during imitations, a study of the semantics of the imitations, a study of the influence of speakers' native language on vocal imitations (Section 5).

But before diving into this detailed progress report, the rest of this section begins with summarizing and discussing the main results and their contributions to the project.

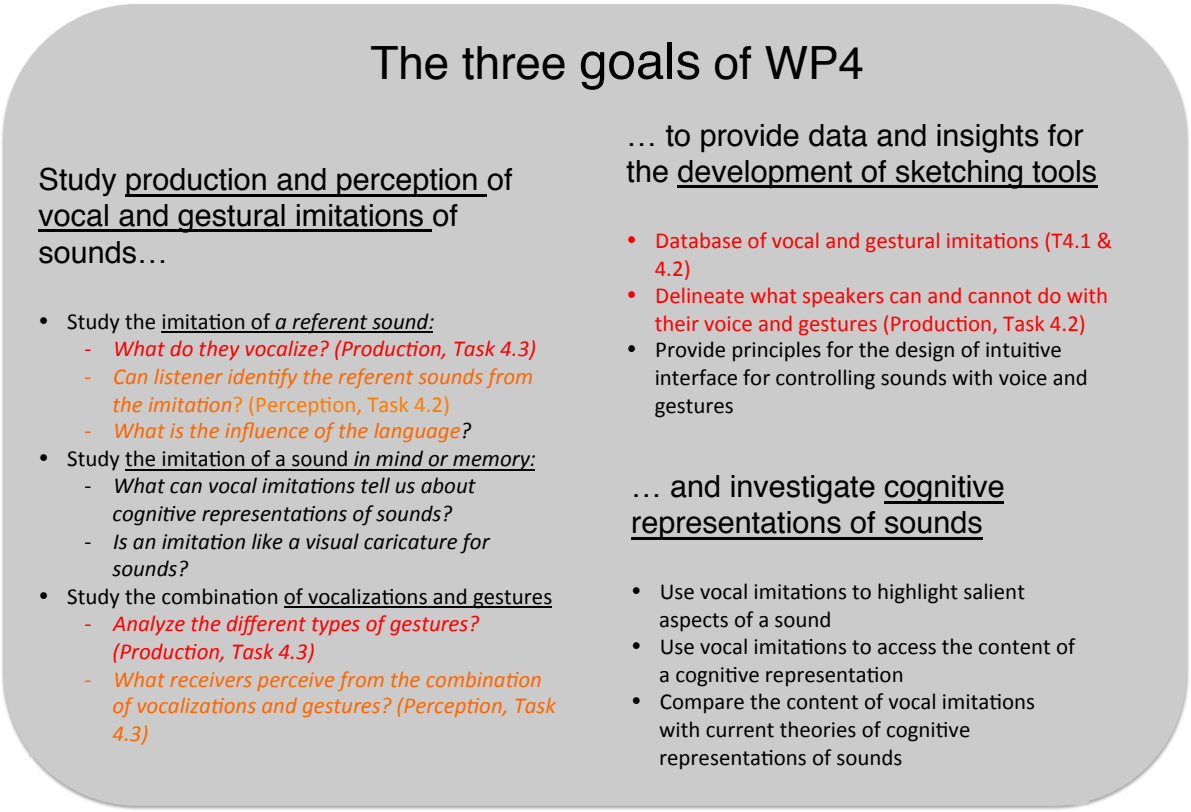

Figure 2: The three goals of WP4. Items in red are completed and form the core of this document (D4.4.1). Items in orange are work in progress.

1.2 Inputs to the project

The work reported in this document has provided the project with two types of inputs: sound databases and new insights into how vocal and gestural imitations could be used for sound sketching.

Sound databases The two databases of 52 referent sounds (see Section 2) and **about 8000 vocal and gestural imitations** (see Section 3) are an unprecedented amount of data to our knowledge[1]. They constitute the first milestone (MS1 "Accumulation of a large enough database of recorded, sorted, and labeled imitations") already initiated by WP2 at month 12. Reaching this milestones has already unlocked numerous studies using the databases. For instance, the databases are already used by WP5 to train classifiers of vocalizations, design new gestural descriptors, and was passed to WP3 for annotation. But more than just data, the studies reported in this document provide the project with new, sometimes unexpected insights into how vocalizations and gestures could be used to design intuitive sound sketching

[1] Cartwright and Pardo (2015) recently published a database of 4429 vocal imitations (audio only) recorded by 185 imitators across the world. The poor quality of the recordings and the lack of control in the procedure make them unfortunately unusable for our purpose.

tools.

Regularities and idiosyncrasies in the database of imitations The analysis of the imitations in Section 4.1 shows **common strategies** to produce imitations across the speakers and families of referent sounds. For example, there is a clear distinction between imitations of Mechanical interactions and Machines: imitations of Mechanical interactions are more noisy whereas imitations of the Abstract and Machine families are more tonal. Imitators have also imitated referent sound using repetitive patterns contrasting with more stable sounds. Specific sounds like impulsive sounds are also well reproduced with brief imitations. These results are encouraging (and somehow expected) as they clearly confirm that speakers can produce imitations that encompass a large variety of environmental sounds and that they are able to **reproduce the main aspects of the sounds** (tonalness, temporal patterns, etc.). These aspects are well captured by the descriptors developed in WP5. Furthermore these results now allow us to select referent sounds for the identification studies by balancing what we now know to be important features for vocalizations (see Section 5.3).

Focusing on imitators' strategies, one striking result is that **we did not identify different strategies across imitators**. The results strongly suggest that the imitators used very similar acoustic strategies (to the extent of what we measured). We only occasionally found a few outliers who produced, for example, longer imitations than the other imitators, or tonal imitations for the Mechanical interactions whereas the vast majority of imitators produced noisy imitations. Therefore, these speakers were not considered in the identification studies (see Section 5.3).

How to design intuitive gestural interfaces? The qualitative analysis of video recordings in the database of imitations (Section 4.2) revealed a somewhat unexpected result. Our initial expectation (based on Caramiaux et al. (2014)) was that imitators would either mime the physical source of the sound, mark articulations in the vocalizations, or "trace" the evolution of some acoustical feature of the referent sound that the voice could not reproduce (descriptive gestures). Whereas we observed the descriptive behavior in a few cases, we also observed a lot of **iconic gestures**: gestures that signified some aspect of the referent sound (e.g. "noisy") by a particular gesture that was not necessarily following any feature of the referent sound (e.g. shaking the hand rapidly). Another unexpected result (confirmed by our current studies, see Section 5.2) is the **poor performance to precisely reproduce the rhythm** of a referent sound with gestures "in the air" (whereas performance are obviously much better when the imitators produce a rhythm by hitting a surface). These results and observations have led WP6 to **rethink the role of gestures for the control of sound synthesis**. For instance, the design of the MIMEs (see corresponding work packages) is inspired by these considerations: gestures are not used to precisely control the production of the sketches, but to manipulate them.

How accurate can you get? One of the core findings that motivated the SkAT-VG project was that vocal imitations are accurately associated with their referent sounds (Lemaitre et al., 2011; Lemaitre and Rocchesso, 2014). But does this necessarily mean that speakers can reproduce accurately the basic auditory features of the referent sounds? The study presented

in Section 5.1 suggests that naive and expert speakers are fairly accurate at reproducing pitch, rhythm, but only moderately at reproducing the sharpness and attack time of artificially created sounds when these features are systematically varied and controlled (Lemaitre et al., 2015a). Since it is unlikely that the good performance observed by Lemaitre and Rocchesso (2014) were only due to a good reproduction of pitch and rhythm, this means that speakers use more specific, ad hoc strategies to convey the identity of a referent sound event. We are now considering several alternative hypotheses: listeners may be able to recognize the referent sources not because of the precise features of imitations but because they recognize the overall articulatory mechanisms (a sort of iconic-ish strategy); imitating may also amount not so much in an accurate reproduction of features but in a selection and **exaggeration of certain salient features akin to visual caricatures**. These ideas will be investigated in the coming year.

Of importance is also the fact that we did not observe any outstanding difference between the performances of expert singers and lay participants. Initial observations and informal experiments made with KTH (comparing French, Italian, and Swedish speakers) also suggest that the native language of the imitators has little influence on vocal production than, compared to what happens in speech production. If this is true, this will be a very exciting result: people could escape the contraints of their native language once they are not placed in a linguistic context. Of course, this intuition deserves a more careful examination that has just started.

The semantics of imitations So far, our work has studied how vocal imitations are similar to their referent sounds (Lemaitre and Rocchesso, 2014). The positive answer to this question confirms that users would be able to use their voice to control the parameters of sound synthesis. However, at this stage, we have only little evidence that imitations can elicit the semantic content associated with the referent sounds to the same extent as the referent sounds themselves, even if Lemaitre et al. (2011) showed that the classification of vocal imitations of kitchen sounds followed the same hierarchical structure as the classification of the referent sounds. It is however important to assess the semantics of imitations, if only to evaluate the relevance of using automatic classifications of imitations as an initial stage of the design tools envisioned by the project. WP4 is therefore currently conducting a new set of experiments to investigate the **identification of vocal imitations** (see Section 5.3). We are considering methodologies based on free verbalizations, forced-choice tasks, and semantic priming.

In the context, we are introducing an exciting new idea: so far, we have considered imitations as sketches made human beings. But there are other ways to create sketches, based on sparsified representations of the signals. Our goal is therefore to compare human-made imitations with **mathematically derived sketches**. More generally, this idea is important to think about the concept of sketch in the project.

Do imitators adjust their imitations when provided with feedback? In what we have done so far, imitators produced an imitation (vocal or gestural), and the only **feedback** they got was a playback of the audio or audio-video recording of their imitations. Things may be different in the context of an actual communication between two persons, such as those described by Lemaitre et al. (2014). In addition, users of the SkAT-VG tools may also **learn how the system behaves**, and learn how to adjust their vocal and gestural production to

reach their sound design goals goal. This question will be addressed during the coming year (see Section 5.5).

Imitation or vocal/gestural production of an idea? So far, our work has focused on *imitations* of a *referent sound*. This paradigm is necessary because it allows us to know exactly what it is that the imitators are trying to vocalize or gesticulate. However, the situation may be actually different when there is no referent sound but the **memory or the idea of a sound**. These situations are also closer to a real sound design case study. This new line of research will be conducted during the coming year (see Section 5.6) .

1.3 What have we learned so far? Reframing the results in a more general plan

Even if this report focuses on the creation and the analysis of the database of vocal and gestural imitations, we nevertheless would like to reframe this work in the general plan of the questions addressed by WP4. Figure 3 further details these questions.

At its root, WP4 is based on a set of general hypotheses. These hypotheses are grounded in initial observations and pilot experimental work (see for instance Lemaitre et al., 2011). It is however important to stress their hypothetical status: the overarching goal of WP4 is precisely to investigate these hypotheses by collecting and analyzing vocal and gestural imitations of a variety of referent sounds, and provide the project with precise and principled insights into how vocal and gestural imitations can be used in the context of sound design. These hypotheses are:

- H1. Imitating a referent sound allows the *communication* of this sound: the listener will eventually understand what referent sound the speaker intends to communicate, at some level of generality.
- H2. An imitation consists of a selection of *relevant* features of the referent sound. "Relevant" means that the features are necessary for the listener to recover the referent sounds.
- H3. An imitation *reproduces and transforms* these features within the constraints of a speaker's capabilities. These constraints are both the general constraints of the human vocal apparatus and related to idiosyncratic speakers' abilities and skills.
- H.4 Most speakers can easily control pitch, voicing, formant frequencies, amplitude modulation, at least, because these features are used in spoken languages.
- H.5 Temporal variations of these features are very important for sound recognition.
- H.6 Speaker invariance is obtained at the level of the articulatory features, not necessarily at the level of superficial acoustic features.
- H.7 Listeners can recover the referent sound, and the source of referent sound (what has produced the sound).
- H.8 The selection of features is optimized to correspond to how listeners represent and memorize the sounds.

So far, our work has focused on what can imitators reproduce when they are required to reproduce a referent sounds, and the accuracy of this reproduction (touching upon H1, H3, H4, and H5). The next studies planned in WP4 (imitations of a sound in memory, what

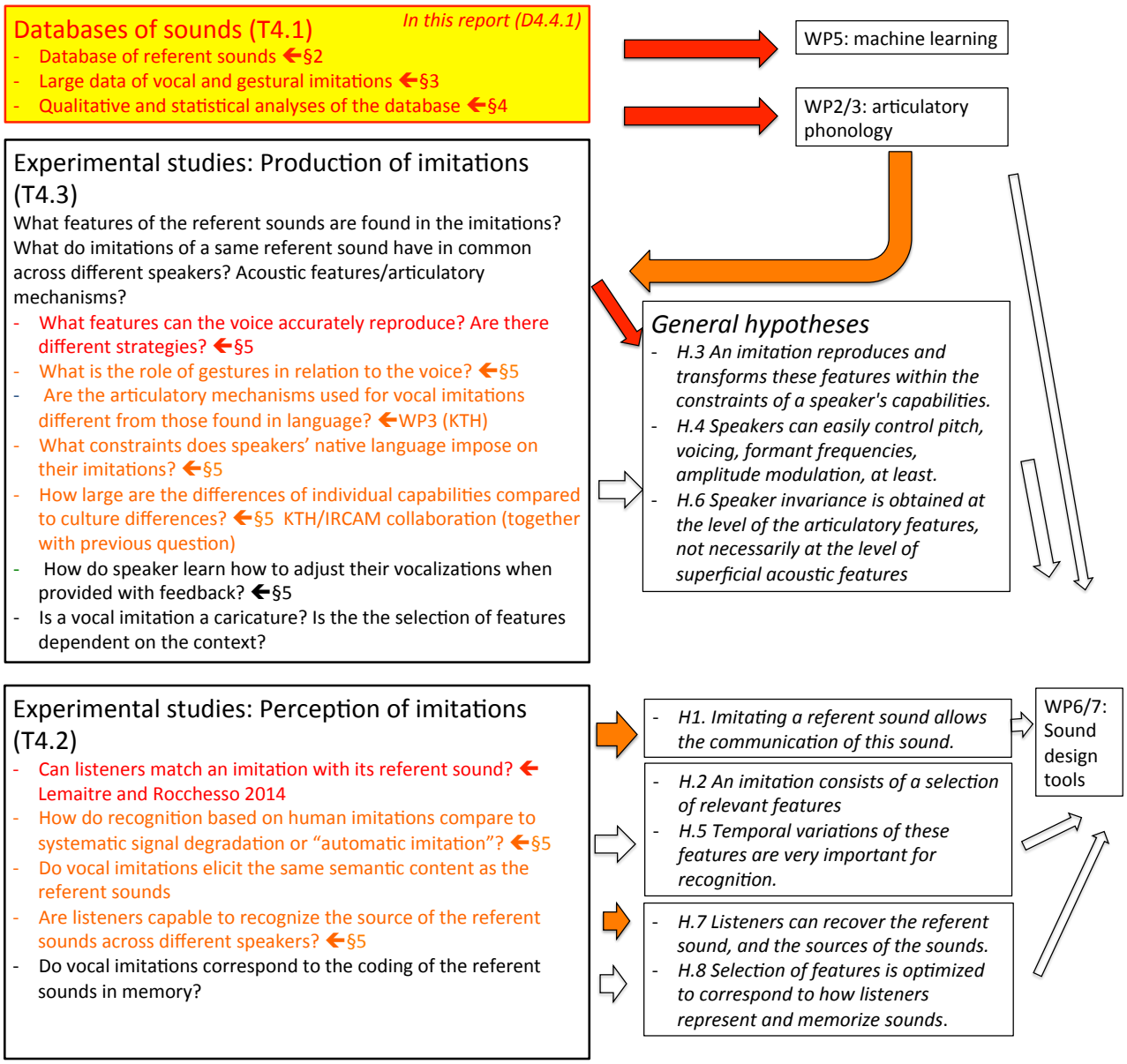

Figure 3: Questions and items studied by WP4. Items in red are completed, items in orange are works in progress, items in black will addressed during last year. Items in red are reported in this document (D4.4.1).

listeners identify from an imitation, influence of native language etc.) will address hypotheses H2, H6, H7, and H8. But certainly, and hopefully, new interesting questions will also pop up along the way. An indicative schedule of these studies is represented in Figure 4.

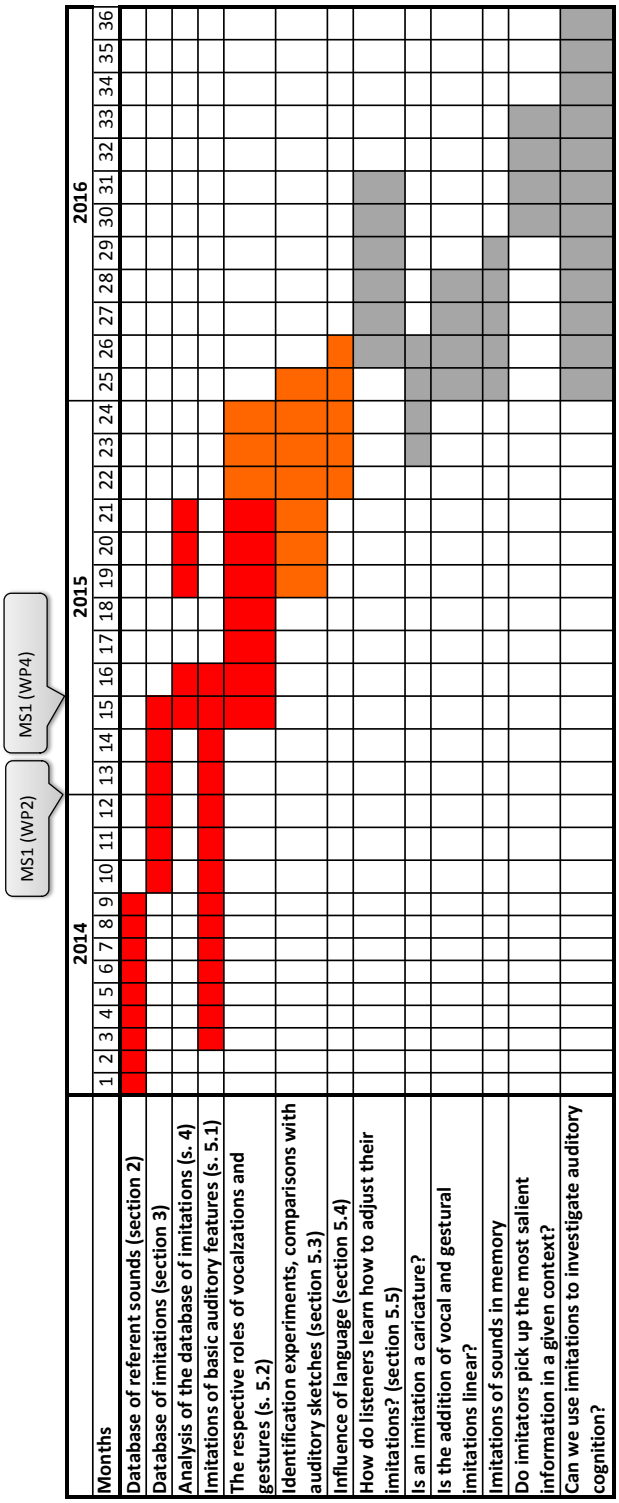

Figure 4: Indicative schedule of WP4. Items in red are completed, in orange are in progress, and in gray are planned.

2 The database of referent sounds (Task 4.1)

This section describes the methodology and the selection of a set of 52 referent sounds that have been used to elicit vocal imitations in Section 3. This section and the following include a comparison with the methods used at KTH in WP2 and WP3. In fact, the recordings made at Ircam in WP4 were mirrored at KTH in WP2 and WP3, but with different constraints and thus different methods and material.

2.1 General principles

The selection of the referent sounds was constrained by three criteria that are detailed below It must however be already noted that the goal was not to create an exhaustive ontology of sound production, but to provide the projects with a good sample of the major applications of product sound design.

2.1.1 Three criteria to select the sounds

The selection of referent ounds was constrained by three criteria:

- C1. The categories must cover the major applications of product sound design. This criterion was enforced by considering previous studies of sound design, sound quality, and with interviews with sound designers.
- C2. The categories must be perceptually relevant. This means that the sound categories used to select the sounds should be meaningful to the listeners (they correspond to stable representations in listeners) and non-overlapping. In practice, this amounts in selecting sounds that are easily and accurately categorized by the users in the selected categories. This criterion was enforced by selecting the categories on the basis of previous perceptual studies and running an identification experiment (see below).
- C3. The imitations of the sounds in the categories must be balanced as regards the vocal production mechanisms. This criterion was enforced by pilot experiments and a priori assumptions made by KTH.

2.1.2 Three points of view and three families of sounds

To reach a good coverage of the major applications of product sound design, we adopted three partially overlapping points of view on the universe of product sounds. In practice, these three points of view consisted in three families of referent sounds. Browsing through scientific publications in product sound quality highlights three types of sources: road vehicles (cars, motorcycles, buses, etc. see for instance Parizet et al., 2003, 2004, 2008; Cerrato, 2009; Lemaitre et al., 2015b), domestic appliances (refrigerators, air-conditioning, etc. see for instance Ih et al., 2003; Susini et al., 2004; Sato et al., 2007; Jeon et al., 2007; Penna Leite et al., 2009) and alarm sounds (see for instance Stanton and Edworthy, 1999; Lemaitre et al., 2007, 2009; Suied et al., 2008). The selection of referent sounds therefore used a first point of view: it sampled the sounds of a variety of industrial products, focusing on vehicles, domestic and office appliances, and alarms. This resulted in the first family: *Machine sounds*.

In consequence, this first family of sounds is defined in terms of the *products* that produce the sounds. It mainly consists of mechanical or electromechanical sounds sources, and is biased towards certain types of sound production mechanisms: combustion engines, electrical motors, gears and rotating mechanisms, etc. Such sounds have a strong tonal character, and pilot experiments showed that participants imitate them with some kind of myoelastic vibration (vocal folds, etc.). To balance the variety of the articulatory mechanisms potentially elicited by these sounds (criterion C3), we also selected a number of product sounds with no tonal character: air-conditioning units, sounds of non-motorized tools, etc.

However, nothing guarantees that these categories of sound products are perceptually relevant, or even discriminable (criterion C2). For instance, without any context, it is very hard to distinguish between the sounds of an air-conditioning unit and the sound of car engine inside a car running at fast and constant speed. Özcan and van Egmond (2007, 2012) have studied how listeners categorize and memorize product sounds. Their results have singled out categories that are not defined in terms of products, but in terms of the basic phenomena at the source of the sounds: air, impacts, liquids, motors, etc.. Thus, the selection of referent sounds adopted a second point of view and a second family of referent sounds, based on the *basic mechanical interactions* that produce the sounds. It sampled trough sounds produced by the interaction of solid objects, liquids and gases and used the categories resulting from the classification experiments conducted by Lemaitre et al. (2010); Houix et al. (2012). Furthermore, the selection also focused on sounds produced by human gestures (e.g. hitting, tapping, scraping), as Lemaitre and Heller (2013) have shown that this forms the basic level of the cognitive representations of action sounds across listeners.

The two first families mostly include sounds produced by mechanical sources, or, at least, sounds that listeners identify as produced by some mechanical phenomenon. Thus, we also selected a third family consisting of sounds that listeners cannot associate with any mechanical source: abstract sounds. Such sounds are typically produced by sound synthesis and are very important for video games and and human-machine interfaces (Brewster, 2009). Few studies of how listeners categorize such sounds are however available. The most exhaustive work on this topic was conducted by Schaeffer (1966) (see also Chion, 1983) in the context of electroacoustic music. This work has however no experimental validation and is much too detailed for our purpose. Our selection was based on a somewhat simpler taxonomy of dynamic profiles (morphologies) proposed by Peeters and Deruty (2010). Since, the concept of morphology applies to any kind of sounds, we also took care that the selections of machine sounds and mechanical interactions are also balanced following these morphologies.

In summary, the selection of referent sounds consisted of the three following families:

- **Sounds of machines**. This family corresponds to a sampling of products whose sounds are or may be designed by sound designers and thus correspond to the potential applications of the SkAT-VG tools. The different sounds in these categories are therefore organized chosen after discussing with Genesis, who selected them according to their potential for sound design.

- **Basic mechanical interactions**. This family corresponds to the taxonomy of sounds established by Lemaitre et al. (2010), Houix et al. (2012), and Lemaitre and Heller (2013). In short, this work outlines categories of basic mechanical interactions (sounds

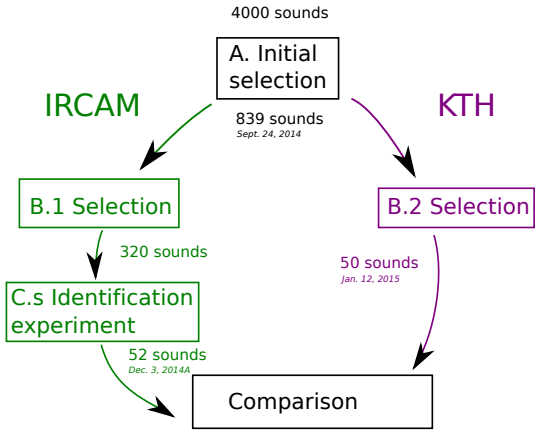

Figure 5: The different steps of the selection of referent sounds.

produced by the interaction between two objects or by the deformation of one object) that are represented in stable cognitive categories across

- **Abstract sounds**. Whereas the two former families correspond to sounds produced by a mechanical phenomenon, the family of abstract sounds includes artificial sounds that listeners cannot identify as produced by any mechanical phenomenon. The sounds were recorded from human computer interfaces (mobile phones, video games, computer operating systems) or synthesized. The categories within this family were inspired by Peeters and Deruty (2010).

It is important to note that the three families are not mutually exclusive. For instance the sound of hammering can be categorized as a tool in the family of machine sounds or as a basic hitting interaction. Rather, they represent three different, overlapping sets.

2.1.3 The different steps of the selection procedure

The selection of categories and sounds resulted from an iterative process whose steps are described in the following section:

- A. First selection of 839 sounds in the three families, organized in coarse categories, based on discussions with Genesis.
- B.1 (IRCAM). Selection of a subset of 320 sounds organized in 32 categories selected after discussions with Genesis and KTH (criterion C1 and C3)
- B.2 (KTH). Refinement of the categories and selection of a subset of 50 sounds organized in the categories of potential articulatory mechanisms (criterion C3).
- C. (IRCAM). Selection of a subset of 52 sounds organized in 26 categories on the basis of an identification experiment (criterion C2).

These different steps are represented in Figure 5.

2.2 Hand-made selections (Steps A and B)

2.2.1 Initial coarse selection (A)

We first selected 839 sounds from an initial selection of about 4000 sounds (delivered to the consortium by ftp on September 24, 2014). This selection is represented in Figure 6. Note that the definition of categories was only indicative at this stage.

Sounds and machines and mechanical interactions The sounds and machines and mechanical interactions were selected in commercial and freely available databases (Hollywood Edge, Blue Box, Sound Ideas, Freesound, etc.). This selection was mainly based on indications by Genesis about potential applications for the SkAT-VG tool.

Abstract sounds The abstract sounds were first selected from three main sources: mobile phone video games, and user interfaces of a number of devices:

- The video games were purchased at the Google Play site for Android smartphones. We chose 44 games across different categories (reflexion, arcade, adventure, strategy, simulation, sport, role playing, etc) in order to scan a wide variety of sounds. We played the different games on a smartphone Nexus S i9023 with the operating system Android 4.3.1 / CyanogenMod 10.2.1-crespo. We adjusted the sound settings with the application Woodoo Sound 3.1.2.2. We recorded the game sounds using the headphone output connected to a soundcard RME 400 directly within the software SoundStudio 4.6.12 on OS X 10.9.5.
- The UI sounds came from different sources: Linux (KDE desktop and Ubuntu), Windows and Mac OSX operating systems, Android, WindowPhone and iOS operating systems (sounds of the user interface), devices (Apple TV, Jambox, Tivo, Wii) and applications (Skype, Facebook) but also IHM from Renault trucks (Genesis database).

We first set aside vocalizations (especially for games) and also musical melodies. We then used categories defined by the morphological profiles from the Schaeffer (1966): resonant and friction grains, trame, iterative, impulse and complex. A first selection consisted of choosing examples in each of the categories. A second sorting used categories defined by the simpler profiles of Peeters and Deruty (2010): increasing, decreasing, increasing-decreasing, decreasing-increasing, modulated, stable, repeated, and impulse profiles.

Finally, we completed some categories by synthesizing sounds with Cecilia5 (`http://ajaxsoundstudio.com/software/cecilia/`). We used additive, pulsar and different granular synthesis by automating the parameters (fundamental frequency, base pitch of the grains, resonant frequency, etc.) to match the different morphological profiles.

A first selection resulted in 173 sounds.

2.2.2 IRCAM's refined selection (B.1)

IRCAM refined the selection by discussing with Genesis to select only the most relevant categories of applications (criterion C3), and the best exemplars in each category. The selection of sounds was made by informal listening sessions and by pilot testing. This resulted in updating some of the categories. In particular, the categories used for the family of abstract sounds were completely changed.

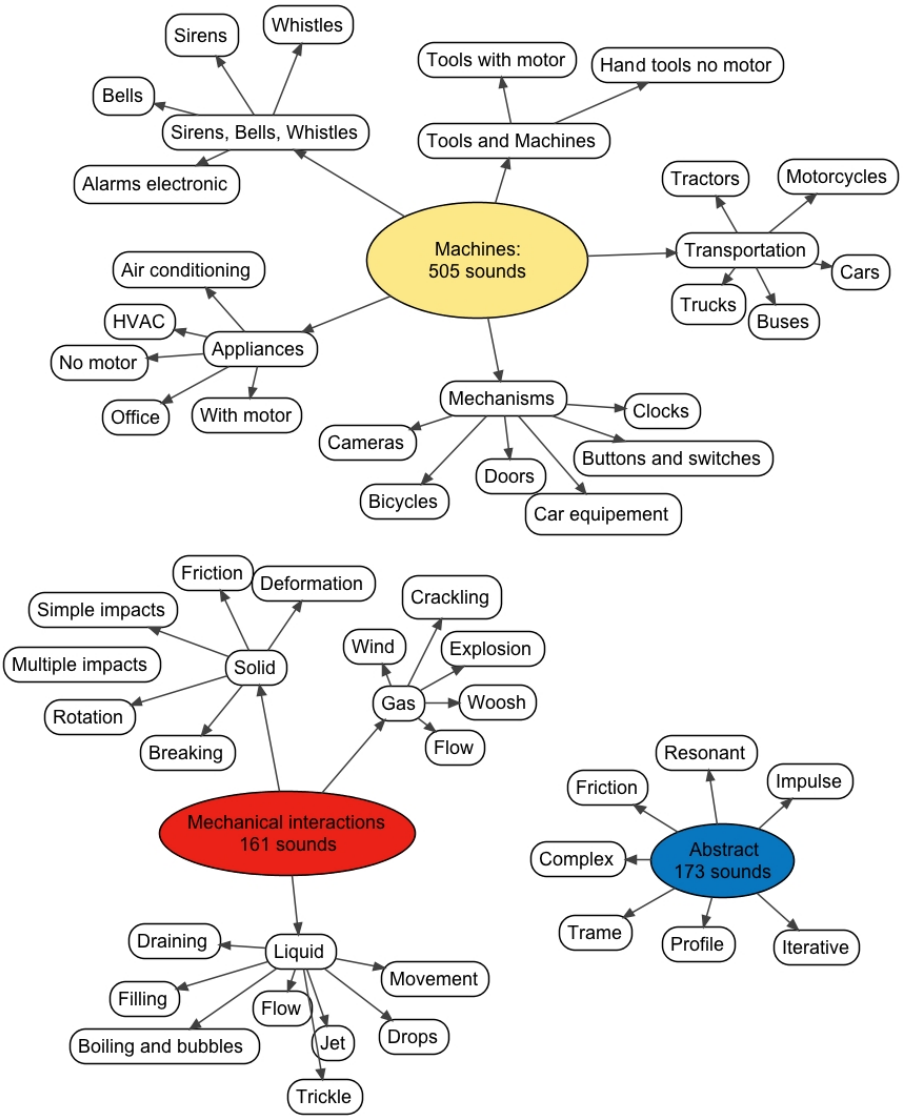

Figure 6: The initial selection (A) of 839 sounds.

This resulted in a selection of 320 sounds and 32 categories (10 sounds in each category). There were 12 categories of machines, 12 categories of mechanical interactions and 8 categories of abstract sounds.

2.2.3 KTH's refined selection (B.2)

KTH selected 50 sounds from the initial 839 sounds, as well as 10 animal sounds. The criterion used to select the sounds was criterion C3. The imitations of these sounds should be balanced as regards the articulatory mechanisms. The balance was done using the following 10 categories: affricate-like (6 sounds), dynamic friction noise (4 sounds), stable friction noise (5 sounds), slow (supralaryngeal) myoelastic vibration (6 sounds), stops, clicks and percussives (6 sounds), intermittent stops (4 sounds), supraglottal laryngeal vibration (4 sounds), dynamic voicing (6 sounds), stable voicing (5 sounds), whistling (4 sounds). This selection was shared on OwnCloud on January 12, 2015.

2.3 Selection based on an identification experiment (C).

Whereas previous selection steps were based on informal listening sessions and discussion, the final step involved a formal experiment with hired participants. The aim of this experiment was to select, for each family and associated categories, the best identified two sounds. These sounds can be considered as "prototypes" of their category.

2.3.1 Methods

Stimuli We used the 320 selected sounds (B.2) from the 3 families (80 abstract sounds, 120 machines, and 120 mechanical interactions) and their associated categories. The experimenters set the level of each sound individually, by agreeing on a level that they deemed acceptable and consistent with the sound source. Levels averaged across sound duration varied from 34 to 78 dB SPL for the machine sounds, from to 46 to 77 dB SPL for the mechanical interactions, and from 55 to 79 dB SPL for the abstract sounds.

Participants Twenty-four participants (13 women and 11 men) volunteered as listeners and were paid for their participation. The participants were between 19 and 48 years old of age (median: 28 years old). All reported normal hearing and were native French speakers. The participants were non-expert or lay participants (no professional musician or sound engineer, etc.). Due to a crash during the experiments, the analyses only used the results of 22 participants for the machine sounds, and 23 participants for the mechanical interactions. Five expert participants (sound engineers, computer music producer, psychoacoustician) also did the experiment for the abstract sounds.

Apparatus The sounds were played by a Macintosh Mac Pro (Mac OS X v10.6.8) workstation with a RME Fire Firewire 800 sound card. The stimuli were amplified over a pair of Yamaha MSP5 loudspeakers. Participants were seated in a double-walled sound isolation booth. The software used to run the experiment and to implement the graphical interface was Matlab 2008b using the Psychotoolbox 3.0.8.

Procedure The procedure was an n-alternative forced choice. It had three parties, corresponding to three families of 320 sounds (120 machine sounds, 120 mechanical interactions, 80 abstract sounds). For each family (abstract, mechanical interactions and machines), a list of descriptions was proposed corresponding to the different categories. We asked participant to simply indicate the most appropriate description for each sound. They selected the different descriptions using the "up"arrow and "down" key. The descriptions for the machines sounds were the following (in French):

- "Une alarme électronique, un buzzer, une sonnerie" (an electronic alarm, buzzer, or bell)
- "Une personne qui appuie sur un interrupteur, un bouton ou une touche" (someone presses a switch, a button, or a key)
- "Une porte qui se ferme" (closing a door")
- "Une personne qui scie ou lime ou objet à la main" (hand sawing of filing an object)
- "Le bruit du réfrigérateur en marche" (the noise of the refrigerator)
- "Une ventilation, l'air conditionné qui fonctionne" (the noise of the air conditioning)
- "Un robot ménager, un mixeur, un hachoir électrique" (electric food processor, mixer, or grinder)
- "Une imprimante ou fax qui imprime des pages" (a printer or a fax printing out pages)
- "Un essuie-glace qui fonctionne" (windshield wiper)
- "Une voiture ou une moto qui vous passe devant" (a car or a motorcycle passing by)
- "A l'extérieur d'une voiture ou une moto qui rugit, à l'arrêt" (outside a car or a motorcycle revving up)
- "A l'intérieur d'une voiture qui accélère" (inside a car accelerating)

The descriptions for the mechanical interactions were the following (in French):'

- "Souffler, expirer" (blowing)
- "Fouetter, cingler dans l'air" (whipping)
- "Tirer avec une arme à feu, une explosion" (shooting)
- "Déformer, écraser, froisser un objet" (crumpling)
- "Un objet qui roule sur une surface" (rolling on a surface)
- "Grincer, crisser, couinement" (squeaking)
- "Gratter, racler, frotter un objet" (scraping)
- "Frapper, taper, heurter, cogner un objet" (hitting)
- "Une ou plusieurs gouttes qui tombent" (dripping)
- "Eclabousser, asperger" (splattering)
- "Remplir un récipient avec un liquide" (filling a recipient with liquid)
- "De l'eau qui coule, un jet d'eau" (gushing)

The descriptions for the abstract sounds were the following (in French):

- "Un son qui monte" (rising)
- "Un son qui descend" (decreasing)
- "Un son qui monte puis descend" (rising and then decreasing)
- "Un son qui descend puis monte" (decreasing and then rising)
- "Un son impulsif, très court" (an impulsive sounds)
- "Un son constitué de répétitions d'éléments très courts" (repetition of very short elements)

- "Un son qui fluctue régulièrement" (a regularly fluctuating sound)
- "Un son stable, stationnaire" (stable, stationnary)

2.3.2 Results and selection

Analyses were based on confusion matrices and hit rates for each sound. The goal of the selection was to select a maximum of 10 categories for each family, and two sounds per category. We selected categories for which there was no systematic confusion overall with another category. For each category, we selected the two sounds with the best hit rates. In case of ties, we selected two sounds that were the most different to maximize the variety of the selection. In rare cases when the two sounds with the best hit rates were too similar, we selected a second sound with a slightly lower hit rate. In every case, we also tried to balance the morphologies. We distinguished between continuous and discrete sounds, and then we distinguished between single impulse, repeated discrete sounds, continuous and stable sounds, continuous and dynamic or complex sounds. We reasoned that these morphological aspects would be the easiest thing to reproduce with voice, and thus the selection should be balanced according to these criteria. Another aspect that is particularly important is the presence and the strength of tonal components, since we expected that it would determine whether speakers use voiced or unvoiced production mode. However, the tonalness of the sounds is not orthogonal to the types of sounds: mechanical interactions are mostly made of noises, and sounds of machines have strong tonal components, since they very often include engines, electrical motors, and other rotating elements.

Machines The top panel of Figure 7 represents the confusion matrix for the machine sounds. Overall, identification was precise for these sounds, with hit rates close to 100%. Confusion mainly occurred between the sounds of fridges and HVACs on the one hand, and between the three categories of vehicle sounds on the other hand. We decided to remove the category of HVAC sounds because discussions with Genesis indicated that they were less interested by these sounds. We also decided to remove the category of vehicles passing by: these sounds have a strong stereo effect, which is not the case for any other category. We therefore decided to leave out the particular case of spatialized sounds. Figure 8 represents the final selection of machine sounds together with the hit rates of each sound.

This selection has also the advantage that it balances different morphological aspects of sounds: impulses, repetitions, continuous-stable morphologies (i.e. quasi stationary), continuous-dynamic sounds, and sounds composed of a complex mixture of events. It also balances sounds with mainly made of tonal components, sounds without tonal components, and sounds mixing both types of components. Table 1 summarizes this selection.

Mechanical interactions The middle panel of Figure 7 represents the confusion matrix for the mechanical interactions. Overall, hit rates were fairly good, yet weaker than for the machine sounds.

Most confusion occurred between scraping and squeaking sounds. We decided to keep the scraping category and leave out the squeaking sounds, since scraping is an important component of Iuav's developments. Many splattering sounds were also confounded with other liquid sounds. We therefore decided to leave this category out of the selection. Figure 9

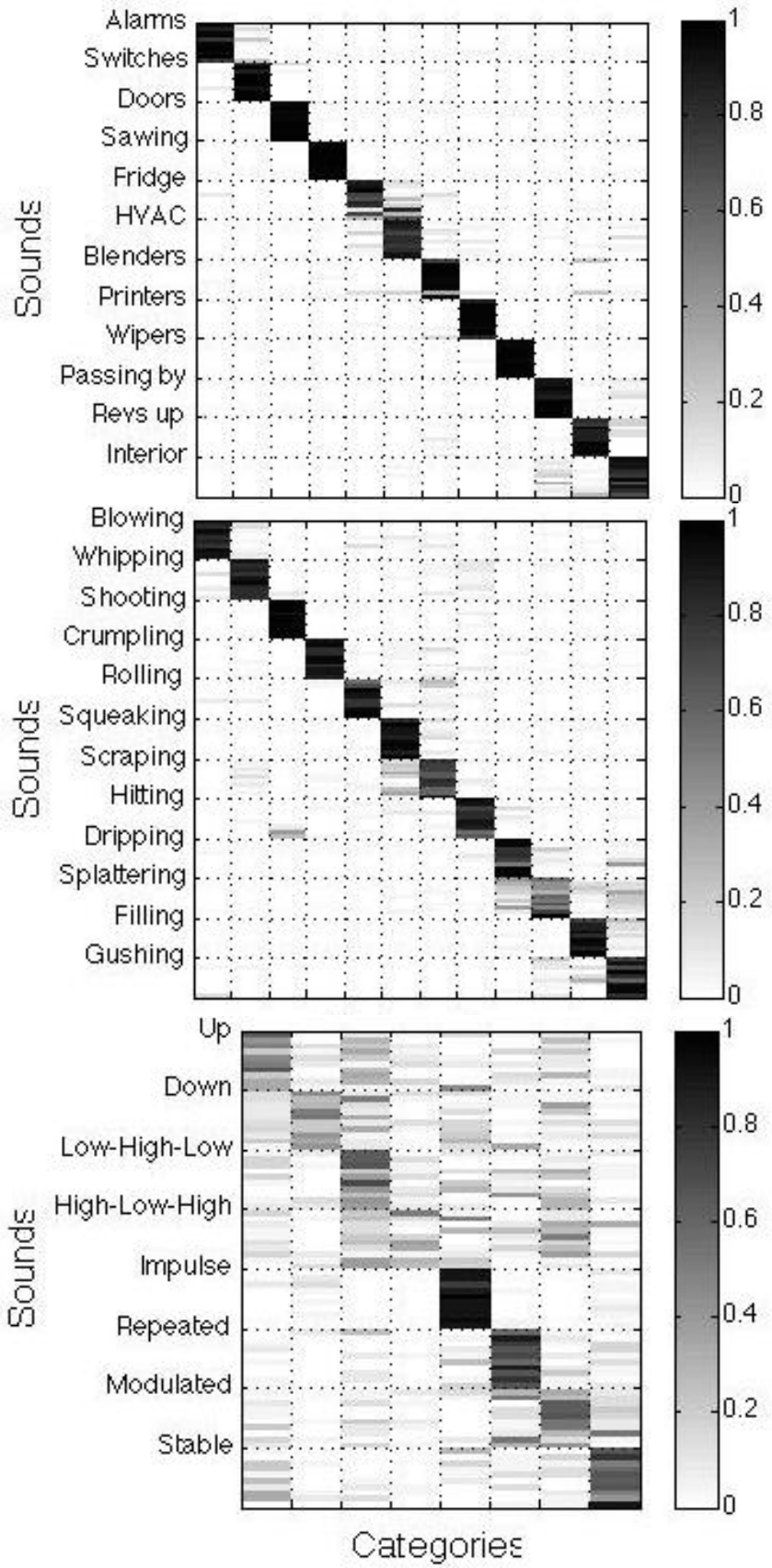

Figure 7: Confusion matrix for the machine, interaction, and abstract sounds.

Category	Sound	Short name	Hit rate
Alarms	Machines#SirensBellsHornsWhistles_AlarmsElectronic#_#bb_ED#6-1_electric_alarm_clock_multi	Machine02	100%
	Machines#SirensBellsHornsWhistles_AlarmsElectronic#Alarm_Burglar__BurglarAlarmGoingOff#si_ED#si_10_04-1	Machine01	100%
Buttons and switches	Machines#Mechanisms_ButtonsAndSwitches#si#si_28_15-6	Machine04	100%
	Machines#Mechanisms_ButtonsAndSwitches#FromToyo_Stylos#Gene#Espace13	Machine03	100%
Doors	Machines#Mechanisms_Doors#Doors_Close_Cars#ss#CARDOOR1	Machine06	100%
	Machines#Mechanisms_Doors#Doors_Close#he_EDITED#18_Heavy Screen Door 1	Machine05	100%
Hand tools	Machines#ToolsAndMachines_HandToolsNoMotor#_#bb_ED#2-2_saw_long	Machine08	100%
	Machines#ToolsAndMachines_HandToolsNoMotor#_#bb#5-8_saw	Machine07	100%
Fridges	Machines#Appliances_WithMotor#Fridge#ss_ED#FRIDGE 1	Machine10	100%
	Machines#Appliances_WithMotor#Fridge#FreeSound_ED#175243_rhythmdriver_fridge-running	Machine09	91%
Blenders and mixers	Machines#Appliances_WithMotor#Mixer_Hand_Electric_StartRunStopLowSpeed#si_ED#si_21_29-1	Machine12	100%
	Machines#Appliances_WithMotor#Mixer_Hand_Electric_MixInBowlHighSpeed#si_ED#si_21_30-2	Machine11	100%
Photocopiers	Machines#Appliances_Office#Photocopier__Med_FourOriginalsSixCopiesEachSortInTrays#si_ED#si_28_79-1	Machine13	100%
	Machines#Appliances_Office#Printer_LaserPrinter#si_ED#si_28_47-1	Machine14	100%
Windshield wipers	Machines#WindshieldWipers#_#si#si_07_84-1	Machine15	100%
	Machines#WindshieldWipers#_#si#si_08_54-2	Machine16	100%
Revs up	Machines#RoadTransportation_Cars_ExteriorStatic#Volume29#he_EDITED#02_Cobra	Machine17	100%
	Machines#RoadTransportation_Motorcycles_Static#Motorcycle_Suz1100_StationaryRevsShort_LongGroupOfFourRear#si_EDITED#si_26_29-1	Machine18	100%
Car interior	Machines#RoadTransportation_Cars_Interior#Auto_90FordEscort_Int_StartIdleAccelerateQuicklyToHighSpeed#si_ED#si_05_65-1.wav	Machine19	95%
	Machines#RoadTransportation_Cars_Interior#FromRenaultDM#Gene_ED#BMW 330 CI	Machine20	95%

Figure 8: Final selections of machine sounds together with the hit rates of each sound.

Morphology		Category
Discrete	Impulsive	Buttons and switches; Doors closing
	Repeated	Alarms; Sawing and filing; Windshield wipers
Continuous	Stable	Fridge hums; Blenders
	Dynamic/complex	Vehicle interior; Revs up; Printers

Table 1: The selection of machine sounds, classified in morphologies. Items in blue are noisy, items in red have strong tonal components; items in purple mix noisy and tonal components.

represents the selection of mechanical interactions, together with their hit rates. Table 2 shows that most interaction sounds are noisy signals (in blue). Conversely to the selection of machine sounds, the selection of mechanical is not balanced into different sound morphologies.

Morphology		Category
Discrete	Impulsive	Whipping; Shooting; Hitting 1, Hitting 2, Scraping 1
	Repeated	Scraping 2; Dripping
Continuous	Stable	Blowing; Gushing
	Dynamic/complex	Filling; Crumpling; Rolling

Table 2: The selection of mechanical interaction sounds, classified in morphologies. Items in blue are noisy, items in red have strong tonal components; items in purple mix noisy and tonal components.

Abstract sounds The bottom panel of Figure 7 represents the confusion matrix for the abstract sounds. Hit rates were overall much weaker for this family than for the other two families. To guide our choices, we therefore also ran five expert participants on the same sounds (results not reported here). All our decisions were based on a comparison between the confusion matrices for the lay participants and the experts.

The category of "modulated sounds" was the most problematic category: several participants indicated that they used this category as a "garbage category". When they did not know how to categorize a sound, they put it in the modulated category. The "high-low-high" category was also problematic, since very few sounds were categorized in this category. We therefore decided to remove these two categories, resulting in a total of six selected categories. Table 3 reports these categories and their morphologies.

Figure 10 represents the selection of abstract sounds, together with their hit rates.

2.4 Conclusion and comparison with KTH's selection

This selection procedure described above resulted in two selections: three families, 26 categories and 52 sounds (2 sounds in each category) for IRCAM, and 50 sounds in the three families for KTH (plus 10 animal sounds). Ircam selection included 10 machine sounds, 10 sounds of mechanical interactions and 6 abstract sounds in IRCAM's selection. The procedure insures that these selections provide a good coverage of products whose sounds may potentially

Blowing	Mi#Gas_Wind#_#he#22_Medium Eerie wind, shutter bangs	Interaction02	96%
	Mi#Gas_Wind#GaverTaxonomy#ALab#02_blowing2	Interaction01	96%
Whipping	Mi#Gas_Whipping#_#FS_ED#243468_mark-ian_whoop	Interaction03	91%
	Mi#Gas_Whipping#_#FS_ED#51755_erkanozan_whip-01	Interaction04	96%
Shooting	Mi#Gas_Explosion#_#ALab#Shooting2	Interaction05	100%
	Mi#Gas_Explosion#_#he#10_KentuckyRifle	Interaction06	100%
Crumpling	Mi#Solid_Continuous_Deformation_Rigid_material_crushing#can_crush#si#6020_24-1	Interaction08	96%
	Mi#Solid_Continuous_Deformation_Crumpling #ALab#Crumpling5_65	Interaction07	96%
Rolling	Mi#Solid_Continuous_Rolling#PipeRollingDownARamp #ALab#Rolling_Wood_Small_067	Interaction09	91%
	Mi#Solid_Continuous_Rolling#_#FreeSound#186965_robinhood76_01433-rolling-metal-piece-1	Interaction10	100%
Scraping	Mi#Solid_Continuous_Friction_Rigid material_Rubbing#sanding#si_ED#6012_18-1.wav	Interaction12	96%
	Mi#Solid_Continuous_Friction_Rigid material_Rubbing#_#ALab#Scraping_Plastic_Big_065.wav	Interaction11	83%
Hitting	Machines#SirensBellsHornsWhistles_Bells#_#he#09_Bell ring, boxing 2	Interaction13	96%
	Mi#Solid_Discret_Simple impact#_#ALabTapping4_79	Interaction14	91%
Dripping	Mi#Liquid_Discret_Drop#waterdrip#si_ED#si_21_69-2	Interaction15	100%
	Mi#Liquid_Discret_Drop#waterdrip#si_ED#si_39_82-1	Interaction16	100%
Filling	Mi#Liquid_Continuous_Filling a small container#_#ss#64-WINE_BOTTLE-OF-WINE-4	Interaction17	96%
	Mi#Liquid_Continuous_Filling#_#si_ED#si_20_66-4	Interaction18	100%
Gushing	Mi#Liquid_Continuous_Jet#_#apsel_ED#AP12-Household,-Clocks,-Tools_56-Water-Bucket	Interaction19	100%
	Mi#Liquid_Continuous_Jet#water_sink#si_ED#6021_76-3	Interaction20	96%

Figure 9: Final selections of sounds of basic mechanical interactions together with the hit rates of each sound.

Up	AdditiveSynth_up1_cecilia5.wav	Abstract01	62%
	abstract#game#_#worms3#rocketpowerup.ogg	Abstract02	67%
Down	abstract#game#_#superbrothers#Slow down1	Abstract04	54%
	abstract#game#_#clash#pekka_death_06	Abstract03	50%
Low-High-Low	abstract#game#_#tinythief#Robot_MagnetLoop-1	Abstract05	67%
	abstract#hci#icad2003#pds_ircam#Alarm_03+	Abstract06	71%
Impulse	abstract#hci#hci_alarms_renault_truck#Gene#alarme1_pulse	Abstract07	100%
	abstract#os#notifications#android#Plastic_Pipe.ogg	Abstract08	92%
Repeated	abstract#hci#icad2003#pds_ircam#Alarm_01	Abstract09	87%
	abstract#os#camera#android#selftimer_2sec_.ogg	Abstract10	87%
Stable	FeedbackLooper_glass_cecilia5_001	Abstract11	79%
	abstract#game#_#tinythief#Robot_MagnetMove.gran1	Abstract12	92%

Figure 10: Final selections of sounds of abstract sounds with the hit rates of each sound.

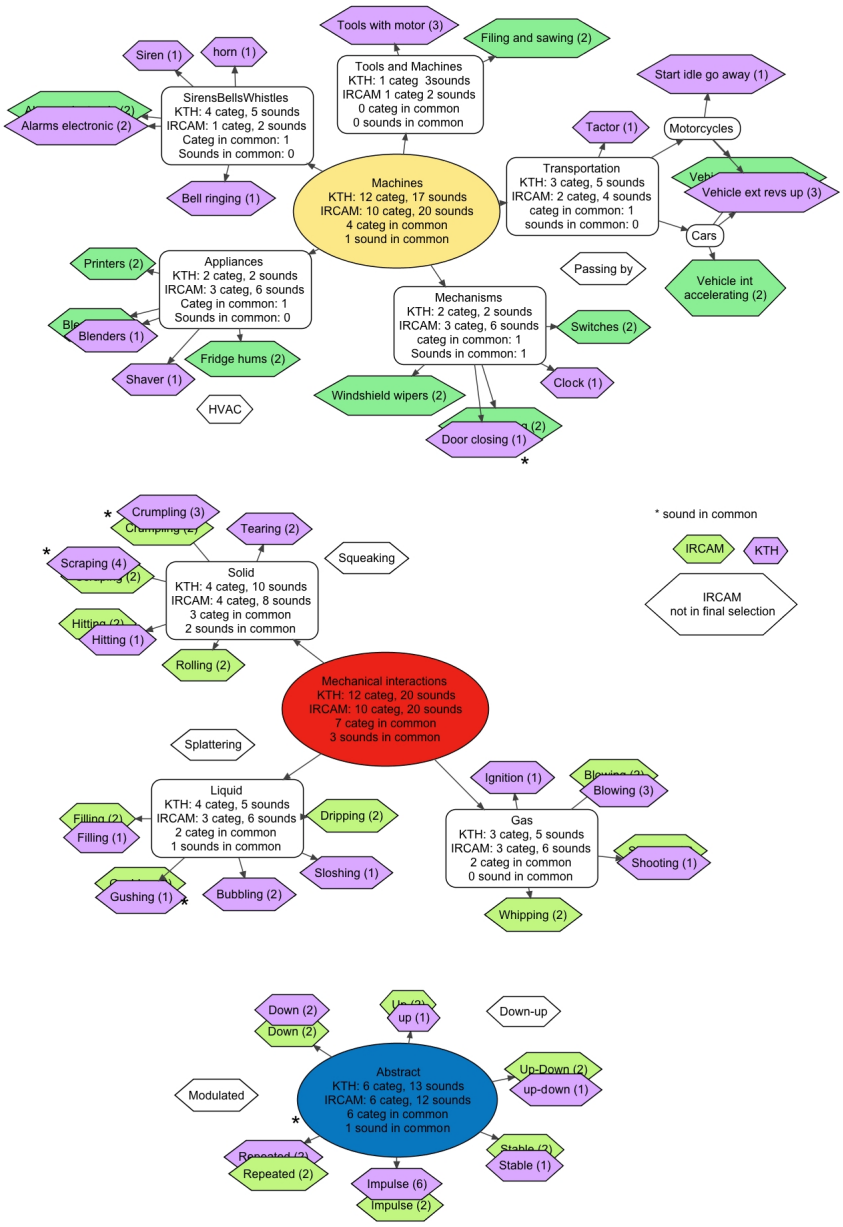

Figure 11: The final selection of referent sounds in the three families. Green boxes correspond to IRCAM's selection. Purple boxes correspond to KTH's selection. White boxes correspond to categories that were dropped after the identification experiment. Overlapping boxes correspond to categories common to KTH's and IRCAM's selections. Stars indicate sounds common to KTH's and IRCAM's selection.

	Morphology	Category
Discrete	Impulsive	Impulse
	Repeated	Repeated
Continuous	Continuous-stable	Stable 1; Stable 2
	Continuous-dynamic/complex	Up; Down; Up-Down 1; Up-Down2

Table 3: The selection of abstract sounds, classified in morphologies. Items in blue are noisy, items in red have strong tonal components; items in purple mix noisy and tonal components.

be designed (criterion C1), and of categories that are perceptually relevant (criterion C2). It also tried to balance different articulatory mechanisms (criterion C3). These selections are represented in Figure 5. IRCAM's selection was shared on Owncloud on December 3, 2014. In particular, Iuav used these categories to refine the SDT tools: the goal is that the tools developed in WP6 and 7 can synthesize these sounds.

Figure 11 represents the categories selected by Ircam and KTH. Figure 12 represents a schematic view of the intersection of these two selections. Note that the sounds selected by KTH were categorized by informal listening. This intersection is further detailed in the following table. What is important to note is that, in total, 9 categories of sounds selected by IRCAM are not included in KTH's selection, and there are only 5 sounds in common. The divergence between IRCAM's and KTH's selection resulted from two distinct objectives: cover the potential applications of sound design and cognitive categories of sounds for IRCAM, balance between potentially elicited mechanisms for KTH.

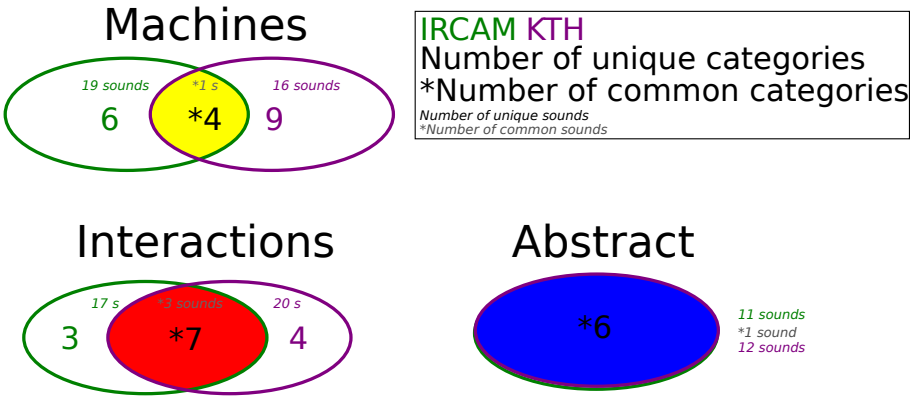

Figure 12: Comparing IRCAM's and KTH's selections.

3 The database of vocal and gestural imitations (Task 4.2)

This part describes the methodology and the recordings of the vocal and gestural imitations of the referent sounds with 50 participants. Whereas there exists a similar attempt at collecting a large amount of imitations of sounds using on-line procedures (Cartwright and Pardo, 2015), our approach has two innovative aspects. First, it controlled precisely the recording procedure and the quality of the recordings. Second, it also used high speed video recordings and accelerometers and we recorded imitators using expressive gestures. In fact, we had already observed in previous work that imitators use a lot of expressive gestures when vocally imitating sounds (Lemaitre et al., 2013). Therefore, we wanted to able to compare vocal imitations with and without gestures to study the role of gestures.

3.1 Recording setup and procedure

Overall, the procedure consisted for the participants to listen to each of the 52 referent sounds and record a vocal or a gestural imitation.

Participants Fifty participants (21 male, 29 female) aged from 18 to 47 (average age 28.3 years old) took part to the recording sessions. All reported normal hearing and were native speakers of French. None of them had received formal training in music, audio, dance, or theater, expect one person who was a professional actress.

Procedure Figure 13 represents the structure of a recording session. The recording session had two *conditions*: participants recorded vocal imitations (i.e. no gesture) in the first part (V condition), and vocal and gestural imitations in the second part (V+G condition)[2][3]. There were three blocks for each condition (i.e. six blocks in total), each block corresponding to one of the three families of referent sounds (machines, mechanical interactions, and abstract sounds). Participants always began with the V condition followed by the V+G condition, but the order of the three families within each condition was randomized for each participant.

For each condition, participants used an custom-made Max/MSP user interface (represented in Figure 14, see below for technical details). The interface consisted of a number of cells (20 for the machines and the mechanical interactions, 12 for the abstract sounds), with each cell corresponding to one referent sound. Each cell allowed the participants to listen to the referent sound, record and play back an imitation (audio playback in part V, audio video playback of the webcam in part V+G part, see below). There was a limit of five trials for each recording. Participants could record an imitation only if they had listened to the referent sound at least once.

Participants were autonomous during the experiment to enable maximum creativity without being intimidated by the presence of the experimenter. They were instructed to provide an imitation in such a way that someone listening and watching them would be able to identify

[2]Participant 01 did not do the V+G condition

[3]Pilot experiments suggested that a gesture-only condition made little sense to the participants.

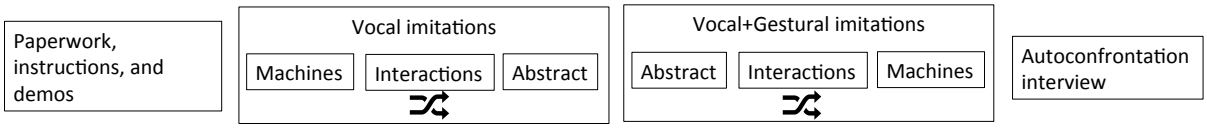

Figure 13: Structure of the recording sessions.

the sounds within the family. Participants were instructed not to use any conventional onomatopoeia. The order of the sounds on the interface was randomized for each participant. The experimental interface presented all referent sounds and imitations on the same interface, so that participants could compare the different sounds. The participants were strongly encouraged to compare and evaluate the quality of their imitations, and to compare their imitations with the referent sounds.

We changed and adjusted the instructions for the V+G condition over the first few participants, based on how they behaved. Initially, the instructions did not specify what type of gestures the participants should perform ("free protocol"). Qualitative analysis of the first results revealed that many subjects actually *mimed* the situation in which they thought the referent sound was produced[4]. Since this is not what we expected, we modified the instructions for subsequent subjects, and specifically instructed them to describe the referent sound itself ("directed protocol"). Fifteen participants (01 to 07, 09, 12, 13, 15, 17, 19, 21 23) used the free protocol. The remaining 35 used the directed protocol.

Before the actual recording sessions, participants signed a consent form to participate to the recording session and an authorization to exploit the audio/video recordings of their performance[5]. Then, they read the written instructions that the experimenter further repeated. The experimenter then demonstrated how to use the interface, for the V condition only (no gesture nor use of video playback at this stage of the procedure).

After the recording session, the experimenter reviewed all the recordings with the participants and the participants were invited to comment on their video and explain their strategy (autoconfrontration). The autoconfrontation interview was recorded.

The sessions lasted on average three hours and the participants were compensated 30 euros.

Setup Participants were seated in a double-walled IAC sound isolated booth. The setup was both located inside and outside the booth (see Figure 15). Basically, the setup consisted of a microphone and an audio interface to record the vocal imitations, a fast camera (Gopro) providing a close-up video recording of the participants' face (these video recordings are especially important for the phonological analyses at KTH), motion sensors and a depth camera for the gesture analyses, and a webcam for monitoring purposes.

[4]For instance, one participant imagined that a broadband referent sound was produced by static noise on a TV set. Thus, she outlined a living room and a TV set with her hands and mimed the action of tuning up an old TV receiver

[5]The data were anonymized: each participant was identified by a code, and the correspond between codes and participants' identity is secured in a separate file.

Figure 14: Interface for the recording of imitations (part V+G in this example)

In more detail, the setup inside the booth consisted of:

- A computer display presenting the user interface and controls,
- A computer mouse on a raised stand located next to the participant's dominant hand,
- A pair of studio monitors (Yamaha MSP5) ;
- A microphone headset (DPA d:fine omni)
- Two inertial motion units (Ircam's "Musical Objects": MO), fixed on the participants' wrists; Each IMU contains 3D accelerometers and 3-axis gyroscopes and transmits the data wirelessly with a latency of about 5-10 ms (Rasamimanana et al., 2011),
- A depth camera (Microsoft Kinect, v. 2),
- A high-definition video camera (Gopro Hero4),
- A low-definition webcam (Logitech HD1080p) ;
- A LED located in the cameras' field of view (to be used as a synchronization signal),
- A micro-controller (Arduino) controlling the LED,
- An Apple MacMini desktop computer (Intel Quadcore 3.2 GHz) running MacOS 10.9, with SSD mass memory (so as to be as silent as possible) controlling the Gopro camera through wifi and the Arduino board through USB;

In addition, the setup outside the booth consisted of:

- A desktop computer (Apple Mac Pro with Intel Dualcore 2.6 GHz, running MacOS 10.6.8), controlling the user interface, audio recording and playback, and the webcam (it also displayed video monitoring to the experimenter outside the booth via the webcam),
- An audio interface (RME Fireface 800) connected to the Mac Pro,
- A desktop computer (Asus PC with Intel Quadcore 3.2 GHz, running Microsoft Win-

dows 8), for motion capture (gesture and motion sensors data-recordings from the Kinect and the IMU sensors),
- An pair of headphones for audio monitoring
- A desktop computer (Apple Mac Pro Dualcore 2.6 GHz running McOS 10.6.6) for data back up.

The user interface and data stream management (IMU sensors, HD video recorder, webcam, etc.) were fully developed in Max/MSP v.6.1 (Ircam/Cycling74) on both stations, using Dale Phurrough's Max/MSP external for the Kinect[6] and Ircam's Mubu Max/MSP externals for the IMUs[7]. We used Harald Meyer's GoPro Camera Control 2 for the wifi control of the Gopro camera[8]. All computers shared a dedicated ethernet Gigabyte Local Area Network (LAN), communicating with delays lower than 1 ms, using the Open Sound Control TCP/IP (UDP) protocole.

The audio was recorded at a sampling rate of 64 kHz, in 16 bits PCM WAV files, the video from the webcam at 25 frames per second (640 x 480 pixels), the HD video at 120 frames per second (1920 x 1080 pixels). Data from the IMUs were collected at 100 Hz.

The audio and video files were automatically named with a code including the subject ID, the session order and condition (V or V+G), the name of the referent sound, the actual time (year, month, day, hour, minute, second) and, finally, the trial number.

To allow a precise synchronization of the different files during post-production, a multimedia synchronization signal was generated at the same time, 1500 ms after the participant initiated the recording[9], with the LED flashing a red signal, a sampled "clap" sound fed to the audio tracks, and a vector of arbitrary numbers for the motion and gesture data

User interface The user interface consisted in a single full-screen window presenting a number of cells (see Figure 14). Each cell corresponded to one of the referent sounds. Each cell consisted of: a green button to play the referent sound, a red button to record the imitation, a smaller blue button to play back the last recorded imitation and, finally, a counter showing the number of remaining possible trials.

The window was dynamically generated for each participant and each set of sounds. The cells were randomly — but regularly — placed within the main window. To avoid mistakes, only the record and play back buttons of the actual imitation were active just after having played the referent sounds. Then, even if the recording of the imitations started as soon as the subject clicked on the record button, the participants was requested to start imitating only when the background color of the user interface changed from white to orange.

During the V+G part, the video recorded with the webcam was played together with the audio tracks when participants hit the play back button.

[6] `http://hidale.com`, last retrieved on August 17, 2015

[7] `http://forumnet.ircam.fr/fr/product/mubu/`, last retrieved on August 17, 2015.

[8] `http://www.tequnique.com/gopro`, last retrieved on August 18, 2015

[9] Such a delay for the clap was required since the WiFi remote control of the Gopro Hero 4 HD video recorder presents some irregular delay to start recording, up to half a second.

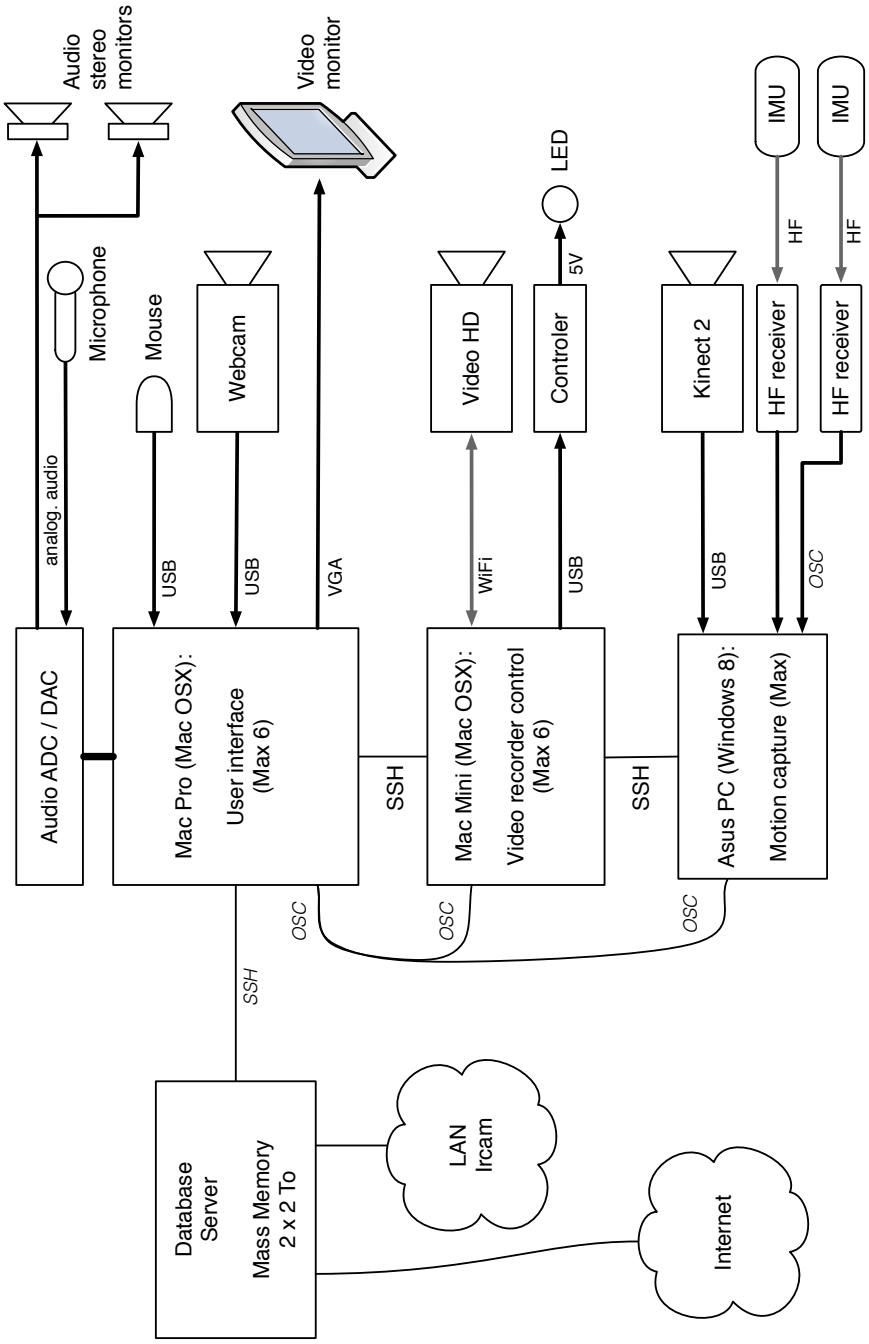

Figure 15: Setup for the recording sessions.

3.2 Data screening, labeling, and tallying

The fifty participants produced a total of 7929 imitations (4410 in the V condition, 3519 in the V+G condition), making an average of 1.5 imitation per participant and per referent sound (i.e. on average, participants made no more than a few trials). This makes a total of about one terabyte worth of data.

The data were first manually screened by listening to each audio track and watching the videos. At this stage, 536 imitations (7%) were rejected for technical reasons (resulting in a total of 7393 files, 4062 in the V condition, 3331 in the V+G condition). Most issues resulted from truncated recordings when the participants stopped the recording before the end of their imitation. Twenty-one participants produced at least one good audio recording for each referent sound in the V condition, 29 in the V+G condition.

Some other data files were also missing in some cases. In fact, the wireless connections were sometimes slow, resulting in the Gopro camera not starting up properly. In addition the IMU data were not recorded for the first 8 participants. In the V condition, there were 3586 imitations with both the audio and Gopro data files. In the V+G condition, there 2726 imitations for which the audio, Gopro, Kinect, and IMU files were recorded properly. In total, this results in 6312 usable imitations (i.e. with all data collected; 20% of the imitations missed at least one of the data files). All things considered, 19 participants produced at least one imitation with all data files collected for each referent sound in the V condition, 21 in the V+G condition.

3.3 Delivery to the consortium

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

3.4 Comparison with KTH's recording sessions

KTH used the same Max/MSP recording interface provided by IRCAM (there were in fact slight differences between the two, which were adjusted by each group to their own requirements). The comparison is summarized in Table 4. Whereas Ircam recorded naive speakers, KTH recorded professional actors. In addition to audio and video recordings, Ircam recorded the acceleration and position of the participants' wrists. KTH recorded EGG signals.

Equipment	IRCAM	KTH
Number of referent sounds	52	50 (+10)
Number of imitators	50 French-speaking, naive (no expertise in music, voice, audio, theatre, dance, etc.)	4 Swedish-speaking, professional actors
Procedure	Imitate so that someone else can recognize the referent sound	Imitate in a manner analogous to sketching
Location	Sound-proofed booth with acoustic padding, black background	Sound-proofed booth with white background
User Interface	Max/MSP	Max/MSP
Data collection interface	Max/MSP	Cockos Reaper
Microphone	DPA headset (d:fine 4066 omni)	DPA headset (d:fine 4066 omni)
EGG	No	yes
High-speed video	Gopro Hero 4 (120 fps)	Gopro Hero 3 (100 fps)
Additional video	Logitech webcam	Canon Legria
Accelerometers on wrists	Yes	No
Audio interface	RME Fireface 800, 64 kHz	RME UFX, 44.1 kHz
Loudspeakers	Yamaha MSP5	Genelec 1031A

Table 4: Comparison between the setups for recording imitations at IRCAM and KTH

4 Analysis of the database of imitations (Task 4.3)

We conducted two analyses of the database of imitations: a statistical analysis of the acoustic properties of the imitations and a qualitative analysis of the gestures, based on a manual annotation of the video recordings.

4.1 Statistical analysis: vocalizations

The goal of the statistical analyses of the imitations is to describe the acoustical invariants or regularities across families and categories but also to reveal specific strategies through the speakers. To reach this goal, we used two sets of descriptors (or audio features). One set is common to the three families (Interactions, Machines and Abstract sounds), the second one is specific to the Abstract family. The descriptors have been developed for the classification tasks (WP5). Their computations is detailed in the deliverables of WP5. Here we focus on the interpretation of these descriptors, in order to describe the imitations.

We computed statistical representations based on hierarchical clustering: we computed distances between imitations or between speakers and performed a hierarchical clustering analysis to reveal invariant structures.

4.1.1 Sound descriptors

This section summarizes the different acoustical descriptors developed for WP5. We also report a qualitative description of each descriptor to facilitate the interpretation of the statistical analyses.

Descriptors common to the three families Ten different descriptors were developed in order to describe the different sound profiles (stable, iterative, impulsive, ...) but also the timbre and tonal parts.

NumActive (Number of non-silent distinct regions in the signal) and *RelDuration* (Relative Duration as the ratio between average active region duration and total signal duration) are tailored to be the characterization of signals with more than one active (non silent) region. When then signal is made of small fragments, *numActive* is higher than 1 (usually more than 4) and *RelDuration* is small. Impulsive signals, by contrast, have small *RelDuration* and usually *NumActive* equals 1.

AbsDuration (Duration of active regions) then, coupled with *RelDuration*, discriminates these impulsive signals from longer, steady signals.

Noisiness (Median noisiness value) and *Zerocross* (Median number of zero crossings) find similar information: a noisy signal will have high *Noisiness* and *Zerocross* values. By contrast, a strongly pitched signal has a low *Zerocross* value, but could still have a high *noisiness* value in case of background turbulence noise (a rather common situation in the dataset, actually). The *PitchStrength* (Median pitch reliability/strength) finds equivalent information: a pitched signal is expected to have a high *PitchStrength* value and a low *Zerocross* value, and vice versa.

The *SlopeAmplitude* expresses the evolution in time of the loudness/energy of the signal: usually it is close to zero (steady signals), but often the imitators tend to reduce loudness along

the phonation. Coupled to this descriptor there is *ModAmplitude*, which measures the amount of variation in the signal loudness along its duration. Slope and modulation together give a first insight into the signal content. Similarly, *SlopeFrequency* and *ModFrequency* express informations about the frequency content of the signal: the former is the pitch/centroid evolution along the time, and the former tells about magnitude of variations.

Morphological descriptors for the Abstract familly The abstract family is based on six different categories, corresponding to six different morphological profiles (see Figure 10 and Table 3):

- **Up:** Sounds which have an increasing profile in terms of spectral content and/or loudness, thus expressing a kind of rising;
- **Down:** Opposite of the previous one, these stimuli present a downward profile;
- **Up/down**: Sounds with non-monotonic profiles: they can be described as a combination of the previous two profiles. The profile moves upward and then downward;
- **Impulse:** This class contains very short sounds, like sound impulses and sharp attack and decay;
- **Repeated:** Sounds that are composed by the repetition of short and almost impulsive elements, with varied rhythmic patterns;
- **Stable:** Longer sounds, with almost flat pitch and loudness profiles.

Different descriptors have thus been developed to describe the specific morphologies. Each imitation is described with eight different acoustical descriptors.

The first two pattern descriptors *meanDC* and *sdtDC* are defined as the mean and the standard deviation of non-silent regions duty-cycles. With these descriptors, impulse and repetition categories have small values of *meanDC*. Conversely, the other categories (stable, up, down, up/down) have a single long active region with a large value of *meanDC*. *SdtDC* measures the regularity of the repeated patterns.

The third descriptor *numActReg* is the number of active (non-silent) regions. For single-region signals, this descriptor is equal to 0, whereas for signals with three or more regions like repetition patterns, *numActReg* is above 0.7.

Descriptors *numActReg* and *mainRegDC* (the duty cycle of the main region) focus on the main regions with non silent signal. *NumActReg* discriminates impulse / repetition and stable categories.

MainRegDC improves the discrimination between categories that have flat or non-flat evolution, such as up/down vs stable. The *Stable* descriptor, computed on the loudness time series, improves the discrimination between categories that have flat or non-flat evolution, such as up/down vs stable.

AbsDuration, defined as the sum of the lengths of active regions enhances the discrimination between impulse and stable (or repetition) categories.

The *slope1* and *slope2* morphological descriptors have been developed to measure the slopes of the signal and are based on the Spectral-peak-min (Marchetto and Peeters, 2015). Spectral-peak-min is based on the energy spectrum(computed as the square DFT). The 5 most important frequency bins are selected. The Spectral-peak-min is defined as the lowest frequency among these 5 frequencies; it is thus measured in Hz. Slope1 corresponds to the

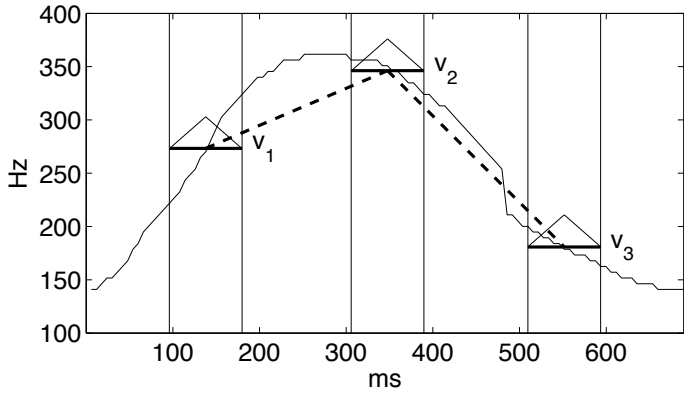

Figure 16: Computation of descriptors *slope1* and *slope2* on an up/down imitation. Spectral-peak-min is showed (thin line), with 3 windows centered at 1/5, 1/2 and 4/5 of its total length. *Slope1* and *slope2* are represented by the dashed bold line.

slope of the first part of the signal, slope 2 of the second part of the signal (see Figure 16). They discriminate between upward, downward (and upward/downward) profiles by combining them.

4.1.2 Statistical analyses

We used these two sets of descriptors to analyze the imitations across families and categories, and across speakers. We used hierarchical clustering (Legendre and Legendre, 1998).

Cluster analysis reveals meaningful structures within data without hypothesis and statistical test. For each descriptor, the distance between imitations or speakers is calculated with an Euclidean distance from the raw data. The raw data is a rectangular matrix (50 x 52) corresponding to the values of a specific acoustical descriptor calculated on the imitations produced by 50 speakers for all the 52 referent sounds (see Section 2 for the details of the referent sounds).

For distances between imitations, Euclidean distances are computed between the imitations of the 52 referent sounds across the speakers. In the case of distances between speakers, Euclidean distances are computed between the speakers across the imitations of the 52 referent sounds.

The hierarchical clustering analysis is based on distances between imitations or between speakers. The hierarchical cluster analysis fits a dendrogram to the distance matrix. The distance within the dendrogram is calculated with the Ward method minimizing the sum of squares of any two possible clusters that can be formed at each step.

4.1.3 Results

General description of imitations First, if we focus on the imitations, the cluster analysis shows a distinction between the Abstract + Machines families on the one hand and the Interactions family on the other hand based on noisiness and pitch strength. For instance, Figure 17 shows two distinct clusters based on pitch strength: one corresponding to imitations

of the Abstract and Machine families and the other one corresponding to the Interaction family. If we consider the values of these descriptors, imitations of Abstract and Machine sounds are less noisy and more tonal than the imitations of the Interaction sounds, which is consistent with our expectations: tonal referent sounds elicit tonal imitations, noisy referent sounds elicit noisy imitations. This distinction is also found for the *vocal and gestures* condition.

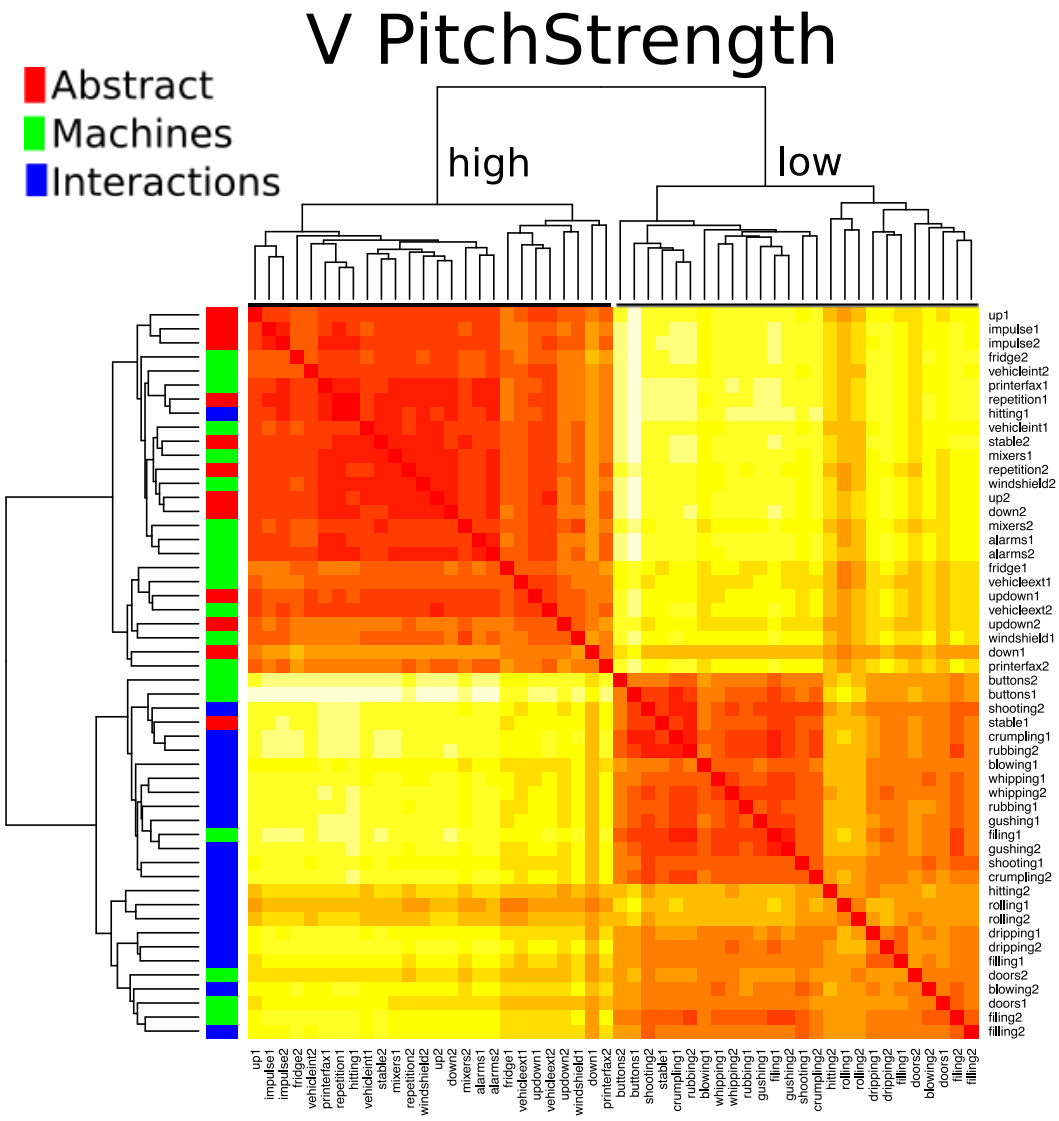

Figure 17: Euclidean distances between imitations for the *pitch strength* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map. The associated colors (red, green and blue) respectively indicate Abstract, Machines and Interactions families.

Considering the acoustical descriptor Relative Duration, Figure 21 (in Appendix) shows a distinction between short or repetitive imitations, like impulses, doors or alarms sounds (with small values) and more continuous imitations (high values) like mixer or crumpling sounds. The *NumActive* descriptor give additional details (see Figure 22 in Appendix) highlights the

imitations of repetitive sounds like windshield, alarms, repetitions but also filling sounds.

Slope of amplitude descriptor is also interesting because of the discrimination of the impulsive sounds without repetition, like whipping, door, shooting or impulsive sounds with a brief variation of amplitude (see Figure 23 in Appendix).

In summary, this first general analysis shows three different strategies to imitate the sounds that are shared between the different families. First, there is a clear distinction between noisy imitations and the others that can be also related to less tonal imitations. Second, repetitive and continuous imitations that are clearly discriminable on the basis of the acoustic descriptors. Third, short and impulsive imitations are also a specific strategy to imitate sounds with this profile.

Speakers If we now consider the speaker, the cluster analysis did not highlight groups of speakers with specific strategies. Depending on the acoustical descriptor, we found marginal strategies that concern only a few speakers. For example some speakers produced imitations longer than the other speakers that are not necessarily related to the length of the referent sounds (see Figure 24 in Appendix). Considering now the *pitch strength* descriptor, some speakers (numbers 34, 37, 4, 8, 44, 33, 45) produced more tonal imitations for the interaction family that the other speakers (see Figures 18 and 25 in Appendix).

Abstract family Specific acoustical descriptors have been developed in order to discriminate the imitations of the categories Up or Down, Up-down, Stable, Impulsive or Repeated imitations. First we saw a clear distinction between Impulsive, Repeated, and Stable categories and other imitations (Figures 28 and 26 in Appendix) by looking at the different acoustical descriptors, respectively absDuration (sum of active regions lengths), the number of non-silent/active regions and stable. This confirms that speakers were able to produce imitation that reproduce the temporal patterns of the referent sounds. Figure 27 shows a specific strategy used across speakers in order to imitate Up, Down and Up-down profiles. Speakers seemed in fact to produce imitations with an accentuated upward profile (spectral variations) when the referent sounds started with an upward slope, to discriminate them from referent sounds starting with a downward slope or no slope at all. The descriptors *slope1* and *slope2* provide a good discrimination between Up/Up-down profiles and other profiles. In fact, *slope1*, and *slope2* show that Up profiles, including the first part of Up-down profiles, are well described with these descriptors. This result indicates that imitators deliberately accentuate Up profiles when they imitate referent sounds with these profiles.

4.1.4 Conclusion

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

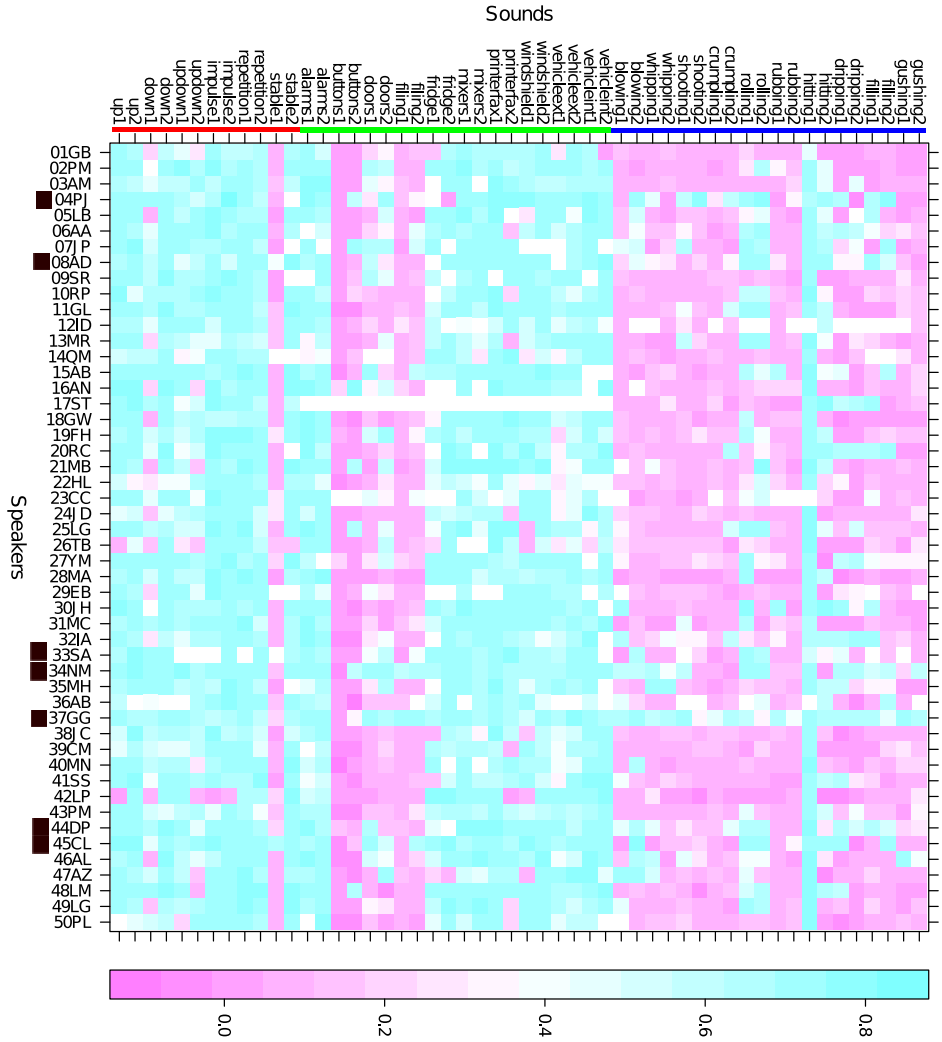

Figure 18: *Pitch strength* calculated on imitations of 52 referent sounds produced by 50 speakers (blue represents tonal imitations, pink noisy associations). The associated colors (red, green, and blue) respectively indicate Abstract, Machine and Interaction families. Speakers 4, 8, 33, 34, 37, 44 and 45 producing tonal imitations specific for when other speakers produce noisy imitations are marked with black squares.

to reproduce the main aspects of the referent sounds (tonalness and temporal profiles), and that these aspects are well captured by the descriptors developed in WP5.

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

4.2 Qualitative analysis: gestural strategies

Lemaitre et al. (2014) have shown that people use a lot of gestures when describing sounds. The role of these gestures is however not clear. Are they redundant with the vocalizations

or do they convey additional pieces of information that the voice is unable to express? Do gestural imitations focus on certain specific aspects of the sounds ? Many examples found in the aforementioned study show people imitating the gestures that produce the sounds (e.g. beating eggs). Other studies have shown that imitative gestures somehow follow the temporal envelope of the sounds when they are not identifiable (Caramiaux et al., 2014). One can also imagine that, faced with the task of reproducing many features of a referent sound, a participant may choose to convey certain features with the voice and certain others with gestures. The present experimental study aimed at studying theses different ideas by doing a qualitative analysis of the database of vocal and gestural imitations previously described. This study was carried out in part during Hugo Scurto's master thesis (March-August 2015). These results will be presented at the Meeting of the Acoustical Society of America in Jacksonville, FL in November 2015.

The first step consisted of a qualitative analysis of the video recordings of the whole set of imitations of 15 randomly-selected participants to identify different strategies and gestures. The second step consisted used an analysis grid focused on a selection of ten participants and eight referent sounds.

4.2.1 Initial observations

We first reviewed the video footage to get a sense of the process of imitation.

Stimuli We selected the recordings of 15 participants at random in the database of vocal and gestural imitations (V+G) of the 52 referent sounds. Five participants (2 male, 3 female; average age 26.2 years old) had followed the *free gesture protocol*, ten (6 male, 4 female; average age: 24.6 years old) had followed the *directed protocol*. We also compared the vocal imitations in the V condition with the V+G condition.

Procedure We first created an analysis grid. It consisted in:

1. *Describing* the gestural imitation in one sentence,
2. Noting if the gesture *reinforces* an aspect of the vocalization, or the opposite (i.e. if gesture has a proper and distinct meaning from vocalization)
3. Noting if adding a gesture to the act of vocalizing *modify* the vocalization.

This analysis is highly subjective as it was conducted by one experimenter only. It was also based on the autoconfrontation interviews, which somehow reduced ambiguity.

Results As expected from the literature (Caramiaux et al., 2014), each of the five participants in the free gesture protocol mimicked the sound source or the action that could have produced the referent sound.

A few cases of mimicry were still observed for the ten participants in the directed protocol: six sounds out of 52 were mimicked by two or three participants out of ten. It is important to note that these six referent sounds (abstract impulse sound, closing door, sawing, rubbing, hitting and whipping) are in fact very hard to imitate without mimicking the sound source since they are human-triggered sounds. In other cases, gestural imitations thus tended to express the

referent sound itself rather than its cause.

Globally, imitations were very diverse. The greater the referent sound complexity is, the more diverse imitations are. On the one hand, basic interactions such as whipping or switching a button were imitated in the same way; on the other hand, crumpling a can brought about several imitations that were unique to the imitator. Another interesting finding is that gestures that we initially expected to be very common (e.g. raising hands for a rising pitch) were not observed.

An aspect of gestural imitation held our attention: we noticed that noisy stable sounds were gestured by shaking hands and fingers. For stable abstract sounds and a blowing sound for instance, seven participants out of ten made a stable noisy vocalization while shaking their hands. Also, for complex sounds such as filling a glass with water, six participants out of ten made a gesture that seems to convey another information that was not vocalized. Lastly, for the fridge sound, eight participants out of ten made a stable vocalization while shaking their hands.

Finally, we did not find any consistent differences between vocalizations alone (V) and vocalizations with a gesture (V+G). Some participants did exactly the same vocalizations in the two conditions, some other did changed them. In the first case, it is possible that the participants had remembered their previous vocal imitation during the second condition (V+G).

4.2.2 Analysis grid

The previous analysis provided us with first observations. We then decided to refine them by focusing on a subset of ten imitators and eight referent sounds, and analyzing the audio, video (including slow-motion), and acceleration data with a grid derived from the initial observations.

Procedure The analysis grid consisted of:

1. Describing a potential synchrony between gesture and vocalization,
2. Extracting information (if any) that is specific to gesture on the one hand, and specific to vocalization on the other hand,
3. Noting the presence/absence of preparatory and/or recovery gestures in the gesture unit,
4. Noting the main direction of the gesture (if any), and
5. Characterizing the possible distorsion between imitation and stimulus,

For each of these items, we associated a potential description with a code. It is important to underline that even if this analysis is more precise, it is still subjective to say that in some cases, information could be peculiar to gesture (or to vocalization). We tried to minimize this subjectivity by focusing on a reduced set of referent sounds.

Stimuli We used the recordings of ten randomly selected subjects (five males and five females). They all followed a *directed gesture protocol* (i.e. they were required not to mimic the imagined sound source).

We selected eight referent sounds on the basis of their acoustic complexity:

- **Simple sounds.** These are sounds for which only one acoustic characteristic (e.g., tonal component, periodicity) evolves. Five sounds were selected: stable noise, repetitive noise, a closing door (human impact), pitch going up, pitch going down.

- **Complex sounds.** These are sounds for which several acoustic characteristics vary at the same time (vertical complexity) or in time (horizontal complexity). Three sounds were selected: a "humming fridge", a "printer" and "filling a recipient with a soda". The "humming fridge" consists in a tonal stable sound plus stable noise and random bubble sounds (vertical complexity). The "printer" sound has two distinct parts (horizontal complexity): the first part consists in a tonal repetitive sound plus random paper sounds and stable noise, while the second part is just stable noise. Finally, the "filling" sound has both vertical and horizontal complexity. Its first part is the impact of the soda in the recipient; its second part is noise plus two tonal components whose pitches evolve in an opposite way; its third part is noise plus a higher pitch going up.

These categories are somehow overlapping. They however provide use with an useful selection of sounds with simple and complex variations.

Aspects of imitations shared across participants For 90% of the imitations, vocalization and gesture begin and end at the same time. Preparation and recovery gestures are present in the same percentage of the imitations (only one subject made clear pauses at both the beginning and the end of his imitations).

Simple sounds. There were basic similarities among the imitations:

- For the stable noise, eight participants out of ten vocalized a noise while shaking the hands without any specific direction;

- For the repetitive noise, ten participants out of ten vocalized a repetitive noise while moving their hands in rhythm in a specific direction;

- For the impact sound, every participant made a noisy and decreasing vocalization while underlining the impact with their gesture;

- For pitched sounds, nine participants out of ten vocalized the evolution of the pitch while reflecting it with their gesture. What is interesting is that they seemed to emphasize either the beginning or the end of their imitation. Six out of ten emphasized the end of their "pitch going up" imitation and nine out of ten the beginning of their "pitch going down" imitation.

However, despite these high-level similarities in the imitations of elementary sounds, several specificities are observed at a lower-level. Three participants out of ten tried to imitate the random aspect of the stable noise by modulating their formants. For the pitched sounds, the main direction of the gestures, while including the up/down aspect, is not purely up or down : in most cases, it is coupled with a backward/forward or left/right direction. The same aspect is present in the repetitive noise and the impact sound : there is no specific direction in gesture across participants.

Complex sounds One can identify basic similarities among the imitations:

- For the "humming fridge"", seven participants out of ten made a stable tonal vocalization while shaking their hands;
- For the "printer," every participant tried to vocalize the repetitive tonal aspect while underlining it with their gesture. Interestingly, most of them did not imitate the second part of the sound.

For the "filling a recipient" sound, there were too many different imitations to be able to draw up basic similarities. We will discuss this point later.
There were even more singularities for these complex sounds than for the elementary sounds, particularly for horizontally complex sounds. For the printer, almost every participant underlined the repetitive aspect in a different manner; vocalizations were also variable.

Separation of vocalization and gesture The analysis of the recordings suggests that gesture always reflects at least one aspect of the vocalization. In some other cases, gesture may communicate a feature of the referent sound that is not present in the vocalization. In this section, we will precisely focus on these cases, i.e. on cases in which gesture gives an additional information about the imitation the vocalization does not give.

Elementary sounds. First, it is important to notice the presence of such a separation in some imitations of elementary sounds. For example, in a third of the cases, a constant movement complements the imitation of noisy sounds, perhaps standing for the temporality of the sound. Participants who used both their hands sometimes made them come apart or closer, which is not clearly related to an acoustic property. It is the case for pitched sounds.

Complex sounds. The "filling" sound is particularly interesting to study the separation of vocalization and gesture since it has both horizontal and vertical complexity. Here are some interesting examples :

- One participant vocalized an upward sweep while shaking his fingers;
- Three participants made a noisy formant-modulated vocalization while moving their hands up;
- One participant made an upward noisy vocalization while moving his hands down;
- One participant made a stable noisy vocalization while moving his hand down;
- One participant vocalized a upward rough sweep while moving his hand down.

It is however difficult to say if these global movements stand for the evolution of one pitch component, or just for the temporality of the sound. Another interesting point is that four of the gestural imitations ended after the vocalization, as if it was standing for the third part of the sound.

Other separations between gesture and vocalization are observable for the two other complex sounds:

- For the "humming fridge", as seen before, seven participants out of ten made a stable tonal vocalization while shaking their hands;
- For the "printer", four participants out of ten made a shaking movement with their hands.

Comparing vocalizations without (V) and with gestures (V+G) In some cases, an energetic gesture may modify the vocalization. For example, a vibrating gesture with the hand may make one's chest vibrate, thus making the vocalization vibrate. In these cases, gesture and vocalizations share a common part.

Besides, adding a gesture to a vocalization may modify it in two different ways: (1) gesture can push him to vocalize in a different way, and (2) gesture may help the participant embody the sound he has to imitate.

Change in vocalization. There were some cases in which vocalization was totally different when completed by a gesture. For example, a participant who imitated the closing door with a trembling tonal vocalization turned the latter into a noisy vocalization when adding a trembling gesture to it. The same participant turned a going up noisy vocalization for the filling sound into a stable noisy vocalization when completed by a going up gesture. Another participant who imitated the stable noise with a rough vocalization transformed it into a noisy vocalization when adding a trembling gesture. In an interview, he stated that as he could not reproduce some aspects of the sound with his voice, he had to make them with his gestures.

Embodying the sound. Another change that gesture seemed to trigger is the implication of participants in their imitations. In some cases, their global imitations seemed more accurate when they add a gesture to their vocalization. For example, a participant made a more complex and convincing vocalization of the stable noise when he added a gesture to it. A relevant phenomenon is that a lot of participants tended to use a gesture even when they are asked to perform a vocalization only. This suggests that body movement helps them imitate sounds more confidently.

4.2.3 Conclusions and hypotheses

The first important conclusion of this qualitative analysis of the vocal and gestural imitations is that we did not observe the phenomena that we had initially expected. For instance, we expected that participant would imitate a rising pitch with an upward movement. This was the case for only some subjects: some others performed a downward gesture (as if on the neck of a string instrument), or brought their hands close to their chest (as in something approaching). The role of gestures during imitations of sounds is therefore more subtle than just drawing the time envelope of some parameters in the air. Instead, we observed several phenomena that occurred regularly across listeners:

- Executing gestures with arms and torso seem to facilitate the production of expressive vocal imitations.

- Vocalization and gestures are usually not synchronous, because of phase and frequency differences. When participants imitate rhythmic sounds, vocalizations are generally more precisely locked to the rhythm of the referent sounds. This suggests that biomechanical constraints may limit the bandwidth of gestures performed in the air.

- Participants very often express the presence of noise in the referent sounds by rapidly shaking their hands. This gesture is not really descriptive (frequencies do not correspond) but seems to be used as shared powerful metaphor.

- In some cases, participants were able to communicate different pieces information about the referent sounds with their hands and gestures separaltey. They seemed to use their voice to imitate either tonal aspects of sounds or the most salient aspect of sounds.

These observations are very important, and inform us on the possible strategies that users could use with the SkAT-VG tools. First, an initial idea was that participants could precisely draw the temporal evolution of sound parameters in the air. This does not seem to be intuitive nor possible for most of the non-expert participants that we studied. If a precise gestural control is required, we should probably consider object manipulation instead of gestures in the air. Second, this analysis suggests that participants are more likely to use gestures to signify features of the referent sounds in a metaphoric way: a repeated gesture (whose rhythm is unrelated to the referent sound) to express a rhythmic sound, rapidly shaking hands to signify a noisy component. Similarly, the SkAT-VG tools could exploit the ability to use two communication channels at the same time. Third, the results also suggests that participants can use these metaphorical gestures to add something to their vocalization that they are not able to vocalize. Similarly, the SkAT-VG tools could exploit the ability to use two communication channels at the same time.

These conclusions should however be considered with care, as they are based on qualitative observations of the database. A more precise study has been conducted to test two of the aforementioned ideas. It used a new set of specifically created stimuli, new recordings of the imitations, gestural descriptors, and a controlled experimental setup. Since this work does deal with the database itself, we will report in D4.4.3.

5 Completed, ongoing and future work to appear in D4.4.2

We report in this sections studies (completed, in progress, or planned) that do not belong to D4.4.1 proper, but are nevertheless important to frame the results in the general context of WP4. The details of the descriptions of the following paragraphs reflect on our progress toward completing these studies.

5.1 Vocal imitations of basic auditory features: what is the human voice able to reproduce?

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

5.2 Experimental study: what are the respective roles of voice and gestures?

During Year 2, IRCAM studied more precisely the role of gestures during imitations of sounds. An experimental study tested the following hypotheses:

- Vocalizations reproduce more precisely rhythmic sequences than gestures
- Imitators use "shaky gestures" to express that the referent sound has a noisy component
- When imitators imitate sounds made of different layers, they use different strategies, one of which consists of conveying one layer with the voice and one with the gestures.

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

5.3 Identification experiment: can listeners access the semantic content of vocal imitations?

At the root of the SkAT-VG project is the idea that imitating a sound is similar to drawing a sketch: it simplifies the referent sound that is imitated (within the constraints of human voice production) to effectively convey it to the listener, who can identify what is being imitated. The goal of this study is twofold: i) to compare how effectively listeners can *identify* the source of the referent sound; and ii) to compare vocal imitations produced by human speakers to another type of sounds sketching, *auditory sketches* based on sparsified representations of the signal (Suied et al., 2013).

In our previous work (Lemaitre and Rocchesso, 2014), we had evaluated the effectiveness of an imitation by presenting each vocal imitations to the participants with a list of potential referent sounds. The listeners then had to select the referent sound corresponding to the imitation (N-response classification task). Thus, the effectiveness of the imitation was defined as the *similarity* of the imitation to the referent sound, within a context defined by the other potential referent sounds.

In this study, our aim is to go beyond the mere similarity of the imitations, and investigate the semantic content of the imitations. There are different methods to study the semantic representations associated with a stimulus. One such method (free verbalizations) consists of having participants freely write down or verbalize what they identify (Ballas, 1993; Lemaitre et al., 2010; Houix et al., 2012). This method provides rich and complex information but is also particularly difficult to analyze since it requires a linguistic analysis of the verbal productions. Another type of methods in the context of the Signal Detection Theory (SDT) consists of presenting the participants with a list of labels potentially describing the stimulus (yes-no, N-response classification, rating tasks). This method uses textual descriptions of stimuli, but the analysis is more straightforward than a method based on free verbalizations, since various accuracy scores can be easily computed on the basis of the confusion matrix. Finally, another method consists of using the stimulus to prime a lexical decision ("Is this string of letters a word?") (van Petten and Rheinfelder, 1995; Lemaitre and Heller, 2013). Reaction times decrease when the prime stimulus and the target word are semantically related (semantic priming effect), and the size of the priming effect reflects the strength of the semantic relation. Such a method provides really strong information about the semantic representation elicited by a stimulus but requires a large number of trials per stimuli to get reliable reaction times. This limits in practice the number of stimuli that can be tested in an experimental session.

We had several criteria to select an experimental method. First, we wanted that the participants did not compare the imitations with the referent sounds. Instead, we wanted to use verbal descriptions of the referent sounds, which is possible with whichever of the three aforementioned methods. Second, we wanted to control the potential bias of the participants toward certain responses. In fact, since we are evaluating rather unusual stimuli (e.g. we are asking whether a sound that is clearly produced by a human voice could be the sound of machine), we were anticipating strong biases in participants. For instance, a participant may be strongly biased toward saying that none the the human-made imitations can be the sounds of a machine. SDT and priming methods are immune to biases: SDT metrics actually separate response sensitivity and bias, and semantic priming does not rely on participants' voluntary decisions. Third, we wanted to test a significant number of conditions (referent sounds and

imitators). Semantic priming methods can only test a few cases (because it uses reaction times, a large number of repetitions is required), and requires a tight control over the duration of the stimuli. SDT methods also require of lot of trials to evaluate response bias, but we can use a larger variety of stimuli. We therefore chose this kind of methods.

SDT methods have however the disadvantage that the measured identification scores are completely determined by the chosen list of stimulus descriptions that participants choose from. The accuracy scores for a given stimulus have therefore to be interpreted in relation to the accuracy scores of some references. Here, we compared vocal imitations produced by human participants to "auditory sketches" computed on the basis of sparsified mathematical representations of the referent signals (Suied et al., 2013). This sketches are scalable (i.e. the faithfulness of the sketch to the referent sounds can be controlled and measured) and based in part on models of auditory processing. They are therefore a very interesting comparison for human vocal imitations.

5.3.1 Creating "auditory sketches" as comparison points

We created auditory sketches based on the method proposed by Suied et al. (2013). It consists in three parts: 1. Computing a time-frequency representation of the signal inspired by models of peripheral auditory processing ; 2. Selecting the most important elements of the representation based on a given criterion; 3. Inverting the representation. Based on the results of Suied et al. (2013), we used the auditory spectrogram proposed by Chi et al. (2005)[10] and a simple maxima-peaking algorithm[11] to select the most important elements of the representation. To produce the auditory spectrogram, the acoustic signal is analyzed by a bank of constant-Q cochlear-like filters. The output of each filter is processed by a hair cell model followed by a lateral inhibitory network, and is finally rectified and integrated to produce the auditory spectrogram. The inversion of the auditory spectrogram is approximated by the convex projection algorithm proposed by Yang et al. (1992).

On the one hand, this method gives results in good results for sounds containing salient tonal contents and transients that concentrate energy in localized parts of the spectro-temporal representation, but also create audible artifacts for broadband sounds without tonal components or localized transients. On the other hand, a simple method to approximate broadband noisy signals consists of approximating the spectral envelope of the noise with the transfer function of an all-pole filter with p poles via linear predicting coding (LPC) and applying the resulting filter to a white noise (Schwarz et al., 1999). Since the referent sounds that we use include harmonic sounds (e.g. electronic alarms), broadband noises (e.g. water flowing) and sounds consisting of a mix of tonal and noisy components (e.g. engines), it is important that the model can handle these types of sounds. Therefore, our method consisted in: 1. Separating tonal and noisy components; 2. Applying the method of Suied et al. (2013) to the tonal components to create a sketch of the tonal components; 3. Applying the LPC method to the noisy components to create a sketch of the noisy components; 4. Mixing the two sketched components. This method is summarized in Figure 19.

[10] We used the NSL toolbox for signal representation and inversion `http://www.isr.umd.edu/Labs/NSL/Software.htm`, last retrieved on September 15, 2015.

[11] We compared this method to the peak picking method used by Suied et al. (2013): simply selecting the bins with the maximum absolute values creates less artefacts that the peak-picking method.

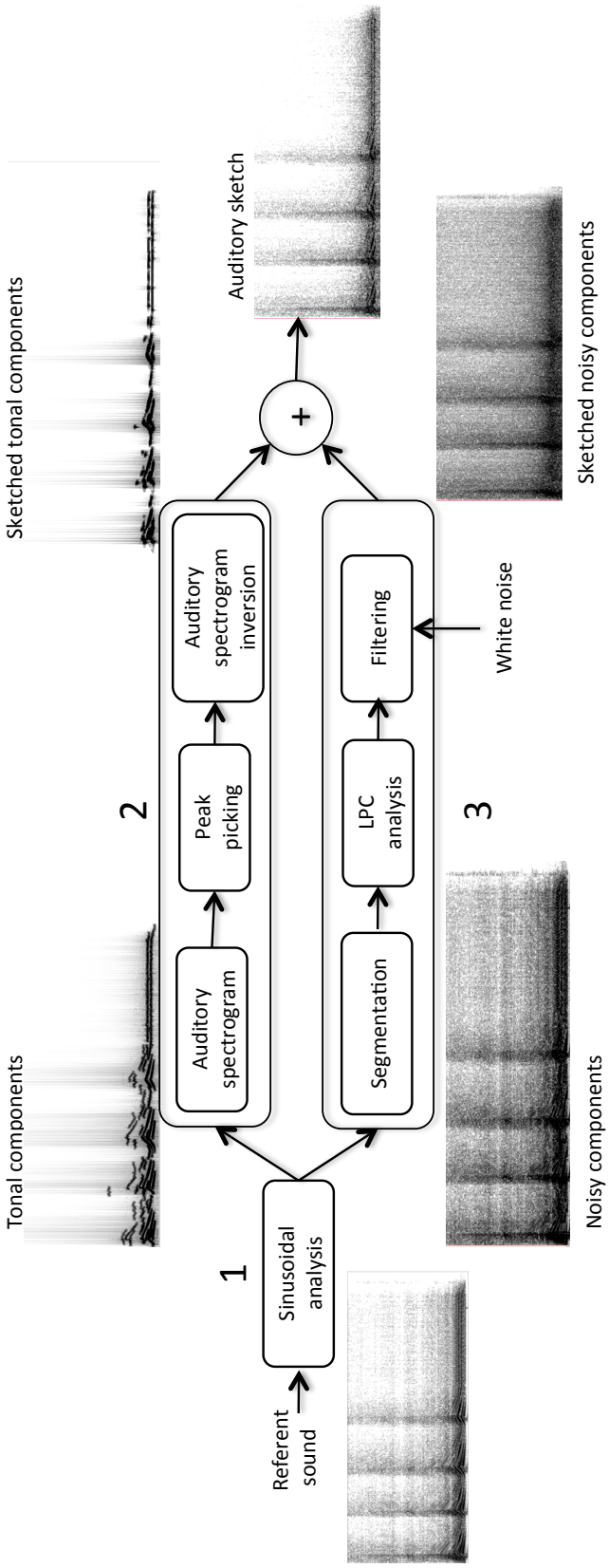

Figure 19: Method to create auditory sketches.

In practice, we used Ircam's pm2 algorithm to track the tonal components of each referent sound and separate them from the noisy components (Roebel, 2008). The parameters of the algorithm were adjusted for each referent sound to ensure good separation of tonal and noisy components. The auditory spectrogram used a of 8-ms frame length and a 128-ms time constant. The auditory spectrogram used 128 filters between 90 and 3623 Hz (referent sounds were first down sampled to 8 kHz before entering the model; tonals components were therefore considered only in the 0-4 kHz range; the remaining components were merged into the noisy components).

The other parameters of the tonal model were adjusted to produce sketched tonal components with different qualities. These qualities were measured by computing the number of coefficients per second used to model the signal. For instance, the complete auditory spectrogram uses 16000 coefficients per seconds. As a starting point, we adjusted the threshold in the maxima-picking algorithm to keep 4000 coefficients per second (Q3, 25%). Pilot tests showed that these parameters produce sketches that are reasonably close to the referent sounds. We also created two other sketches with lower quality by dividing the number of coefficients by 5 at each step, with 800 coefficients per second (Q2, 5%) and 160 coefficients per second (Q1, 1%).

We used the same method for the noisy sketches. However, the quality of the sketched noisy components is controlled by two parameters: the temporal resolution (hop size) and the number of LPC coefficients. As a starting point we used 36 LPC coefficients and a 9 ms temporal resolution (i.e. 4000 coefficients per second), which produced reasonable sketches for most sounds. Just as the maxima-picking method selects portions of the auditory spectrograms by sampling both the temporal and frequency dimensions, we decided to decrease the temporal resolution and the number of LPC coefficients equivalently: we multiplied the temporal resolution and divided the number of LPC coefficients by $\sqrt{5}$ between each step of quality. In practice, this amounted in using 16 LPC coefficients and a 20-ms temporal resolution (Q2, 800 coefficients per second), and 7 LPC coefficients and a 44-ms temporal resolution (Q1, 160 coefficients per second). The segmentation used a overlap of 75% whatever the temporal resolution.

Parameters	Q1	Q2	Q3
Coefficients per second	160	800	4000
Temporal resolution (LPC model)	44 ms	20 ms	9 ms
LPC coefficients (LPC model)	7	16	36

Table 5: Parameters used to synthesized the sketches.

It is important to note that the selection of parameters is a compromise. For instance, for stationary sounds (e.g. a fridge hum), using a slower time resolution improves the modeling, whereas the opposite is true for sounds with a high density of events (e.g. crumpling a piece of paper). Similarly, the modeling of tonal components focuses on the 90-4000 Hz range, because most of the sounds (but not all) have their partials in this range. In consequence, this model is more effective for certain sounds than for other sounds. Our selection of referent sounds balancing between different morphologies and textures ensured that we addressed all different cases for which the sketching method will be more or all effective. This makes the comparison

with vocal imitations even more interesting: are speakers adjusting their vocalization strategies to each sound?

5.3.2 Method - version yes/no

Stimuli We used the eight categories of machine sounds (16 referent sounds) and eight categories of mechanical interactions (16 referent sounds). Half of the referent sounds were used as targets, half as lures. The selection of target and lures categories was based on the morphologies identified in Tables 1 and 2. For each target, we selected the lures in the same morphological category, to maximize the difficulty of the task. The selected categories are represented in Tables 6 and 7. As expected, the selected machine sounds have a strong tonal character and the interaction sounds a strong noisy character.

Morphology	Targets	Lures
Impulsive	Buttons and switches	Doors closing
Repeated	Windshield wipers	Sawing and filing
Continuous-stable	Fridge hums	Blenders
Continuous-dynamic/complex	Vehicle interior	Revs up

Table 6: The selection of machine sounds used in the identification experiment, classified in morphologies. Items in blue are noisy, items in red have strong tonal components; items in purple mix noisy and tonal components.

Morphology	Targets	Lures
Impulsive	Hitting	Whipping
Repeated	Scraping	Dripping
Continuous-stable	Blowing	Gushing
Continuous-dynamic/complex	Crumpling	Rolling

Table 7: The selection of mechanical interaction sounds used in the identification experiment, classified in morphologies. Items in blue are noisy, items in red have strong tonal components; items in purple mix noisy and tonal components.

We decided not to use the abstract sounds because the results of the identification experiment (see Section 2.3) had shown that it was difficult to describe abstract sounds with words.

We also selected the vocal imitations (V condition) of ten participants (five male and five female) from the database of vocal imitations. These ten participants were randomly drawn from the database, after rejecting participants who used onomatopoeia and for there was some technical problems with the audio files (e.g. saturation, troncation, etc.). We also took care to select participants for whom we had good video recordings to ensure collaboration with KTH (phonetic analysis). This amounted in a total 320 vocal imitations.

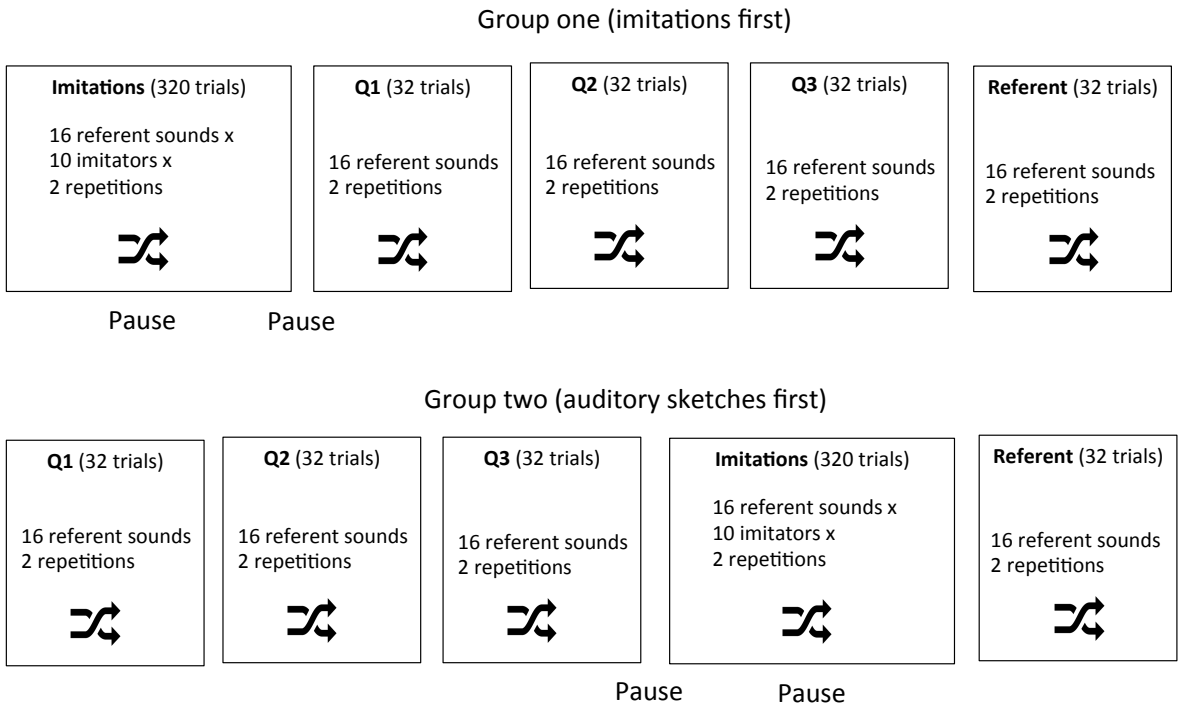

Figure 20: Structure of the yes-no identification experiment.

Finally we used the auditory sketches (Q1, Q2, Q3) of the 16 referent sounds. In total, we therefore used 448 different sounds sounds (32 referent sounds, 320 imitations, and 96 auditory sketches). Each sound was played at two different levels.

Procedure There were two groups of participants, one for each family (machine or interaction).

The main procedure consisted in a series of yes/no tasks for each sounds. Participants read the description of the target category and indicated whether they felt the sound corresponded to the description. Within each family, there four possible yes/no tasks.

We used a blocked design with five blocks (one block for the vocal imitations, one block for each quality of auditory sketch, one block for the referent sounds). To control the possibility that the identification of imitations could be influenced by the presentation of the auditory sketches and vice versa, we used two orders, and presented the block of the referent sounds always at the end of the session. Half of the participants started with the vocal imitations, half with the auditory sketches. The auditory sketches were always presented in order Q1, Q2, Q3. The order of the sounds in each block was also randomized. There was a pause between the blocks of imitations and the blocks of auditory sketches, and within the block of vocal imitations. Each sound was presented twice, to ensure correct calculation of the statistics (448 trials for each group in total). The two repetitions were played at a different level (akin to the roving level procedure). The structure of the blocks is represented in Figure 20.

The description of each target and lure category are reminded in Tables 8 and 9.

Morphology	Categories	Descriptions
Impulse	Switches Doors	Une personne qui appuie sur un interrupteur, un bouton ou une touche
Repeated	Sawing Windshield wipers	Une personne qui scie ou lime ou objet à la main
Continuous-stable	Blenders Fridges	Un robot ménager, un mixeur, un hachoir électrique
Continuous-dynamic /complex	Revs up Car interiors	A l'extérieur d'une voiture ou une moto qui rugit, à l'arrêt

Table 8: The correct descriptions of machine sounds in the identification experiment.

Morphology	Categories	Potential descriptions
Impulse	Hitting Whipping	Frapper, taper, heurter, cogner un objet
Repeated	Scraping Dripping	Gratter, racler, frotter un objet
Continuous-stable	Blowing Gushing	Souffler, expirer
Continuous-dynamic /complex	Rolling Crumpling	Un objet qui roule sur une surface

Table 9: The correct descriptions for the family of mechanical interactions in the identification experiment.

5.3.3 Results

The experiment is scheduled to begin early October.

5.4 Experimental study: Imitations across languages

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

Based on discussions with KTH at Ircam last June (meeting of WP5), we agreed on the following plan:

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

This work could result in a nice publication, either in a linguistic journal or in e.g. JASA.

5.5 How do listeners learn how to adjust their imitations when provided with a feedback?

In what we have done so far, imitators produced an imitation (vocal or gestural), and the only feedback they got was a playback of the audio or audio-video recording of their imitations. The goal of this procedure was to put the participants in charge of the quality of the recording: the technical quality, but also the "communication" quality. The instructions specified that the imitators had to assess whether their imitations could help an hypothetical fellow receiver identify the referent sound based on their imitations.

Things may be different in the context of an actual communication between two persons, such as those described by Lemaitre et al. (2014). Imitators may adapt their imitations in response to the feedback of their counterpart until successful communication. In addition, this behavior may also occur for users using the SkAT-VG sketching tools. If a user produces a vocalization and that the system outputs a sound that does not correspond to what he or she has in mind, he might adjust his or her production until reaching the desired output. In other words, users may learn how the system behaves, and learn how to adjust their vocal and gestural production to reach their goal.

It is therefore important to study such a phenomenon in collaboration with WP6 and WP7. The procedure has not been decided yet. The study is planned for the third year of the project.

5.6 Imitations of sounds in memory

So far, our work has focused on *imitations*. There is a *referent sound* that imitators can listen to (as many times as necessary), and imitators are required to "reproduce" them with their voice and their gestures. This paradigm is necessary because it allows us to know exactly what it is that the imitators are trying to vocalize or gesticulate. However, the situation may be actually different when the referent sound is not physically present at the time of the imitation (is in *memory*), or because there is no referent sound but the *idea* of a sound. These situations are also closer to a real sound design case study, and introduce a new question: do imitations correspond to how people remember or imagine sounds?

Our initial plan to address this question is to use paradigms in which we separate in time the referent sounds and the imitations, both for production and for recognition of the imitations. In both cases, we will first start by teaching a set of referent sounds to the participants, until they have memorized them with a very good accuracy (tested by an old/new paradigm for instance). For the case of *production* of imitations, we will ask them to come back a few days or weeks later, and *then* imitate the memorized referent sounds. For the case of *perception* of imitations, we will ask them to come back a few days or weeks later, and *then* test how well they can recognize the referent sounds from the imitations without providing them with the actual referent sounds. We will used the methods proposed in Section 5.3, and, in particular, compare human-made imitations with automatic auditory sketches.

References

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):250–267.

Brewster, S. (2009). Non-speech auditory outputs. In Sears, A. and Lacko, J. A., editors, *Human-computer interaction: fundamentals*, chapter 13, pages 223 – 240. CRC Press, Boca Raton, FL.

Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., and Susini, P. (2014). The role of sound perception in gestural sound description. *ACM Transactions on Applied Perception*, 11(1).

Cartwright, M. and Pardo, B. (2015). Vocalsketch: vocally imitating audio concepts. In *Proceedings of CHI 2015*, Seoul, Republic of Korea.

Cerrato, G. (2009). Automotive sound quality - Powertrain, road and wind noise. *Sound and Vibration*, pages 16–24.

Chi, T., Ru, P., and Shamma, S. (2005). Multiresolution spetrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2):887–906.

Chion, M. (1983). *Guide des objets sonores*. Buchet/Chastel, Paris, France.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52–80.

Ih, J.-G., Lim, D.-H., Shin, S.-H., and Park, Y. (2003). Experimental design and assessment of product sound quality: application to a vacuum cleaner. *Noise control engineering journal*, 51(4):244–252.

Jeon, J. Y., You, J., and Chang, H. Y. (2007). Sound radiation and sound quality characteristics of refrigerator noise in real living environments. *Applied acoustics*, 68:1118–1134.

Legendre, P. and Legendre, L. (1998). *Numerical ecology*. Developments in Environmental Modelling. Elsevier, Amsterdam, The Netherlands, second english edition.

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

Lemaitre, G. and Heller, L. M. (2013). Evidence for a basic level in a taxonomy of everyday action sounds. *Experimental Brain Research*, 226(2):253–264.

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

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

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

Lemaitre, G., Rocchesso, D., Susini, P., Lambourg, C., and Boussard, P. (2013). Using vocal imitations for sound design. In *Proceedings the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Marseille, France. LMA.

Lemaitre, G., Susini, P., Rocchesso, D., Lambourg, C., and Boussard, P. (2014). Non-verbal imitations as a sketching tool for sound design. In Aramaki, M., Derrien, O., Kronland-Martinet, R., and Ystad, S., editors, *Sound, Music, and Motion. Lecture Notes in Computer Sciences*, pages 558–574. Springer, Berlin, Heidelberg, Germany.

Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., and McAdams, S. (2007). The sound quality of car horns: a psychoacoustical study of timbre. *Acta Acustica united with Acustica*, 93(3):457–468.

Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., and McAdams, S. (2009). The sound quality of car horns: Designing new representative sounds. *Acta Acustica united with Acustica*, 95(2):356–372.

Lemaitre, G., Vartanian, C., Lambourg, C., and Boussard, P. (2015b). A psychoacoustical study of wind buffeting noise. *Applied acoustics*, 95:1–12.

Marchetto, E. and Peeters, G. (2015). A set of audio features for the morphological description of vocal imitations. In *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway.

Parizet, E., Brocard, J., and Piquet, B. (2004). Influence of noise and vibration to comfort in diesel engine cars running at idle. *Acta Acustica united with Acustica*, 90:987–993.

Parizet, E., Guyader, E., and Nosulenko, V. (2008). Analysis of car door closing sound quality. *Applied acoustics*, 69:12–22.

Parizet, E., Hamzaoui, N., Ségaud, L., and Koch, J. R. (2003). Continuous evaluation of noise uncomfort in a bus. *Acta Acustica united with Acustica*, 89:900–907.

Peeters, G. and Deruty, E. (2010). Sound indexing using morphological description. *IEEE-Transactions on audio, speech, and language processing*, 18(3):675 – 687.

Penna Leite, R., Paul, S., and Gerges, S. N. Y. (2009). A sound quality-based investagation of the HVAC system noise of an automobile model. *Applied Acoustics*, 70:636–645.

Rasamimanana, N., Bevilacqua, F., Schnell, N., Guédy, F., Maestracci, E. C., Zamborlin, B., Frechin, J., and Petrevski, U. (2011). Modular musical objects towards embodied control of digital music. In *Proceedings of Tangible Embedded and Embodied Interaction Conference (TEI), Funchal, Portugal*, New York City, NY. ACM.

Roebel, A. (2008). On sinusoidal modeling of nonstationary signals. *The Journal of the Acoustical Society of America*, 123(5):3803–3803.

Sato, S., You, J., and Jeon, J. (2007). Sound quality characteristics of refrigerator noise in real living environments with relation to psychoacoustical and autocorrelation function parameters. *Journal of the Acoustical Society of America*, 122(1):314–325.

Schaeffer, P. (1966). *Traité des objets musicaux*. Seuil, Paris, France.

Schwarz, D., Rodet, X., et al. (1999). Spectral envelope estimation and representation for sound analysis-synthesis. In *Proceedings of the International Computer Music Conference (ICMC), Beijing, China*, pages 351–354, San Francisco, CA. International Computer Music Association.

Stanton, N. A. and Edworthy, J., editors (1999). *Human Factors in Auditory Warnings*. Ashgate Publishing Ltd.

Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. (2013). Auditory sketches: sparse representations of sounds based on perceptual models. In Aramaki, M., Barthet, M., Kronland-Martinet, R., and Ivi Ystad, S., editors, *From Sounds to Music and Emotions, 9^{th} International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers*, volume 7900 of *Lecture Notes in Computer Science*, pages 154–170. Springer, Berlin/Heidelberg, Germany.

Suied, C., Susini, P., and McAdams, S. (2008). Evaluating warning sound urgency with reaction times. *Journal of Experimental Psychology: Applied*, 14(3):201–212.

Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. (2004). Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, 65(8):763–790.

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

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *Information Theory, IEEE Transactions on*, 38(2):824–839.

Özcan, E. and van Egmond, R. (2007). Memory for product sounds: the effect of sound and label type. *Acta Psychologica*, 126:196–215.

Özcan, E. and van Egmond, R. (2012). Basic semantics of product sounds. *International Journal of Design*, 6(2):41–54.

A Appendix: Hierarchical clustering of imitations

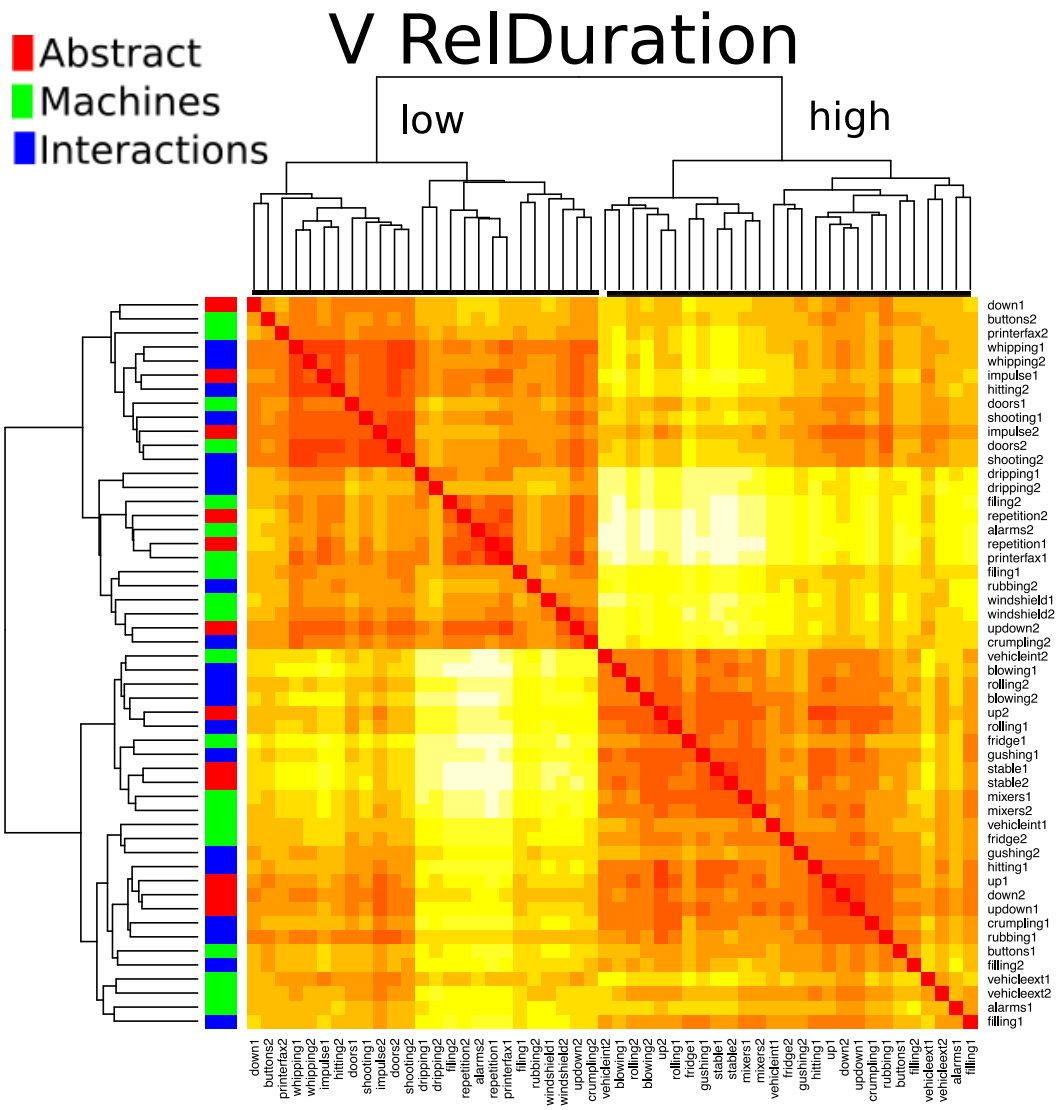

Figure 21: Euclidean distances between imitations for the *relative duration* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map. The associated colors (red, green, and blue) respectively indicate Abstract, Machine and Interaction families.

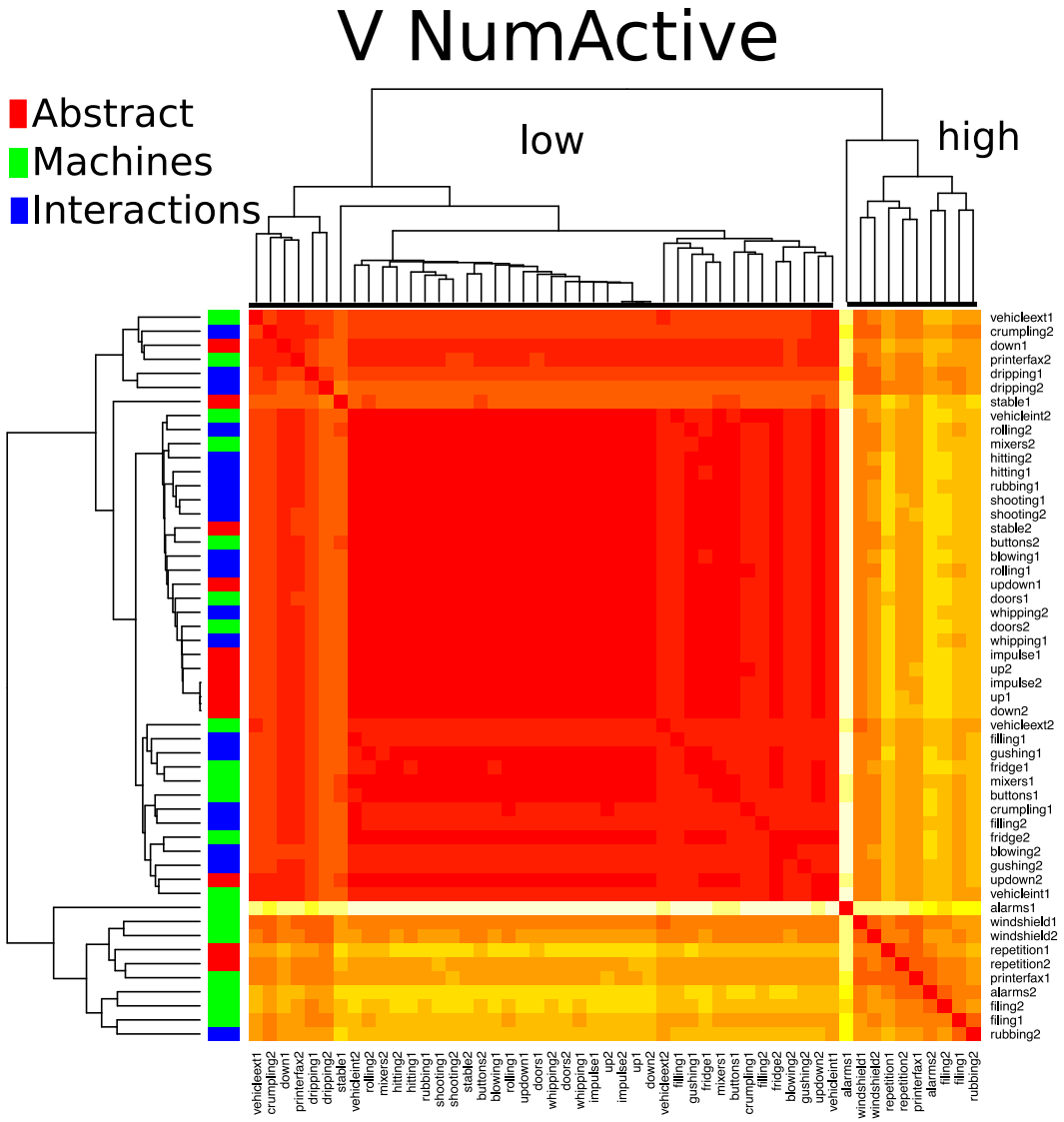

Figure 22: Euclidean distances between imitations for the *number of active regions* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map. The associated colors (red, green, and blue) respectively indicate Abstract, Machine and Interaction families.

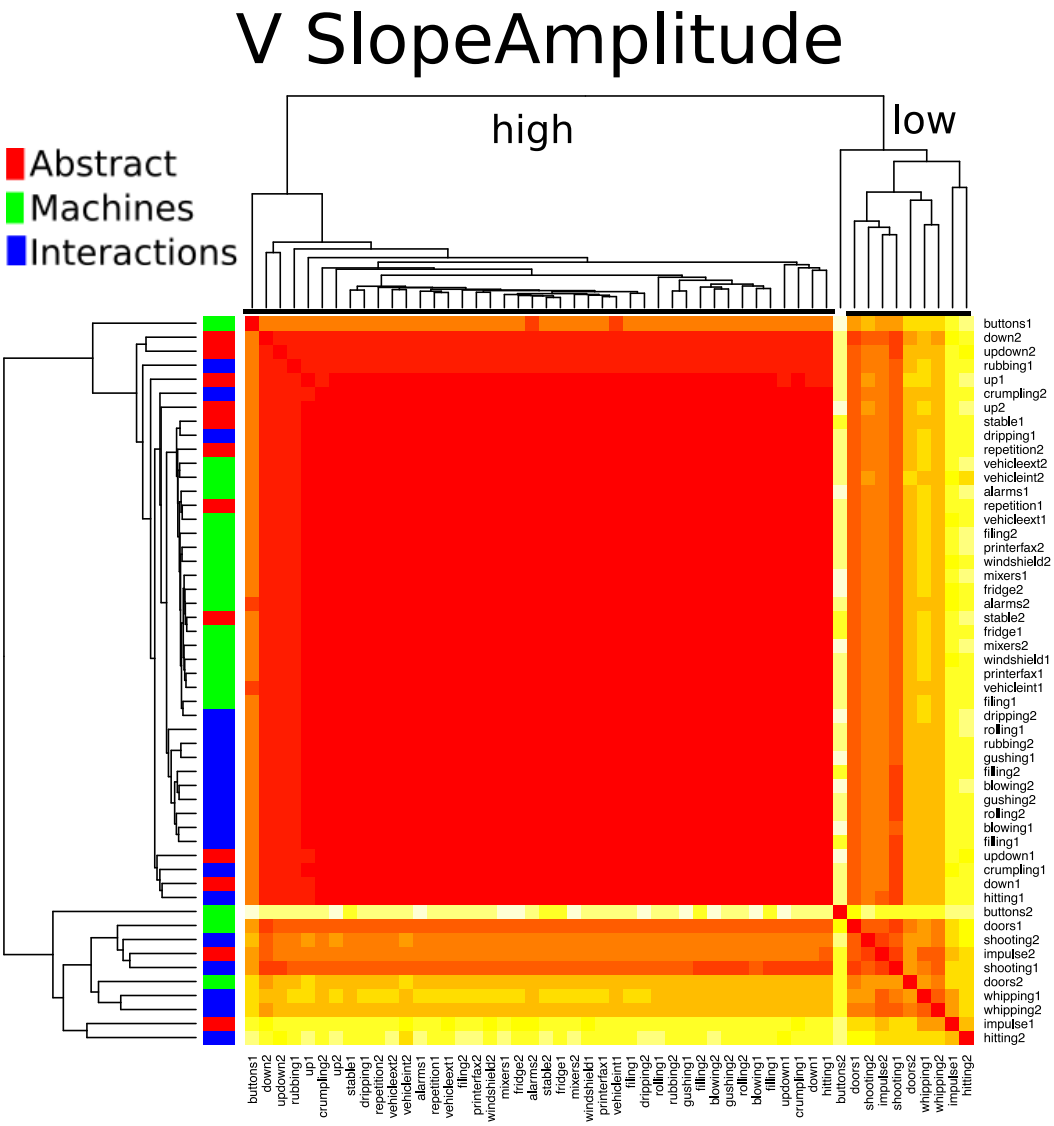

Figure 23: Euclidean distances between imitations for the *slope of amplitude* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map. The associated colors (red, green, and blue) respectively indicate Abstract, Machines and !nteractions families.

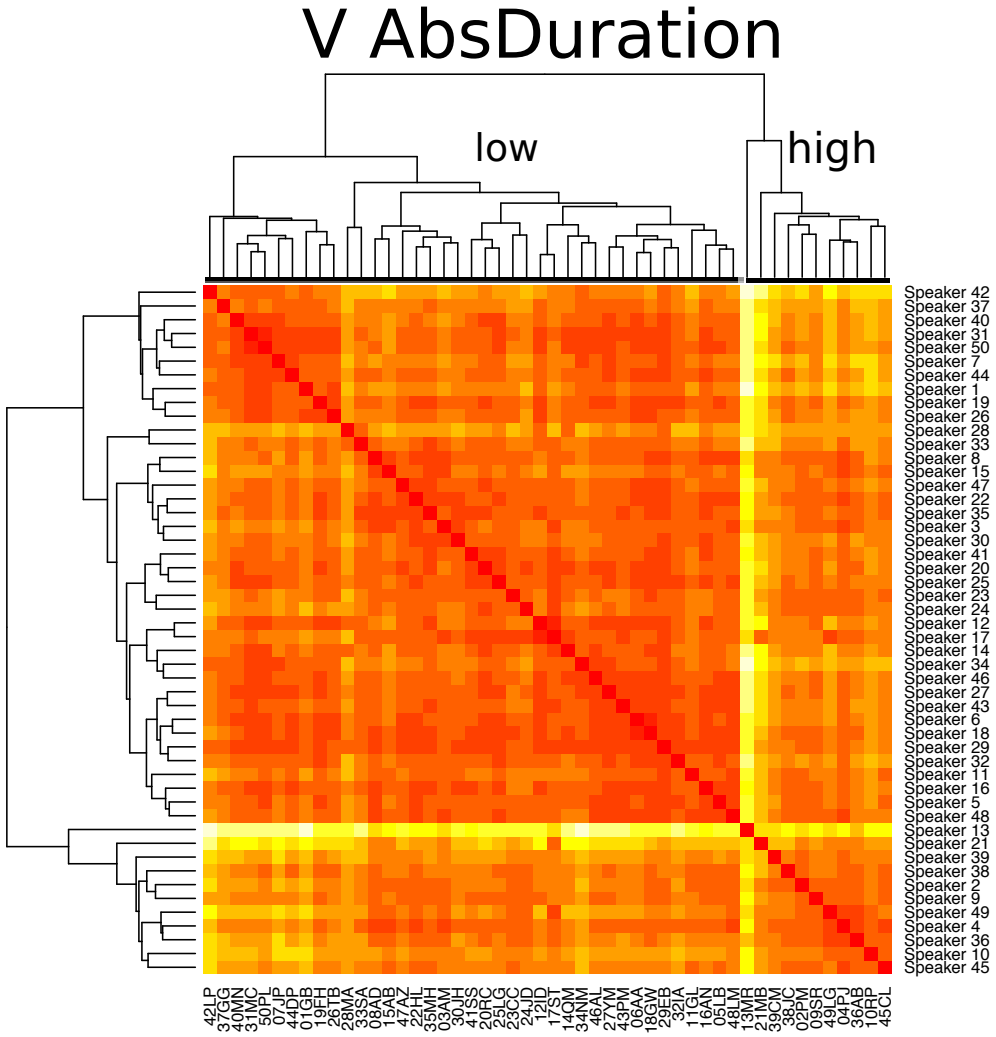

Figure 24: Euclidean distances between speakers for the *absolute duration* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map.

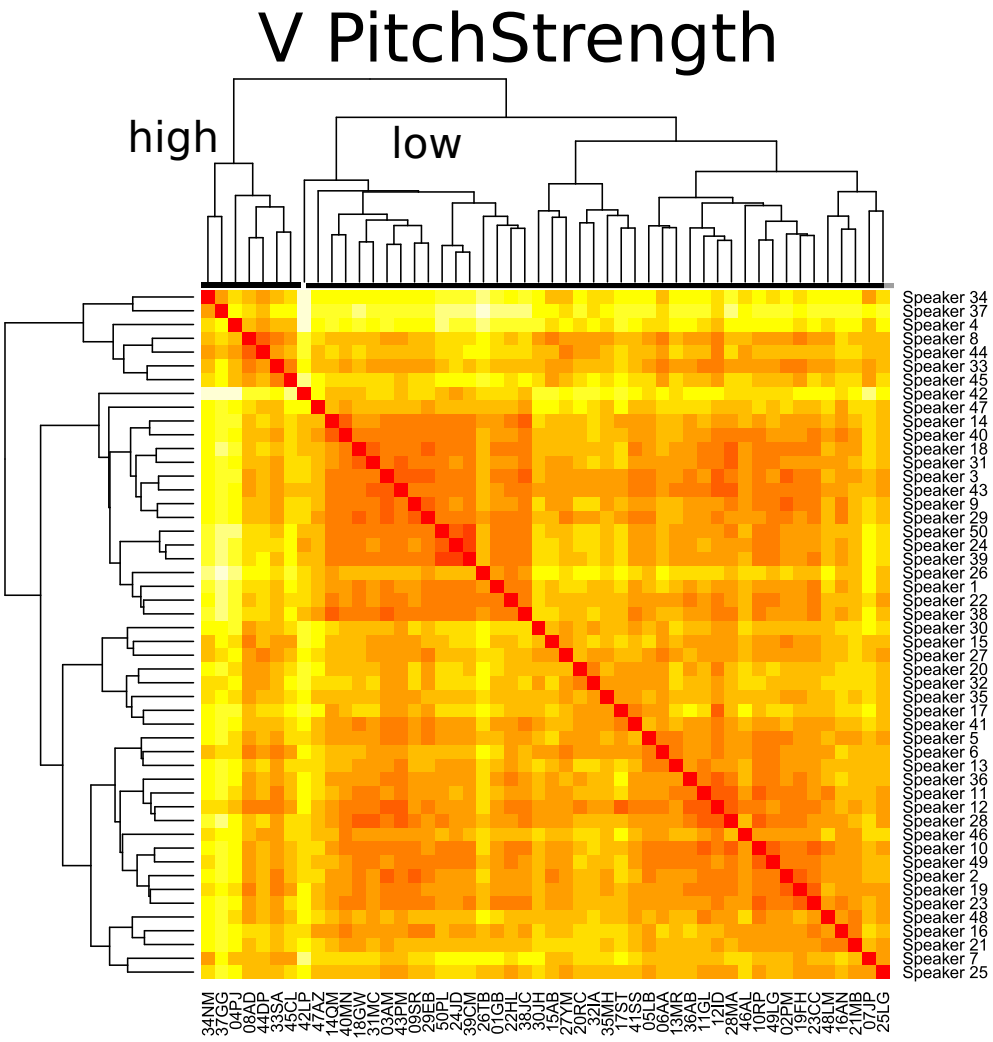

Figure 25: Euclidean distances between speakers for the *pitch strength* descriptor (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map.

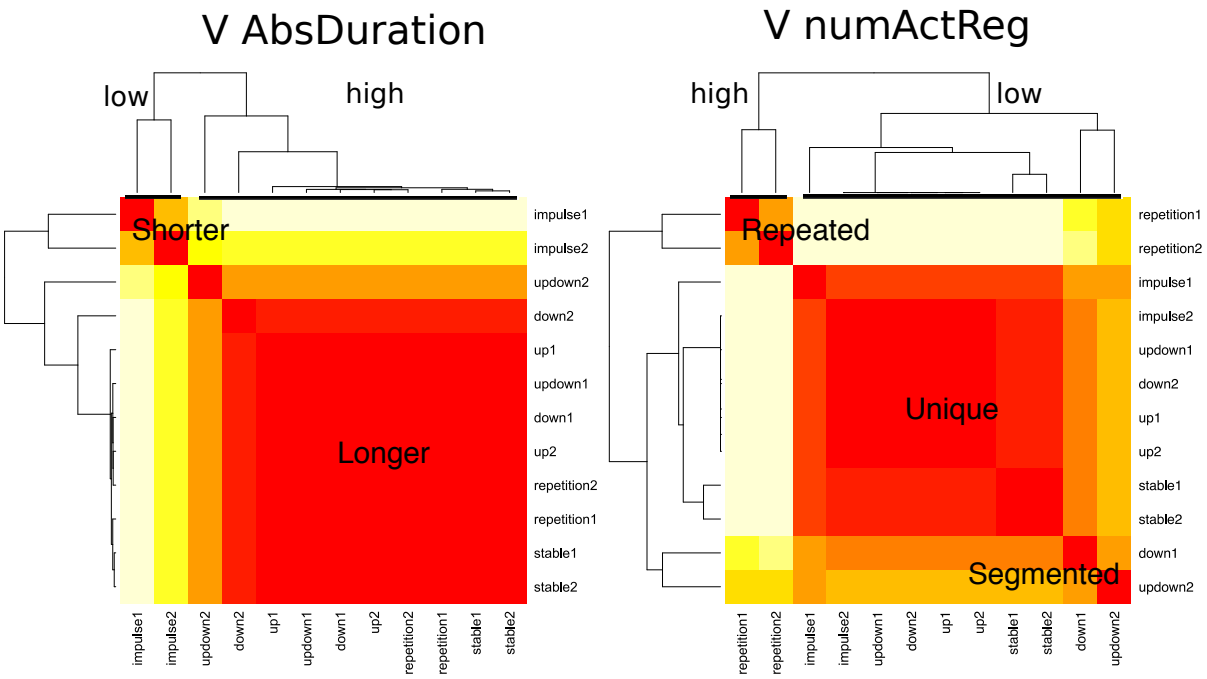

Figure 26: Abstract family. Euclidean distances between imitations for the *AbsDuration* (Left) and *slope 2* (Right) descriptors (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map.

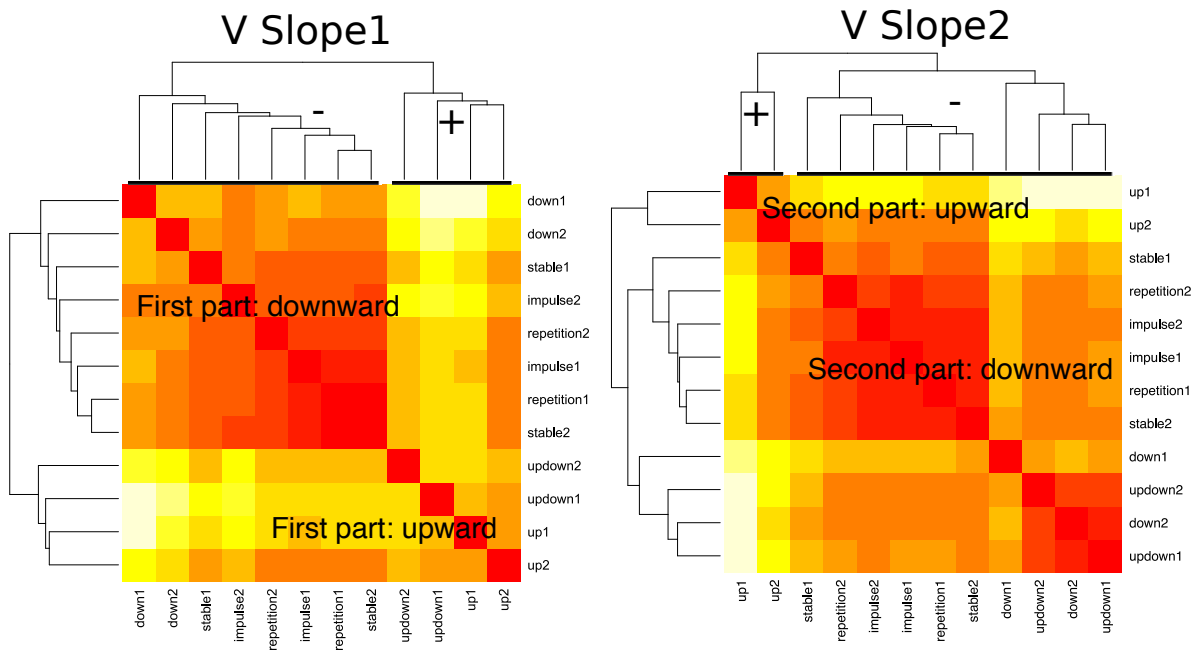

Figure 27: Abstract family. Euclidean distances between imitations for the *slope 1* (Left) and *slope 2* (Right) descriptors (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map.

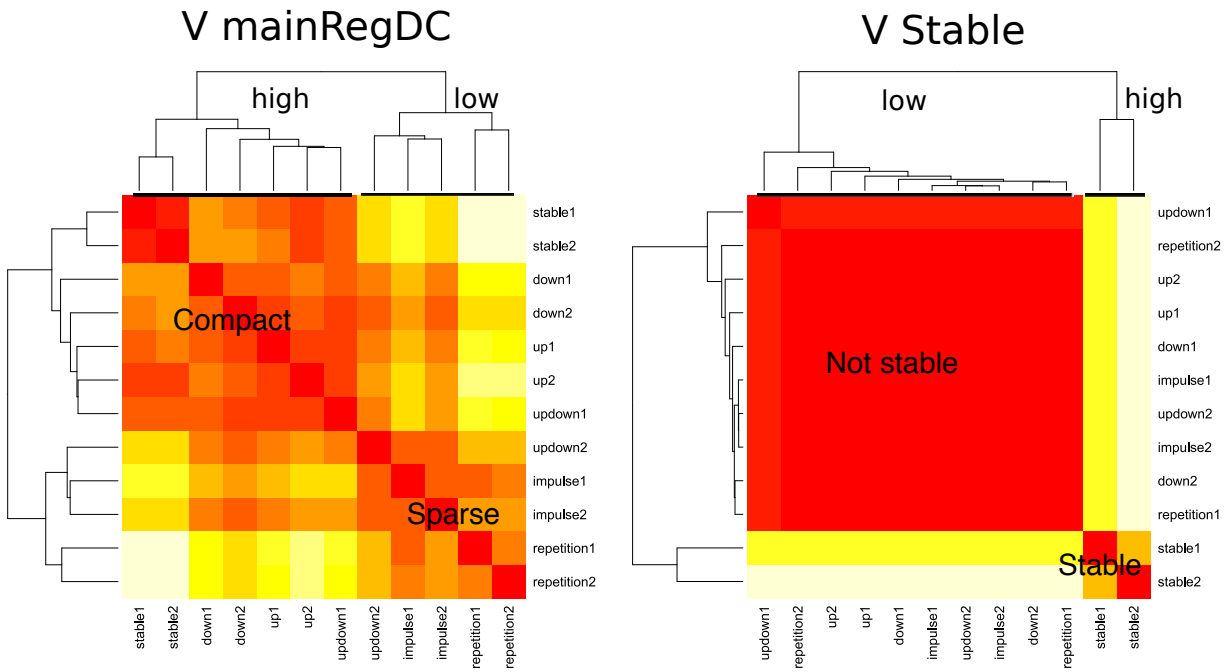

Figure 28: Abstract family. Euclidean distances between imitations for the *mainRegDC* (Left) and *stable* (Right) descriptors (red values are small distances, yellow values high distances). Hierarchical clustering is displayed along the heat map.