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

SkAT-VG:
Sketching Audio Technologies using Vocalizations and Gestures

D2.2.2
Extensive set of recorded imitations

First Author	Sten Ternström
Responsible Partner	KTH
Status-Version:	Draft-1.0
Date:	January 12, 2015
EC Distribution:	Consortium
Project Number:	618067
Project Title:	Sketching Audio Technologies using Vocalizations and Gestures

Title of Deliverable:	Extensive set of recorded imitations
Date of delivery to the EC:	12/01/2015

Workpackage responsible for the Deliverable	WP2
Editor(s):	Sten Ternström
Contributor(s):	Davide A. Mauro
Reviewer(s):	Davide A. Mauro
Approved by:	All Partners

Abstract	The scope of the current deliverable is to present the results of task T2.2: Newly recorded imitations of action primitives, covering all types of sounds that the SkAT-VG system will be able to handle.
Keyword List:	sound collection

Disclaimer:

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

The SkAT-VG Consortium consists of the following entities:

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

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

Document Revision History

Version	Date	Description	Author
First draft	06/11/2014	Import from .tex template	DAM
0.1	06/01/2015	Contributions from KTH	SteT

Table of Contents

Index of Figures

List of Acronyms and Abbreviations

DoW Description of Work

EC European Commission

PM Person Months

WP Work Package

1 Introduction

This deliverable describes the output of Task 2.2, a corpus of multiple-media recordings of skilled imitators imitating four categories of sounds. The bulk of the report is on the methods for data acquisition, with a description of the current status of the database at the end. The acquisition of recordings will continue until the requirements of WP5 for a machine-learning training corpus are met.

2 Data acquisition

2.1 Elicitation

A great deal of discussion and consideration was devoted to the issue of how to elicit the imitations from the imitators. In fact, the effect of the mode of elicitation on the resulting imitation could be a research topic in itself. If the imitator is asked to imitate the sound of a named object or phenomenon, as it was done in Task 2.1, then the reference sound (being the imitator's internal idea of what the object sounds like) is not known, and it becomes impossible to rate the success of the sonic imitation using objective acoustic methods. With listening tests, though, it would still be possible to rate the imitator's success in evoking the idea of the object in the listener. For Task 2.2, the sound to imitate was played as an acoustic reference for the imitator. Here, an interesting question is whether or not the referent sound should also be identified on presentation: "a fan" or "a car engine"; and whether or not such "informed" imitation would lead to different imitation strategies. For the Task 2.2 recordings, the sounds were not identified to the participating imitators.

2.2 Materials (referent sounds)

50 referent sounds were drawn from the 3 major classes identified by the WP4 team: (1) basic mechanical interactions (with solids, liquids and gases as subclasses); (2) abstract sounds; and (3) engine sounds. We strived for a balance in the number of major articulatory mechanisms used by imitators, especially the use of voiced vs. turbulent (fricative) articulations. However, given the unpredictable nature of the imitations, such a balance could not be guaranteed. Therefore, in addition to these 50 referent sounds, 10 referent sounds from the animal sound category were included that were likely to invoke voiced imitations. For machine learning, a data set with a factorial design is the most appropriate, as this will avoid overrepresented types or structures in the data. In other words, one should strive for a similar type/token ratio for all types in the data set. From the point of view of KTH, types imply "vocal primitives", i.e. the set of articulatory descriptors used for labeling the data. For WP4, types may rather be the classes that are to be identified using the informed and blind classifiers.

2.3 Rationale for continuous recording

Although the sounds of imitations in themselves each are quite brief, we chose to record each imitation session in its entirety, and then to use the ELAN tool to index the (small) portions of interest. This preserves the context in which each imitation was made, and gives a record of any incidental events or mishaps that may occur of the course of the recording. It also dramatically reduces the file count, from hundreds to tens. The disadvantage is the large size of the resulting files.

ABSTRACT SOUNDS VoSt_Abs-Fric_os-ui-ios-ussd.wav StCP_Abs-Impu_os-ui-android-popup.wav StCP_Abs-Impu_os-ui-wii-Item-selected.wav Affr_Abs-Impu_os-ui-windowsphone7-menu-pop.wav Affr_Abs-Impu_pica-concha.wav StCP_Abs-Impu_plastic-pipe.wav VoSt_Abs-Iter_game-alice-sonido-coche.wav MyoE_Abs-Iter_game-tinythief-robot-magnet-loop.wav Whis_Abs-Iter_os-notifications-android-12-tweeters.wav MyoE_Abs-Prof_game-superbrothers-marker-28.wav FrDy_Abs-Prof_os-ui-ios-mail-sent.wav VoDy_Abs-Prof_os-ui-ios-received-message.wav Whis_Abs-Tram_game-olo-ho-mainmenu-l2.wav
MACHINE SOUNDS SupG_Ma-Ap-Mo_blender.wav VoSt_Ma-Ap-Mo_electric-shaver.wav StIn_Ma-Ap-NM_coffee-perking.wav FrSt_Ma-Ap-NM_large-pepper-mill.wav StCP_Ma-Me_Do_car-door.wav StCP_Ma-Me-Cl_clock.wav VoDy_Ma-Me-Do_door-squeak.wav Whis_Ma-Si-Al_smoke-detector.wav SupG_Ma-Si-Al_x-ray-buzzer.wav MyoE_Ma-Si-Be_clock-alarm-ring.wav VoSt_Ma-Si-Be_fog-horn-whistle.wav VoDy_Ma-Si-Si_wail-siren.wav SupG_Ma-To-Mo_drill-hammer.wav MyoE_Ma-To-Mo_lawn-mower.wav SupG_Ma-To-Mo_wood-lathe.wav VoDy_Ma-Tr-Ca_honda-accord.wav MyoE_Ma-Tr-Mo_honda-650.wav VoDy_Ma-Tr-Mo_motorbike.wav VoDy_Ma-Tr-Tk_bulldozer.wav MyoE_Ma-Tr-Tr_tractor.wav
MECHANICAL INTERACTION SOUNDS Whis_Mi-Gas-C_cold-gusts-howling.wav Affr_Mi-Gas-C_flame-thrower.wav FrSt_Mi-Gas-C_flow-air.wav

StCP_Mi-Gas-D_explosion.wav Affr_Mi-Gas-D_match-strike-and-ignite.wav StIn_Mi-Liq-C_boiling-bubbles.wav Affr_Mi-Liq-C_filling-a-container-with-spritzer.wav FrSt_Mi-Liq-C_jet-of-water-in-sink.wav Affr_Mi-Liq-D_sloshing-in-red-glass-vase.wav StIn_Mi-Sol_C_crushing-a-can.wav StIn_Mi-Sol-C_cracking-an-egg.wav FrDy_Mi-Sol-C_glass_scrape.wav FrDy_Mi-Sol-C_knife-sharpener.wav FrSt_Mi-Sol-C_paper-rip.wav FrDy_Mi-Sol-C_rip-cloth.wav FrSt_Mi-Sol-C_sanding.wav VoSt_Mi-Sol-D_strike-an-anvil.wav
ANIMAL SOUNDS (51-60) Animal_common-raven.wav Animal_cow-mooing.wav Animal_dog-bark-growl.wav Animal_donkey-braying.wav Animal_elephant-trumpeting.wav Animal_horse-whinny-blow.wav Animal_lion-growl.wav Animal_mosquito.wav Animal_pig-snort.wav Animal_songbird-singing.wav

Table 1: Referent sounds, to be imitated. The list of referent sounds is arranged into the four main classes of sounds proposed by the IRCAM team (animal, abstract, machine and basic mechanical interaction). The referent sounds were further classified by the KTH team according to principal articulatory mechanisms that imitators were likely to use in their imitations. For this data set, we aimed for a balance in the number of imitations using a myoelastic (Vo, Myo, Sup) mechanism vs. turbulent mechanism (St, Fr, Affr), as well as some representation of whistled sounds (Whis). All sounds are quite short, ten seconds or less in duration.

2.4 Participants

To the current date, four imitators have been recorded, and we are aiming for about ten in total. These four imitators were all improvisational actors, recruited through an agency and paid for their participation. The imitators were aged 20-40, two male and two female. All were interested in the task, and performed at a high level.

2.5 Setup

The recordings were carried out in a sound-proofed but not anechoic booth, approximately 5 x 3 x 2.4 m, with a reverberation time of 0.1 s. During the recording the operator sat outside the booth and could communicate with the subject visually through a window, and orally using a talkback microphone (see Figure 1). The two computers were located outside the booth.

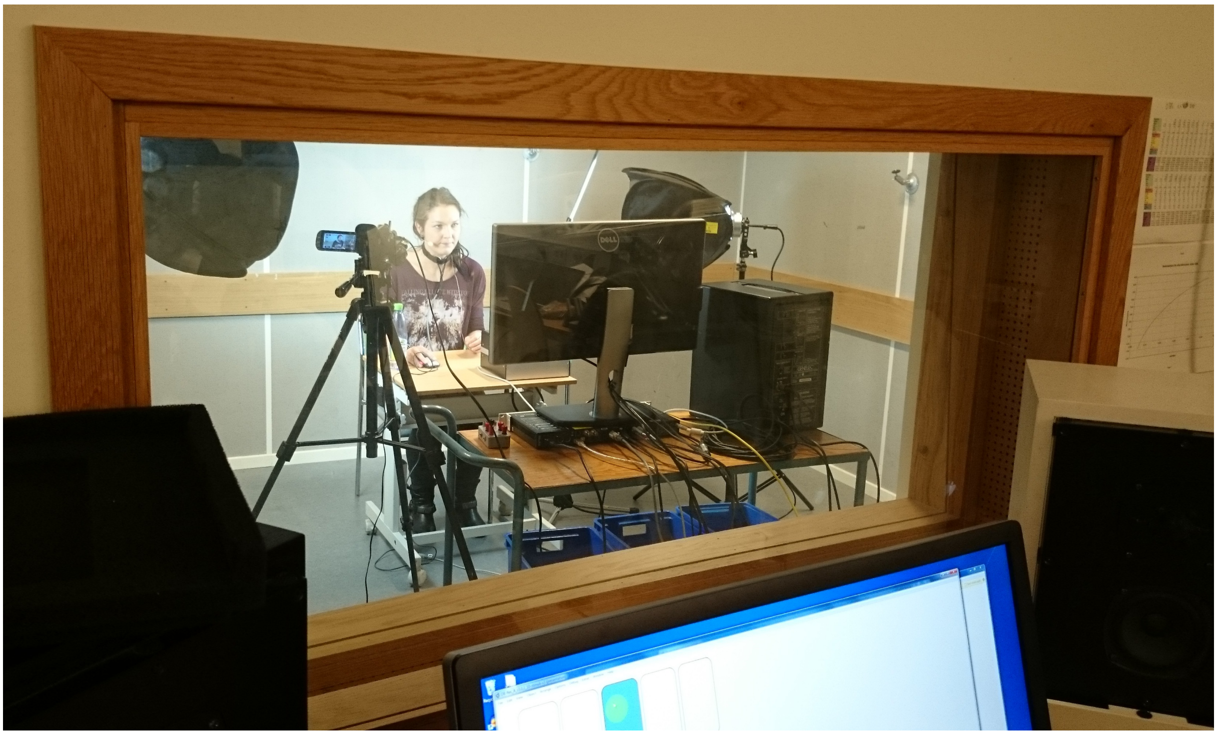


Figure 1: Imitator participant as seen from the control room.

The imitator participants wore (a) an elastic neck strap with the EGG electrodes placed across the prominence of the thyroid cartilage (the "Adam's apple") and (b) a miniature boom microphone at about 7 cm to one side from the mouth (model 4066, DPA Microphones, Allerød, Denmark). They sat at a small table with a flat screen monitor and a mouse. The referent sounds were presented to the imitators through an active monitor-grade loudspeaker (Genelec, model 1031A) at a distance of approximately 2 meters. This loudspeaker presented also the imitator's own efforts for comparison. The voice of the experimenter outside the booth could also be played on this loudspeaker.

Although both video cameras can and did record the audio, for quality reasons the main audio recording was made separately, using a pro-quality digital audio interface (model Fireface UFX, RME, Germany, `www.rme-audio.de`) running at 24-bit sample resolution and 48 kHz sampling rate. The UFX device contains also microphone preamps, a router of analog and digital audio signals, and a backup recorder for USB-compatible media. Three channels of audio-frequency sound were recorded: ch1 = microphone audio; ch2 = EGG signal; ch3 = sync signal + referent sounds (both from the control software in Max/MSP). The gain of the microphone was adapted to the vocal loudness of the subject. The sound level was not calibrated for, as this was not part of the research question. The electroglottograph (model MC2-1, Glottal Enterprises, Syracuse, NY) was a dual-channel device (with upper and lower electrode pairs). Its "Average" output was used, for a normal single-channel representation. The gain of the EGG signal is not critical; a five-fold gain option was selected when necessary. The dialogue between imitator and experimenter, where present, is clearly audible on channel one.

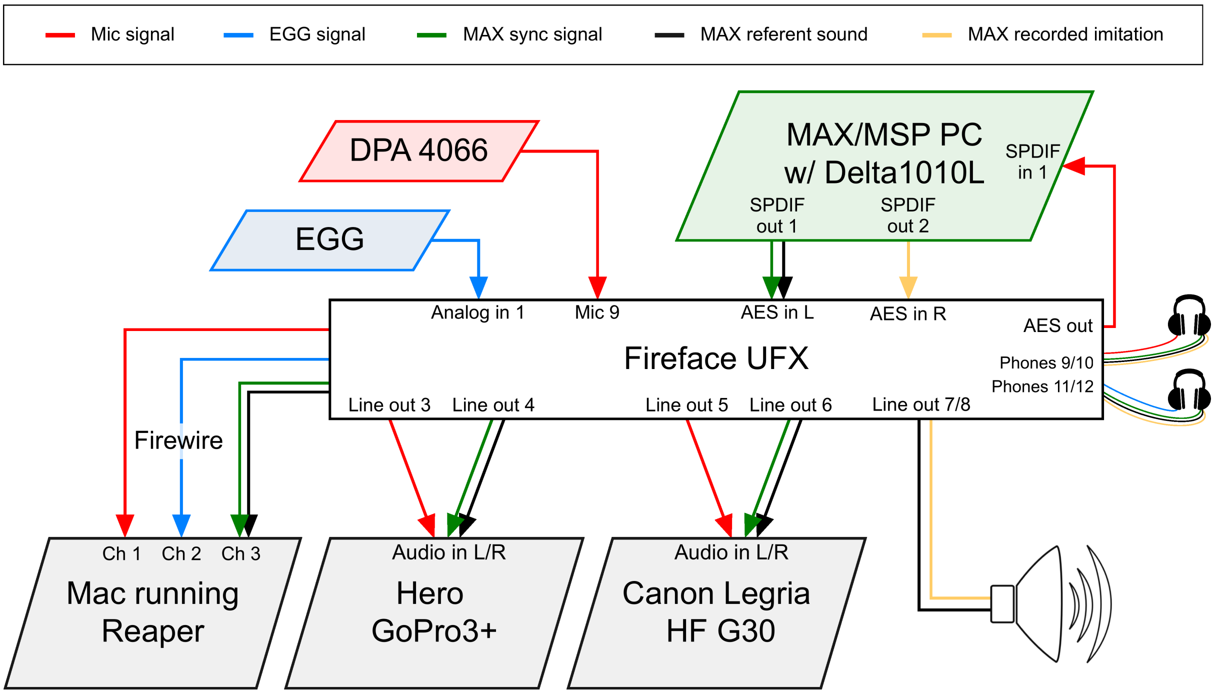


Figure 2: Block diagram of audio connections.

Two video streams were recorded. The first was shot using a full-size video camera (Canon model Legria HF G30, recording on a SD/XC memory card), with the imitator in an anterior view, slightly offset to capture the right side of the imitator's face (Figure 3 (a)). The resolution was set to 1920 x 1080 pixels, and the frame rate was set to 50 fps. For synchronization purposes, the front camera recorded also the microphone audio and the sync signals on separate audio channels. In this camera's view, a fair amount of space was allowed around the imitator, so as to capture also spontaneous gestures, if any, for possible subsequent use by WP4.

The second video stream was shot using a miniature video camera (Hero, model GoPro3+, recording on a micro SD memory card) that recorded the imitator with the lens at neck height, at an angle of approximately 45°, thus capturing the left side of the imitator's face fully, with the right side partly obscured (Figure 3 (b)). The resolution was set to 1280 x 720 pixels, the frame rate was set to 100 fps and the 'Narrow' view option was selected. The side camera, too, recorded the microphone audio and the sync signals on its two separate audio channels.

Type	Manufacturer and Model	Purpose
Digital Audio Interface	RME Fireface UFX	Mic pre-amps, routing of audio signals, talkback
Movie Camera	Hero GoPro3+ Black Edition	Side View
Movie Camera	Canon Legria HF G30	Front View
Daylight Lamps	2 Tristar Magic Square	Lighting
Computer 1	Apple Macbook Pro, OSX 10.6.8	Control the digital audio interface, makes contiguous audio+EGG recordings
Computer 2	Dell Optiplex 750 PC, Windows 7	Run experiment control software
Microphone	DPA model 4066 omnidirectional condenser boom-mounted	Record voice
Electroglottograph	Glottal Enterprises Model MC2-1	Distinguish episodes of glottal phonation (as opposed to other sources of vibration)
Loudspeaker	Genelec 1031A active monitor	Play audio to imitator, talkback from operator
Control Software	Custom patch (in MAX/MSP v.6, cycling74.com)	Present stimuli and record responses, produce trigger signals
Recording Software	Reaper v 4.62 (Cockos Inc., New York, N.Y.)	Record microphone & EGG signals and prompts/interactions

Table 2: List of equipment.

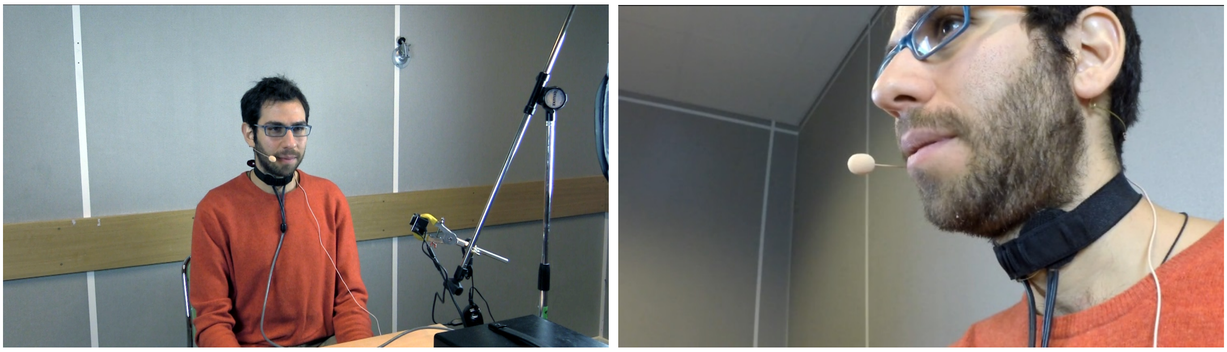


Figure 3: (a) Front camera view (example), (b) Side camera view (example).

2.6 Experiment control

For the presentation of referent sounds, a Max/MSP program script written at IRCAM (Guillaume Lemaitre, modified slightly by Pétur Helgason) was used. The program presented 10 referent sounds at a time, starting with the set of ten animal sounds, but otherwise in randomized presentation order. That is, the order was randomized across the 50 referent sounds from the categories 'abstract', 'machine', and 'mechanical interaction'; with a separate randomization for the initial group of 10 referent animal sounds. The imitators selected and played back sounds from the computer screen using the mouse. Each referent sound had an associated playback button and record button, as well as a button to play back the recorded imitation, see Figure 4. Imitators could play referent sounds at will and record the ten presented referent sounds in any order.

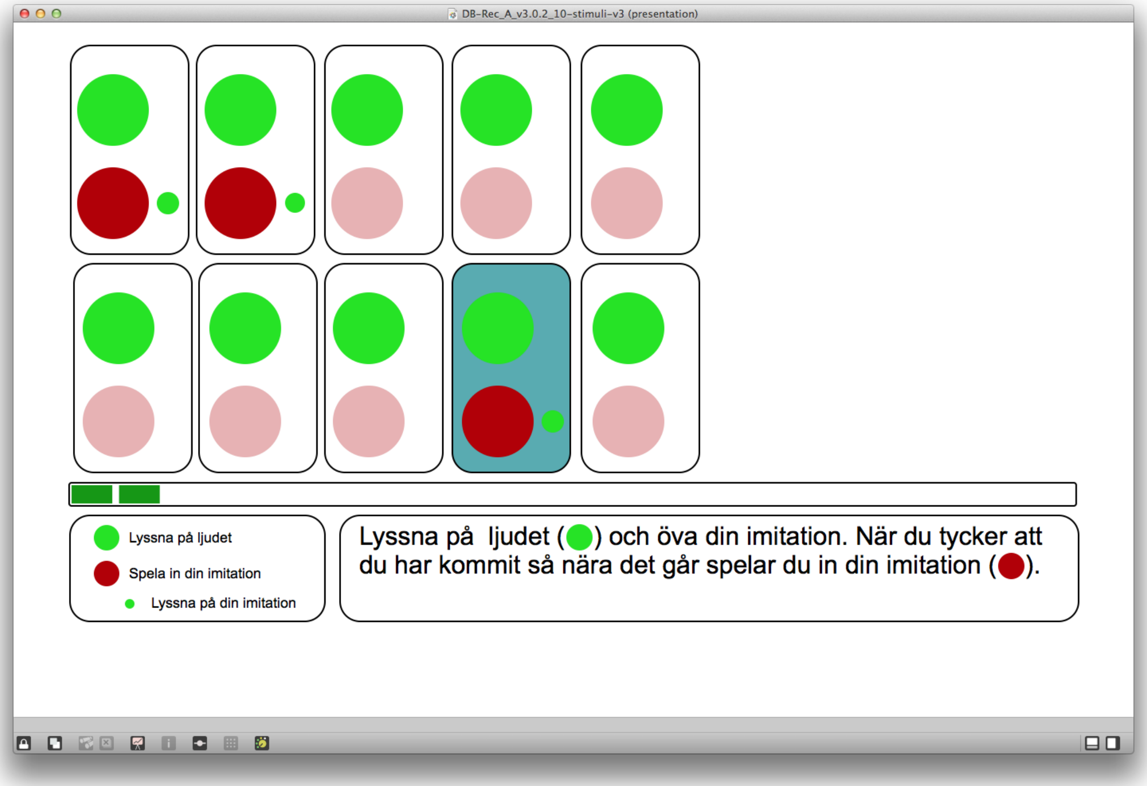
The Max/MSP script generated a synchronization signal at the start of the experiment and at the start of each recorded imitation. It also generated label files with time stamps to facilitate the subsequent annotation work in WP3. It created a separate WAV file for each recorded imitation (mono, at 44.1 kHz sampling rate, 16 bits.)

2.7 Procedure

Imitator participants arrived at the KTH studio by appointment through an impresario agency (Improvisationsstudion AB, Stockholm, Martin Geijer). The imitators were briefly informed that the goal of the experiment was to study how people imitate non-verbal sounds.

Directions were given to the imitators to attempt to give the impression of a sound in a manner analogous to sketching. The analogy was presented to the imitators that, if given a photograph of a horse, they would not try to copy the photograph, but rather sketch the horse. They were asked to try to apply that mindset in their imitations.

The experiment was run in six sets of ten referent sounds. For each set, first the video and audio recordings were started, and then the Max/MSP interface (Figure 4) was invoked. A typical session took two to three hours, with short breaks when requested by the imitator. After the recording session, each imitator signed a Consent Form (see 3) approving only the scientific use of the recordings, including possible presentation at conferences, etc.



● Lyssna på ljudet

● Spela in din imitation

● Lyssna på din imitation

Lyssna på ljudet (●) och öva din imitation. När du tycker att du har kommit så nära det går spelar du in din imitation (●).

Figure 4: The GUI presented to the Swedish imitators (coded in Max/MSP). The instruction is, "Listen to the sound (green circle) and practice your imitation. When you think you are as close as you can get, record the imitation (red circle)." Here, items 1, 2 and 9 have already been recorded and the subject has yet to listen to the remaining referent sounds.

2.8 Post-processing

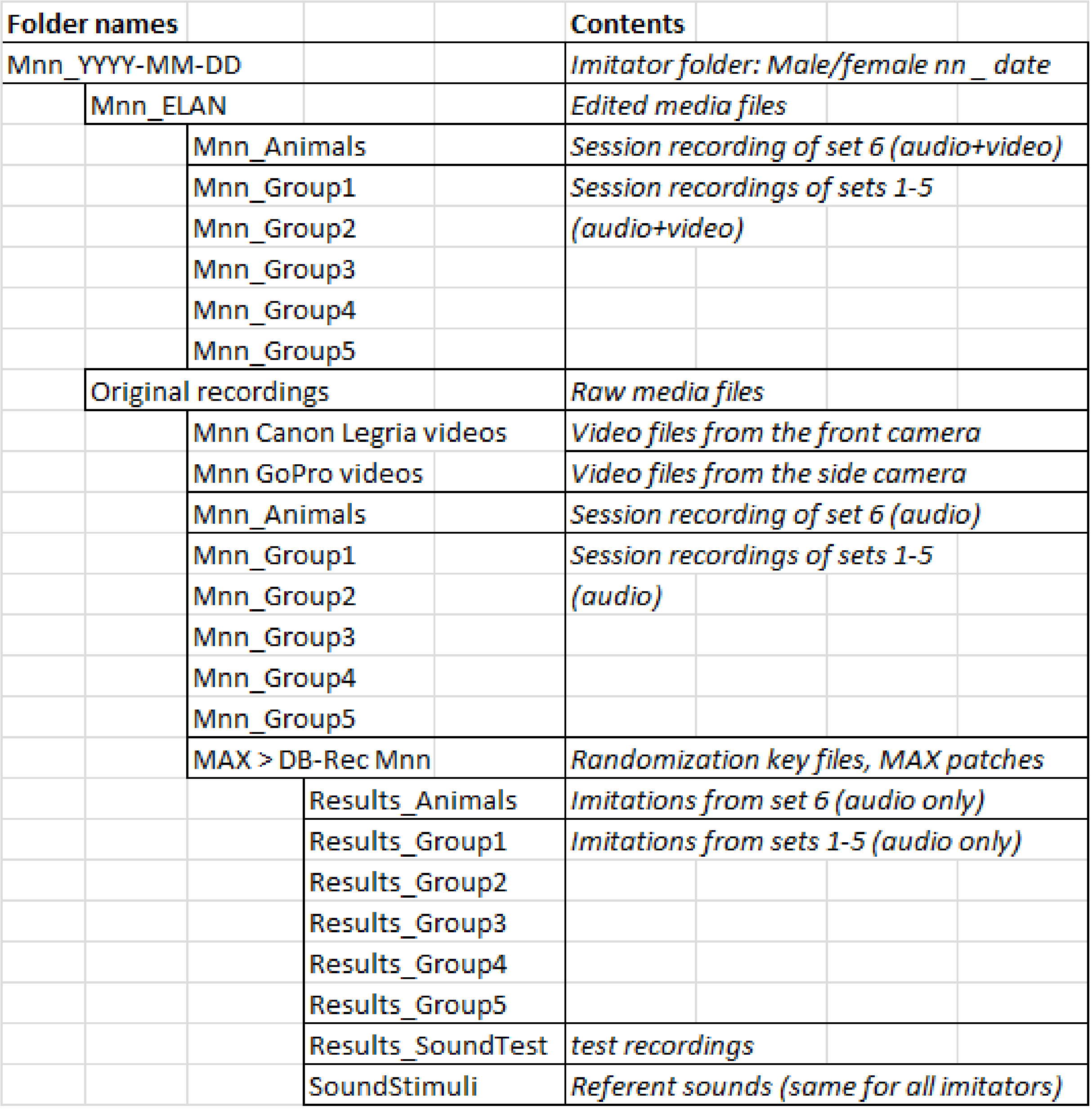
The entire recording sessions were captured in contiguous files, which could become very large. After recording, the media files were transferred from the two camera memory cards and from the Macbook Pro to the central file server. A potential problem is the lack of time synchronization in hardware. The audio and two video recordings could drift by up to 100 ms per hour relative to each other, with the drift being slightly different from recording to recording. We did consider the possibility of using electronic synchronization, using the industry-standard SMPTE protocol, but found that it would incur too much technical overhead. The audio recordings were therefore time-base corrected manually (by minuscule amounts) to stay in sync over the entire recording. This was made possible thanks to using a time-scaling function in the Audacity sound editor, with no audible loss of quality. The ELAN software can manage only 16-bit audio files, so the reference audio recordings also had to be truncated from 24 to 16 bits resolution. Also, the GoPro3+ video camera divides recordings larger than 2 GB into multiple files. These were manually concatenated. Both the original files and the files processed as described were saved on the central server. From there, the processed files were mounted into an ELAN workspace and synchronized manually, using the sync signal that was recorded on one audio track of both cameras and track 3 of the audio-only recording. Once synchronized, the media files are ready for annotation.

2.9 Results

Task 2.2 has resulted in a database of recordings of four skilled imitators. This database will continue to grow in months 13-15. The file system layout of this database is shown for one imitator (Figure 5). Since the entire sessions are recorded, each imitator generates about 100 GB of edited video and audio files, plus 100 GB of raw original files.

2.10 Database location

The recordings of skilled imitators are stored on a file server centrally managed by the KTH IT-support division, under a specific contract with the department of Speech, Music and Hearing. The storage is backed up. Access is currently authorized only to WP2 collaborators with KTH login credentials. We are investigating how to make the database accessible to WP5 across the Internet.



Folder names				**Contents**
Mnn_YYYY-MM-DD				*Imitator folder: Male/female nn _ date*
	Mnn_ELAN			*Edited media files*
		Mnn_Animals		*Session recording of set 6 (audio+video)*
		Mnn_Group1		*Session recordings of sets 1-5*
		Mnn_Group2		*(audio+video)*
		Mnn_Group3		
		Mnn_Group4		
		Mnn_Group5		
	Original recordings			*Raw media files*
		Mnn Canon Legria videos		*Video files from the front camera*
		Mnn GoPro videos		*Video files from the side camera*
		Mnn_Animals		*Session recording of set 6 (audio)*
		Mnn_Group1		*Session recordings of sets 1-5*
		Mnn_Group2		*(audio)*
		Mnn_Group3		
		Mnn_Group4		
		Mnn_Group5		
		MAX > DB-Rec Mnn		*Randomization key files, MAX patches*
			Results_Animals	*Imitations from set 6 (audio only)*
			Results_Group1	*Imitations from sets 1-5 (audio only)*
			Results_Group2	
			Results_Group3	
			Results_Group4	
			Results_Group5	
			Results_SoundTest	*test recordings*
			SoundStimuli	*Referent sounds (same for all imitators)*

Figure 5: Structure of the database of results, for one imitator.

3 Appendix: Consent Form

SkAT-VG
Sketching Audio Technologies
using Vocalizations and Gestures

Consent form

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

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

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

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

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

Signature

Date

Printed name

□

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