MASTER'S THESIS

February 9, 2015 — August 7, 2015

Combination of gesture and vocalization in the imitation of sounds

HUGO SCURTO

Supervisors:

GUILLAUME LEMAITRE[1]

FRÉDÉRIC BEVILACQUA[2]

JULES FRANÇOISE[2]

PATRICK SUSINI[1]

ircam Centre Pompidou

1 place Igor Stravinsky

[1]Sound Perception and Design team *75004 Paris*

[2]{Sound Music Movement} Interaction team *France*

Résumé

Communiquer sur les sons n'est pas chose aisée pour qui n'a pas le vocabulaire technique adapté : l'usage de vocalisations non-linguistiques et de gestes est alors souvent privilégié. De précédents travaux ont indépendamment étudié la description gestuelle de sons [Caramiaux et al., 2014] ainsi que l'efficacité des vocalisations pour communiquer un son [Lemaitre and Rocchesso, 2014]. Cependant, des études en communication suggèrent un lien plus intime entre voix et geste [Kendon, 2004]. Notre but est donc de comprendre le rôle du geste dans l'imitation de sons, lorsqu'il est utilisé conjointement aux vocalisations.

Le travail de ce stage a été de : (a) formuler des hypothèses à partir d'une analyse qualitative d'une base de données, (b) construire une expérience ainsi que des mesures adaptées afin de tester statistiquement ces hypothèses, et (c) d'appliquer de nouveaux outils d'analyse du geste à la construction d'un classificateur de gestes.
Nous avons tout d'abord analysé qualitativement des données d'imitations vocales et gestuelles de plusieurs sons (données audio, vidéo et de mouvement). En émergent trois hypothèses : (1) la voix est plus précise que les gestes dans l'air pour communiquer des informations rythmiques, (2) des aspects de texture sont communiqués par des gestes tremblants, et (3) deux sons peuvent être imités simultanément en utilisant geste et voix. Ces hypothèses sont validées au cours d'une nouvelle expérience, dans laquelle 18 participants ont imité 25 sons synthétisés pour l'occasion : bandes de bruits rythmées, textures granulaires, et sons simultanés. Des analyses statistiques comparent les caractéristiques acoustiques des sons synthétisés à celles des vocalisations, ainsi qu'à de nouvelles caractéristiques du geste extraites d'une représentation en ondelettes de données d'accélération. Les données collectées nous permettent enfin de construire un classificateur pour gestes tremblants utilisant la méthode des k plus proches voisins, avec 79% de reconnaissance.

Mots-clés: perception sonore, cognition incarnée, geste, voix, k plus proches voisins.

Abstract

Communicating about sounds is a difficult task without a technical language, and naive speakers often rely on different kinds of non-linguistic vocalizations and body gestures. Previous distinct works have been done on gestural description of sounds [Caramiaux et al., 2014] and vocal imitation effectiveness to communicate a sound [Lemaitre and Rocchesso, 2014]. However, speech communication studies suggest a more intimate link between the two processes [Kendon, 2004]. Our study thus aims at understanding the role of gesture when combined with vocalization in sound imitation.

The work of this internship was to: (a) extract hypotheses from a qualitative analysis of a database, (b) construct an experiment to test these hypotheses as well as adapted measures to make statistical analyses, and (c) apply newly created gesture analysis tools to build a gesture classifier.
We first used a large database of vocal and gestural imitations of a variety of sounds (audio, video, and motion sensor data). Qualitative analysis of gestural strategies resulted in three hypotheses: (1) voice is more precise than air gestures for communicating rhythmic information, (2) textural aspects are communicated with shaky gestures, and (3) concurrent streams of sound events can be split between gesture and voice. These hypotheses were validated in a new experiment in which 18 participants imitated 25 specifically synthesized sounds: rhythmic noise bursts, granular textures, and layered streams. Statistical analyses compared acoustic features of synthesized sounds, vocal features, and a set of novel gestural features based on a wavelet representation of the acceleration data. Collected data finally allowed us to build a k-nearest neighbor-based classifier for shaky gestures with 79% recognition accuracy.

Keywords: sound perception, embodied cognition, gesture, voice, k-nearest neighbors.

Acknowledgements

I would like to thank all of my supervisors (Guillaume Lemaitre, Frédéric Bevilacqua, Jules Françoise and Patrick Susini) for having trusted me in the carrying out of this work, thus allowing me to bring my research work closer to my profound interests. A special thank for Frédéric Voisin for having shared his graceful ethnomusicological memories with me. I also thank all the members of the Sound Perception and Design team for the pleasant atmosphere we shared these six months. Lastly, real love for all the other interns I met, with whom I learned as much as in the course of my work.

Table of contents

Abstract v

Acknowledgements vii

1 Introduction 1

2 Background: gesture and voice 3
- 2.1 On gesture . . . 3
  - 2.1.1 Gesture in speech . . . 4
  - 2.1.2 From musical gestures to embodiment in music . . . 4
  - 2.1.3 Motion descriptors . . . 5
- 2.2 On vocalization . . . 6
  - 2.2.1 Spoken and sung voice . . . 7
  - 2.2.2 Non-verbal vocalization . . . 8
  - 2.2.3 Audio descriptors . . . 8
- 2.3 Combining gesture and vocalization . . . 9

3 Background: sound perception and imitation 11
- 3.1 The listening process . . . 11
  - 3.1.1 Interpretating auditory information . . . 11
  - 3.1.2 Auditory imagery . . . 12
- 3.2 The imitation process . . . 13
  - 3.2.1 Cognitive approaches . . . 13
  - 3.2.2 Music and imitation . . . 14
- 3.3 Sound imitation . . . 15
  - 3.3.1 Using gesture . . . 16
  - 3.3.2 Using vocalization . . . 16
- 3.4 Conclusion — Methodology . . . 18

4 Imitating sounds: first study 19
- 4.1 Data collection . . . 19
  - 4.1.1 Method . . . 19
  - 4.1.2 Experimental setup . . . 20
- 4.2 Preliminary analysis . . . 21
- 4.3 Focused analysis . . . 22
  - 4.3.1 Shared aspects of imitations across participants . . . 23
  - 4.3.2 Separation of vocalization and gesture . . . 24
  - 4.3.3 Accuracy of the imitations . . . 25
  - 4.3.4 Modifying the vocalization with gesture . . . 26
- 4.4 Conclusion: drawing up hypotheses . . . 28

5 Combining gesture and vocalization 29
5.1 Designing a new experiment . . . 29
5.1.1 Creating abstract sounds . . . 30
5.1.2 Method . . . 33
5.2 Analysis: rhythmic sounds . . . 34
5.2.1 Tempo tracking . . . 34
5.2.2 Rhythmic pattern reproduction . . . 36
5.3 Analysis: textural sounds . . . 40
5.3.1 Vocal strategies . . . 40
5.3.2 Gestural behaviour . . . 42
5.4 Analysis: layered sounds . . . 43
5.4.1 Global analysis . . . 43
5.4.2 Strategies' specifications . . . 46
5.5 Discussion . . . 47

6 A classifier for shaky gestures 51
6.1 Classifier specification . . . 51
6.1.1 Database description . . . 51
6.1.2 Computed features . . . 52
6.2 Evaluation . . . 52

7 Conclusion 53

Bibliography 55

Appendix A

1

Introduction

Sound design, as the process of specifying, acquiring, manipulating or generating audio elements, is one of the growing cultural practice of our time. The SkAT-VG European project (*Sketching Audio Technologies with Vocalizations and Gestures*), in which this study is incorporated, aims at enabling designers to use their voice and hands, directly, to sketch the auditory aspects of an object, thereby making it easier to exploit the functional and aesthetic possibilities of sound. People's natural use of vocalizations and gestures to communicate sounds thus needs understanding ahead of any thinking of the technology involved.

Our study will focus on the combination of gesture and vocalization in the imitation of sounds among French people*. With an experimental study, we aim at *understanding the role of gesture* when combined with vocalization in sound imitation, in order to technically integrate interaction design ideas in a sound sketching tool as part of the SkAT-VG project.

To address this issue, we decided to explore different scientific fields by way of state of the art. In a first part, we define *gesture* and *voice* along with technical means to describe them: we thus have an insight of these two streams of communication. In a second part, we investigate the *sound imitation* process, dividing it into two processes: listening to a sound, and imitating it. We conclude our state of the art by extracting our study's metholodogy from it.

In a third part, we *analyze qualitatively* a database of vocal and gestural imitations of a variety of sounds: three hypotheses emerge from this analysis. In a fourth part, we *construct a new experiment* to test our hypotheses as well as *adapted measures* to make statistical analyses. Finally,

the previously collected data allows us to build a *gesture classifier* that we specify in a fifth part.

We took the decision to tackle several issues rather than one with the underlying idea of raising research prospects. As a consequence, issues we have raised are general, but still offer a wealth of information. Our experimental approach is also not conventional, but it does not prevent it from being a research subject in itself. Finally, analyses being not similar between raised issues is a consequence of the latter being original and creative.

* The cultural factor is essential in understanding communication. A population's vocal habits and abilities are shaped by its mother tongue; and vocabulary as body language are shaped by its cultural practices. Communicating about sounds is thus also shaped by cultural practices. Trying to grasp a possible universality in body language and vocalization is not our intention here: yet we want to underline that sound imitation, as being rooted in communication, is rooted in culture. As previously mentioned, we decided to focus only on French people's sound imitation.

2

Background: gesture and voice

This chapter provides the reader with an overview on gesture and voice studies, starting from basic definitions to specific aspects involved in speech and music. We also expose techniques to acquire and analyze them. Finally, we present fields situated at the confluence of these two streams of communication, letting us suggest a more complex coupling between them.

2.1 On gesture

Defining what a gesture is in its globality is not an easy task. There seems to be as many descriptions as there are fields in which it appears. We can try to summarize these by roughly saying that gesture is a *visible bodily action that has a meaning*. In the frame of man-machine communication, [Cadoz, 1994] asserts that one of the particularity of the 'gestural channel' — as he calls it — is that it is both a means of *action* in the physical world and a means of *communication*. Definitions arising from speech and music research add a *metaphorical* viewpoint to gesture, as metaphor can be involved when gestures work as concepts that project physical movement, sound, or other types of perception to cultural topics [Godøy and Leman, 2010]. We will mainly focus on these two fields of research (*I.e.* speech and music), as they are intimately related to our study.

2.1.1 Gesture in speech

Speech research has identified different phases in gesture used during communication. This identification is necessary to understand the nature of the coordination between gestural and speech components, as we will see in section 2.3.

When a person engages in gesturing, the body parts that are employed undertake a movement excursion. This excursion, from the moment the articulators begin to depart from a position of relaxation until the moment when they finally return to one, is referred to as a *gesture unit* [Kendon, 2004]. The phase of the movement excursion when the 'expression' of the gesture is accomplished is called the *stroke*. The phase of movement leading to the stroke is termed the *preparation*. The phase of movement that follows, as the body part is relaxed, is referred to as the *recovery*. The stroke may sometimes be followed by a phase in which the articulator is sustained in the position at which it arrived at the end of the stroke. This is referred to as the *post-stroke hold* and spares the speaker to relax between two different strokes. We can thus define a *gesture phrase* as a package containing a preparation, a stroke and a post-stroke hold. This is summed up in figure 2.1.

To sum up, the *gesture unit* is the entire excursion of the articulator of the gestural action. This excursion may contain one or more *gesture phrases*. It is, generally speaking, the *strokes* of such gesture phrases that are picked out by casual observers and identified as 'gestures'.

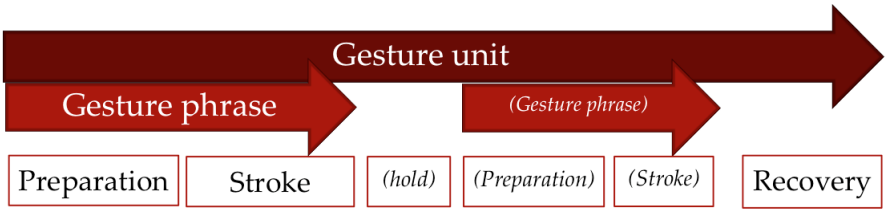

Figure 2.1: Gesture unit as defined in [Kendon, 2004].

2.1.2 From musical gestures to embodiment in music

During the performance of a musical piece (as during listening or during dancing activities), we engage with musical sounds through body movement. The underlying concept of *musical gesture* implies an interdiplinary approach, the main issues of which are resumed in the following paragraphs.

Musical gestures

A musical gesture is a human body movement that goes along with sounding music. We can divide musical gestures into four functional categories [Godøy and Leman, 2010]:

- *Sound-producing gestures*: those that effectively produce sound, such as pushing a piano key (also called instrumental gestures by [Cadoz, 1994]);
- *Communicative gestures*, as [McNeill et al., 1990,Kendon, 2004] consider it (see section 2.1.1), are used to communicate with other performers or perceivers (for example, pointing the audience);
- *Sound-facilitating gestures*, often called ancillary gestures, support the sound-producing gestures in various ways (for example, moving one's head while playing the violin);
- *Sound-accompanying gestures* are not involved in the sound production itself, but follow the music [Godøy et al., 2006] (sound-tracing or mimicry of sound-producing gestures, see section 3.3.1).

In all of this, it is important to remember that most musical gestures (if not all) may have multiple significations, ranging from the more physical to the more metaphorical.

Embodiment in music

The previous definitions connects well with recent approaches in *embodied music cognition* [Leman, 2008]. This theory, hypothesizing that the nature of musical communication is rooted in a particular relationship between musical experience (mind) and source energy (matter), sees the human body as the mediator in this two-way cognitive process. It notably stipulates that *intentionality* can be conceived of as an emerging effect of motor resonances (corporeal articulations), the latter linking the complexities of the physical world to our personal experiences. Intentionality would thus be grounded in the coupling of action and perception. Through this coupling, the human brain creates an action-oriented ontology[1] of the world that would form the basis of musical communication.

2.1.3 Motion descriptors

Human beings seem to have little or no problem with perceiving and understanding the expression of them [Godøy and Leman, 2010]. What

[1]Ontology is the philosophical study of the nature of being, becoming, existence, or reality.

is easy for us to understand may be very difficult to measure for machine-based systems of gesture recognition. On the paper, the processes required for it are quite the same as those for audio signals (see section 2.2.3 for more details): a *sensor* measures some signal from which some *features* are extracted, the latter being *recognised* and converted into a symbol. However, relevant features for gesture are not as well-established as for audio signals. Also, choosing a sensor (camera, leap-motion, Kinect, accelerometers and gyrometers for example) already influences what we will be able to measure.

[Camurri et al., 2004] propose algorithms and computational models for real-time analysis of expressive gesture in full-body human movement, based on computer vision. Starting from body silhouettes and tracking information, they extract features such as what they call Quantity of Motion, which can be considered as an overall measure of the amount of detected motion, involving velocity and force.

Françoise (unreleased work) developed novel gestural features based on wavelet representation in the frame of the SkAT-VG project. Acceleration data (or position data) goes through a simple wavelet filter bank that allows offline approximation of the continuous wavelet transform. We thus obtain a scalogram, which is the equivalent of a spectrogram for wavelets. This scalogram is then averaged across time in order to obtain an image of gesture's spectral distribution. Lastly, we compute the first four statistical moments of this distribution, which gives us information about gesture's spectral content. This method functions well to measure low-frequency gestures, as we will see in chapter 5.

2.2 On vocalization

The classical source-filter model [Sundberg et al., 1977] describes the voice organ as an instrument consisting of a *power supply* (the lungs), an *oscillator* (the vocal folds) and a *resonator* (larynx, pharynx and mouth forming the vocal tract).

Pitch differences are caused by varying the rate of vibration of the vocal folds, two small bands of muscles in the larynx. Tensing the vocal folds makes them vibrate faster, so that the pitch increases. *Loudness* is produced by the speaker pushing more air out of the lungs. The shape of the vocal tract is determind by the positions of the lips, the jaw, the tongue and the larynx. The vocal tract is a resonator that has four or five important resonances, called *formants*.

We will see how human beings can play with it to produce different kinds of sounds, from language sounds to non-verbal vocalizations.

2.2.1 Spoken and sung voice

Nowadays, there are about 200 different vowels in the world's languages and more than 600 different consonants. Let's see in a nutshell how we can characterize these sounds acoustically.

Spoken voice

In normal speech, the voice fundamental frequencies of male and female adults center on approximately 110 Hz and 200 Hz, respectively, and generally do not exceed about 200 Hz and 350 Hz, respectively [Sundberg, 1999]. Speech sounds can be divided in two groups, called vowels and consonants. The acoustic vowel space can be considered to be an area bounded by the possible ranges for the frequencies of the first two formants. The range of frequency of the two latter are 300-800 Hz and 700-2200 Hz, respectively. The third and higher formants contains information about the identity and the intonation of the speaker.

The consonants of English can be divided into stops, approximants, nasals, fricatives and affricates. Consonants from the three first groups have their own formant pattern that resemble vowels' formant pattern, except that noise bursts are added. On the other hand, fricatives and affricates have energy over a wide range of higher frequencies: these consonants are called *voiceless consonants* and can be modeled by a semi-random noise with its center frequency and amplitude. For example, the energy in [s] is mostly in the high frequency range from about 3,500 Hz upwards, and [ʃ] ('sh') has most energy lower, around 3,000 Hz.

To sum up, there are nine principal components of speech sounds [Ladefoged, 2001] : (1-3) the frequencies of the first three formants, (4-6) the amplitudes of the first three formants, (7-8) the frequency and amplitude of the voiceless components, and (9) the fundamental frequency of voiced sounds.

Sung voice

Singing can be defined as producing musical sounds with the voice. All the elements and functions of the voice previously described are common to singers and nonsingers alike; but singers can play their voice in different ways. The highest pitches for soprano, alto, tenor, baritone, and bass correspond to fundamental frequencies of about 1400 Hz, 700 Hz, 523 Hz, 390 Hz and 350 Hz, respectively [Sundberg, 1999]. Singers learn to move the frequency of the first formant close to that of the fundamental, this by wide opening their jaw. Thus, singers tend to change their jaw opening in a pitch-dependent manner rather than in a vowel-dependant manner, as in normal speech.

Also, the partials falling in the frequency region of 2,500-3,000 Hz, approximately, are much stronger in sung vowels than in spoken vowels. This peak is generally referred to as the *singer's formant*, and can be achieved by a lowering of the larynx.

2.2.2 Non-verbal vocalization

Aside from sounds produced to convey verbal language, there are other sounds that humans can make with their vocal organ. In comparison to voice, nonconventional vocalizations have been rarely studied.

Whistling is produced by the compressed air in the cavity of the mouth, forced either through the smallest hole of the vocal tract or between fingers. A study on whistled languages [Meyer, 2008] shows that whistlers make the choice to reproduce definite parts of the frequency spectrum of the voice, ranging approximately from 1,000 Hz to 3,500 Hz.

Beatboxing techniques use non-syllabic patterns and inhaled sounds to produce an audio stream in which language-like patterns are suppressed, so that it makes the illusion of a non-vocal sound source(s) [Stowell and Plumbley, 2008].

Screams are produced by a nonlinear oscillatory regime and can be seen as a modulation of the normal voice [Arnal et al., 2014].

2.2.3 Audio descriptors

There is a large literature about audio descriptors extraction. We will only give a brief insight of the methods we chose, since it is not the aim of this study. We report the reader to the references for more detailed information.

The *Timbre Toolbox* [Peeters et al., 2011] provides a set of descriptors, mostly created for music information retrieval. Sounds are first analyzed using different input representations such as the short-term Fourier transform, the energy envelope, auditory models and timbre models of [McAdams et al., 1995]. Audio descriptors are then computed from these representations. As we will analyze vocalizations, we chose to use the spectral centroid, which is the frequency center of gravity of the energy envelope.

In addition to this toolbox, we used a specific algorithm to compute time-varying sound *loudness* [Glasberg and Moore, 2002]. We also use the *YIN* algorithm [De Cheveigné and Kawahara, 2002] to estimate the fundamental frequency of our signals. This algorithm is based on an autocorrelation method that contains some modifications to make it more robust, and is particularly adapted to voice signals.

2.3 Combining gesture and vocalization

Now that we have globally understood what is at stakes with gesture and vocalization, we would like to focus on the combination of both. As we said previously, hardly any research has been done on it; however, some fields provides us with the insight that gesture and vocalization, as two streams of commnication, may merge into each other to become one.

Communication conduct

In spoken language analysis, an utterance is a unit of speech. When a speaker speaks, the speech is organized into a series of packages, tending to correspond to units of meaning that are at a level above the lexical level, and which may be referred to as 'idea units'.

Kendon states that the previously defined *gesture phrases* coincide with and tend to be semantically coherent with 'idea units' [Kendon, 2004]. He observed that speakers are able to *orchestrate* the gesture and speech components of an utterance, changing these orchestrations in relation to the momentary demands of the communication moment or shifts in the speaker's aim. The gestural component is under the control of the speaker in the same way as the verbal component, making it possible for him to accomplish a more complete expression, this being true in terms of deployment in the utterance as in terms of referential meaning.

Figure 2.2: Continuum of gestural movements relative to speech, as proposed by [McNeill et al., 1990].

To go further in this sense, [McNeill et al., 1990] proposed to place gestural movements along a continuum reflecting their *relationship to speech* (see figure 2.2). At one end of this continuum, gesture is used in conjuction with speech and users are marginally aware of their use of it. At the other end, gesture is used independently of speech. In between, he places 'mime', which can be used in alternation with speech, and 'emblems' which are standardized gestures which can function as complete utterances in their own right but which do not constitute the components of a language system, as is the case with signs. With this classification, he made it clearer that one should not think of 'gesture' as

one and unique notion, but that there are different kinds and that each should be dealt with theoretically in a rather different way.

Movement practice

Another example comes from movement practice. A recent study by [Françoise et al., 2014] suggests that interactive sound feedback based on the practice of vocalizing while moving could allow dancers to access a greater range of expressive movement qualities. Similarly, [Françoise, 2015] shows synchrony between a tai-chi expert gestures and vocalizations, which let us suggest that both could have been produced in the same intention.

Conclusion

Gesture and voice are two streams of communication, both having their own advantages and constraints. These two streams seem to merge into one when used in communication conduct and movement practice. Such interrelations could be measured with the help of motion and audio descriptors. Now, one question arises: how do we use them in *sound imitation*?

3

Background: sound perception and imitation

In the previous chapter, we defined gesture and vocalizations as two distinct streams that are interrelated when used for speech communication. In the frame of the SkAT-VG project and as a following of [Lemaitre and Rocchesso, 2014] works, we would like to understand how they could be combined to communicate about sounds. Imitating sounds can roughly be separated into two phases: listening to a sound, and then imitating it. This chapter will first give an overview of the listening process, from interpretating sounds to auditory imagery. We then describe the imitation process in a general case and apply its consequences in the field of music. Lastly, we review results on sound imitation, which allows us to draw up the key-issue of our study.

3.1 The listening process

The first thing that one has to do before imitating a sound is to listen to it. In this section, we give results on auditory information interpretation, plus an insight of auditory imagery.

3.1.1 Interpretating auditory information

Different listening modes exist, and they do not transmit the same information about sounds. [Gaver, 1993] defines *musical listening* and *everyday listening*. In the first case, we are sensible to perceptual attributes (such as pitch or loudness) that have to do with the sound itself and are

also those used in the creation of music. In the second one, we listen to events that created the sounds rather than to the sounds themselves. Quoting him, "the distinction between everyday and musical listening is between experiences, not sounds".

Sound identification

Studies have provided some insights into factors that influence sound identification. Identification time and causal uncertainty (which quantifies alternative causations for a given sound) are highly interrelated, and both are related to ecological frequency (which is the occurrence frequency of a given sound) and the presence of harmonics and similar spectral bursts [Ballas, 1993]. Rating data suggested that sound identifiability is related to the ease with which a mental picture is formed of the sound, context independence, the familiarity of the sound, the similarity of the sound to a mental stereotype, the ease in using words to describe the sound, and the clarity of the sound.

Expertise in categorization

Another study covering listener expertise and sound identification showed that expert participants (for instance, sound engineers or musicians) tended to categorize sounds on the basis of their acoustical similarities, whereas non-experts tended to base the categorization on causal similarities [Lemaitre et al., 2010]. Experts have developed specific listening techniques allowing them to focus on specific aspects of a sound. They are also capable of using a technical vocabulary that is not available to non-experts.

3.1.2 Auditory imagery

As we hear a sound or a melody, we can recall it 'in our mind' quite accurately, and that this is the case regardless of levels of musical training. This phenomenom is called *auditory imagery*. It is the occurence of a perception sensation in the absence of the corresponding perceptual input.

Empirical findings

Recent studies suggest that auditory images reflect considerable interpretation and are not uninterpreted sensory experiences; rather, auditory images contain both depictive information and descriptive information [Hubbard, 2010]. It involves semantically interpreted information and expectancies. It can be related to musical ability or experience, although the mechanisms of that relationship are not clear. It is

often but not necessarily influenced by subvocalization (i.e., the internal speech made when reading a word).

Auditory imagery appears to draw on most of the neural structures used in auditory perception [Kosslyn et al., 2001], although activation is stronger during perception than in imagery.

Musical imagery

Another study dealing with acoustically complex and ambiguous sounds in terms of pitch showed that beyond a certain point of complexity, listening has to rely on some kind of simplification of the sound material [Godøy and Jørgensen, 2001]. That is, we have to 'overrule' the acoustic material and make an 'idealised' or 'stylised' image of the musical sound by filtering out features which would otherwise lead to ambiguous images.

In a more recent paper, it is suggested that images of sound-related actions may be a significant component of sonic images [Godøy, 2010]. The fact that images of sound-related actions can trigger sonic images in our minds leads us to the idea of action imagery in musical imagery. This may be subsequently understood in the perspective of embodied music cognition, as we will see in the next section.

3.2 The imitation process

Imitation is an advanced behavior whereby an individual observes and replicates another's behaviour. This phenomenom has been studied in many scientific domains. In the frame of our study, we briefly review the process itself within cognitive approaches, then making connections with musicology and musical practice.

3.2.1 Cognitive approaches

We focus on two fields of cognitive science that are relevant for our study: developmental psychology (imitation as social learning) and cognitive neuroscience (imitation involves the human mirror system).

Developmental psychology

Acquiring the capacity of imitation enables the human child to economically acquire a great variety of practical skills, as much as conceptual knowledge given by his relatives and the scholar system [Proust, 2002]. Thanks to the 'imitative language' (which is firstly sensory-motor), newborns learn social coordination, by switching from being the imitator to being imitated. For the baby imitator, imitation is then a means of understanding other people mental states. Finally, the child imitates

consciously, which leads him towards language-based communication mode.

Cognitive neuroscience

Motor cognition is a field that deals with the role of actions in building the self. The concept of imitation intervenes when it comes to discussing action observation and learning by observation.

Mirror neurons. It appeared that the representations we build about the actions we observe are not originating exclusively from sensory signal, but also implicates the participation of motor mechanisms [Jeannerod, 2006]. More concretely, this means that seeing someone performing a gesture activates the same motor areas as we ourselves would perform the same gesture. The neurons involved in these motor areas are called *'mirror neurons'*: they encode both an action and its goal, irrespectively of the agent who performs it[1].

The activation of mirror neurons by an observed action allows consequently a simulation of that action — that is to say, an imitation.

Defining imitation. One can make a distinction between *mimicry* (the ability to duplicate observed movements) and *true imitation* (the ability to understand the goal of one's action and to re-enact that action to achieve the intended goal) [Jeannerod, 2006].

Mimicry is a primitive form of imitation that is also referred to as 'resonance behavior': the only detection of some typically human behaviour can cause the observer to engage in a replication of the observed behaviour (emotional contagion and yawning are typical examples). It starts very early in life, as seen in the previous paragraph, and is normally inhibited by social constraints.

On the other hand, in *true imitation*, the imitated action deals with the goal of the action and not only with its form, and can be replicated after some delay. For example, when we try to copy the voice and the gestures of a famous person, we try to be accurate in copying the physical action but we also try to suggest the supposed mental content of the character we imitate.

The question of whether true imitation could be discriminated from more primitive forms by brain activity remains an open one.

3.2.2 Music and imitation

As seen in section 2.1.2, corporeal articulations have been defined as expressions of the attribution of intentionality to music. They are indi-

[1] As an observation, action understanding is impaired in autistic children.

cations of synesthetic and kinesthetic action processes, and have a predictive and anticipatory character. Motor aspects of imitation that we described in section 3.2.1 can be discerned in relation to these corporeal articulations.

Corporeal articulations as imitation

[Leman, 2008] states that involvement with music is based on the mirroring process that rules imitation. It would be realized in the coupling of action and perception and would allow the attribution of intentionality to music. Distinctions can be made between imitation of skills, imitation of musical figures, imitation of symbols, imitation of moving sonic forms (corporeal imitation), and imitation of group behavior (allelo-imitation). We will not discuss these distinctions.

Degrees of empathic musical involvement and imitation

In [Leman, 2008] again, corporeal imitation is analyzed in terms of *synchronization*, *attuning*, and *empathy*. They all involve imitation, but in different degrees of participation and identification. *Synchronization* (e.g. tapping the beat) is an aspect of resonance behaviour (mimicry): thus, it can be seen as something that the subject largely undergoes, such as a sensation.

In contrast, *empathy* seems to involve the emotional system. Within music, subjects may attribute aspects of their own expressive intentionality (such as affects and feelings) to physical energy. This attribution would be an effect of the mirroring processes which allow subjects to translate moving sonic forms into components of their action-oriented ontology. These processes may call on the emotional system so that human subjects become emotionally involved with music.

Attuning occupies the middle position between synchronization and empathy. Attuning brings the human body into accordance with a particular feature of music. It can be seen as navigation with or inside music. The activity of the subject is in harmony with a particular aspect of the music, such as singing along or moving in time to the music. Although it is a kind of participation, attuning may be less involved with identification.

3.3 Sound imitation

So far we have reviewed the literature about auditory perception and imitation. Embodied music cognition states that we engage through music as a vector of intentionality by imitation. Then, how people would imitate sounds?

3.3.1 Using gesture

The issue of 'gesturing a sound' can appear to be complex, since, as a consequence of what we saw in section 2.1, gesture never produces sound without interacting with extrinsic elements. However, people do gesture sounds, as reviewed before. We will thus focus on studies trying to relate clues about how people gesture sounds.

Sound-tracing experiments

In an exploratory work on sound tracing (meaning tracing gestures that listeners make with a pen in response to a sound), [Godøy et al., 2006] found that an ascending pitch is mostly traced as an ascending curve, and a percussive onset followed by a long decay will be traced as a steep slope followed by a long descent.

Another experiment was then carried out where participants were asked to move a rod in the air, pretending that moving it would create the sound they heard [Nymoen et al., 2011]. The presence of a distinct pitch in the latter seems to influence how people relate gesture to sound: there is a very strong correlation between pitch and vertical position. There might also be nonlinear correspondences between motion features and other sound features (such as brightness and loudness).

Gestural sound description

In a recent paper, [Caramiaux et al., 2014] underline the role of sound source in gestural sound description. They show that for the sounds where causal action can be identified, participants mainly mimic the action that has produced the sound. In the other case, when no action can be associated with the sound, participants trace contours related to sound acoustic features. They also found that the interparticipants' gesture variability is higher for causal sounds compared to noncausal sounds. In the first case, sound causality as action is represented by an iconic gesture that can be performed under distinct forms (depending on the participant's habits in doing the action). In the second case, participants perform a metaphoric gesture that follows the acoustic energy contour of the sound (the common reference is the sound itself).

3.3.2 Using vocalization

It has been observed that vocal imitations are spontaneously used when participants have to communicate a sound that they just have heard [Lemaitre et al., 2009,Wright, 1971]. We can distinguish two types of vocal imitations: onomatopoeias and nonconventional vocalizations. Imitations of the former type are close to words: their meaning is associated to the word through a symbolic relationship. Unlike onomatopoeias,

a nonconventional vocalization is a creative utterance intended to be acoustically similar to the sound or the sound produced by the thing to which it refers. We are interested in the process of nonconventional vocalization, that we will call vocalization from now on.

Vocalizations and sound event identification

Studying these vocalizations could be useful to understand sound event identification. For speakers, vocally imitating a sound consists of conveying the acoustic features they deem important for recognition [Lemaitre et al., 2011]. A cluster analysis of vocal imitations of everyday sounds revealed that the listeners have only used a limited number of simple acoustic features to cluster the imitations. These features did not imply any complex characteristic but apparent simple characteristics: continuous versus rhythmic sounds, tonal versus noisy, short versus long, and so on. These coarse features were sufficient for the listeners to recover the types of sound production.

Vocalizations are more effective than verbal descriptions

Following this work, another study compared the effectiveness of vocal imitations and verbalizations to communicate different sounds [Lemaitre and Rocchesso, 2014]. These sounds were selected on the basis of participants' confidence in identifying the cause of the sounds, ranging from easy-to-identify to unidentifiable sounds. Participants were first asked to describe these sounds with words, then to vocalize them. Recognition accuracy with verbalizations dropped when identifiability of the sounds decreased. Conversely, recognition accuracy with vocal imitations did not depend on the identifiability of the referent sounds and was as high as with the best verbalizations.

Discussion

These works suggest that the phenomenon of vocal imitation corresponds to a form of caricature of the original sounds. Participants had to vocalize within the constraints of human vocal production (that is, as a consequence of the properties presented in section 2.2, periodic or noisy signals, essentially monophonic and limited in pitch): thus, they selected acoustic features they deemed relevant to communicate the idea of the sound and, by this way, maximized the probability that it will be recognised.

3.4 Conclusion — Methodology

The previous review has made it possible for us to understand the process of gestural and vocal imitation that we are going to deal with in this study. In a nutshell:

Chapter 2

Gesture and voice are *two streams of communication*, both having their own advantages and constraints. These two streams seem to *merge into one* when used in communication conduct and movement practice.

Chapter 3

— **Sections 3.1 & 3.2**

Interpretating sound information depends on both our listening attitude and sound expertise. *Musical* imagery can be related to *action* imagery.
True imitation (not to be mistaken with mimicry): (1) is innate, (2) is well-developed in humans, and (3) fosters learning. Involvement with music may be based on the mirroring process that rules imitation.

— **Section 3.3**

Nonconventional *vocalizations* can be efficiently used to communicate a sound. They put emphasis on the most salient acoustic feature, acting as caricatures of sounds.
In *gestural* sound description, gesture is used either to describe the sound source, or to trace acoustic features.

Methodology

Our study aims at understanding the role of gesture when combined with vocalization in sound imitation. As speech studies show it [Kendon, 2004], one of our expectations is that gesture would give metaphorical information while vocalization would be more precise in sound imitation.

We will proceed in two steps: first, we will analyze qualitatively a database of vocal and gestural imitations of a variety of sounds, in order to come up with hypotheses.
We will then construct a new experiment to test these hypotheses, as well as adapted measures to make statistical analyses. We will extract the experiment's method from our state of the art.

4

Imitating sounds: first study

This chapter reports on a database collected before this Master's Thesis, during the SkAT-VG project. Our aim is to use this database to draw up hypotheses that will be tested in the experimental study reported in section 5. We start by describing the aim and methodology of the database collection; then, we analyze data qualitatively, first providing a global view on imitations' strategies, then analyzing it more deeply. This analysis will finally result in three hypotheses.

4.1 Data collection

The SkAT-VG project collected a database of vocal and gestural imitations prior to this work. This section reviews its method and experimental setup.

4.1.1 Method

The experiment asked participants to imitate referent sounds so that somebody else could recognize them only by listening and watching their imitations. They were not allowed to use onomatopoeias. About gesture, participants either were free in imitating the sounds (*free gesture protocol*), or were asked not to mimic the source that could have produced the sound, but to concentrate on the properties of the sound itself (*directed gesture protocol*). The free gesture protocol aims at showing that people instinctively mimic the imagined sound source, as shown in [Caramiaux et al., 2014].

The recording session was divided into two steps: during the first step,

participants had to record vocalizations of referent sounds, whereas the second step consisted in performing vocal and gestural imitations of referent sounds.

Participants

Fifty persons (21 male, 29 female), from 19 to 47 years old (mean 28.3) volunteered as participants. All reported normal hearing and were native speakers of French. None of them had either musical or dancing expertise. Ten participants followed a free gesture protocol; the forty remaining ones followed a directed gesture protocol.

Stimuli

Fifty-two referent sounds were used, classified into three families: sounds of machines (20 sounds), basic mechanical interactions (20 sounds) and abstract sounds (12 sounds). Each of these families are further organized in categories (see appendix B). Two sounds are selected for each family and associated category, according to the results of a preliminary identification experiment held with 320 sounds and 24 participants.

Procedure

A graphical user interface (GUI) allowed participants to listen to the referent sounds and to record their imitations (see appendix A)[1]. Each step was subdivised into three phases, corresponding to the sound families.

For example, in a given phase, every sound of one family were disposed at random on a GUI. Participants first had to listen to every referent sound before performing their imitation. They could listen to each referent sound as many times as they wanted to. They also could train themselves to imitate referent sounds without recording themselves as long as they wanted to. When they felt ready, they could record their imitation. They had five record trials. The last trial was considered as their best trial. The phase order was random.

4.1.2 Experimental setup

For each imitation, the cartesian coordinates of their articulations are recorded using a kinect device. The acceleration and angular velocity of their wrinkles is recorded using "inertial measurement units" (IMUs) . Finally, video recordings are made using both a GoPro camera (HD 1080p, 120 fps) and a webcam (see appendix E).

[1]The GUI was conceived by Frédéric Voisin and Guillaume Lemaitre.

4.2 Preliminary analysis

We first reviewed the webcam footage, in order to get a sense of the process of imitation.

Procedure

For 15 randomly selected participants, we made an analysis grid. For each of the 52 referent sounds, it consisted in:

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

This analysis is subjective; however, listening to interviews made after the recording of imitations sometimes allowed us to confront our vision of their imitation to what participants were actually thinking of doing.

Results

Among the 15 selected participants, five followed a *free gesture protocol* (2 males, 3 females; mean 26.2). As expected, each of these 5 participants mimicked the sound source or the action that could have produced the referent sound. These participants were removed from the next analyses.

The 10 remaining participants followed a *directed gesture protocol* (6 males, 4 females; mean 24.6). Despite this instruction, a few cases of mimicry were observed, but were not significant: 6 sounds out of 52 were mimicked by 2 or 3 participants out of 10. Also, these 6 referent sounds (abstract impulse sound, closing door, sawing, rubbing, hitting and whipping) are very hard to imitate without mimicking the sound source since they are human-triggered sounds. In other cases, gestural imitations thus tend to express the referent sound itself rather than its cause.

Globally, imitations are very diverse. The greater the referent sound complexity is, the more diverse imitations are. On one hand, basic interactions such as whipping or switching a button are imitated in the same way; on the other hand, crumpling a can brought about several personal imitations that are very difficult to analyze.

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, 7 participants out of 10 made a

stable noisy vocalization while shaking their hands. Also, for complex sounds such as filling a glass with water, 6 participants out of 10 made a gesture that seems to convey another information that was not vocalised. Lastly, for the fridge sound, 8 participants out of 10 made a stable vocalization while shaking their hands.

Lastly, analyzing differences between vocalizations alone and vocalizations with a gesture did not come up with a result. Some participants were rather consistent in their vocal imitations, and some were not. The firsts perhaps remembered their previous vocal imitation during their vocal and gestural imitation.

4.3 Focused analysis

The previous analysis shed light on the complexity of the imitation process. In order to make hypotheses, we focused our analysis on a reduced set of 8 referent sounds, and also used other data collected during the experiment.

Procedure

Ten participants (5 male, 5 female; mean 30.4) were randomly selected. They all followed a *directed gesture protocol*. We analyzed both slow-motion recordings and computed descriptors qualitatively. We made an analysis grid consisting in:

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.

Each item was coded by a number standing for a potential verbal description (see appendix C for transcription specifications). It is important to underline that even if this analysis is more precise, it is still subective to say that in some cases, information could be specific to gesture (or to vocalization). We tried to minimize this subjectivity by focusing on a reduced set of sounds.

Selecting the sounds

We focused on 8 referent sounds selected from the initial set of 52 sounds used in the experiment. One can classify them in two categories:

Elementary sounds. Only one acoustic characteristic of these sounds (e.g., tonal component, periodicity) evolves. We selected five referent sounds: stable noise, repetitive noise, a closing door (human impact), pitch going up, pitch going down.

Complex sounds. Several acoustic characteristics of these sounds vary at the same time (vertical complexity) or in time (horizontal complexity). We selected 3 referent sounds: 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, whereas 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.

We made these categories before the analysis to make it easier for us to identify bevahiours. However, these categories are overlapping: indeed, some elementary sounds are not that elementary, and one can argue participants may perceive that many acoustic aspects vary. It is also important to keep in mind that complexity may not be the only key factor in the imitation process. The causality of the sound may be important when it comes to imitating it.

4.3.1 Shared aspects of imitations 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).

Elementary sounds. There are basic similarities among the imitations:

- For the stable noise, 8 participants out of 10 vocalized a noise while shaking the hands without any specific direction;
- For the repetitive noise, 10 participants out of 10 tried to vocalized a repetitive noise while moving their hands in rhythm in a specific direction;

- For the impact sound, every participant made a noisy and loudness-decreasing vocalization while underlining the impact with their gesture;
- For pitched sounds, 9 participants out of 10 tried to vocalize the evolution of the pitch while reflecting it with their gesture. They seemed to emphasize either the beginning or the end of their imitation. Six out of 10 emphasized the end of their "pitch going up" imitation and 9 out of 10 the beginning of their "pitch going down" imitation.

However, despite these high-level similarities in the imitations of elementary sounds, we observed several singularities at a lower-level. Three participants out of 10 imitated 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 (agreeing with [Nymoen et al., 2011]), 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. There are basic similarities among the imitations:

- For the humming fridge, 7 participants out of 10 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. Most of them did not imitate the second part of the sound.

For the filling sound, there are too many different imitations to draw basic similarities. We will discuss this point later.

There are 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. This variability let us analyze the different roles of vocalization and gesture in the imitations.

4.3.2 Separation of vocalization and gesture

The analysis of the recordings suggests that gesture always reflects at least one aspect of the vocalization. In some cases, gesture may complete 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 one third of the cases, a constant movement completed 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 for studying the separation of vocalization and gesture since it has both horizontal and vertical complexity. Here are some interesting examples:

- One participant vocalized a going up tonal sound while shaking his fingers;
- Three participants made a pitch-oscillating vocalization while moving their hands upward;
- One participant made a going up noisy vocalization while moving his hands downward;
- One participant made a stable noisy vocalization while moving his hand downward;
- One participant vocalized a going up rough tonal sound while moving his hand downward.

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 were observable for the two other complex sounds:

- For the humming fridge, as seen before, 7 participants out of 10 made a stable tonal vocalization while shaking their hands;
- For the printer, 4 participants out of 10 made a shaking movement with their hands.

4.3.3 Accuracy of the imitations

So far, our analysis has focused on the specific roles of gesture and vocalization in the imitations. What about the actual accuracy (that is to say, their being similar to the referent sound) of these imitations?

Generally, accuracy of imitations is very dependent on both listening and vocal skills of the participants. Since they are non-experts, most of them have difficulties in controlling both their vocalization and their

gesture precisely. For example, it is not easy to produce a vocalization with stable energy and pitch. This observation suggests that imitations are not to be perfect in any case. Another important distorsion is that the duration of imitations may differ from the duration of referent sounds. We did not study this temporal contraction/dilatation effect.

Elementary sounds. One can identify basic distorsions among the imitations:

- For the stable noise, 3 participants out of 10 made a pitch-oscillating noisy vocalization;
- For the repetitive noise, 4 participants out of 10 did not produce synchronous vocalization and gesture in the repetition process;
- For pitched sounds, most participants emphasized the beginning or the end of their imitation, as we have seen before. More than a third of the participants vocalized both a tonal and a noisy component. For the "pitch down" sound, five participants did not imitate the evolution of the fundamental frequency properly.

Complex sounds. One can identify basic distorsions among the imitations:

- For the humming fridge, 5 participants out of 10 made a vibrato and four participants made a tremolo;
- For the printer, 7 participants out of 10 omitted to imitate the second part of the sound;
- For the filling sound, 3 participants out of 10 omitted to imitate the last part of the sound

These observations suggests that the accuracy of the imitation should not only be viewed as dynamically correlated attributes of imitations and sounds, but more as its expressivity in communication, as [Lemaitre et al., 2011] suggests it. Imitations of complex sounds also suggests that an acoustic property of a vocalization may stand for another attribute of the sound (e.g., a vibrato standing for a random aspect of the sound). It can also be related to a simplification of the sound material, as [Godøy and Jørgensen, 2001] suggests it.

4.3.4 Modifying the vocalization with gesture

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 the participant to vocalize in a different

way, and (2) gesture may help the participant embody the sound he has to imitate. We compared vocal and gestural imitations to vocalizations alone.

Change in vocalization. There are some cases in which vocalization is 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 seems to trigger is the implication of participants in their imitations. In some cases, their global imitations seems 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: this agrees with [Leman, 2008] views on musical involvement.

4.4 Conclusion: drawing up hypotheses

This qualitative analysis has shed light on how complex the imitation process can be. We decided to pick up three different phenomena that we now want to study more systematically.

The first one is about the repetitive sound. We expected that voice and gesture would be synchronous in the process of imitation; however, 4 out of 10 participants had a **phase difference**. This suggests that biomechanical constraints in such air gestures (that we could put as "emblem" in [McNeill et al., 1990] continuum) may prevent participants from imitating rhythmic information well, and push them to turn their gesture into a metaphorical gesture.

The second one is about the stable noise and the humming fridge. 7 out of 10 participants made a *shaky gesture* while vocalizing, either making their hands or their fingers vibrate. This gesture (that we could put as "gesticulation" in [McNeill et al., 1990] continuum) may reinforce a **textural aspect** of the stimulus that could be less satisfying to evoke with voice only.

The third and last one is about layered sounds. Participants were able to communicate different information about the sound using their **voice and gesture separately**. They seemed to use their voice to imitate either tonal aspects of sounds or what they deemed to be the most salient aspect of sounds.

We are thus able to make hypotheses on three aspects of sound imitation:

1. Voice is more effective than air gestures to imitate *rhythmic information* precisely.
2. *Textural aspects* can be evoked by a shaky gesture.
3. *Vertical complexity* of sounds can be addressed by separating the roles between gesture and voice. In particular, the voice imitates the most salient aspect of the referent sound.

As we saw, there is a great diversity in the imitation strategies of complex sounds: this suggests that participants are able to combine gesture and vocalization in order to make a complete imitation. Yet, even if we selected referent sounds for which features evolve quite distinctly, we did not control them precisely. Therefore we constructed an experimental study to test the forementioned hypotheses with controlled referent sounds.

5

Combining gesture and vocalization

In the previous chapter, we drew up three hypotheses about how people combine voice and gestures when they imitate certain types of sounds. The following chapter aims at testing these hypotheses in controlled conditions. We first describe our experiment for which we created abstract referent sounds. Then, we present a quantitative analysis of the collected data. Finally, we discuss these results and suggest research prospects that could follow this work.

5.1 Designing a new experiment

In order to test our hypotheses, we first need to create a set of referent sounds. A first criterion was to prevent participants from mimicking the sound source. Thus, we created *abstract sounds*, as sounds that do not have an identifiable cause. Obviously, we cannot prevent participants from imagining a sound source when they listen to a sound: however, using abstract sounds dismisses action-triggered sounds. One could also argue participants would imitate computer-produced sounds, which is a kind of sound source itself. Yet computer sounds do not produce sounds by a movement: it is thus the best way to evaluate how people imitate basic (or say "neutral") features of sounds, thus triggering a *musical listening* [Gaver, 1993].

Creating new sounds is also a way to control what participants will hear. We can control the *acoustical features* of the referent sounds themselves, as well as the *number of layers* for polyphonic sounds.

5.1.1 Creating abstract sounds

We created 25 new sounds that we distributed among three families: *rhythmic* sounds, *textural* sounds and *layered* sounds. Each sound family aims at testing one of our three hypotheses.

Rhythmic sounds

There were 9 rhythmic sounds, splitted in two groups.

Repetitive sounds. We created 5 repetitive noisy sounds containing two phases: a repetitive noise, followed by a burst of noise (impulse), the latter being preceded by a short crescendo. Their respective periods are constructed regarding the period of the repetitive noise (250 ms) studied in the previous chapter (see figure 5.1).

> *We expect vocalization to be* more effective *to track high tempi than air gesture. Gesture and vocalization would desynchronize from 250 ms but would* resynchronize *for the impulse.*

Rhythmic patterns. We synthesized 4 sounds which consists in rhythmic sequences of short tones. 3 of them are rhythmic patterns; the last one is a random pattern. The tempo of the three first as the number of tones is increasing with their index (see figure 5.2).

> *We expect vocalization to be* more precise *than air gesture in reproducing rhythm. Gesture would only underline rhythmic patterns'* main pulse*, but would underline most random pattern' impacts.*

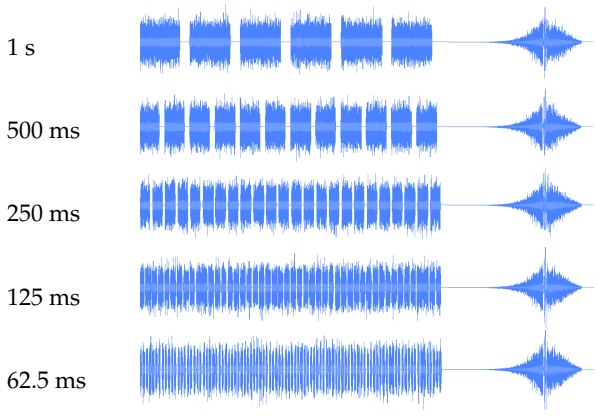

Figure 5.1: Repetitive sounds.

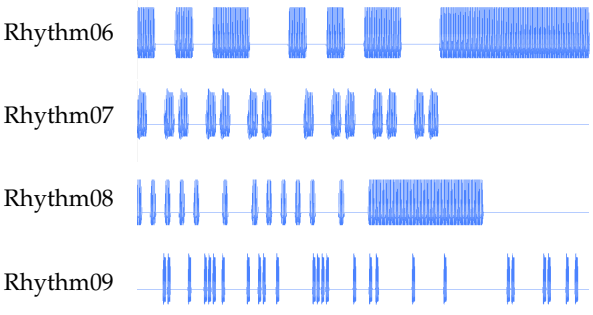

Figure 5.2: Rhythmic sounds. The three first are on the same temporal scale. The last one is longer.

Textural sounds

There were 8 textural sounds, splitted in two groups.

Stable sounds. We synthesized 4 stable sounds. The first one is a harmonic tone; the 3 others are respectedly a granulated noise, a granulated tone and an extremely granulated noise (see figure 5.3).

> *We expect stable granular textures to be imitated with a* shaky gesture, *while the stable harmonic tone would trigger a stable gesture. Vocalizations would be stable in every case, trying to convey either a tonal or a noisy texture.*

Dynamic sounds. We used the same synthesis parameters as the 4 previous sounds, and added a dynamical aspect: a frequency sweep for tonal sounds and a spectral centroid sweep for noisy sounds (see figure 5.4).

> *We expect gesture to follow the* dynamical aspect *of sounds rather than the previous textural aspects. Vocalizations would follow the dynamical evolution in every case, trying to convey either a tonal or a noisy texture.*

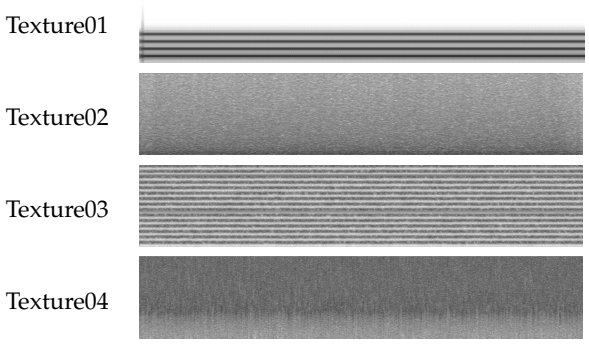

Figure 5.3: Stable sounds. Texture01 and Texture03 have an audible pitch.

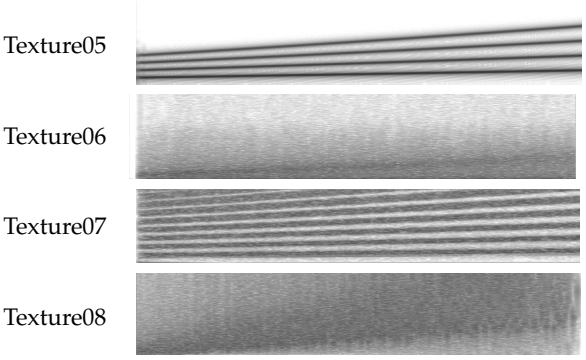

Figure 5.4: Dynamic sounds. Texture05 and Texture07 have an audible pitch.

Layered sounds

There are 8 layered sounds, splitted in two groups. We aim at combining an impulsive layer with a sustained layer in order to exhibit dissociations between gesture and vocalization. Based on our previous analysis, we would not add *more than two aspects*, so that participants could be able to imitate every aspect of the sound (as they have two streams of communication at their disposal).

Repetitive sounds. There are 4 repetitive sounds. With these sounds, we aim at studying what feature of the sound is more often vocalised than gestualised. We subdivised them in two groups:

	Stable layer		***Dynamic layer***
Layer01	Repetitive noise + stable noise	Layer03	Repetitive noise + dynamic noise
Layer02	Repetitive tone + stable tone	Layer04	Repetitive tone + dynamic tone

We expect participants to separate the roles *between gesture and vocalization for stable layer sounds.*
On the other hand, dynamic layer sounds would allow us to observe different imitation strategies.

Melodic sounds. With using a melody, we introduce an emotional process in the imitation process, which can be discussed. However, it can be an effective way to push participants to separate tasks between their gesture and vocalization. There are 4 melodic sounds, splitted in two groups:

	Stable layer		***Dynamic layer***
Layer05	Melodic tone + stable noise	Layer06	Melodic tone + dynamic noise
Layer07	Melodic noise + stable tone	Layer08	Melodic noise + dynamic tone

We expect participants to vocalize the tonal melody *for both stable and dynamic layer sounds .*
On the other hand, melodic noise sounds would allow us to observe different imitation strategies.

Sound synthesis

We created a Max/MSP patch to synthesize our referent sounds, based on additive synthesis, noise filtering and granular synthesis. The granular aspect was generated by sogs~, a smooth overlap granular synthesizer (Ircam)[1].

For textural sounds, fundamental frequencies as sweep parameters were chosen regarding our vocal tract abilities [Sundberg, 1999, Ladefoged, 2001]. Finally, we equalized our sound set in loudness using [Glasberg and Moore, 2002] model.

[1] http://forumnet.ircam.fr/fr/product/max-sound-box/.

5.1.2 Method

Participants imitated referent sounds so that somebody else could recognize them only by listening and watching its imitation. They were not allowed to use onomatopoeias. About gesture, they were only allowed to use their dominant hand and arm; also, they were not allowed to mimic the imagined sound-producing action. By this way, we wanted to trigger *true imitation* [Jeannerod, 2006].

The experiment was divided into three steps: during the first step, participants performed vocal and gestural imitations of sounds, whereas they respectively recorded vocalizations and gestualizations of sounds during the second and third steps. Data from these two lasts steps was not exploited.

Participants

Eighteen persons (10 male, 8 female), from 18 to 45 years old (mean 26.6), volunteered as participants. All reported normal hearing and were native speakers of French. None of them have either musical or dancing expertise.

Stimuli

For each step, we used the 25 previously described referent sounds, classified into three families.

Procedure

The experiment used the same interface as previously. Each step was subdivided into three phases, corresponding to the family of referent sounds.

In the first phase, the GUI presented all rhythmic sounds. The position of each referent sound was randomly chosen for each participant. Participants first listened to every referent sound before performing their imitation. They could listen to each sound as many times as they wanted to. They also could practice without recording themselves as long as they wanted to. When they felt ready, they recorded their imitation. There was a maximum of five trials. The last trial was considered as their best trial. The phase order was: rhythmic sounds, texture sounds, and layered sounds.

At the end of the experiment, we recorded an interview with the participant, looking over each imitation of the first step (voice and gesture step). The interview grid is shown in appendix D.

Experimental setup

We used the same experimental setup than the previous experiment, i.e. a microphone for audio data, a webcam and a GoPro for video data, an inertial measurement unit (IMU) for wrist's acceleration and a kinect for skeleton position (see appendix E). Qualitative analyses exploited video and interview data; statistical analyses exploited audio and IMU data.

5.2 Analysis: rhythmic sounds

In this section, we present the results of the first phase of the experiment: vocal and gestural imitation of rhythmic sounds. We first analyze how vocalization and gesture reproduce different tempi; then, we analyze rhythmic pattern reproduction precision in vocal and gestural imitations of sounds. The analysis consisted in first defining a measure of the phenomenon; then, we submitted this measure to an analysis of variance (ANOVA). The latter were subjected to a Geisser-Greenhouse correction due to a possible violation of sphericity when necessary ; p-values are reported after correction. Planned contrasts used Pillai's test. In all figures, vertical bars represent the 95% confidence interval.

Before analyzing data, we segmented it by hand, in respect with the gesture unit definition [Kendon, 2004]: each imitation was divided into a preparation phase, one or two stroke phases, and a recovery phase.

5.2.1 Tempo tracking

To study tempo tracking of the imitation, we focused on imitations of repetitive sounds (five first referent sounds).

Measure

For each vocal imitation, we computed the onsets of the audio track, first using Super VP and then correcting possible errors by hand. We then computed inter-onset intervals (IOI), which are *period values*. We divided these period values by the period of the referent sound and finally took the mean of the distribution. If the vocal imitation reached the good tempo, the measure should be equal to 1.

For each gestural imitation, we computed the scalogram of the IMU data (see section 2.1.3 and appendix F). We then estimated the time-varying frequency of the gesture with a ridge-tracking algorithm (scalogram maximum estimation adjusted with statistical moments). We converted these frequencies into period values, divided them by the period of the referent sound and finally took the mean of the distribution.

Again, if the gestural imitation reached the good tempo, the measure should be equal to 1.

Analysis

One participant was excluded from this analysis since he did not imitate one of the periods. For the 1 s period, 9 participants out of 17 made a gesture the period of which was two times smaller (3 out of 17 for the 500 ms period): for analysis, we took their period modulo the period of the stimulus (we discuss it in section 5.5). Results are shown in figure 5.5.

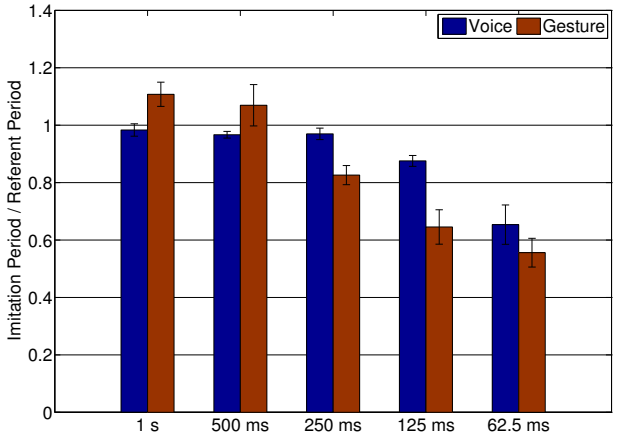

Figure 5.5: Estimated period of the imitations relative to the period of the referent sound, averaged across participants. (1 = same tempo.)

Voice and gesture period ratios were respectively submitted to two one-way ANOVAs with period as the within-subject factor. The effect of sound was significant for both voice and gesture (respectively $F(4,64)=11.4$, $p<0.05$ and $F(4,64)=43.4$, $p<0.05$). On the one hand, planned contrasts showed that voice period ratio is not significantly lower for a 250 ms period than for 1 s and 500 ms periods (0.97 vs 0.98, $F(1,16)=0.86$, $p=0.37$). On the other hand, planned contrasts showed that gesture period ratio is significantly lower for a 250 ms period than for 1 s and 500 ms periods (0.83 vs 1.09, $F(1,16)=23.0$, $p<0.001$).

5.2.2 Rhythmic pattern reproduction

We first studied synchrony between voice and gesture for a single impulse that was presented at the end of the five previous repetitive sounds; then, we studied imitations of the 4 remaining rhythmic referent sounds, that can be seen as several impulses following a temporal pattern.

Rhythm06, Rhythm07 and Rhythm08 can be considered as sorted by order of *"complexity"*. Their tempo as their number of impulses is increasing with their index. One can finally distinguish Rhythm09 from the three other stimuli. Rhythm09 is a *random pattern*: thus, one will not study the reproduction of the pattern itself, but more the reproduction of a random pattern.

Single impulse

Measure. For each vocal imitation, we computed the onset of the impulse the same way as we did for the previous sounds. For each gestural imitation, we computed the time-varying energy of the scalogram of the IMU data. We defined the impulse of a gesture as the instant where the scalogram energy is maximum. We finally computed *time difference* between voice and gesture impulse and averaged it across participants.

Analysis. Two participants were excluded from this analysis since they did not imitate the single impulse. Results are shown in figure 5.6.

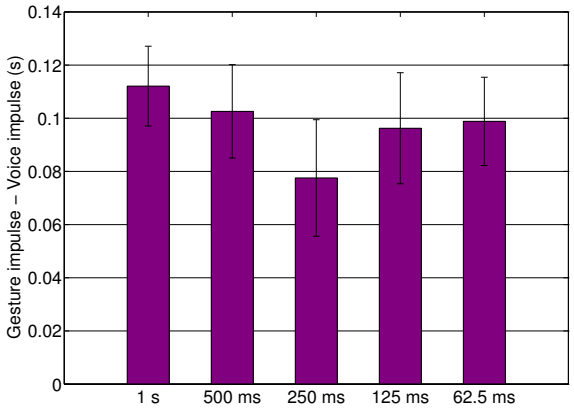

Figure 5.6: Time difference between voice and gesture in the imitation of a single impulse, averaged across participants.

Time differences were submitted to a one-way ANOVA with referent sounds as the within-subject factor. The effect of sound was not significant ($F(4,64)=0.82$, $p=0.50$).

Rhythmic patterns

We computed voice and gesture's onsets the same way as we did for the study of the impulse. There are different techniques that have proven to be useful in the study of rhythm, such as dynamic time warping or IOI dendrograms. However, differences between imitations and referent sounds pushed us to use 2 simpler measures, which are well-adapted to our study.

Measure 1: number of onsets. First, if we compute voice and gesture onset vectors' lengths, and divise both by the length of the onset vector of the stimulus, we surely do not know if rhythmic imitation is well reproduced, but we could know which communication stream reproduce the *correct number of onsets* (ratio = 1). Results are shown in figure 5.7.

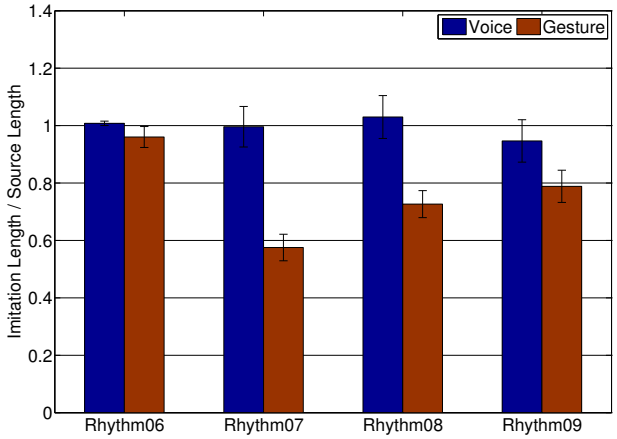

Figure 5.7: Relative length (between imitations and referent sounds), averaged across participants.

Analysis. Voice and gesture relative lengths were respectively submitted to two one-way ANOVAs with referent sounds as the within-subject factor. The effect of sound was not significant for voice ($F(3,51)=0.44$, $p=0.65$) whereas it was significant for gesture ($F(3,51)=15.6$, $p<0.05$). In addition, Figure 5.7 shows that relative length was systematically close to 1 for voice whereas it was smaller for gesture: participants produced the correct number of onsets with the voice whereas they produced fewer onsets with gesture.

Measure 2: average IOI. The average IOI is another measure of the *accuracy of the pattern reproduction*. Again, this is not a precise nor a perfect way to study pattern reproduction: yet, as our database consists in very basic rhythmic patterns the imitations of which are qualitatively different from one referent sound to another, we can reasonably consider analyzing such feature to seize a tendency. Thus, we computed the average IOI of the imitation and finally divided it by the average IOI of the referent sound. Results are shown in figure 5.8.

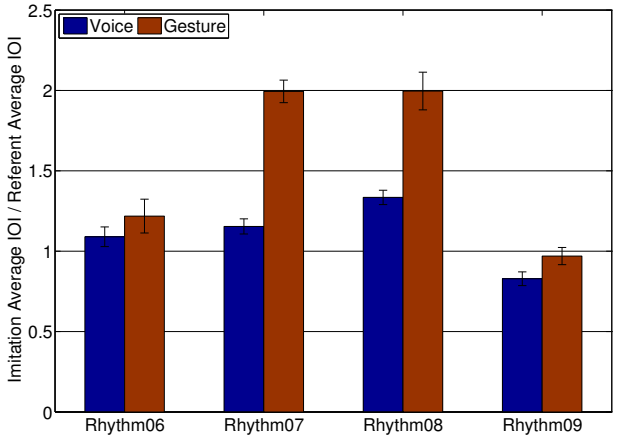

Figure 5.8: Average relative IOI (between imitations and referent sounds), averaged across participants.

Analysis. Voice and gesture average relative IOI were respectively submitted to two one-way ANOVAs with referent sounds as the within-subject factor. The effect of sound was significant for both voice and gesture (respectively $F(3,51)=16.1$, $p<0.05$ and $F(3,51)=58.3$, $p<0.05$), which reveals nothing new for gesture but indicates that voice may sometimes not be able to accurately reproduce a pattern. Planned contrasts compared the average relative IOI between rhythms 6 and 7. Average relative IOI were not significantly different for the voice, whereas they were for gesture (1.09 vs 1.15, $F(1,17)=0.44$, $p=0.52$ for the voice; 1.22 vs 1.99, $F(1,17)=103.2$, $p<0.001$ for gesture). Planned contrasts compared then the average relative IOI between rhythms 6 and 8. Average relative IOI were significantly different for both the voice and gesture (1.09 vs 1.33, $F(1,17)=10.3$, $p<0.01$ for the voice; 1.22 vs 2.00, $F(1,17)=55.2$, $p<0.001$).

As a remark, gestural average IOI for Rhythm07 equals 1.79, which is quite near from Rhythm07 ratio between tempo and its average IOI (1.85). This will be discussed in section 5.5.

Synchrony between gesture and voice

So far, we compared gesture to referent sounds on the one hand, and vocalization to referents sounds on the other hand. An interesting observation emerges when we compare *gesture to vocalization*.

Measure. We computed length ratios between gesture avec voice, as the average relative IOI between gesture and voice. Results are showed in figure 5.9.

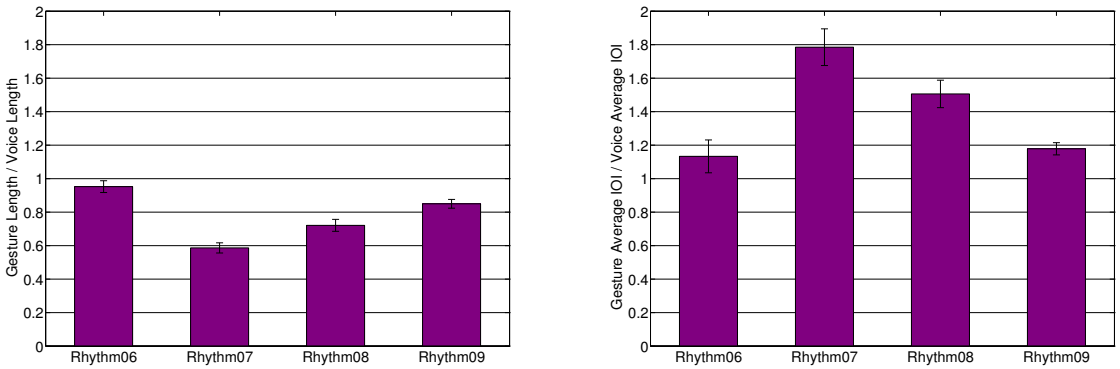

Figure 5.9: Left: Relative length between gesture and voice. Right: Average relative IOI between gesture and voice. Averaged across participants.

Analysis. Both measures were submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant for both measures (respectively $F(3,51)=28.0$, $p<0.05$ and $F(3,51)=14.0$, $p<0.05$). Planned contrasts showed that the relative length (respectively the average relative IOI) was significantly higher (respectively lower) for Rhythm06 and Rhythm09 than for Rhythm07 and Rhythm08 (0.90 vs 0.65, $F(1,17)=93.5$, $p<0.001$ for relative length; 1.15 vs 1.65, $F(1,17)=45.5$, $p<0.001$ for average relative IOI).

5.3 Analysis: textural sounds

The second phase of the experiment was about imitating different sound textures. We first present high-level descriptions of participants' vocal strategies, and then study their gestural behaviour.

5.3.1 Vocal strategies

We focused on high-level descriptions of participants' imitations. We thus decided to study the amount of aperiodicity in their vocalizations, as the reproduction of the stable/dynamic characteristic of the referent sounds. As a reminder, even texture referent sounds were noisy referent sounds while the odd ones were texture referent sounds with a distinct tonal pitch.

Amount of aperiodicity

Measure. For each vocal imitation, we computed the time-varying aperiodicity provided by the YIN algorithm [De Cheveigné and Kawahara, 2002], which is similar to *signal-to-noise* ratio. We then took the average value of it. Results are shown in figure 5.10.

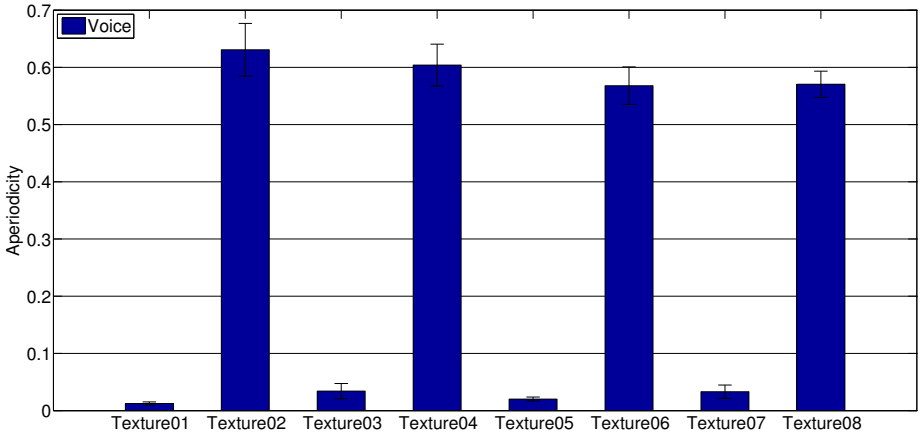

Figure 5.10: Amount of aperiodicity averaged across participants.

Analysis. Aperiodicity was submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant ($F(7,119)=126.9$, $p<0.05$). Planned contrasts showed that aperiodicity was higher for even referent sounds than for odd referent sounds (0.59 vs 0.02, $F(1,17)=1035.9$, $p<0.001$).

Dynamic attributes

Measure. For vocal imitations of odd stimuli, which are voiced vocalizations, we computed the time-varying fundamental frequency estimator provided by the YIN algorithm; we then made a linear regression of it and took the ratio of the last value against the first value.

For vocal imitations of odd stimuli, which are voiceless vocalizations, we applied the same computation to IrcamDescriptor's spectral centroid [Peeters et al., 2011].

Both these measures indicate if participants made a stable vocalization (ratio = 1) or a dynamic vocalization (here, ratio > 1): we called them *pitch increase*. We deliberately did not take the gradient value since we waned to free ourselves from duration difference between participants. Also, such a measure allows us to study vocalization regardless of the differences in participants' vocal ranges. Results are shown in figure 5.11.

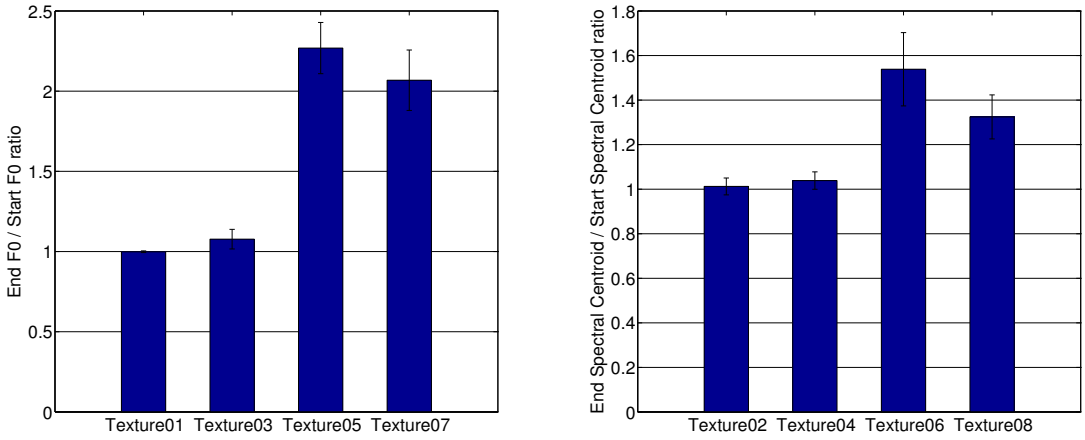

Figure 5.11: Left: Pitch increase for voiced imitations, based on f0 computation. Right: Pitch increase for voiceless imitations, based on spectral centroid computation. Averaged across participants.

Analysis. Both ratios were submitted to a one-way ANOVA with referent sounds as the within-subject factors. The effect of sound was significant for both ratios ($F(3,51)=28.9$, $p<0.05$ for f0 ratio; and $F(3,51)=6.71$, $p<0.05$ for spectral centroid ratio). Planned contrasts showed that pitch increased more for dynamic sounds than for stable sounds (2.17 vs 1.04, $F(1,17)=53.1$, $p<0.001$ for f0 ratio; 1.43 vs 1.03, $F(1,17)=11.8$, $p<0.01$ for spectral centroid ratio).

5.3.2 Gestural behaviour

Now, we would like to know how people gesture these sounds as they are making the previously studied vocalizations.

Measure. For each gestural imitation, we computed the scalogram of the acceleration data provided by the IMU. We then took the frequency distribution of the scalogram and computed its *centroid* (see appendix F). This measure should provide us with an insight of the presence of shaky gesture (high-frequency centroid, i.e. a lower scale value), or stable gesture (low-frequency centroid, i.e. a higher scale value). Results are shown in figure 5.12.

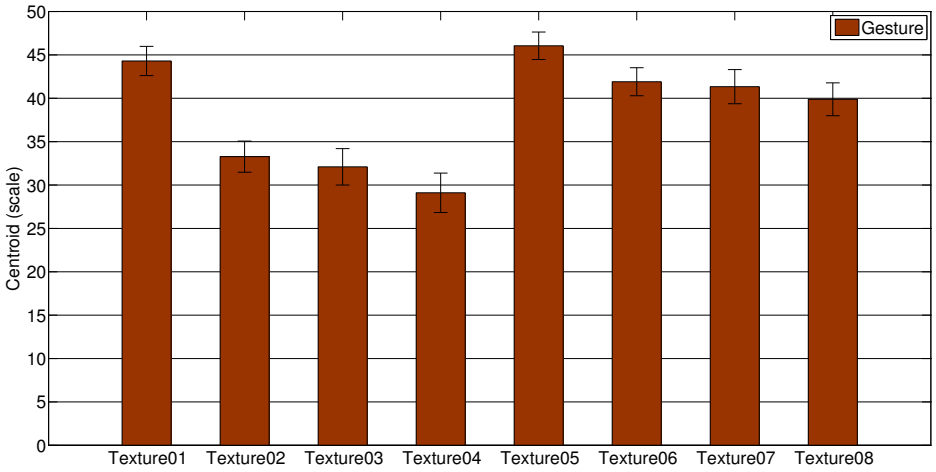

Figure 5.12: Gesture scale distribution centroid averaged across participants.

Analysis. Centroid was submitted to a one-way ANOVA with referent sounds (Texture01 to Texture08) as the within-subject factor. The effect of sound was significant ($F(7,119)=12.7$, $p<0.05$). Planned contrasts showed that the centroid was lower for stable granular sounds (Texture02-03-04) than for the other ones (31.5 vs 43.4, $F(1,17)=45.0$, $p<0.001$).

Stable textures

Centroid was also submitted to a one-way ANOVA with stable referent sounds (Texture01-02-03-04) as the within-subject factor. The effect of sound was also significant ($F(3,51)=15.1$, $p<0.05$). Planned contrasts

showed that the centroid was lower for granular sounds (Texture02-03-04) than for the harmonic sound (Texture01) (31.5 vs 44.3, $F(1,17)=33.2$, $p<0.001$).

Granular textures

We also submitted centroid to a one-way ANOVA with granular sound stimuli (Texture02-03-04 and Texture06-07-08) as the within-subject factor. Again, the effect of sound was significant ($F(5,85)=10.2$, $p<0.05$). Planned contrasts showed that the centroid was lower for stable granular sounds (Texture02 to Texture04) than for dynamical granular stimuli (Texture06 to Texture08) (31.5 vs 41.0, $F(1,17)=27.2$, $p<0.001$).

Tonal textures

Finally, centroid was submitted to a one-way ANOVA with tonal stable referent sounds (Texture01 and Texture03) as the within-subject factor. The effect of sound was significant ($F(1,17)=28.1$, $p<0.05$). Planned contrasts showed that the centroid was lower for the stable tonal granular sound (Texture03) than for the stable harmonic sound (32.1 vs 44.3, $F(1,17)=28.1$, $p<0.001$).

5.4 Analysis: layered sounds

The third and last phase of the experiment consisted in imitating layered sounds. This was the most exploratory part of our work: we thus proceeded to a qualitative analysis of participants' strategies. We first review global descriptive statistics of the whole data set, and then analyze participants' behaviours in specific strategies.

5.4.1 Global analysis

For each referent sound, we first asked the participant how many sounds he heard. All participants heard two layers (lay1 & lay2) for each referent sound, meaning that their imitation was made being conscious that they have to imitate these two layers. In order to analyze participants' behaviours when imitating layered sounds, we reviewed both their video and interview data. This reviewing allowed us to fill an analysis grid (see appendix G).

We identified 4 different strategies :

1. **Separation of roles between voice and gesture [lay1/V lay2/G]:** participants decided to imitate one layer with their voice, and the remaining one *simultaneously* with gesture;
2. **One after the other [lay1/V+G], [lay2/V+G]:** participants first decided to imitate one layer with both their voice and gesture, and *in a second time* the second layer with both their voice and gesture;
3. **Only one layer [lay1/V+G]:** participants decided to imitate *only one layer* with both their voice and gesture;
4. **Merging the two layers [lay1&2/V+G]:** participants *mixed* the two layers in a creative way.

A first look at the global strategy distribution of the whole imitation data set (see table 5.1) let us assume that *separation of roles* has a little advantage over the three other strategies, which are slightly equally distributed. This could be set up in the SkAT-VG project.

[lay1/V lay2/G]	[lay1/V+G], [lay2/V+G]	[lay1/V+G]	[lay1&2/V+G]
40.3% (58)	20.8% (30)	21.5% (31)	17.4% (25)

Table 5.1: Strategy distribution across imitations for 8 sound stimuli and 18 participants, i.e. 144 imitations. (In brackets: number of imitations.)

We can also see on figures 5.13 and 5.14 that 15 participants out of 18 favoured one strategy more than half the time, whereas 1 referent sound out of 8 triggered one strategy more than half the time. This let us suggest that strategies tend to be *more consistent for a participant* than for a given sound.

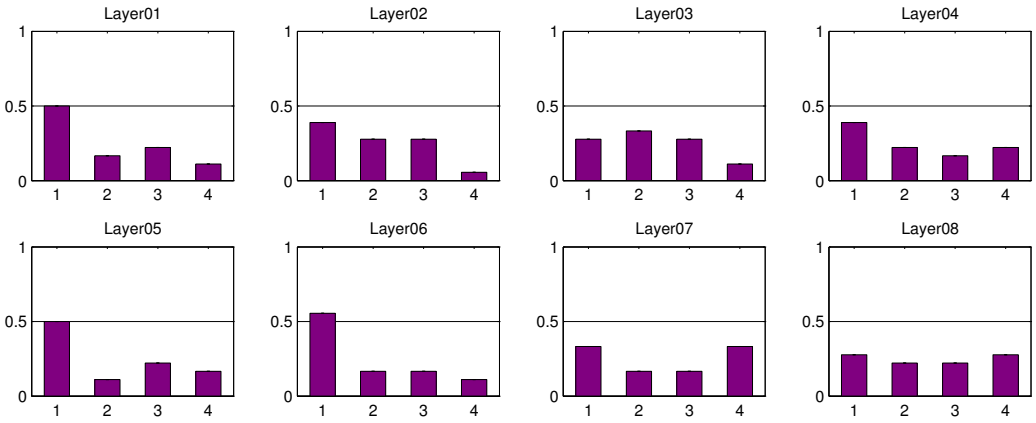

Figure 5.13: Imitation strategies for each referent sound. 1=[lay1/V lay2/G]; 2=[lay1/V+G], [lay2/V+G]; 3=[lay1/V+G]; 4=[lay1&2/V+G].

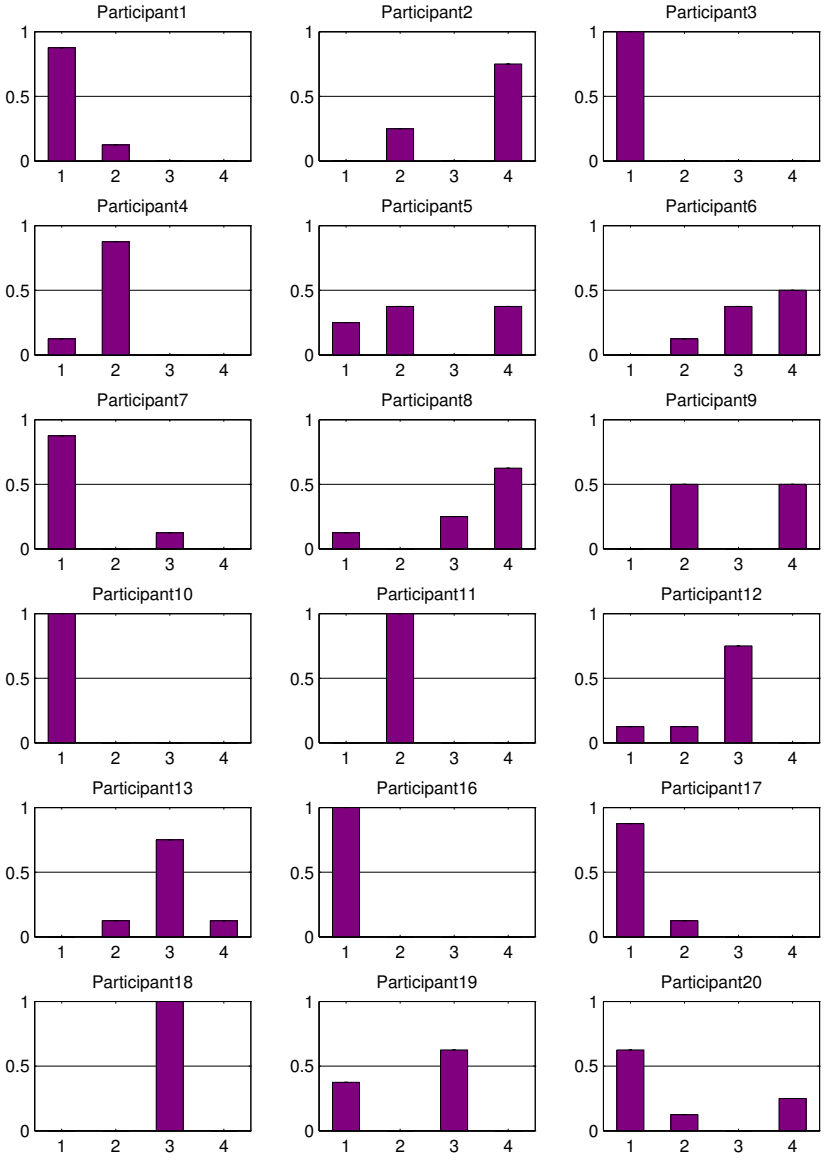

Figure 5.14: Imitation strategies for each participant. 1=[lay1/V lay2/G]; 2=[lay1/V+G], [lay2/V+G]; 3=[lay1/V+G]; 4=[lay1&2/V+G].

Tonal melodic impulsive layer stimuli (Layer05 and Layer06) seems to trigger most of the [lay1/V lay2/G] strategy. It is also interesting to observe that 5 participants out of 18 (subjects 3, 10, 11, 16 and 18) were 100% consistent in their strategy, and 3 out of these 5 (subjects 3, 10 and 16) used the [lay1/V lay2/G] strategy.

5.4.2 Strategies' specifications

We may now look deeper into these strategies. For a given strategy, we tagged additional information:

- For the **[lay1/V lay2/G]** strategy, we tagged which of the two layers was imitated with the voice;
- For the **[lay1/V+G], [lay2/V+G]** strategy, we tagged which of the two layers was first imitated;
- For the **[lay1/V+G]** strategy, we tagged which of the two layers was imitated.

These additional tags would allow us to see if the layer type (impulsive or sustained) had an influence on participants' strategies. Figures are shown in table 5.2.

[lay1/V lay2/G] strategy

What is interesting is that participants who used the [lay1/V lay2/G] strategy mainly imitated the *impulsive layer with the voice* while imitating the sustained layer with gesture (50 times out of 58). The 8 remaining times are mostly caused by one participant. During the interviews, participants reported that impulsive layers were "easier" to reproduce with the voice than with gesture, or that gestualizing impulsive layers was not "satisfying", hence their choice. This may agree with what was found in section 5.2 about rhythmic sound imitation.

[lay1/V lay2/G]	[lay1/V+G], [lay2/V+G]	[lay1/V+G]	[lay1&2/V+G]
40.3% (58)	20.8% (30)	21.5% (31)	17.4% (25)
Impulsive with voice	Impulsive first	Impulsive	-
86.2% (50)	66.7% (20)	96.8% (30)	-

Table 5.2: Additional information on impulsive layer imitation across different strategies over 144 imitations. (In brackets: number of imitations.)

Another information would be to know if there is a distinction between noisy and tonal impulsive layers. Figures in table 5.3 let us suggest that there is no distinction between tonal and noisy impulsive layers. These results are to be taken with care since there are not that many representative imitations.

Other strategies

Table 5.2 also gives us interesting figures about other strategies that goes in line with the previous observation about impulsive layers. For [lay1/V+G] strategy, the *impulsive layer is the only imitated layer* 30 times

Layer01	Layer02	Layer03	Layer04
8 (9)	6 (7)	5 (5)	4 (7)
Layer05	Layer06	Layer07	Layer08
9 (9)	9 (10)	5 (6)	4 (5)

Table 5.3: Impulsive layer imitated with the voice across SR strategy.

out of 31. Participants that have used this strategy either decided to imitate only one layer since they felt "not capable" to imitate both, or they just "forgot" to imitate the second layer. Yet, in both cases, they mainly decided to imitate the impulsive layer.

Another information is that the *impulsive layer was first imitated* 20 times out of 30 for the [lay1/V+G], [lay2/V+G] strategy. Participants that used this strategy reported that they felt like the "have to vocalize" each layer to be satisfied with their imitation, hence their separation in time. One could interprete this order as an importance ranking, since participants also qualified the impulsive layer as the "first sound", and the sustained layer as the "other sound", or sometimes the sound "behind".

5.5 Discussion

In this section, we discuss the previous analyses and expose possible research prospects.

Rhythmic sound imitation

Tempo tracking. The analysis in section 5.2.1 suggests that *voice is more precise than air gestures to communicate tempo information*. It also shed light on a desynchronization between voice and gesture that occur when imitating a sound the period of which is higher than 250 ms. When crossing this value, gesture appears to become *metaphorical rather than precise*.

During this task, for 1 s and 500 ms periods, some participants made a gesture the period of which was two times smaller than the period they had to imitate. It is as they gestualized noise bursts' onsets and offsets. As our analysis only treated onset reproduction, we took the modulo, assuming that they gestualized onsets well. As a research prospect, one could then study duration reproduction: it is possible that voice would again be more precise than gesture.

Rhythmic pattern reproduction. The analysis in section 5.2.2 suggests that *voice and gesture are synchronous in the imitation of a single impulse*, regardless of the phase desynchronization that may have happened before this impulse. Synchrony between gesture and voice analysis suggests that both desynchronize when it comes to imitating a complex rhythm (Rhythm07 and Rhythm08) but tend to be synchronous when it comes to imitating a "simple" pattern (Rhythm06) and random patterns (Rhythm09), the latter resembling to separated impulses. This agrees with the previous result on tempo tracking (a slow tempo standing for a "simple" rhythmic pattern, and faster ones standing for more "complex" rhythms). Voice is thus *more precise than air gestures to reproduce a rhythmic pattern*, but has also its limits, since average IOI analysis suggests that voice does not imitate Rhythm08 as well as Rhythm06 and Rhythm07, depending from participants' musicality and biomechanical constaints.

One can thus legitimally wonder about gesture's usefulness. It is interesting to notice that in the rhythmic pattern reproduction task, participants sometimes gesture a regular subdivision of the tempo. That was mostly the case in the imitation of Rhythm07, which was synthesized in the idea of triggering such a beat pattern for gesture and rhythm reproduction for voice. In this case, gesture beating the tempo may *help participants to vocalize the accurate pattern*.

In this case, gesture does contain rhythmic information; voice just contains more precise rhythmic information. One can also argue that adding a gesture help participants to better remember such rhythmic information. Comparing the previously analyzed data with data collected during the two other phases of the experiment (respectively voice only and gesture only) would be an interesting and relevant following to this study.

Texture sound imitation

The analysis in section 5.3 suggests that participant *vocally imitates the acoustical features of the referent sounds*. They imitate tonal sounds with a voiced vocalization, and noisy sounds with a voiceless vocalization; they also vocalize the presence of a pitch dynamic in the referent sound they imitate.

About gesture, the analysis suggests that *shaky gestures appear when imitating stable granular textures*. Thus, a stable harmonic tone is imitated with a stable gesture, while a stable granulated tone is imitated with a shaky gesture. When imitating dynamic granular textures, this shaky gesture tends to disappear in favour of a stable aspect. Gesture thus may stand for the *most relevant metaphorical aspect of a sound*.

It is important to notice that some participants had their gesture contain two different aspects: a stable aspect (standing for a high scale value) and a shaky aspect (standing for a lower scale value). Computing the centroid allowed us to take into account both aspects, respectedly weighted by their amount of energy. Figure 5.15 shows an example of such more complex gesture. Centroid then appears as a measure of gesture's main component.

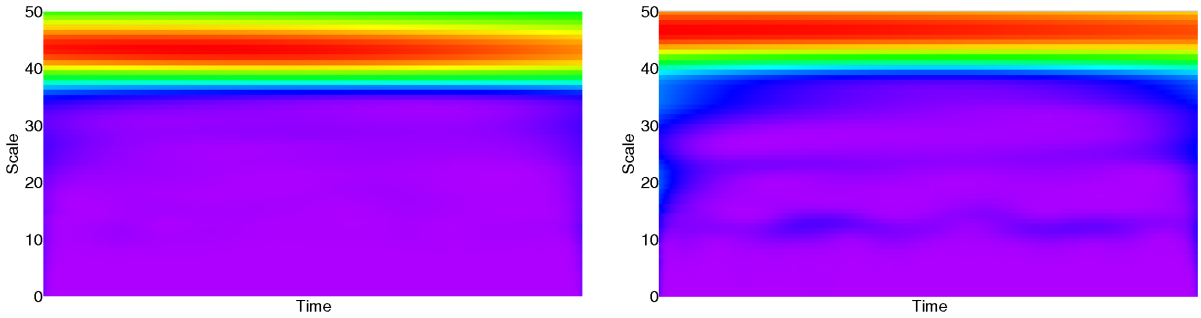

Figure 5.15: Left: Scalogram for a "stable" gesture (simple scale distribution). Right: Scalogram for a "stable" and "shaky" gesture (multiple scale distribution). In red: high amplitude; in purple: low amplitude.

These observations as our whole study focused on gesture. Yet it is important to notice that many participants made use of *hand postures* in the imitation of such textures. For example, participants sometimes raised their forefinger to imitate harmonic sounds (judging them "precise"), while opening their hand wide to imitate noisy sounds (judging them "large"), or even clenching their fist because they felt like a sound was "stronger" than others. Such subjective judgements were also rendered by favouring one given direction in their gesture. All these observations are as many interesting research prospects that would need studying.

In our test, the choice of stimuli's parameters was made in order to validate our hypotheses. It would be interesting to study sound imitation with only two sound parameters (such as pitch and duration). A possible case would be that long sounds as high-pitched sounds would trigger oscillating gestures, as their metaphorical content would differ from short and low-pitched sounds.

Layered sound imitations

The analysis of layered sound imitation was the most exploratory part of our study. That was why we chose many different parameters to study. We saw that in most cases, *the impulsive layer was vocalised while the sustained layer was gestualised*. One should treat this result with caution: we did not define what sound feature made the impulsive layer

more salient than the other. It could be indeed because of its impulsive nature, but also because of its relative loudness compared to the sustained layer loudness. During the synthesis process, we tried to synthesize equalized layers for each sound, judging it by ear. Hovewer, this is not a perfect way to realize this.

Conclusion

All of this let us suggest that gesture and vocalization, as two streams of communication, should not be treated equally in sound imitation. While vocalization would imitate sounds as precisely as it can acoustically speaking, gesture would communicate metaphorical information that seems really hard, or even not relevant, to link with acoustical features directly. When asking participants about what abstract referent sounds made them think about, we harvested very different points of view. For example, a dynamic harmonic sound was described as "speeding up"; its stable counterpart was described as "taking all the space". This kind of metaphorical verbalization is transcribed into gesture. It would thus be totally wrong to claim that gesture is of no use: *gesture is the reflection of intentionality*. Decoding gesture with the help of scientific tools would be a fascinating advance in the frame of the SkAT-VG project.

6

A classifier for shaky gestures

Gestural data collected during textural sound imitation was roughly divided into two classes: "stable" gestures and "shaky" gestures. The statistical analyses reported in in the previous chapter showed that the results of our experimental study were consistent with our hypotheses; now, we wonder if the same data could be used in another scientific purpose: the building of a classifier for shaky gestures. We first present our classifier's specifications, and finally evaluate its quality.

6.1 Classifier specification

We decided to study a *k-nearest neighbor classifier*. Despite its relative theoretical simplicity, this kind of classifier can prove to be very powerful, provided that we are able to use relevant features for our case study.

6.1.1 Database description

Gestural imitations of textural sounds constitute the 160 observations of our classifier. Classes were defined relatively to our hypotheses: 100 observations were tagged as "stable" (imitations of Texture01, Texture05, Texture06, Texture07 and Texture08), and the 60 remaining were tagged as "shaky" (imitaitons of Texture02, Texture03 and Texture04).

These tags do not necessarily correspond to the observed behaviour: for example, a participant who imitated Texture01 with a shaky gesture (and thus with shaky gesture features) would still be tagged as "stable", as our hypothesis suppose it.

6.1.2 Computed features

For each of these observations, we computed three *statistical moments* of the frequency distribution of their IMU acceleration scalogram (centroid, variance and kurtosis), and added another feature related to *gesture's energy* (the logarithm of the average energy of the frequency distribution of the scalogram). We centered each of these features by substracting their means, and then divided them by the maximum of the modulus of the centered value. This computation made each feature vary between -1 and +1, which is necessary to allow a good scaling, since k-nearest neighbors is based on a euclidean distance computation. A representation of the observations is shown in figure 6.1.

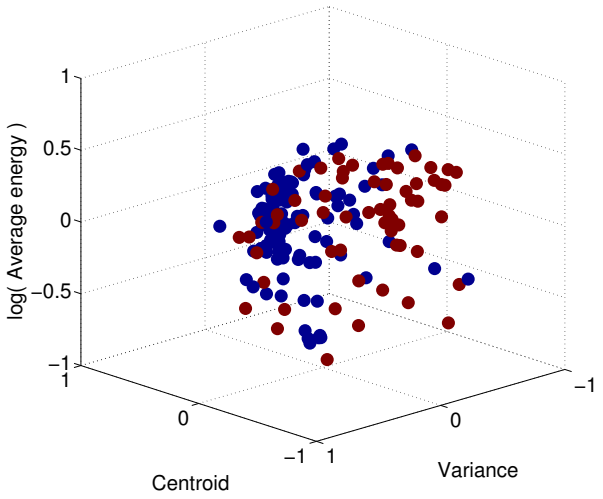

Figure 6.1: Observations for three features: centroid, variance and log(average energy). In red: "shaky" class; in blue: "stable" class.

6.2 Evaluation

We first train our k-nearest neighbor; then, we compute the cross-validation loss, which is the average loss of each cross-validation model when predicting on data that is not used for training. We chose the previous features (centroid, variance, kurtosis and log(average energy)) so that the cross-validation loss would be the smallest with the fewer neighbors.

For the leave-one-out cross-validation, the cross-validation loss is 21% with k = 5 neighbors (being 79% recognition accuracy). Such a classifier is an example of practical application of our study that could be integrated in the SkAT-VG project.

7

Conclusion

Based on a qualitative analysis of a data collection, we were able to draw up hypotheses about the combination of gesture and vocalization in the imitation of sounds.

The results of our study show a *quantitative advantage of voice over gesture* in sound imitation for communicating rhythmic information: voice can track higher tempi than gesture, and is more precise when imitating rhythmic patterns than gesture. They also exhibit the use of *shaky gestures* to communicate stable granular textures. Finally, they show that some people are able to *imitate two sounds at the same time*, using their voice and their gesture simultaneously. Additionally, data collected in texture imitation allowed us to build a classifier for shaky gestures with a few spectral features.

Moreover, our study shed light on the *metaphorical function of gesture* when combined with voice during sound imitation. Such a function should not be seen as less relevant than voice's acoustic features, but as equally relevant.

This explanatory work opens up many different research prospects. One could study hand postures during sound imitation, or study the influence of sound duration and frequency on gestural imitation. Layered sound imitation encourages us to study salience more deeply in order to understand the prevalence of voice over gesture. It would also be great to study expert participants behaviour in such experiment: we could expect more precise gestures and vocalizations.

We will present a poster of this work during the 170th Meeting of the Acoustical Society of America that would be held in early November 2015. We also plan to submit a paper to PLOS ONE and to submit another poster to the Seventh Conference of the International Society for Gesture Studies that would be held in June 2016.

Bibliography

[Arnal et al., 2014] Arnal, L., Flinker, A., and Poeppel, D. (2014). Screams occupy a privileged spectro-temporal acoustic region.

[Ballas, 1993] 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.

[Cadoz, 1994] Cadoz, C. (1994). Le geste canal de communication homme/machine: la communication" instrumentale". *Technique et science informatiques*, 13(1):31–61.

[Camurri et al., 2004] Camurri, A., Mazzarino, B., and Volpe, G. (2004). Analysis of expressive gesture: The eyesweb expressive gesture processing library. In *Gesture-based communication in human-computer interaction*, pages 460–467. Springer.

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

[De Cheveigné and Kawahara, 2002] De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

[Françoise, 2015] Françoise, J. (2015). *Motion-Sound Mapping by Demonstration*. PhD thesis, Université Pierre et Marie Curie.

[Françoise et al., 2014] Françoise, J., Fdili Alaoui, S., Schiphorst, T., and Bevilacqua, F. (2014). Vocalizing dance movement for interactive sonification of laban effort factors. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 1079–1082. ACM.

[Gaver, 1993] Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29.

[Glasberg and Moore, 2002] Glasberg, B. R. and Moore, B. C. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society*, 50(5):331–342.

[Godøy, 2010] Godøy, R. I. (2010). Images of sonic objects. *Organised Sound*, 15(01):54–62.

[Godøy et al., 2006] Godøy, R. I., Haga, E., and Jensenius, A. R. (2006). Exploring music-related gestures by sound-tracing: A preliminary study.

[Godøy and Jørgensen, 2001] Godøy, R. I. and Jørgensen, H. (2001). *Musical Imagery*, volume 5. Taylor & Francis.

[Godøy and Leman, 2010] Godøy, R. I. and Leman, M. (2010). *Musical gestures: Sound, movement, and meaning*. Routledge.

[Hubbard, 2010] Hubbard, T. L. (2010). Auditory imagery: empirical findings. *Psychological bulletin*, 136(2):302.

[Jeannerod, 2006] Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. Number 42. Oxford University Press.

[Kendon, 2004] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

[Kosslyn et al., 2001] Kosslyn, S. M., Ganis, G., and Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9):635–642.

[Ladefoged, 2001] Ladefoged, P. (2001). Vowels and consonants. *Malden, Mass.: Blackwell*.

[Lemaitre et al., 2009] Lemaitre, G., Dessein, A., Aura, K., and Susini, P. (2009). Do vocal imitations enable the identification of the imitated sounds? In *Auditory Perception and Cognition Meeting (APCAM)*.

[Lemaitre et al., 2011] Lemaitre, G., Dessein, A., Susini, P., and Aura, K. (2011). Vocal imitations and the identification of sound events. *Ecological Psychology*, 23(4):267–307.

[Lemaitre et al., 2010] 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.

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

[Leman, 2008] Leman, M. (2008). *Embodied music cognition and mediation technology*. Mit Press.

[McAdams et al., 1995] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192.

[McNeill et al., 1990] McNeill, D., Pedelty, L. L., and Levy, E. T. (1990). Speech and gesture. *Advances in Psychology*, 70:203–256.

[Meyer, 2008] Meyer, J. (2008). Typology and acoustic strategies of whistled languages: Phonetic comparison and perceptual cues of whistled vowels. *Journal of the International Phonetic Association*, 38(01):69–94.

[Nymoen et al., 2011] Nymoen, K., Caramiaux, B., Kozak, M., and Torresen, J. (2011). Analyzing sound tracings: a multimodal approach to music information retrieval. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 39–44. ACM.

[Peeters et al., 2011] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916.

[Proust, 2002] Proust, J. (2002). Imitation et agentivité. *Imiter pour découvrir l'humain,*, pages 189–216.

[Stowell and Plumbley, 2008] Stowell, D. and Plumbley, M. D. (2008). Characteristics of the beatboxing vocal style. *Dept. of Electronic Engineering, Queen Mary, University of London, Technical Report, Centre for Digital Music C4DMTR-08-01*.

[Sundberg, 1999] Sundberg, J. (1999). The perception of singing. *The psychology of music*, 1999:171–214.

[Sundberg et al., 1977] Sundberg, J. et al. (1977). *The acoustics of the singing voice*. Scientific American.

[Wright, 1971] Wright, P. (1971). Linguistic description of auditory signals. *Journal of applied Psychology*, 55(3):244.

Appendix

Graphical User Interface

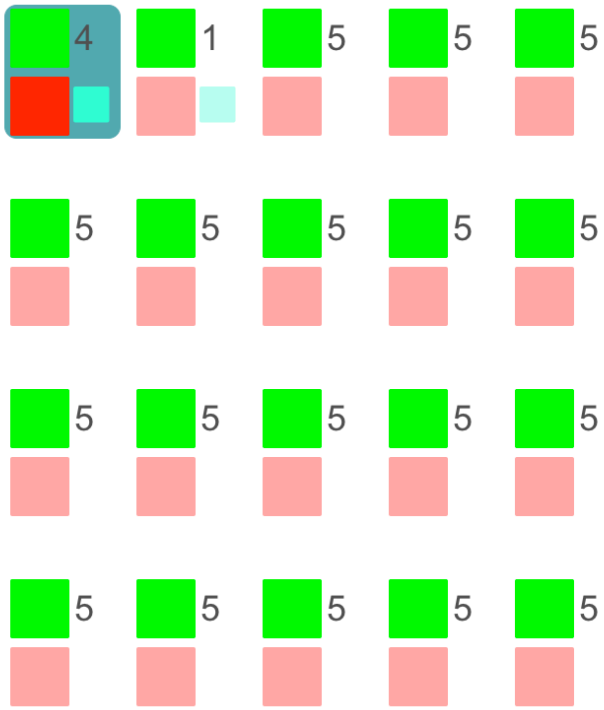

Figure 7.1: GUI used for data collection and experiment. Here, 20 sounds are listenable with clicking on the green button. The red button allows the participant to record an imitation. The blue button allows him to watch his imitation.

First data collection list of referent sounds

Short file name	Category	Short file name	Category
Machines1.wav	Alarms	Interactions1.wav	Blowing
Machines2.wav	Alarms	Interactions2.wav	Blowing
Machines3.wav	Buttons and Switches	Interactions3.wav	Whipping
Machines4.wav	Buttons and Switches	Interactions4.wav	Whipping
Machines5.wav	Doors closing	Interactions5.wav	Shooting
Machines6.wav	Doors closing	Interactions6.wav	Shooting
Machines7.wav	Filing and sawing	Interactions7.wav	Crumpling
Machines8.wav	Filing and sawing	Interactions8.wav	Crumpling
Machines9.wav	Fridge hums	Interactions9.wav	Rolling
Machines10.wav	Fridge hums	Interactions10.wav	Rolling
Machines11.wav	Mixers and blenders	Interactions11.wav	Rubbing and scraping
Machines12.wav	Mixers and blenders	Interactions12.wav	Rubbing and scraping
Machines13.wav	Printers Fax and Xerox	Interactions13.wav	Hitting and taping
Machines14.wav	Printers Fax and Xerox	Interactions14.wav	Hitting and taping
Machines15.wav	Windshield wipers	Interactions15.wav	Dripping and trickling
Machines16.wav	Windshield wipers	Interactions16.wav	Dripping and trickling
Machines17.wav	Vehicles exterior revs up	Interactions17.wav	Filling
Machines18.wav	Vehicles exterior revs up	Interactions18.wav	Filling
Machines19.wav	Vehicles interior accelerating	Interactions19.wav	Gushing
Machines20.wav	Vehicles interior accelerating	Interactions20.wav	Gushing

Short file name	Category
Abstract1.wav	Up
Abstract2.wav	Up
Abstract3.wav	Down
Abstract4.wav	Down
Abstract5.wav	UpDown
Abstract6.wav	UpDown
Abstract7.wav	Impulse
Abstract8.wav	Impulse
Abstract9.wav	Repetition
Abstract10.wav	Repetition
Abstract11.wav	Stable
Abstract12.wav	Stable

Figure 7.2: The 52 referent sounds used for the first data, divided into three families and categories.

Transcription grid for analysis

Code	Gesture/vocalization synchrony
1	kinect follows f0
2	mo-energy follows energy
3	Gesture ends after vocalization
4	Gesture and vocalization desynchronize
5	kinect-energy follows energy
6	kinect follows temporal evolution
7	Gesture and vocalization are synchronous
8	mo-energy follows f0
9	kinect follows a formant
A	kinect follows énergie

Code	Preparation/recovery gestures
0	No (make pauses)
1	Preparation
2	Recovery

Code	Gesture's peculiar information
0	No
1	Discrete sign at the end
2	Constant vibration
3	Hands moving apart
4	Hands opening
5	Global horizontal movement
6	Hands going up
7	Hands oscillating
8	Hands going back and forth
9	Hands going down
A	Alternating hands
B	Hands moving closer

Code	Vocalization's peculiar information
0	No
1	Noisy componant
2	Tona componant
3	Breathe in and breathe out
4	Evolving formants
5	Roughness

Code	Direction
0	No favoured direction
1	Left
2	Right
3	Up
4	Down
5	Forwards
6	Backwards

Code	Distorsion imitation/stimulus
0	No
1	Emphasized end
2	Emphasized beginning
3	Different energy shape
4	Different rhythm
5	Phase difference
6	Different f0 shape
7	Different formant evolution
8	A phase of the stimulus is not reproduced
9	Pause(s) during imitation
A	No tonal component
B	Emphasized middle

Figure 7.3: Subjective verbal descriptions for each item, with their associated number.

Interview grid

Repetitive sounds:

- What did the sound evoke to you?
- What are you globally doing?;
- If phase difference: why?

Rhythmic patterns:

- What did the sound evoke to you?
- What are you globally doing?

Stable textures:

- What did the sound evoke to you?
- Why are you gesturing this way?

Dynamical textures:

- What did the sound evoke to you?
- Why are you gesturing this way?
- If no shaky gesture: why?

Layered sounds:

- What did the sound evoke to you?
- How many sound sources did you hear?
- What was your strategy?
- If separated roles: why do you vocalize one layer rather than the other?
- If not: why?

Example of measurement

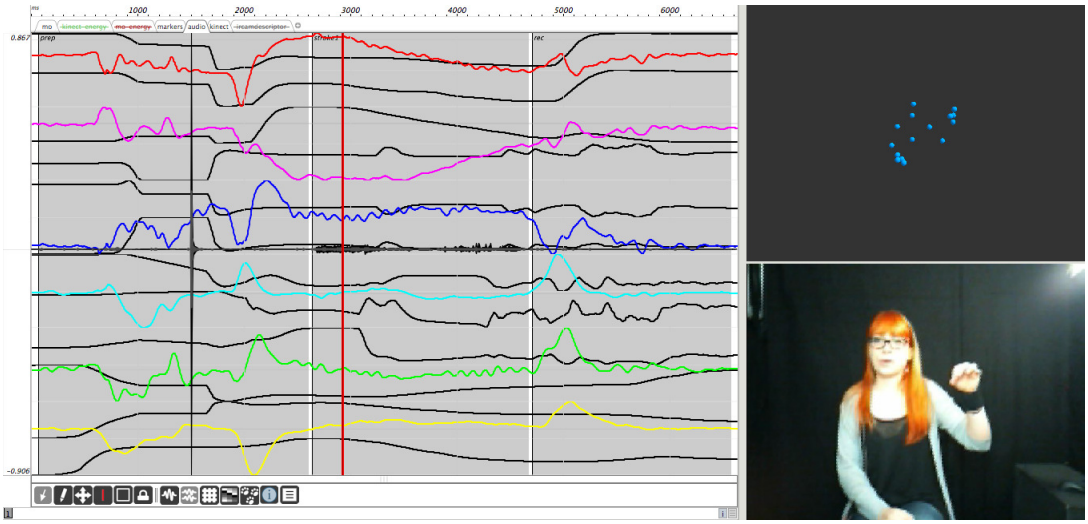

Figure 7.4: Visualization of the collected data. In black: audio and kinect data. In colors: IMU data. In grey: data segmentation. Right: kinect skeleton and webcam recording.

Example of scalogram

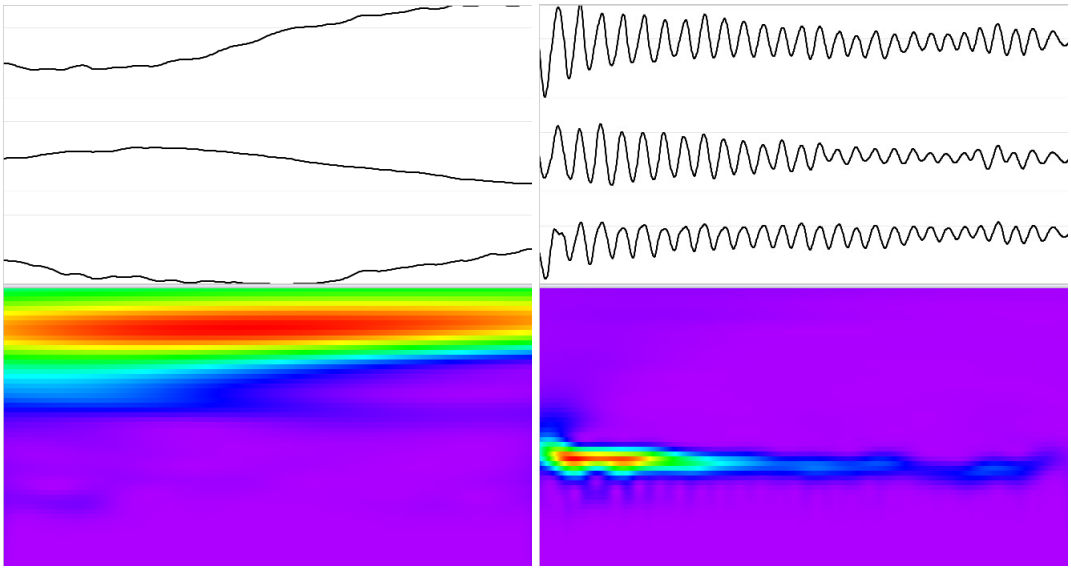

Figure 7.5: Acceleration data from IMU (first line; time on the x-axis, amplitude on the y-axis) and their associated scalogram (second line; time on the x-axis, scale on the y-axis, amplitude in colorscale). Left: stable gesture; right: shaky gesture.

Analysis grid for layered sounds

Code	Strategy	Code	Sub-strategy
0	One after the other	0,B	Impulsive before sustained
1	Separation of roles	1,G	Sustained with gesture
2	Merging the two layers		
3	Only one layer	3,B	Impulsive

Figure 7.6: Transcription grid for analysis of layered sounds.