UPMC
SORBONNE UNIVERSITÉS

PARCOURS MASTER 2
ATiAM
Parcours multi-mentions du Master Sciences et Technologies
Université Pierre et Marie Curie - Paris 6
en collaboration avec TELECOM ParisTech et l'Ircam

ircam
Centre
Pompidou

Master's thesis

Towards a gestural control of environmental sound texture synthesis

Aymeric MASURELLE

supervised by

Diemo SCHWARZ   Baptiste CARAMIAUX   Norbert SCHNELL

August 5, 2011

Acknowledgements

Thanks to *IRCAM* for its creative working atmosphere.

Thanks to the *Real Time Musical Interaction team* for their interesting researches and applications which motivate me to do this project.
Especially, thanks to *Diemo Schwarz* for making this project possible, for his work, his interest, his curiosity and his calm; thanks to *Baptiste Caramiaux* for sharing his experience, providing pertinent comments; thanks to *Norbert Schnell* for his idea and conceptual view on the project. And thanks to *Tommaso Bianco* and *Sarah Fdili Alaoui* for being helpful with me.

Thanks to the *Perception and Sound Design team*, for letting me put my working spot in their balcony, lending me one of their audiometric booth for a while, spreading calm and relax ambiance, and especially thanks to *Nicolas Misdariis* for pertinent advises to help me creating a perceptual experiment.

Thanks to *Jérôme Nika, Benjamin Cauchi, Pierre Beauguitte, Thomas Humbert, Jules Françoise, José Echeveste, Julien Colafrancesco, Louis Delebecque, Van Phan Quang, Aurélie Frère, Sandrine Bonnot, Laure Cornu* for those moments spent on The Passerelle, your laughter and the time that you passed to help me in this work.

Thanks to *Sylvie Benoit, Youcef Bensaid, Alessandro Saccoia* for those assistances which make work easier.

Thanks to *Carlos Agon* for your smile, your good mood, your patience.

Abstract

Through this project, a development of a methodology to guide the creation of intuitive and pertinent mappings between one-hand gesture and environmental sound texture (rain, wind, wave) descriptors is shown. The proposed methodology is composed of two complementary key stages. First, a perceptual investigation on sound descriptors controlling a corpus-based concatenative synthesis process is carried out to figure out the relevant ones. Afterwards an experiment where participants perform gestures while listening to sound texture morphologies is achieved to create a sound and gesture-related descriptor dataset. Then by applying a canonical correlation analysis on those sound and gesture-related data, one is able to determine pertinent gesture features and their correspondences with the retrieved relevant sound features. Thus pertinent clues to create gesture-to-sound mappings are provided by this method.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

To create a realistic synthesis of environmental sound texture (rain, wave, wind, ...) is an important issue for domains such as cinema, games, multimedia creations, installations. Existing methods achieve a certain level of realism in performing this task, but they are not really interactive whereas controlling interactively the synthesis can improve realism due to the inherently multimodal nature of perception.

The present work is part of the ANR project, *Topophonie*, that deals with virtual navigable sound spaces, composed of sounding or audio-graphic objects, [sit11]. It focuses on the development of a methodology for providing a pertinent mapping between gesture and environmental sound texture. Through the following method, one is able to guide the creation of a mapping for controlling interactively an environmental sound texture synthesizer by one hand gesture.

For this purpose, a real-time corpus-based concatenative synthesis system, *CataRT* [SCB08], is used. Its data-driven approach allows to control the sound synthesis by drawing target audio descriptor trajectories in order to build a sound unit sequence out of a grain corpus.

Several investigations, [MMS$^+$10, SMW$^+$04], have already been done to find the most perceptually relevant audio descriptors in environmental sounds (especially industrial sounds). Based on those previous contributions, a perceptual similarity experiment is carried out to retrieve a pertinent set of audio descriptors to drive the corpus-based concatenative synthesis.

A system using inertial sensors (*wii mote*) and a webcam furnishing a rgb image flow with its associated depth (*kinect*) capture the movement of the performer's hand to obtain its positions, rotations, velocities and accelerations.

In order to explore the relationship between gesture and sound, a method deeply inspired by [CBS10] is used. First, subjects perform hand movements while listening to sound examples. The latter are created with *CataRT* by following trajectories of audio descriptors identified as being most meaningful in the above perceptual experiment. Then, based on canonical correlation analysis, the relationship between gesture and audio descriptors is quantified by linear multivariate regression. This method reveals a part of the complex relationship between gesture and sound. This gives pertinent clues to construct a gesture-sound mapping for the efficient gestural control of sound texture synthesis by perceptually meaningful audio descriptors.

Chapter 2

Background

2.1 Sound texture synthesis

2.1.1 Definition of sound texture

To find a pertinent way of defining the term *sound texture*, one can refer to works done by Saint-Arnaud in [SA95, SaP95]: "A sound texture is like wallpaper: it can have local structure and randomness, but the characteristics of the structure and randomness must remain constant on the large scale". Through this definition an interesting characteristic of sound textures regarding to its potential information content over time is highlighted. This particularity can be used as a way of segregating sound textures from other sounds, figure 2.1.

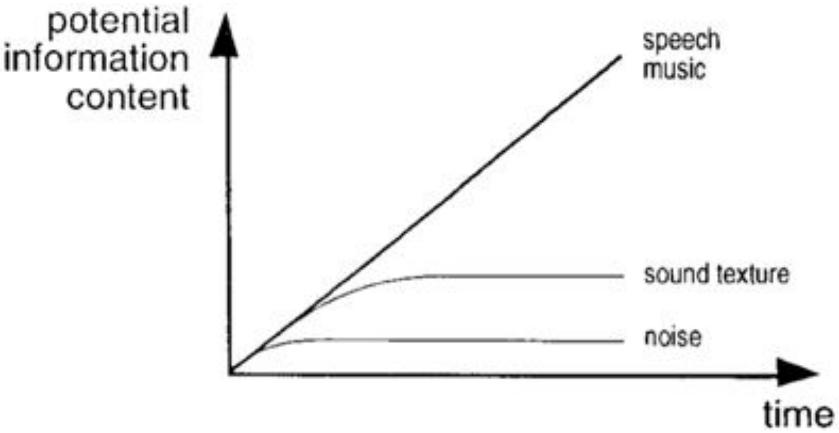


Figure 2.1: Potential information content of a sound texture vs time, [SaP95]

Because of the constant long term characteristics of their fine structure, sound textures can be described by a two-level representation, [SaP95]. The *low level* is constitued of simple atomic elements distributed in time where this distribution in time of those atoms is described by the *high level*. Different ways of distributing atoms have been identified among sound texture: periodic (as engine sound), stochastic (as rain),or both (as wave).

Texture		**Not a Texture**
rain	running water	one voice
voices	whisper	telephone ring
fan	jungle	music
traffic	crickets	radio station
waves	ice skating	single laugh
wind	city ambiance	single hand clap
hum	bar, cocktail	sine wave
refrigerator	amplifier hum	
engine	60 Hz	
radio static	coffee grinder	
laugh track (in TV show)	bubbles	
applause	fire	
electric crackle	whispers	
babble	snare drum roll	
murmur	heart beat	

Table 2.1: Brainstorm - examples of sound textures, [SA95].

Sometimes the term *sound texture* means *non-tonal, non-percussive* sound material, or *non-harmonic, non-rhythmic* musical material, [Sch11].
And, it is important to make the distinction between *sound texture* and *soundscape.* A *soundscape* is a sum of sounds that compose a scene where each sound component could be a *sound texture.*

In order to clarify the concept of *sound texture*, one could refer to examples of sound textures that have been listed by Saint-Arnaud [SA95] through brainstorm session, table 2.1.

2.1.2 Sound texture synthesis

Following the different characteristics of sound textures exposed in the previous part, different synthesis approaches have been developped for sound texture generation.
Schwarz, in [Sch11], proposes a complete state of the art about existing *sound texture synthesis* approaches. The following part is a short listing of different synthesis methods highlighted in this paper.

Substractive synthesis - Noise filtering Using this method, different analysis-synthesis models for texture synthesis have been developped: *6-band Quadrature Mirror filtered noise* in [SaP95], *cascaded time and frequency domain linear prediction* in [AE03, ZW04] and a neurophysically motivated *statistical analysis of the kurtosis of energy in subbands* in [MDOS09]. Those different synthesis techniques seem to be especially suitable for unpitched sound texture like rain, fire, water.

Additive sinusoidal+noise synthesis Through this approach, filtered noise is complemented by oscillators to add sinusoidal partials. It allows to synthesize more complex and well-informed sound texture models such, as traffic noise [Gui08] and large classes of environmental sounds (liquids, solids, aerodynamic sounds) [Ver10], achieving a certain degree of details.

Physical modeling By applying physically-informed models, sound texture synthesis can be carried out, [OSG02, MLS$^+$08, LGDM08, Coo07, Ver10, PTF09]. Even though those synthesis techniques attains realistic performances, this approach presents a principal drawback that for each class of texture sounds (i.e friction, rolling, machine noise, impact,...) a specific model should be developed.

Non-standard synthesis Other synthesis methods are used for sound texture synthesis, such as fractal synthesis or chaotic maps, in order to provide expressiveness through the synthesis process, [Sci99, FAC06].

Wavelets From a multiscale decomposition of a signal, a wavelet coefficient tree can be built for *sound texture synthesis* to model temporal and hierarchic dependencies thanks to its multi-level approaches, [DBJEY$^+$02, BJDEY$^+$99, KP10]. With this method the overall sound of the synthesized textures is recognisable but their fine temporal structure gets lost.

Granular synthesis This method recomposes short segments of original recording, called grains or snippets, with possible transpositions following a statistical model in order to obtain texture sounds, [DBJEY$^+$02, BJDEY$^+$99, PB04, HP01, Hos02, FH09, LLZ04, Str07].

Corpus-based concatenative synthesis Based on *granular synthesis*, *corpus-based concatenative synthesis* proposes a content-based approach for *sound texture synthesis*, [Car04, Sch06, SBVB06, SCB08, SS10, PTF09, Fin09].
Through his work [Sch06, SBVB06], Schwarz et al. propose a real-time corpus-based concatenative synthesis system, `CataRT`. This system allows to drive in real-time the selection and concatenation of sound grains by drawing target audio descriptor trajectories in a predefined audio descriptor space. Those sound units coming from a grain corpus are selected regarding to the proximity between their global[1] audio descriptors and the target ones. This corpus is created by segmenting sounds from a database of pre-recorded audio into units and analysing this units by computing their global audio descriptors.
Thanks to its corpus-based approach, this synthesis technique can perform realistic sounds by feeding the used corpus with a real sound database. Moreover it allows to control the creation of novel timbral evolutions through while keeping the fine details of the original corpus sounds. Then two levels of control are then enabled through this synthesis process: a *low level* and a *high level*. The *low level* of control deals with the local sound structure of the corpus units. Expressed while choosing the sounds to feed the corpus, this control defines the fine sound details of the resulting synthesis. Then an attentive care should be

[1] over the entire unit signal

taken while creating the sound data base in order to match the fine sound characteristics of the wanted synthesis. The other level of synthesis control, the *high level*, is performed by targetting positions (or path) in a predefined audio descriptor space for selecting the sound unit to be played regarding its global audio descriptors.

2.2 Gestural control of sound synthesis

Through recent researches in the domain of neurosciences [KKA^{+}02, MG03] and perception [VTR91, Alv05, Ber97], the inherently multimodal nature of perception has been highlighted by showing the predominant role of action in perception. Then, by providing an interactive control of a sound synthesis process through gesture, one can create a system which improves the realism of the sound synthesis thanks to its multimodal approach. To achieve this goal, the gesture-sound relationship has to be determined in order to find a pertinent way of linking gesture-control data and the sound process, i.e. a gesture-to-sound mapping, allowing musical expression of the produced sound through gestural control.

2.2.1 Mapping strategies

In the literature, several mapping strategies have been developed such as *one-to-one*, *one-to-many*, *many-to-one* and *many-to-many*, [BMS05]. This latter category of mappings despite its complexity can show more satisfying results, after a learning phase, than one-to-one mappings, [Wan02].

2.2.2 Gesture control of granular synthesis

An original and intuitive way of controlling granular synthesis with granular tactile interfaces (*PebbleBox* and *CrumbleBag*, see figure 2.3) is presented by O'Modhrain and Essl in [OE04]. This fully-granular interaction paradigm is based on the analysis of sounds resulting from the manipulation of physical grains of an arbitrary material. This enables an intuitive control of granular synthesis process while maintaining the individual mapping

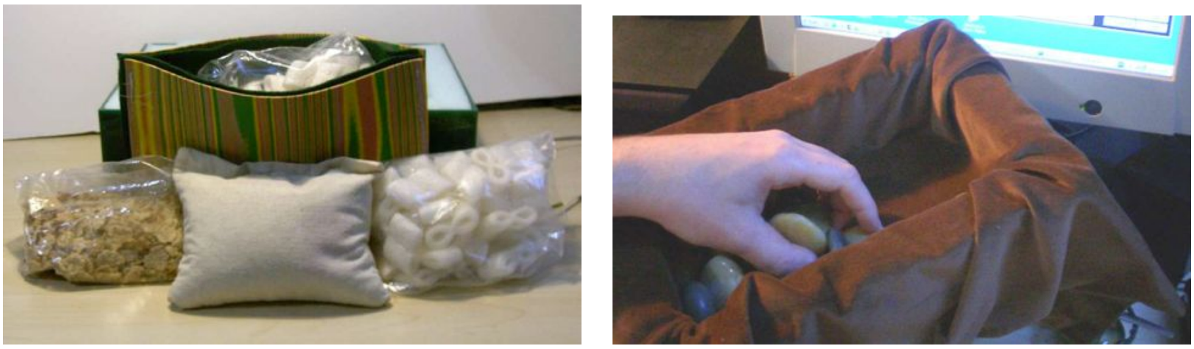


Figure 2.2: Left: the CrumbleBag with cereal, coral and styrofoam fillings. Right: the PeddleBox. [OE04].

of fine-structure temporal behavior (10-100 ms) of granular event to haptic interactions. This analysis extracts parameters as grain rate, grain amplitude and grain density which are then used to control the granulation of sound samples in real-time. This process is illustrated in figure 2.3.

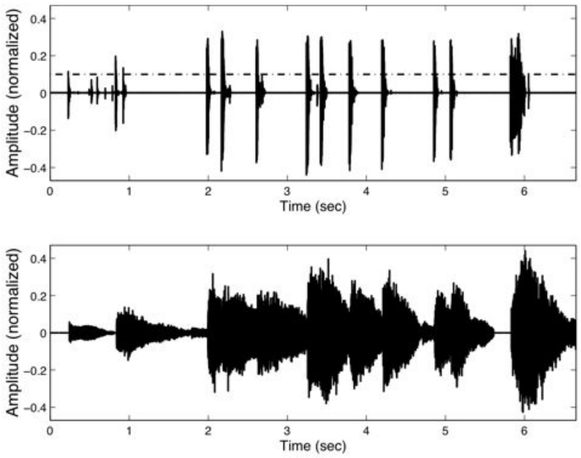


Figure 2.3: Recorded signal of the PebbleBox (top) and granulated response using a Hammer grain (bottom) of the complete granulation process, [OE04].

With this approach, a certain variety of environmental sounds like dropping or shuffling of objects can be performed.

2.2.3 Methodologies for gesture-to-sound mapping design

In a context of artistic performance, Bencina et al. in [BWL08] present a technique for developing new gesture-sound mappings, the *Vocal Prototyping*. With this technique, a movement sound vocabulary are extracted by three exercises:

1. first person vocalise, other find movement that corresponds,
2. first person gesture, other find vocalisations that correspond,
3. each make their own movement/sound pairings.

Thus, from the outcomes of the *Vocal Prototyping*, Bencina et al. "strove to maintain a balance in the relationship between movement and resultant sound that was easy to perceive for audience and performance alike" to create their gesture-sound mappings. This results in a intentionally simple mapping inviting a dramaturgical mode of composition.

Caramiaux et al., in [CBS10], introduce an pertinent and original method for the quantitative multimodal analysis of movement and sound. Using sound and gesture-related data sets from an experiment where subjects performed free hand movement while listening to short sound examples, Caramiaux et al. employed a canonical correlation analysis (CCA) to investigate and highlight mutual shared variance between a set of gesture parameters, like *position, velocity, normal acceleration*, and audio descriptors, as *loudness, sharpness* by a linear multivariate regression. Even though this method cannot exhibit non-linear relations between gesture and sound, it can be used as a selection tool to create relevant gesture-sound mappings.

Chapter 3

A perceptually-motivated synthesis control of environmental sound textures

3.1 First steps

3.1.1 Environmental sound perception

Being interested in environmental sound texture synthesis, one can get interested in perceptual characteristics of those environmental sounds.
Misdariis et al., in [MMS$^+$10], present a perceptual description using audio descriptors of environmental sounds (especially industrial sounds: interior car sounds, air-conditioning units, car horns and closing car doors). In this work, *loudness* and *brightness*[1] have been revealed as the most relevant audio descriptors in the perception of environmental sounds. Afterwards others predominant descriptors have been perceptually retrieved but they are related to specific environmental sound categories: *instrument-like, motor-like, impact.* Nevertheless all of them are in relation with the content and the shape of the signal spectrum.
Based on a study about the characterisation of air-conditioning noise, Susini et al. in [SMW$^+$04] confirm the previous results by showing that the three most relevant audio descriptors for environmental sounds are the *ratio of the noisy part of the spectrum to the harmonic part*, the *spectral center of gravity* and the *loudness.*

3.1.2 Environmental sound texture synthesis

From those previous researches, one can be attracted to develop a synthesis process of environmental sound texture controlled by those perceptually relevant audio descriptors. Thanks to its content-based and corpus-based approach, corpus-based concatenative synthesis method is particularly suitable to perform this attempt in a realistic manner. Indeed regarding the similarities between the two different levels of representation of sound textures (section 2.1.1) and the two levels of control of this synthesis process (section

[1] a linear combination between the *perceptual spectral centroid* values of noisy and harmonic parts of the signal and the *perceptual spectral spread* value of the whole signal, described in [MMS$^+$10]

2.1.2), corpus-based concatenative synthesis systems seem quiet suitable to synthesize global evolutions, also called *morphologies*, of a specific sound textures.
Thus the `CataRT` system is synthesis technique is the one used as real-time audio engine for our project. The handled version is `cataRT-1.2.1`[2] using the library `FTM 2.5-beta17`[3] on `Max 5.1.8`[4].

However to efficiently synthesize environmental sound textures with `CataRT`, a study on finding out the most appropriate set of audio descriptors to drive this synthesis should be carried out.

3.2 Perceptual experiment on synthesis of environmental sound texture morphology

Through this part, a perceptual similarity experiment is presented. Its goal is to evaluate the similarity between original sound texture morphologies and their respective resynthesis driven by different sets of audio descriptors. From the experiment results, the best combinaison of descriptors to control the resynthesis for each proposed sound textures can be revealed. Moreover the overall resynthesis process can be as well validated.

3.2.1 Environmental sound texture morphology resynthesis

To achieve this perceptual similarity experiment, a set of original and resynthesized environmental sound texture morphologies has to be created. Through this part, the procedure to create this sound set is described.

Corpus creation

As seen in section 2.1.1, sound textures have two levels of representation: a low level composed of simple atomic elements and a high level describing the time distribution of those atoms. In section 2.1.2 the two levels of control of the chosen synthesis method (corpus-based concatenative synthesis) have been highlighted. A low level of control achieved by constituting the corpus defines the possible fine sound details for the synthesis and a high one enabling to comtrol the creation of novel timbral evolutions playing sequences of corpus units by drawing trajectories through a predefined audio descriptors space.
Thus to fit the different levels of sound texture representation with the different levels of synthesis control, the used corpus for a specific sound texture should be properly created. First in order to have a realistic illustration of sound textures, recordings of true sound textures (*SoundIdeas* database) have been used to feed the different corpora. In addition, for this synthesis process each sound texture has its associated corpus. For example, the corpus for synthesizing a rain texture is composed of sound units coming from true rain recordings only. Thus the fine atomic details of the sound texture synthesis is realistically illustrated. Then to create an associated corpus for each of those sound textures (rain,

[2] http://imtr.ircam.fr/imtr/CataRT
[3] http://ftm.ircam.fr
[4] http://cycling74.com

wave and wind), recordings of rain[5] (19 sound files, 29.5 min), wind[6] (16 sound files, 24.1 min) and wave[7] (14 sound files, 25.8 min) from the *SoundIdeas* database are respectively feeding the rain, wind and wave corpora.

Moreover the length of the corpus units has to be discussed. About this aspect no literature has been found. For synthesizing rain sound textures with their prototype texture synthesizer ([SS10]), Schwarz and Schnell uses a corpus of rain recording units of length 666 ms. In our case, in order to provide a reactive gestural control of the synthesis the tendancy is to reduce the unit lengt to allow more abrupt timbral changes. But to provide pertinent timbral evolutions of sound textures, each corpus unit should represent a certain time distribution of sound texture atomic elements. Then a unit length of 250 ms seems to be a good compromise between a reactive control and the high level of sound texture representation.

Experiment's audio descriptor sets

According to its data-driven approach, the `CataRT-1.2.1` system proposes 25 different audio features. Those features are related to either *units'segment descriptors*, *unit's category descriptors*, *symbolic and score descriptors*, *signal descriptors*, *perceptual descriptors* or *spectral descriptors.* Regarding to previous contributions presented in section 3.1.1 ([MMS$^+$10, SMW$^+$04]), only *signal descriptors*, *perceptual descriptors* and *spectral descriptors* are pertinent in environmental sounds. Thus to achieve a perceptually relevant synthesis control, this is the sound descriptors provided by `CataRT` that have been selected: *loudness, spectral centroid, spectral flatness, mid-* and *high-frequency energies.* Those descriptors are detailed in appendix A.
From this descriptor selection, five sets of audio descriptors have been built to control global evolutions of environmental sound textures according to their perceptual weight:

1. *loudness* ;
2. *loudness, spectral centroid* ;
3. *loudness, spectral centroid, spectral flatness* ;
4. *loudness, spectral centroid, mid-* and *high-frequency energies* ;
5. *loudness, spectral centroid, spectral flatness, mid-* and *high-frequency energies.*

Morphology resynthesis

In order to have realistic environmental sound texture morphology resynthesis, true morphologies are extracted from real environmental sound texture recordings. From the analysis of those original recording pieces, the trajectories of the selected audio descriptors (section 3.2.1) are utilized as targets to perform their resynthesis. Thus for each

[5] light, medium and heavy rain

[6] aspen wind, storm wind, canyon wind, storm wind, hangar wind

[7] light, medium and heavy waves from ocean and lack

Parameters	**Values**
trigger mode	bow (allows units' overlap)
unit length	250 ms
unit overlap	125 ms
fade-in, fade-out	125 ms through linear ramp
selection radius	0.1
number of closest neightbour units	10

Table 3.1: Parameters' value for environmental sound texture morphology resynthesis

original sound texture morphology a set of five associated resynthesis is created.

In pratice, several resynthesis problems have been encountered. First while concatenating sound units, sound artefacts are appearing at each unit junction. To solve this problem, one can overlap adjacent units. This reduces in a consequent manner sound artefacts but abrupt timbral changes are noticed at the start of each new units. Then a fade-in and a fade-out can be applied on played sound units. This results in a smoother morphology resynthesis.
Through this synthesis process another problem has been revealed in the unit selection. When the evolution of successive target positions in the descriptor space are quasi unchanged, the same unit can be repeated successively which introduces another sound artefact. This problem can be partly solved by defining a selection radius and a number of closest neightbour units to be considered for the unit selection. The selection radius allows to randomly select an unit in a certain interval around the specified target descriptor position. The number of closest neighbour units defines the number of units which are enable to be selected according to their proximity to a target position in the current descriptor space. Those adjustements allows to choose in between several units while staying close enough to a target descriptor position.
Table 3.1 shows the different parameter values for the `CataRT` system allowing to make those adjustements.

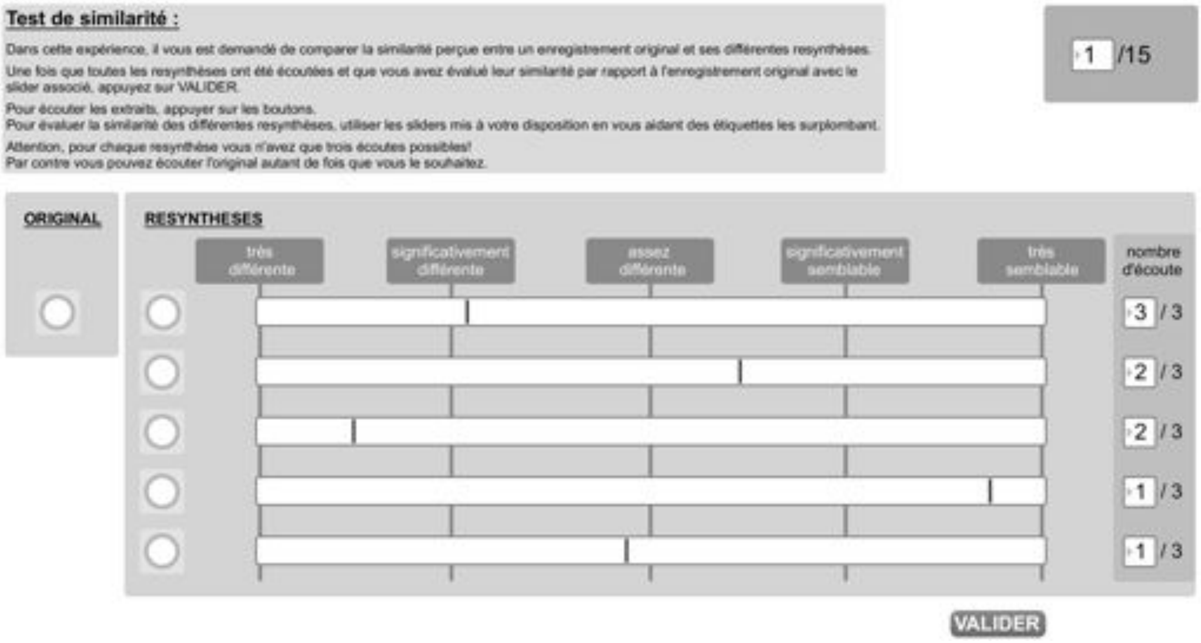
3.2.2 Design of the experiment

Participants

Through this perceptual experiment, fifteen test subjects have been asked to evaluate the similarity between original environmental sound morphologies and their associated resynthesis. Those participants were volunteer and have not been selected regarding to any personal characteristics. Details about them can be found in appendix B.1.

Materials

As previously mentioned, we choose to work with three sound textures: *rain, wind, wave.* For each sound texture, five sets of short audio files (≈ 5-9 s) have been created. Each set



Test de similarité :

Dans cette expérience, il vous est demandé de comparer la similarité perçue entre un enregistrement original et ses différentes resynthèses.

Une fois que toutes les resynthèses ont été écoutées et que vous avez évalué leur similarité par rapport à l'enregistrement original avec le slider associé, appuyez sur VALIDER.

Pour écouter les extraits, appuyer sur les boutons.
Pour évaluer la similarité des différentes resynthèses, utiliser les sliders mis à votre disposition en vous aidant des étiquettes les surplombant.

Attention, pour chaque resynthèse vous n'avez que trois écoutes possibles!
Par contre vous pouvez écouter l'original autant de fois que vous le souhaitez.

Figure 3.1: Interface of the perceptual experiment for environmental sound texture resynthesis.

is composed of six audio files: one original and five different resynthesis. Those resynthesis are built by selecting units within a texture-related corpus according to trajectories of different target descriptor sets from their associated original morphology.

To perform this perceptual experiment without background noise annoyance, participants are seated in a double-walled IAC sound booth. The different sounds are played with a couple of loudspeakers[8]. In order to have a proper sound level through this experiment, a sound level calibration is achieved. This calibration is done by adjusting the sound level of a pure tone (1 kHz) to 84.5 dB SPL. This measure is made through a sound level meter[9] at a potential head position, i.e. at a distance of 50 cm from both speakers and 1.01 m from the floor.

For the purpose of this perceptual experiment, an interface has been implemented in *Max/MSP*[10], see figure 3.1. In this interface a continuous slider for each resynthesis is provided to evaluate the different morphology resynthesis regarding to their associated original morphology. On those scales five different graduations have been added to give categorial marks to test subjects:

very different, significantly different, quite different,
slightly different, identical

This type of scoring furnishes a categorial and continuous score for each audio descriptor set used to resynthesize morphologies. Then the mean and the standard deviation for

[8] YAMAHA MSP5 powered monitor speakers

[9] Brüel & Kjaer 2238 mediator

[10] http://cycling74.com

each of the resynthesis can be calculated over all participants. From those results, one is able to validate or not the overall resynthesis process thanks to the categorial aspect of these results and to choose the best combination of audio descriptors to control the resynthesis of environmental sound textures thanks to the continuous aspect of those results.

Procedure

Before that participants start this experiment, a short training session has been provided for each of them. Through this training session, the interface is first presented. Afterwards participants are asked to work their own strategy out for evaluating morphology similarities through an single training set of morphologies (one original morphology and five associated resynthesis).

In order to keep test subjects highly concentrated and to avoid confusion through the experiment, the number of listening for the resynthesis sounds is limited to three times. On the other hand, the original morphologies can be listened as many times as test subjects want.

To reduce the risk of unwanted effects, the presentation order of the different sound texture resynthesis is randomly chosen: first the presentation order of the different sound textures is randomized, then for each sound texture the presentation order of the different morphologies is also randomized, and finally for each morphology sliders are as well randomly affected to the different resynthesis.

3.2.3 Results of the experiment

Being interested in finding a pertinent set of audio descriptors to drive the synthesis for each environmental sound texture, the mean and the standard deviation over all participant evaluations are calculated separately for each environmental sound texture.

	15 17 18 19 20	15 18 19 20	15 17 18	15 18	15
Rain	144.6 ±40.8	126.4 ±43.0	138.7 ±41.1	104.6 ±56.8	25.3 ±24.7
Wave	108.8 ±48.1	106.2 ±45.1	105.2 ±48.4	100.2 ±47.4	75.2 ±38.6
Wind	94.8 ±39.8	94.9 ±47.3	86.1 ±38.8	76.1 ±37.1	36.9 ±30.4

Table 3.2: Table of the results of the perceptual experiment for each considered sound texture (rain, wave, wind).

Table 3.2 shows the obtained mean and standard deviation over all participants for the rain, wave and wind sound textures on a notation scale going from 0 (*very different*) to 200 (*identical*). In this table, the different sets of audio descriptors are indicated by their descriptor index in `CataRT`:

15 → *loudness* ; 17 → *spectral flatness* ; 18 → *spectral centroid*
19 → *high-frequency energy* ; 20 → *mid-frequency energy*

From table 3.2, one can notice that there is a group constituted by three sets of audio descriptors which seems to be the most perceptually relevant through all considered textures:

- 15 17 18 19 20 → *loudness, spectral centroid, spectral flatness, mid-* and *high-frequency energies* ;
- 15 18 19 20 → *loudness, spectral centroid, mid-* and *high-frequency energies* ;
- 15 17 18 → *loudness, spectral centroid* and *spectral flatness* ;

It seems that the considered textures have the same relevant audio descriptor sets to drive their synthesis. Then a global view of their evaluation over all textures can be interesting, see table 3.3.

	15 17 18 19 20	15 18 19 20	15 17 18	15 18	15
All textures	116.1 ±47.9	109.2 ±47.0	110.0 ±48.2	93.6 ±49.4	45.8 ±38.2

Table 3.3: Table of the results of the perceptual experiment over all considered sound textures (rain, wave, wind).

From table 3.3 the three audio descriptor sets seems to have been evaluated in an equivalent manner. In order to affirm this hypothesis, one can performed a *student's t-test* (see appendix C) to know if their means can be considered as equal. In table 3.4, the results of the performed *student's t-test* of the *null hypothesis* that data are independent samples from normal distributions with equal means and equal but unknown variances, are presented.
With this method, one is able to know if the null hypothesis is rejeted or not at a certain significance level, 5 % in our case. This test is achieved for each possible couple of descriptor sets by using the data over all textures.

From those results (table 3.4), one can see that for the couple where only the three best sets of descriptors are involved in the null hypothesis is accepted. This shows that their means and variances can be considered as equal at the 5 % significant level. Thus one can consider that those three audio descriptor sets are equivalent for the resynthesis process of environmental sound textures. Regarding to the efficienty of the three best descriptor sets, the set 15 17 18 is the best of those three because it uses the fewest number of audio descriptors compared to the two others. Then to take the set 15 17 18 (*loudness, spectral flatness, spectral centroid*) for synthesizing environmental sound textures seems to be a pertinent choice.

couples of descriptors set		**null hypothesis**
15 17 18 19 20	15 18 19 20	**accepted**
15 17 18 19 20	15 17 18	**accepted**
15 17 18 19 20	15 18	rejected
15 17 18 19 20	15	rejected
15 18 19 20	15 17 18	**accepted**
15 18 19 20	15 18	rejected
15 18 19 20	15	rejected
15 17 18	15 18	rejected
15 17 18	15	rejected
15 18	15	rejected

Table 3.4: Table of the T-test results on the perceptual experiment data.

Moreover the mean of the chosen descriptor set over all textures is 110.0 out of 200. This value is above the mean of the notation scale and is situated between the markers *quite different* and *slightly different*. Thus one may conclude that the performed resynthesis process driven by the chosen descriptor set is satisfying.

3.3 Conclusion

After that each participant performed the present experiment, a discussion was undertaken to get comments about this experiment.
The main drawback highlighted is that the different sounds are too long which involves an increase in the difficulty of the experiment. Then shorter sound files (less than five seconds) should be provided. Another interesting comment has formulated by two of the participants: a small and regular sound intensity modulation were perceived along all resynthesis. This problem seems to come from the frequency at which units are launched. Indeed using a linear ramp to achieve the fade-in and the fade-out of each unit can induce this modulation. To avoid this phenomenon a square-root ramp should be used.

An attempt to validate the resynthesis method has been performed regarding to the original morphologies but nothing has been proposed to evaluate the realism of the resynthesis methods. Nevertheless through the 'after-experiment' discussion, participants said that the presented resynthesis are quite realistic in spite of certain sound artefacts introduced by the synthesis process.

Another important improvement should be carried out on the proposed audio descriptors through this presented selection process. In our case we choose to use audio descriptors provided by `cataRT-1.2.1` which seem to be the closest to the ones proposed by perceptual studies ([MMS$^+$10, SMW$^+$04]). However this audio descriptor adaptation is not really pertinent. One should have avoid to adapt the proposed audio descriptors and use them through this selection process. The concerned audio descriptors are:

loudness, brightness, perceptual spectral centroid, perceptual spectral spread, harmonic-to-noise ratio, roughness and *cleaness*. For details about those descriptors, please refer to the Misdariis et al.'s article, [MMS+10].

To achieve this perceptual study in a more conventional way, *multidimensional scaling* (MDS) techniques should have been used to analyze data from the experiment. Those techniques are commonly used in perceptual investigations of complex sound stimuli, [MMS+10, SMW+04, Gre77, MWD+95]. They allow to determine the multidimensional perceptual space and its corresponding physical space (i.e. audio descriptor space) of the considered sound set. Thus perceptually pertinent combinaisons of sound descriptors can be obtained using *MDS*. Details on *MDS* techniques can be found in the articles cited above.

Chapter 4

Towards a gestural control of environmental sound texture

Now that a set of audio features has been found as perceptually relevant to control environmental sound texture synthesis (rain, wave, wind), we want to find out a *gesture-to-sound mapping* in order to control the evolution of those audio descriptors through gestures. For this purpose, the relationships between those audio descriptors and gestural parameters should be revealed.
In this chapter, an experiment, inspired by [CBS10], for bringing out the linear parts of those relationships is presented. Then the results obtained with it are shown and discussed.

4.1 Introduction

4.1.1 Gesture

In our case, *gesture* is considered as the motion of one hand to control environmental sound texture synthesis. This one hand gesture is represented through a set of parameters. In their gesture-sound relationship study, Carmiaux et al. [CBS10] are considering the following gestural observations: *position, velocity, normal acceleration, tangential acceleration, curvature, radius, torsion.* But only the following gesture parameters are found as pertinent:

position, velocity, normal acceleration

Being interested in revealing either the positional (position), kinetic (velocity) or dynamic (acceleration) aspects of gestures for controlling audio descriptor morphologies, the above pertinent gestural observations seem to be a good starting point for our attempt.

In order to capture those different aspects of gesture, two commercial devices are used: the *Wii Remote Plus*[1] and the *Kinect*[2]. The *wii remote* embeds intertial sensors and di-

[1] http://www.nintendo.com/wii/console/controllers

[2] http://www.xbox.com/en-US/Xbox360/Accessories/Kinect/Home

rectly streams reconstructed accelerations and orientations in a relative three-dimension space via bluetooth. This bluetooth communication is then translated into an OSC protocol[3] by the *Osculator* system[4]. This makes possible to use the provided data flow through the software *Max*. In our purpose, this device is used to retrieve the dynamic aspect of the hand motion and the positional and kinetic aspects of the hand rotation. The *kinect* is a webcam which furnishes a rgb image flow with its associated depth maps at a frame rate of 30 Hz via a USB connector. To grab those images from the software *Max*, one can utilize the external object `jit.freenect.grab`[5] which makes use of the OpenKinect project's *libfreenect* library[6]. From the provided depth maps one is able to track the hand positions in the video frames under certain conditions. In our case participants are asked to face the kinect and to keep their hand in front of their body. Moreover their hand should roughly stay in a plane perpendicular to the central axis of the webcam. Under those conditions, the hand position can be tracked by detecting and tracking the closest blob[7] through depth maps provided by the kinect. This tracking process is implemented in *Max* software using the *computer vision for jitter* library, `cv.jit-1.7.2` [8]. This allows to obtain the position and then the velocity of the user's hand in a two-dimensional space. Then this device is used to obtain the positional and kinetic aspect of the user's hand movement.

The following list sums up the different gestural features which can be calculated with the above setup:

- wii mote

  horizontal, vertical, depth- and absolute[9] *hand's acceleration*
  horizontal, vertical, depth- and absolute[10] *hand's angle*
  horizontal, vertical, depth- and absolute[11] *hand's anglular velocity*

- kinect

  horizontal, vertical and absolute[12] *hand's position*
  horizontal, vertical and absolute[13] *hand's velocity*

An illustration of this setup is shown in figure 4.1.

[3] http://archive.cnmat.berkeley.edu/OpenSoundControl

[4] http://www.osculator.net

[5] version 'Release Candidate 3', provided by Jean-Marc Pelletier - http://jmpelletier.com/freenect

[6] https://github.com/OpenKinect/libfreenect

[7] region in images that have different characteristics compared to its surrounding

[8] http://jmpelletier.com/cvjit

[9] $acc_{abs} = \sqrt{acc_{hor}^2 + acc_{ver}^2 + acc_{dep}^2}$

[10] $ang_{abs} = \sqrt{ang_{hor}^2 + ang_{ver}^2 + ang_{dep}^2}$

[11] $ave_{abs} = \sqrt{ave_{hor}^2 + ave_{ver}^2 + ave_{dep}^2}$

[12] $pos_{abs} = \sqrt{pos_{hor}^2 + pos_{ver}^2}$

[13] $vel_{abs} = \sqrt{vel_{hor}^2 + vel_{ver}^2}$

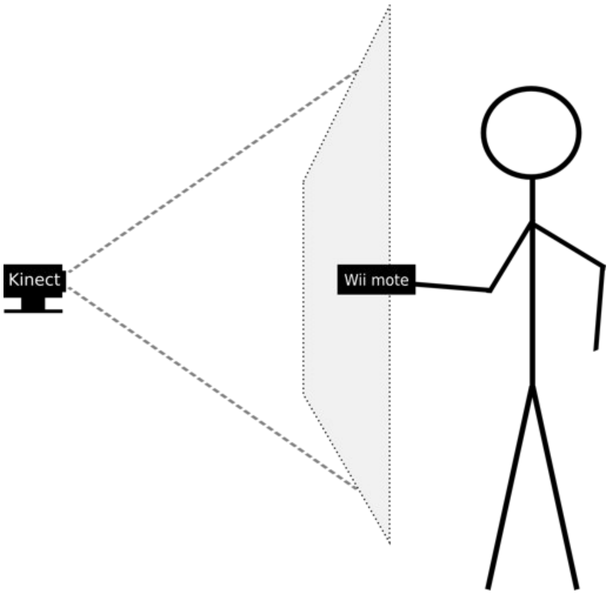


Figure 4.1: Illustration of the setup for capturing the motion of a performer's hand in a plane through an embeded inertial sensor (*wii mote*) and a webcam providing depth maps (*kinect*).

No clear value of the data flow sampling rate for the wii mote has been found, it seems to be between 50 and 100 Hz. Thus in order to have the same number of samples with both devices, the one uses in our setup is the frame rate of the kinect, 30 Hz.

In practice this setup shows a problem with the tracking of the hand: for a short moment the hand position is lost then the tracking system returns the hand's planar coordinates (0,0). To reduce this random punctual problem, a second order lowpass butterworth filter with its cutoff frequency at 5 Hz is applied to those data. Moreover to avoid the phase changes induced by filtering, a zero-phase filter is used. This is done by using the function *butter()* and *filtfilt()* in *MatLab* software[14].

4.1.2 Canonical correlation analysis

In this project, one attempts to find a way of controlling environmental sound texture synthesis by one hand gesture. To provide an intuitive gestural control of this synthesis, the relationship between gesture and sound features should be revealed.
As seen in section 2.2, Caramiaux et al. ([CBS10]) shows an interesting methodology to quantify the linear part of the gesture-sound relationships by investigating the mutual shared variance between the gestural and the audio sets of features using *canonical correlation analysis* (CCA). Proposed by Hotelling in [Hot36], this analysis technique maximises the correlations between two sets of variables by projecting them on a proper basis vectors, eq. 4.1.

$$max_{\mathbf{A},\mathbf{B}}\,[\,corr(\mathbf{XA},\mathbf{YB})\,] \tag{4.1}$$

[14] http://www.mathworks.com

where **A** and **B** are the two projection matrices maximising the correlations between the projected variables. **X** and **Y** are the two matrices representing the two sets of variables with their variables (resp. observations) along the columns (resp. rows). These two matrices must have the same number of observations but they can have different numbers of variables. The projected variables, **XA** and **YB**, are called *canonical variates* and the strength between two canonical variates is given by the *canonical correlation.* From the canonical correlation, one can examine the importance of the relationship between two canonical variates. Then in order to determine the involvement of each original variable in the different associated canonical variates, their *canonical loading* can be expressed. Canonical loadings measure the correlation between variables of the original set and their corresponding projections, eq. 4.2.

$$\mathbf{L_X} = corr(\mathbf{X}, \mathbf{XA}) \tag{4.2}$$

Thus from those values, *canonical correlations* and *canonical loadings*, one can obtained a measure of the linear relationship between gesture and sound features. Even though this techniques shows limitations, it can be used as a selection tool for guiding the creation of a pertinent gesture-sound mapping.

More details about this analysis technique can be found in [HSST04].

4.2 Experiment

Being interested in controlling the synthesis of environmental sound textures by gesture, an experiment allowing to study gesture-sound relationships through CCA is built. This experiment is highly inspired to the one presented in [CBS10].

4.2.1 Experiment protocol

For this experiment eight subjects are invited to perform one-hand movements while listening to nine different sound extracts. Those sound extracts are environmental sound texture morphology resynthesis driven by the chosen set of audio descriptors (*loudness, spectral flatness, spectral centroid*) coming from the perceptual similarity of environmental sound texture resynthesis. For each studied category of environmental sound textures (rain, wave, wind), three sound morphologies have been arbitrary chosen. Their duration is between 5 and 9 seconds. For each presented sound extract participants are asked to imagine an one-hand gesture which could control its sound synthesis process. After an arbitrary number of rehearsals, participants perform three times their imagined gesture listening to the corresponding sound extract. Those final three performances are recorded. This results in a data set of gestural control imitation for each environmental sound texture synthesis.

In order to play the different sound extracts and record the gesture data, a *Max* patch is implemented. An audio-visual signal is provided to performers for helping them to synchronize their gesture with the different sounds. Once the user is ready to perform his/her gesture, he/she launches a metronome (1 Hz) which beeps twice with a red flash before the sound starts. Then, following the same rhythm, a green light (without sound beep) is switched on when the sound is played until its playback is finished.

From this experiment, gesture data sets related to sound are obtained. As seen in table 3.1, a new sound unit of 250 ms is played every 125 ms through the synthesis process. Then the different sounds are analysed through non-overlapping windows of 125 ms to get the global evolutions of its perceptually relevant audio descriptors. Thus the sampling rate of those sound data sets is 8 Hz. For achieving a CCA, the sampling rate of both data sets (sound and gesture-related data sets) should be the same. Then the sound data sets are resampled at the gesture data's sampling rate: 30 Hz.

4.2.2 Experiment results

4.2.3 Choice of gesture descriptors

In CCA, the choice of the input variables can change significantly the obtained results. The more numerous the input variables, the higher the resulting correlation between the two sets of variates. Nevertheless by adding variables, one can easily loose the pertinence of this correlation measure. Then in order to minimise the dimensionality of the gestural data keeping a global representation of performed gestures, one can take only the five absolute parameters of gesture. By taking only absolute parameters, the gesture is assumed to be independent from direction. Since gravity and the distance to the floor can be considered as strong references for performers, we decide to use the *horizontal* and *vertical position* in our analysis instead of the *absolute position.*
Then the chosen set of one-hand gesture features is constituted by *horizontal position, vertical position, absolute velocity, absolute angle, absolute angular velocity* and *absolute acceleration.*

4.2.4 Sound-gesture relationships

Now that sets of audio descriptors (sec.3.2.3) and gesture features (sec.4.2.3) have been chosen, an analysis of the sound-gesture relationships can be achieved.

For this purpose, only performances showing at least one of their canonical correlation coefficients superior to 0.7 through CCA are kept as relevant results. Over the 216 experimental results (72 for each sound texture) only 94 (43.5 %) of them are considered as relevant. Through those 94 relevant performances, 9 (12.5 %) have been performed on the rain sound texture, 42 (58.3 %) on the wave one and 43 (59.7 %) on the wind one. For rain sound texture, the 9 gesture performances expressing a strong correlation were all achieved on the same sound extract. Through wave sound texture, 21 relevant gesture performances were achieved on one of the sound extract and the other 21 ones on another sound extract. For wind sound texture, relevant performances has been done on each sound extracts, 18 on one of them, 16 on another one and 9 on the remaining one.

From this result selection, one can see that results about sound-gesture relationships for rain texture are quite poor, i.e. only few strong correlations between sound and ges-

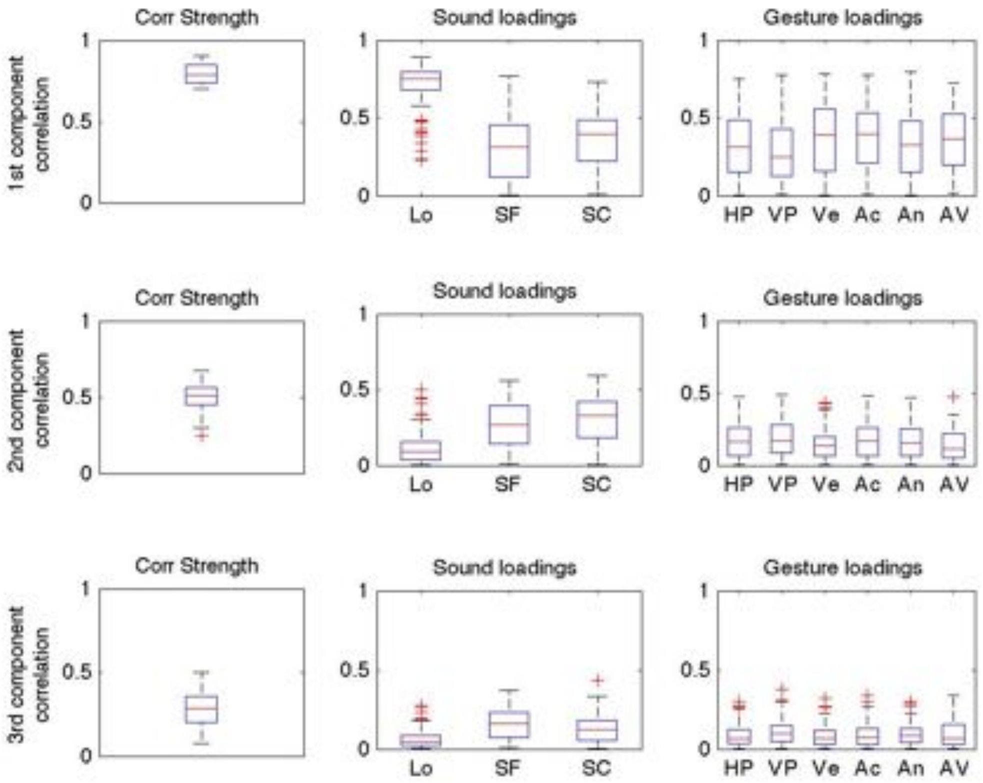


Figure 4.2: Representation of the canonical correlation strength and the canonical loadings for all relevant performances over all sound textures and participants (94 sound extracts). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid.* HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity.*

ture have been revealed. Nevertheless for the two other sound textures, wave and wind, the obtained results are a way better. More than half of them are found as relevant.

As seen previously in section 4.1.2, gesture and sound loadings should be studied to investigate relationships between the different audio and gesture features according to their associated canonical correlation strength. Those values for all relevant performances over all sound textures and participants are shown in figure 4.2.

From figure 4.2, it can be seen that no clear relationship between sound and gesture parameters are highlighted: despite a strong first canonical correlation coefficient and a strong sound loading for *loudness*, all corresponding gestural loadings achieves low correlation and none of them can be distinguished regarding the others. This means that participants are not using the same control strategy through their gesture and/or for different sound textures the strategy of sound control are not similar. Nevertheless one can notice *loudness* is the perceptual audio descriptor that participants are following firstly. Then it can be interesting to specify our study to reveal sound-gesture relationships for

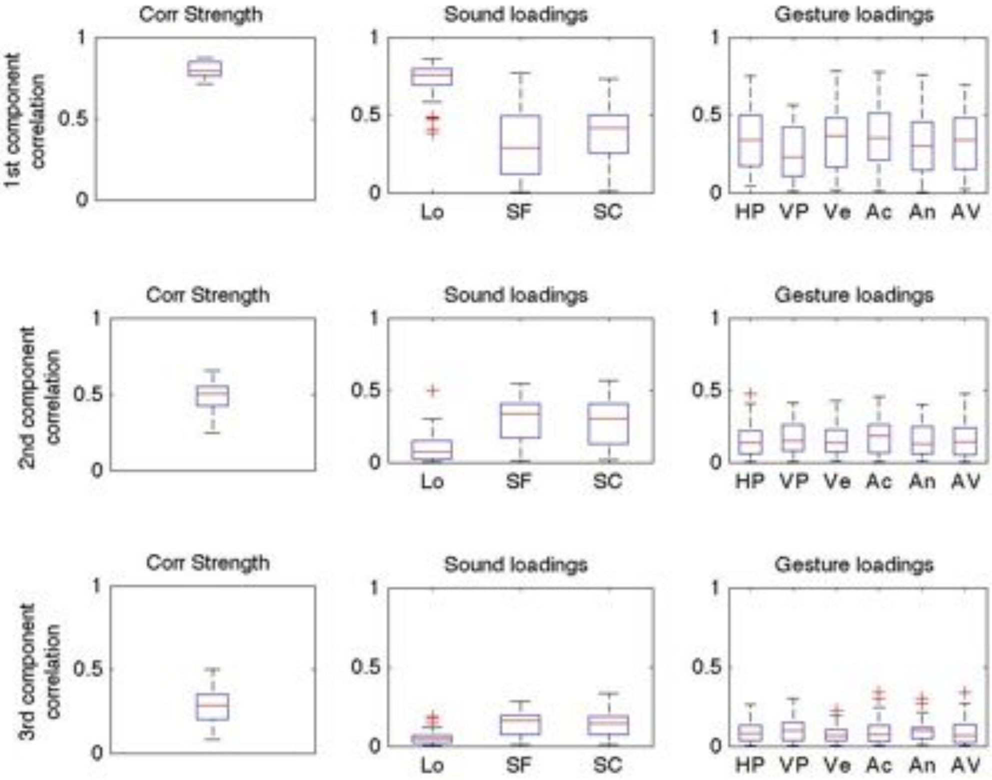


Figure 4.3: Representation of the canonical correlation strength and the canonical loadings for all relevant performances over all wind textures and participants (43 performances). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid*. HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity*.

a particular sound texture.

Figure 4.3 shows the obtained gesture-sound relationships over all participants but just for wind sound texture which is the most successfull among the proposed sound textures with 43 relevant performances. Again, no clear sound-gesture relationship are brought out. The only assumption which can be proposed by watching figure 4.3 is that *loudness* is not related to *vertical position*. This confirms that participants do not perform the same control strategy while listening to a specific sound texture and/or performers change their strategy from one sound extracts to another one.

To continue on this way, figure 4.4 illustrates the obtained gesture-sound relationships over all participants for a single sound extract of wave texture which is one of the most successfull one regarding to all proposed sound extracts (21 relevant performances). From figure 4.4 a tendancy in sound-gesture relationships is shown: *loudness* seems to be rawly related to kinectic and dynamic aspects of the hand's movement (*absolute veloc-*

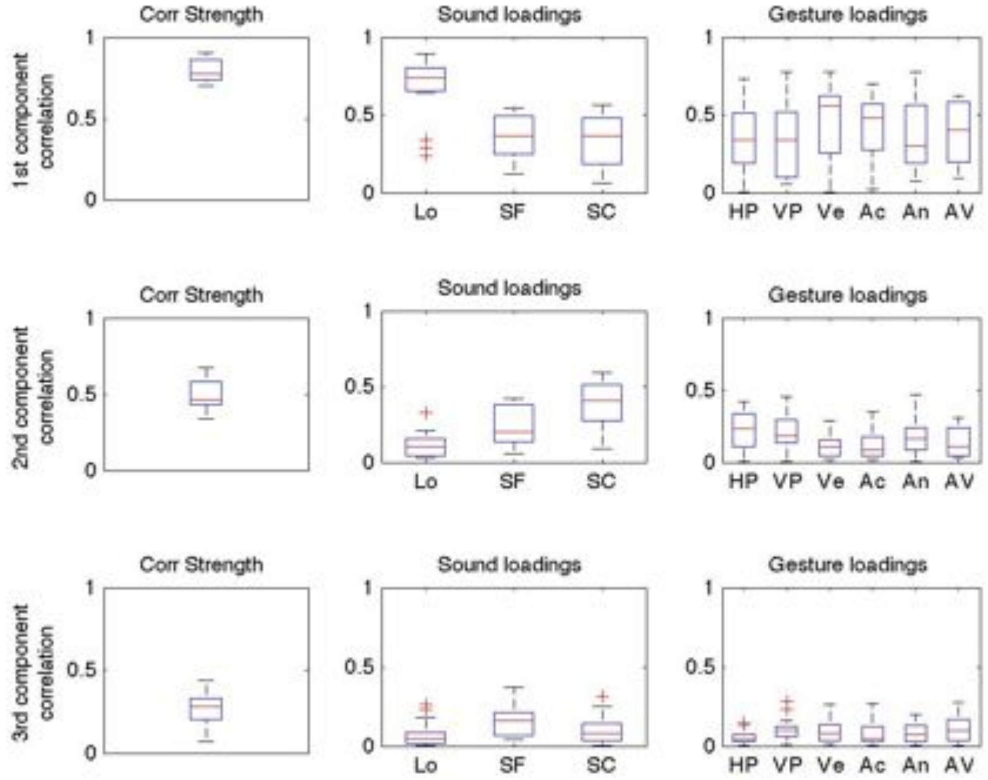


Figure 4.4: Representation of the canonical correlation strength and the canonical loadings for all relevant performances of a single wave sound texture extract over all participants (21 performances). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid.* HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity.*

ity, absolute acceleration and *absolute angular velocity*). Nevertheless figure D.1, placed in appendix D, shows that a positional aspect of gesture (*horizontal position*) could be involved in the control strategy of *loudness.* Then the above tendancy for wave sound texture is not really generalised to others sound textures for all participants.

Another attempt to find a way of showing tendancies in sound-gesture relationships is done by analysing performances of a participant through the different sound extracts of the same sound texture. Figures D.2 and D.3 placed in appendix D, illustrate the revealed gesture-sound relationships over the different proposed sound extracts for the wind sound texture through relevant performances of two participants. From them, it can be seen that the control strategy is changing with the participant. But each participant keeps more or less the same strategy for each sound texture. For the participant 2, the strategy control of *loudness* is essentially done through positional aspect with the *horizontal position* gestural feature. On the other side, the participant 6 chooses mainly the kinetic aspect of gesture for his control strategy of *loudness*: *absolute velocity* of the hand movement.

4.3 Discussion and conclusion

Being interested in controlling environmental sound texture synthesis, an experiment to obtained sound and gesture-related is built. Based on a preliminary study on the relevant perceptive audio features, this experiment consists in inviting participants to perform one-hand's movements while listening to environmental sound texture morphologies (rain, wave, wind). Afterwards CCA technique is used to retrieve the linear part of the complex gesture-sound relationship in two steps: first based on the canonical correlations a selection of pertinent gesture features is achieved, then the analysis of the correlation between the selected gesture features and the audio descriptors is performed.

Through this analysis process, no generalization about sound-gesture relationships over all participants has been achieved. The resulting sound-gesture correlations are specific to each participant and to each sound texture.

Through discussions with participants, imitation of control of rain sound texture has been found as a really hard task. They find this texture too monotonic in a high level of representation with a lot of details in a low one. This were confusing the participants. Thus one of the participant told us that the inertial capture system is too contraining and restrain the expressiveness of the participant because she wanted to use her finger for controlling fine details of the sound texture, especially for rain and wind. Then another capture system letting the hand and its fingers free should be provided for this kind of experiments

By observing the different participants performing this task, numerous problem have highlighted. First participants have been in trouble to synchronize their gesture on the played sound extract. This is due to the length of the different sound extracts, they are too long with too many sound changes in it. Several times participants were surprized by sound events and then tried to hold back or accelerate their gesture to match with the playback sound. Thus performers were more into a gestural illustration strategy of the sound than into a control strategy. To reduce this problem, a second step in the experiment is needed. A proposition for this task is that participants should try the proposed CCA-mapping (from the previous experiment) to reproduce a specific sound texture morphology or just freely and then tune this gesture-sound mapping to fit with what they feel intuitive.

Anyway, the presented method can be used to guide a performer through his/her creation of a personal gesture-to-sound mapping.

Chapter 5

Texture descriptor development

As seen in the previous chapter, rain texture is quite problematic for performers though the gesture-related-to-sound experiment. After discussions with the experiment's participants, one could conclude that rain texture morphologies are monotonic on a high-level of representation. While performing gesture on rain texture almost all participants were shaking their hand. They were more interested in controlling the fine texture details composed of water drop impacts. For this purpose, audio descriptors should be created to describe impact characteristics such as *impact rate*, *impact rate variance*, *impact amplitude*, *impact amplitude variance* in order to be able to discriminate different types of rain, for example light rain, medium rain and heavy rain. Then those descriptors should be perceptually relevant.

Through this chapter an investigation on an *impact rate* descriptor is carried out.

5.1 Impact rate descriptor

An impact is characterized by an abrupt increase in the energy profile of the sound signal. Based on this definition, a method for computing an *impact rate descriptor* is proposed. This method is partly inspired from [OE04].

5.1.1 Method description

In this part the succesive steps to achieve the calculation of our impact rate descriptor are presented.

Instantaneous energy The total instantaneous energy is derived by taking the square of the waveform signal:

$$E_{inst}[n] = (s[n])^2 \tag{5.1}$$

where n the index of the temporal sample.

Envelop The signal envelop can be obtained by low-pass filtering the instantaneous energy. To do so, one can convoluate this latter by a Hamming window of 5 ms duration.

This process describes a low-pass filter with a cut-off frequency at 132 Hz.

$$Env[n] = (E_{inst} * Hamming)[n] \tag{5.2}$$

Framification Afterwards the obtained signal is cut into frames of 250 ms duration with a overlap of 50 % according to the parameter of our synthesis process (see table 3.1).

Impact detection In order to detect impacts in the different frames, peaks are localized through the signal envelop.

$$PicLoc[k,i] = \{(Env[k,i] - Env[k,i-1]) > 0\}\&\{(Env[k,i+1] - Env[k,i]) < 0\} \tag{5.3}$$

where k the frame index, i the sample index of the associated frame with $2 \leq i \leq N$ and N the sample number of a frame.
Once peaks are localized a threshold is applied to keep only consistent ones. This threshold is equal to twice the mean of the associated frame. In order to make this selection more robust to noise another condition should be fullfilled. If a consistent peak is found, at least another peak should be present in the following 0.5 ms. Then to avoid multiple impact detections a retriggering delay, 10 ms, is observed before allowing any new impact detection, [OE04].

Impact rate Thanks to theprevious impact detection, one is able to count the number of impacts per frame and thus to derive a measure of the impact rate.

$$ImpactRate[k] = \frac{N_{impacts}[k] \cdot Fs}{N_{samples}[k]} \tag{5.4}$$

where $N_{impacts}[k]$ the number of retrieved impacts in the k^{th} frame, $N_{samples}[k]$ the number of samples in the k^{th} frame and Fs the sampling frequency of the signal.

5.1.2 Evaluation of impact rate

Creation of evaluation audio files

To evaluate the present method, a set of evaluation audio files is created. First four different single rain drops are extracted from real rain sound extracts (from *SoundIdeas* database). Then from those single rain drops, a matlab program has implemented to create audio files in which those drops are repeated at a certain frequency. This is the different used frequency values: 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 Hz. An additional audio file is created with a frequency repetition varying from 3 Hz to 80 Hz. Moreover in order to see the noise sensibility of the method, a gaussian noise is added to the obtained audio files. Its amplitude is set to have, for each obtained audio files, different signal-to-noise ratio: -10, -5, 0, 5 and 10 dB. In total, a set of 380 audio files is created.

		Signal-to-noise ratio [dB]				
		−10	−5	0	5	10
Frequency Repetition [Hz]	1	100	100	100	100	100
	2	100	100	100	100	100
	3	100	100	100	100	100
	4	100	100	100	100	100
	5	100	100	100	100	100
	7	100	100	100	100	100
	10	100	100	100	100	100
	15	99	100	100	100	100
	20	91	100	100	100	100
	30	34	100	100	100	100
	40	4	99	100	100	100
	50	0	85	100	100	100
	60	0	53	100	100	100
	70	0	17	99	100	100
	80	0	3	91	95	95
	90	0	1	86	100	100
	100	0	0	52	76	75
	110	0	0	26	63	67
	3 − 80	61	100	100	100	100

Table 5.1: F-measure of the presented impact rate measure, in [%], for each combination of frequency repetition and signal-to-noise ratio.

Evaluation

To evaluate the present method, one can derive the F-measure:

$$F_{measure} = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5.5}$$

where *precicion* is equal to $|\{trueImpact\} \cap \{retrievedImpact\}|/|retrievedImpact|$ and *recall* is equal to $|\{trueImpact\} \cap \{retrievedImpact\}|/|trueImpact|$. An interval error of ±2.5 ms is allowed while achieving the intersection between the true impacts and the retrieved ones. Table 5.1 shows the obtained results with this evalutaion technique.

Through the presented tables, one can see that the measure does not induce errors for the frequency rates from 1 Hz to 10 Hz. Afterwards from 15 to 90 Hz proper results are achieved for a signal-to-noise ratio is superior to 0 dB. Then for superior repetition frequencies the results get poor, this is due to the retriggering delay which is equal to 10 ms implying then a limit about 100 Hz regarding the possible detectable impact repetition frequency.

5.2 Conclusion

The presented impact rate descriptor gives satisfying results according to the performed F-measure. Despite the simplicity of this implementation, numerous parameters are involved in it and some of them induces limitations (for example the retriggering delay). Those parameters have been rawly adjusted by a heuristic approach. Then for continuing this audio descriptor investigation, one should find a way to reduce the number of parameters for computing this descriptor and to pertinently adjust the different parameters. Moreover the method to get this descriptor can be seen as an onset detection technique. This similarity between our descriptor and onset descriptor techniques shows that our descriptor expresses more a physical measure than a perceptual one. Then perceptual experiment should be carried out in order to evaluate this impact rate descriptor.

This descriptor can be interesting for characterizing other impact sound textures such as fire, electric crackle, coffee grinder...

Chapter 6

Conclusion

Through this project, a methodology to guide the construction of a mapping for controlling environmental sound textures (rain, wave, wind) through one-hand gesture has been developed.
Based on previous contributions, a perceptual experiment is carried out in order to find an efficient and perceptually relevant set of audio descriptors for controlling corpus-based concatenative synthesis process of environmental sound textures. From this experiment, the audio descriptor set composed of *loudness*, *spectral flatness* and *spectral centroid* are revealed as the most efficient and pertinent set of audio descriptors to use for controlling the environmental sound texture synthesis. Moreover through this perceptual experiment our overall synthesis process for environmental sound textures has been considered as satisfying regarding to the obtained results.
Afterwards an explorative experiment where participants are asked to perform one-hand gestures pretending that their motions would create the sounds they hear. Through this experiment a set of sound and gesture-related features are obtained. Then by applying CCA to those obtained data, the linear part of gesture-sound relationships can be revealed. From this analysis, first, we noticed that this gesture-sound experiment is not adapted to rain textures. For the two others studied sound textures, wave and wind, significant correlations are expressed. Those two studies show that there is one of the audio descriptors, *loudness*, which is mainly followed by the participants through their gesture. Nevertheless no common gesture-sound relationship over all participants and sound textures is highlighted. Then we decide to focuse this analysis technique on particular participant over sound extracts coming from a single sound texture. From this attempt correlations between gesture and sound are figure out. But those correlations are specific to participants. Thus no generalization about gesture-to-sound mappings over all participants and all sound textures can be done. However the presented methodology can be used for guiding the construction of personal gesture-to-sound mappings.

In future works, sets of perceptually relevant descriptors specific to the different categories of sound textures (periodic, stochastic or both) should be retrieved. This should allow more apropriated control of environmental sound texture synthesis. Moreover an extended and less-constraining gesture capture system should be provided to performers. A full 3D hand tracking system allowing to capture the motion of the user's two hands and also his/her finger's motion in order to propose several independent levels of control through gestures should be developed. In addition through our gesture-sound experiment, we noticed that participants were mainly and almost only following the *loudness* contour of the sound extracts. Thus for futur experiments, participants with a strong musical background should be chosen in order to make more than one sound descriptor to be followed by participants.

Bibliography

[AE03] Marios Athineos and Daniel P. W. Ellis. Sound texture modelling with linear prediction in both time and frequency domains. In *Proc. ICASSP*, pages 648–651, 2003.

[Alv05] Noë Alva. *Action in Perception.* Massachusetts Institute of Technology Press, Cambridge, USA, 2005.

[Ber97] Alain Berthoz. *Le Sens du Mouvement.* Odile Jacob, Paris, France, 1997.

[BJDEY$^+$99] Z. Bar-Joseph, S. Dubnov, R. El-Yaniv, D. Lischinski, and M. Werman. Statistical learning of granular synthesis parameters with applications for sound texture synthesis. In *Proceedings of the 1999 International Computer Music Conference (ICMC)*, pages 178–181, 1999.

[BMS05] Frédéric Bevilacqua, Remy Müller, and Norbert Schnell. Mnm: a max/msp mapping toolbox. In *New Interfaces for Musical Expression*, pages 85–88, Vancouver, Mai 2005.

[BWL08] Ross Bencina, Danielle Wilde, and Somaya Langley. Gesture ≈ sound experiments: Process and mappings. In *Proc. of the 2008 Int. Conf. on New Interfaces for Musical Expression (NIME08)*, 2008.

[Car04] M. Cardle. Automated sound editing. Technical report, Computer Laboratory, University of Cambridge, UK, 2004.

[CBS10] Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell. Towards a gesture-sound cross-modal analysis. In *Gesture in Embodied Communication and Human-Computer Interaction: LNAI 5934*, pages 158–170. Springer Verlag, 2010.

[Coo07] P. Cook. Din of an iquity: Analysisand synthesis of environmental sounds. In *Proceedings of the International Conference on Auditory Display (ICAD2007)*, pages 167–172, 2007.

[DBJEY$^+$02] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. Synthesizing sound textures through wavelet tree learning. *IEEE Computer Graphics and Applications*, 22(4):38–48, 2002.

[FA05] J. Filatriau and D. Arfib. Instrumental gestures and sonic textures. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, 2005.

[FAC06] J. Filatriau, D. Arfib, and J. Couturier. Using visual textures for sonic textures production and control. In *Digital Audio Effects (DAFx)*, 2006.

[FH09] Martin Fröjd and Andrew Horner. Sound texture synthesis using an overlap-add/granular synthesis approach. *Journal of the Audio Engineering Society*, 57:29–37, 2009.

[Fin09] N. Finney. Autonomous generation of soundscapes using unstructured sound databases. Master's thesis, MTG, IUA-UPF, Barcelona, Spain, 2009.

[Gre77] J M Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[Gui08] Sandro Guidati. Auralisation and psychoacoustic evaluation of traffic noise scenarios. *Acoustical Society of America Journal*, 123:3027, 2008.

[Hos02] R. Hoskinson. *Manipulation and Resynthesis of Environmental Sounds with Natural Wavelet Grains.* PhD thesis, The University of British Columbia, 2002.

[Hot36] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[HP01] R. Hoskinson and D. Pai. Manipulation and resynthesis with natural grains. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 338 – 341, Havana, Cuba, 2001.

[HSST04] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[KKA$^+$02] Evelyne Kohler, Christian Keysers, M. Aless, Ra Umiltà, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, pages 846–848, 2002.

[KP10] Stefan Kersten and Hendrik Purwins. Sound texture synthesis with hidden markov tree models in the wavelet domain. In *Sound and Music Computing Conference*, Barcelona, Spain, 2010.

[LGDM08] Mathieu Lagrange, Bruno L Giordano, Philippe Depalle, and Stephen McAdams. Objective quality measurement of the excitation of impact sounds in a source/filter model. *Acoustical Society of America Journal*, 123(5):3746, 2008.

[LLZ04] Lie Lu, Wenyin Liu, and Hong-Jiang Zhang. Audio textures: theory and applications. *IEEE Transactions on Speech and Audio Processing*, 12:156–167, 2004.

[MDOS09] J.H. Mc Dermott, A.J. Oxenham, and E. Simoncelli. Sound texture synthesis via filter statistics. In *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk NY*, 2009.

[MG03] Thomas Metzinger and Vittorio Gallese. The emergence of a shared action ontology: building blocks for a theory. *Consciousness and Cognition*, 12(4):549–571, 2003.

[MLS$^+$08] Emma Murphy, M. Lagrange, Gary Scavone, Philippe Depalle, and Catherine Guastavino. Perceptual evaluation of a real-time synthesis technique for rolling sounds. In *Proceedings of Enactive '08*, Pisa, Italy, 2008.

[MMS$^+$10] Nicolas Misdariis, Antoine Minard, Patrick Susini, Guillaume Lemaitre, Stephen McAdams, and Etienne Parizet. Environmental sound perception: Metadescription and modeling based on independent primary studies. *Eurasip Journal on Audio, Speech, and Music Processing*, 2010:1–27, 2010.

[MWD$^+$95] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

[OE04] M. Sile O'Modhrain and Georg Essl. Pebblebox and crumblebag: Tactile interfaces for granular synthesis. In *NIME*, pages 74–79, 2004.

[OSG02] James F. O'Brien, Chen Shen, and Christine M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *The ACM SIGGRAPH 2002 Symposium on Computer Animation*, pages 175–181. ACM Press, 2002.

[PB04] J. Parker and B. Behm. Creating audio textures by exmple: tiling and stitching. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 4, 2004.

[PTF09] Cécile Picard, Nicolas Tsingos, and François Faure. Retargetting example sounds to interactive physics-driven animations. In *AES 35th International Conference on Audio for Games*, 2009.

[SA95] Nicolas Saint-Arnaud. Classification of sound textures. Master's thesis, Massachusetts Institute of Technology, 1995.

[SaP95] Nicolas Saint-arnaud and Kris Popat. Analysis and synthesis of sound textures. In *Readings in Computational Auditory Scene Analysis*, pages 125–131, 1995.

[SBVB06] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, and Sam Britton. Real-time corpus-based concatenative synthesis with catart. In *9th International Conference on Digital Audio Effects (DAFx)*, pages 279–282, Montreal, Canada, Septembre 2006.

[SCB08] Diemo Schwarz, Roland Cahen, and Sam Britton. Principles and applications of interactive corpus-based concatenative synthesis. In *Journées d'Informatique Musicale (JIM)*, Albi, France, Mars 2008.

[Sch06] Diemo Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35-1:3–22, 2006.

[Sch07] Diemo Schwarz. Corpus-based concatenative synthesis : Assembling sounds by content-based selection of units from large sound databases. *IEEE Signal Processing*, 24-2:92–104, 2007.

[Sch11] Diemo Schwarz. State of the Art in Sound Texture Synthesis. In *DAFx*, Paris, France, 2011.

[Sci99] Agostino Di Scipio. Synthesis of environmental sound textures by iterated nonlinear functions. In *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, pages 109–117, 1999.

[sit11] Topophonie's presentation. http://www.topophonie.fr/apropos, 2010-2011.

[SMW$^+$04] P. Susini, S. McAdams, S. Winsberg, I. Perry, S. Vieillard, and X. Rodet. Characterising the sound quality of air-conditiong noise. *Applied Acoustics*, 2004.

[SS10] Diemo Schwarz and Norbert Schnell. Descriptor-based sound texture sampling. In *Sound and Music Computing (SMC)*, pages 510–515, Barcelona, Spain, Juillet 2010.

[Str07] Gerda Strobl. Parametric sound texture generator. Master's thesis, Universität für Musik und darstellende Kunst, Graz; Technische Universität Graz, 2007.

[Ver10] Charles Verron. *Synthèse immersive de sons d'environnement.* PhD thesis, Université Aix-Marseille I, 2010.

[VTR91] Francisco Varela, Evan Thompson, and Eleanor Rosch. *The embodied mind: cognitive science and human experience.* Massachusetts Institute of Technology Press, Cambridge, USA, 1991.

[Wan02] Marcelo Wanderley. Mapping strategies in real-time computer music. *Organised sound*, 7(2), 2002.

[ZW04] X. Zhu and L. Wyse. Sound texture modeling and time-frequency lpc. *Digital Audio Effects (DAFx)*, 4, 2004.

Appendix A

Perceptually relevant audio descriptors for environmental sound

Loudness

$$L[m] = 10 \cdot \log_{10}(\frac{\sum_{n=1}^{N} x[m,n]^2}{N}) \tag{A.1}$$

where x the waveform of the signal, m the frame index, n the relative sample index and N the number of samples in the current frame.

Spectral centroïd

$$SC[n] = \sum_{k=k_1}^{k_2} \frac{f_k \cdot |X[k,n]|}{\sum_{k=k_1}^{k_2} |X[k,n]|} \tag{A.2}$$

where n the current frame index, k the frequency bin index, $X[k,n]$ the complex signal magnitude of the frequency bin k for the current frame and f_k the frequency value of the k^{th} frequency bin. In `CataRT`, the default values for k_1 and k_2 are respectively corresponding to the frequency values 44 and 3014 Hz.

Mid- and High-frequency energy

$$E[n, k_1 \to k_2] = \frac{\sum_{k=k_1}^{k_2} X[k,n]^2}{N_{fft}} \tag{A.3}$$

where n the current frame index, k the frequency bin index, $X[k,n]$ the complex signal magnitude of the frequency bin k for the current frame and N_{fft} the number of frequency bins. In `CataRT`, the default values for k_1 and k_2 are respectively corresponding to the frequency values:

- 44 and 1033 Hz for the mid-frequency energy,
- 4996 and 10034 Hz for the high-frequency energy.

Spectral flatness

$$SF[n] = \frac{(\prod_{k=k_1}^{k_2} |X[n,k]|)^{1/(k_2-k_1+1)}}{\frac{1}{k_2-k_1+1}\sum_{k=k_1}^{k_2} |X[n,k]|} \tag{A.4}$$

where n the frame index, k the frequency bin index, $X[k,n]$ the complex signal magnitude of the frequency bin k for the n^{th} frame and f_k the frequency value of the k^{th} frequency bin. In `CataRT`, the default values for k_1 and k_2 are respectively corresponding to the frequency values 1507 and 8613 Hz.

Appendix B

Details on participants

B.1 For the similarity perceptual study on environmental texture morpholigy resynthesis

test person index	**gender**	**age**	**expert in sound domain**	**musician**	**experiment duration**
0	m	22	yes	yes	41 min
1	m	25	yes	yes	16 min
2	m	26	yes	yes	23 min
3	f	30	no	no	44 min
4	m	27	yes	no	39 min
5	f	25	yes	no	35 min
6	m	22	yes	yes	32 min
7	m	25	yes	yes	44 min
8	m	25	yes	yes	20 min
9	m	22	yes	yes	36 min
10	m	-	yes	yes	29 min
11	m	22	yes	yes	35 min
12	m	31	yes	no	33 min
13	f	27	yes	no	18 min
14	m	27	no	no	31 min

Table B.1: Details on the test persons which participates in the similarity perceptual study on environmental texture morpholigy resynthesis.

B.2 For the gesture-sound relationship study

test person index	**gender**	**age**	**expert in sound domain**	**musician**	**dancer**
0	m	22	yes	yes	no
1	m	22	yes	yes	no
2	f	27	yes	no	yes
3	m	27	yes	no	no
4	m	25	yes	yes	no
5	f	25	yes	no	no
6	m	25	yes	yes	no
7	m	-	yes	yes	no

Table B.2: Details on the test persons which participates in the gesture-sound relationship study.

Appendix C

Two sample t-test

Null hypothesis

In our case the *null hypothesis* is that data in two independent vectors, x and y, of random samples from normal distributions with equal means and equal but unknown variances regarding a certain significance level (5% in our case), against the alternative that the means are not equal.

T-statistic

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{C.1}$$

$$s = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \tag{C.2}$$

where $\bar{x}$ and $\bar{y}$ the respective sample means, n_y and n_x the respective sample numbers, s_x and s_y the respective sample standard deviations of the two sample vectors x and y.

Criterion

From table of critical value of t-distribution, the critical value for our test is 1.0 according to the degree of freedom (i.e. $(n_x + n_y - 2)$, 448 in our case) and the significance level,5 %.

If t is inferior to the critical value, the null-hypothesis is accepted: the means can be considered as equal. Otherwise, the null-hypothesis is rejected: the means are not equal.

Appendix D

More CCA results

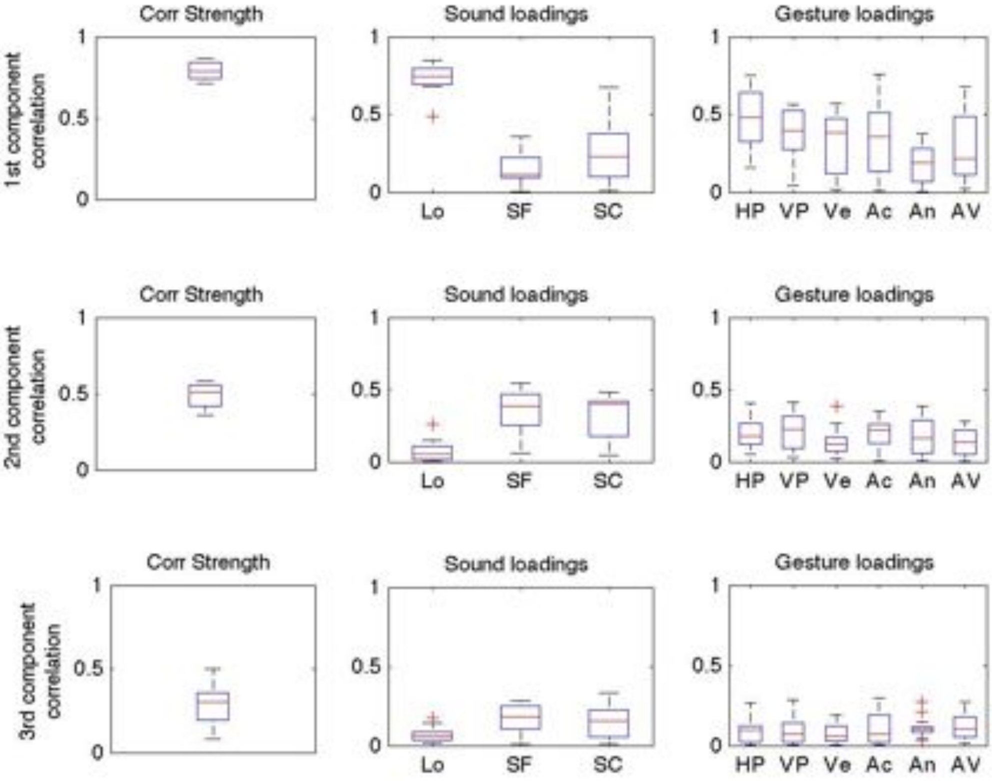


Figure D.1: Representation of the canonical correlation strength and the canonical loadings for all relevant performances of a single wind sound texture extract over all participants (16 performances). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid*. HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity*.

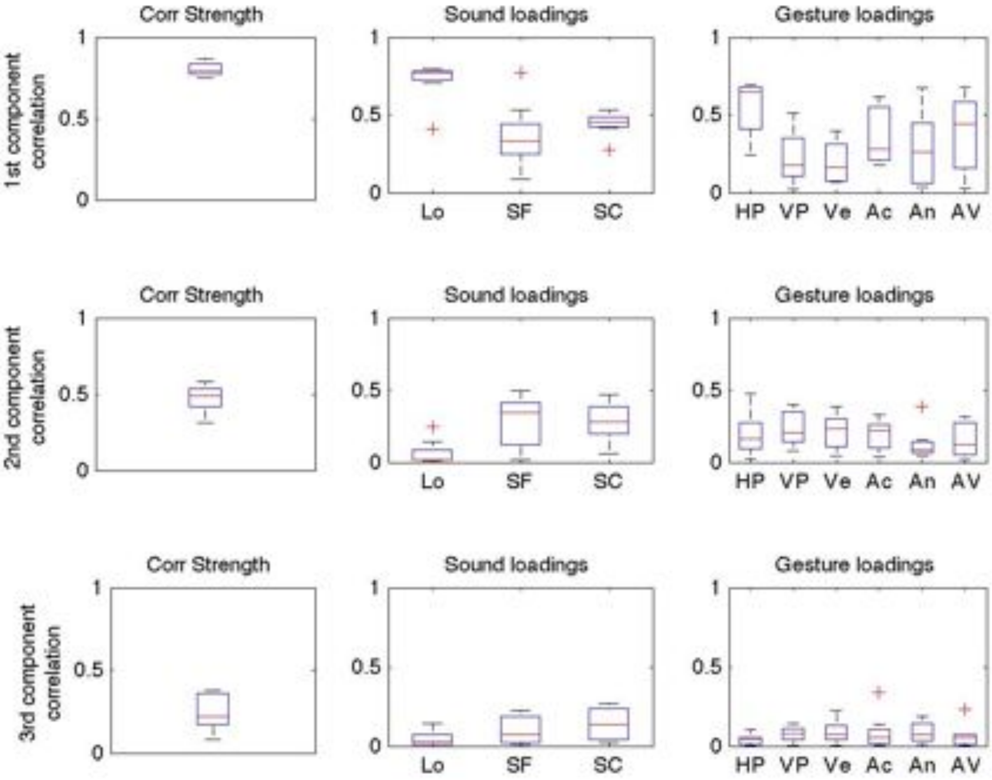


Figure D.2: Representation of the canonical correlation strength and the canonical loadings over all relevant performances (8 performances) on the different sound extracts of wind texture for one participant (subject 2). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid.* HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity.*

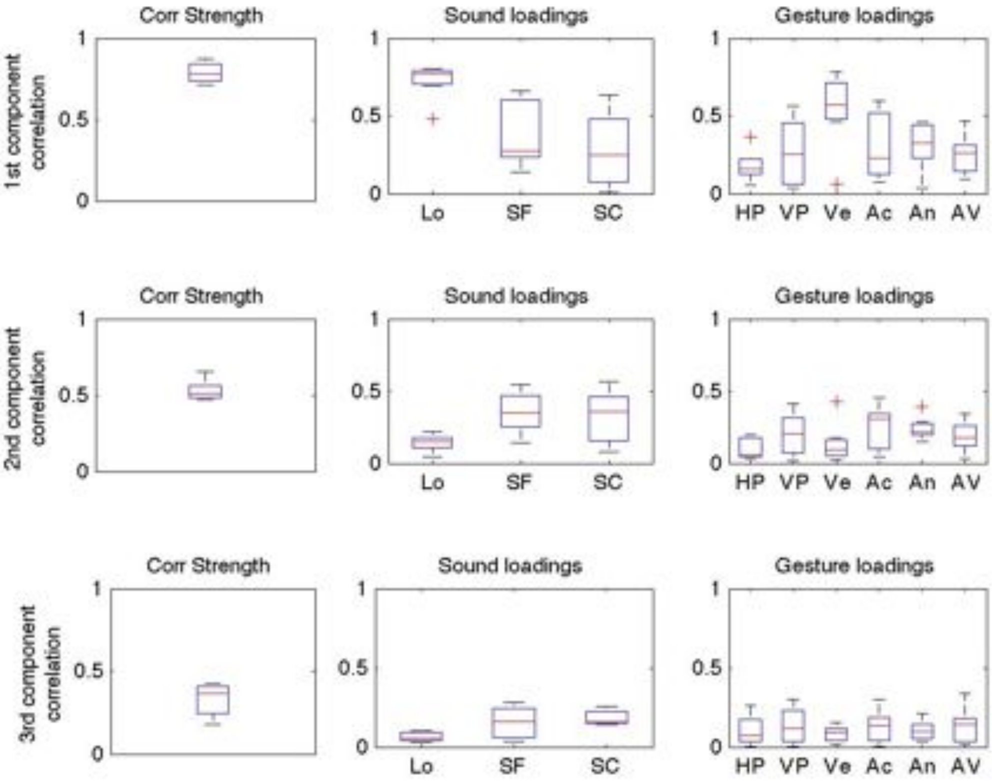


Figure D.3: Representation of the canonical correlation strength and the canonical loadings over all relevant performances (7 performances) on the different sound extracts of wind texture for one participant (subject 6). Lo: *loudness*, SF: *spectral flatness*, SC: *spectral centroid.* HP: *horizontal position*, VP: *vertical position*, Ve: *absolute velocity*, Ac: *absolute acceleration*, An: *absolute angle*, AV: *absolute angular velocity.*