Soundpainting language recognition

EPFL DH Master thesis

By  
Arthur Parmentier

# Introduction

To be written last.

# Table des matières

[I. Introduction 2](#_Toc40431421)

[II. Table des matières 2](#_Toc40431422)

[III. A brief history of Soundpainting 4](#_Toc40431423)

[A. Back in Woodstock 1974: emergence in emergency 4](#_Toc40431424)

[B. Developments 4](#_Toc40431425)

[1. A multidisciplinary language 4](#_Toc40431430)

[2. Fertility in Europe, worldwide spread in modern societies 4](#_Toc40431431)

[IV. Historical and theoretical context 5](#_Toc40431432)

[A. Signs for communication: a long history 5](#_Toc40431433)

[1. A very long time ago… 5](#_Toc40431434)

[2. Middle ages: neumes 5](#_Toc40431435)

[3. Creation of modern sign languages for deafs (?) 5](#_Toc40431437)

[4. Conduction & other contemporary forms of artistic sign languages 5](#_Toc40431439)

[B. Real-time composition/improvisation/generative music in the XXth century 5](#_Toc40431440)

[1. Cage & co 5](#_Toc40431441)

[2. Algorithmic music 5](#_Toc40431442)

[C. Linguistics 5](#_Toc40431443)

[Rousseau, Wittgenstein.. theory of signs: De Saussure (sign theory) 5](#_Toc40431444)

[1. Regular languages 5](#_Toc40431445)

[2. … link with generative music? 5](#_Toc40431446)

[V. (abstract) Mechanisms and concepts of Soundpainting (a theoretical model of SP) 6](#_Toc40431447)

[A. Preliminary remarks 6](#_Toc40431448)

[1. Context and scope of my personal observations 6](#_Toc40431449)

[1. The categorical and prototypical perception of concepts 6](#_Toc40431450)

[A. Mechanism 1: transformations (projections?) of concepts into signs on the physical space of the body 7](#_Toc40431451)

[1. Input space 7](#_Toc40431452)

[2. Transformations and output spaces 7](#_Toc40431453)

[Mechanism 2: 8](#_Toc40431454)

[B. Sign “overloading” 8](#_Toc40431455)

[1. Motivation 8](#_Toc40431456)

[2. Differentiation of signified across disciplines 8](#_Toc40431457)

[3. The difference between discipline (music, theater..) and technical apparatus, or “why we sometimes need to specify how to play a long tone, and sometimes not” 10](#_Toc40431458)

[4. Speculative theory of differentiation of signified across instruments (technical apparatus) of the same discipline 10](#_Toc40431459)

[5. Sign overloading 11](#_Toc40431460)

[C. Personal observations during SP practice 12](#_Toc40431461)

[D. Structural concepts (the grammar of the language, that results in its language) 14](#_Toc40431462)

[1. Identifiers 15](#_Toc40431463)

[2. Content 15](#_Toc40431464)

[3. Modifiers (content parameters) 15](#_Toc40431465)

[4. Mode 15](#_Toc40431466)

[Influence of context (from grammar to performance) 16](#_Toc40431467)

[VI. 16](#_Toc40431468)

[VII. Soundpainting recognition with Max/MSP 16](#_Toc40431470)

[A. A new configuration: motivations, goals, workflow & challenges 16](#_Toc40431471)

[B. The big picture: general description of the system 16](#_Toc40431472)

[C. Part 1: Posenet & gloves feature building 16](#_Toc40431473)

[D. Part 2: Training & data management 16](#_Toc40431474)

[E. Part 3: Classification 16](#_Toc40431475)

[F. Part 4: Grammar parsing 16](#_Toc40431476)

[G. Part 5: Orchestra simulation 16](#_Toc40431477)

[VIII. Learning in SP and numerical tool 16](#_Toc40431478)

[IX. Potential & future of the tool 16](#_Toc40431479)

[A. Topic A: what could be improved and how 16](#_Toc40431480)

[B. Topic B: … 16](#_Toc40431481)

[C. The future 16](#_Toc40431482)

[X. Conclusion 16](#_Toc40431483)

# A brief history of Soundpainting

## Back in Woodstock 1974: emergence in emergency

Emergence of the first SP signs with the orchestra. Emergent in emergency (*Emergent* means beginning to arise and *emergency* means arising unexpectedly)!

<http://www.soundpainting.com/history/>

Origin of “SP” as a term

## Developments

### A multidisciplinary language

SP not limited to music performance. Signs started to be created by W himself from 90+ for multidisciplinary performance.

### Fertility in Europe, worldwide spread in modern societies

Walter gave his first SP workshop in Europe in the late 90’s and found a very fertile ground for Soundpainting in France, which now has probably the largest SP community over the world.

Today, Soundpainting is used for on every continent and several artists created their own signs for the needs of their own group or configuration.

Important point: W wants to make sure that the language stays normalized and universal. Think tanks, glossary/dictionary of signs, community relations…

Modern societies: to what extent can the SP concepts of performance, content, parameters, identifiers… apply to non-modern groups?

Transition: culture to concepts

Write more about the evolution, in order to introduce the concepts later. It should already give an overall broad picture of what SP looks like at this point

Hypothesis: historical link with neums (dev in fr)

# Historical and theoretical context

## Signs for communication: a long history

Chronological view (?)

### A very long time ago…

### Middle ages: neumes

### Creation of modern sign languages for deafs (?)

Lien avec Chant-signes IVT

### Conduction & other contemporary forms of artistic sign languages

## Real-time composition/improvisation/generative music in the XXth century

How to structure this part? By example? Topic?

Musique stochastique xenakis

### Cage & co

### Algorithmic music

## Linguistics

### Rousseau, Wittgenstein.. theory of signs: De Saussure (sign theory)

Emetteur/recepteur;;.

### Regular languages

### … link with generative music?

Check relevancy: Chomsky, hierarchical models, markov models, study of music within linguistic models.

# (abstract) Mechanisms and concepts of Soundpainting (a theoretical model of SP)

Now that we have seen a bit of the development of SP and the theoretical background it lies in, I would like to propose a model of Soundpainting that would explicit its construction mechanisms as a sign language, as well as the implicit operations that makes it an efficient language for art performance.

## Preliminary remarks

### Context and scope of my personal observations

The reader must be aware that my experience with Soundpainting and the observations that I consider in this section:

* Are rather limited in time (3 years span)
* Are very limited in terms of cultural diversity (most of my participating experience was in Lausanne, Switzerland or nearby (France) with performers I was familiar with, who mostly came from European music education institutions; at the exception of one experience in Rio de Janeiro, Brasil where I could both participate and observe SP practices)
* Are rather limited in terms of configuration (mostly groups of musicians, either in weekly sessions with 3-10 performers or workshops with guest soundpainter).

The reader is invited to compare my observation with his and criticize the models and interpretation I give in this section.

### The categorical and prototypical perception of concepts

The human (innate) categorical perception scheme has been studied in music REF and many other fields and plays a very important role in the construction of basic artistic concepts such as note, pitch, scale, line, hit… by constructing discrete categories out of a continuous set of elements.

We also know from research in psychology[[1]](#footnote-1) that a single concept can be modeled as a category of elements around a prototype, considered as the central point of the category. Moreover, people tend to define the concept itself by the characteristic traits of the prototype, whereas in general, it extends beyond such a definition.  
The prototypical scheme rejects the discrete notion of ‘limit’ or ‘border’, replaced by the continuous notions of graded membership (similarity to the prototype) and the fuzzy edges of concepts.[[2]](#footnote-2) But on top of the prototypical scheme, the categorical scheme introduces a rupture by either accepting or rejection an element inside the category based on its similarity with the prototype[[3]](#footnote-3).

*Let’s take the examples of birds… (does it needs an illustration?)*

## Mechanism 1: transformations (projections?) of concepts into signs on the physical space of the body

Key idea: forming a sign/gestural language means creating a mapping between concepts from different sources (fields?) and the physical space of the human person who signs. How is this done in SP?

OVERALL TRANSFORMATION SCHEME (illustrate the scheme of the transformation process: source concepts to signs that represents them, gestures that evoke them…)

### Input space

We can identify several repertoires (sources) of concepts in the input space of this transformation scheme

#### Concepts from artistic disciplines

Long Tone, minimalism…

#### Concepts from oral languages

Logical elements

#### “High/low” cultural representations of quiet/loud (volume), slow/fast (tempo), pitch…

Even though the concepts of volume, tempo or pitch may be universal, their mapping onto a low/high axis is defined culturally in modern societies.

#### …

### Transformations and output spaces

Sign = (symbol, icon, indice).

#### Creation of a symbol

… Describe here more about the process of creating a sign (examples, general rules…)

#### Creation of a icon

#### Creation of an indices

## Mechanism 2: Sign “overloading”

### Motivation

We have seen from the evolution of SP that a single sign can be use to signify a content[[4]](#footnote-4) not only for different instruments of the same discipline (1) but also across discipline (2). This what I call the “overloading” of a sign.[[5]](#footnote-5)

We will try to identify whether the operability of a sign in (1) and (2) involves mechanisms of different nature.

### Differentiation of signified across disciplines

We very commonly observe signs made of one signifier and several signified in oral languages and our everyday life. The signified is understood by an operation of “disambiguation” that depends on the context of the communication. In this section, we will see that a multi-disciplinary sign in SP can indeed have several signified that allow it to operate in several disciplines.[[6]](#footnote-6) Moreover, we will discuss the model of a bijection between signified concepts and disciplines., Finally, we will discuss the construction process of these multi-disciplinary sign to show that their signified are linked by analogies.

#### Existence of several signified concepts

To demonstrate the existence of several concepts (the signified) under a multi-disciplinary sign in SP, let’s take the common example of the LT, that we will use all over this section to demonstrate some of the mechanisms of SP.

For a musician, the LT is a concept preexisting to SP with a specific prototype, whose characteristic traits are “constant volume”, “constant pitch” (among others). But is the concept of the LT for a musician the same as the one for a dancer or a visual artist? To answer this question, let’s first remark that soundpainters often explain how to perform a LT differently for each discipline:

* “A fluid movement, without accent” for dancers
* “A freeze on the first syllable of a word” for actors
* “A note with constant volume and constant pitch over time” for most musicians
* …

They also often illustrate those descriptions with a prototypical example.

By looking at the description themselves, we can see that they involve different concepts: a “movement”, “roll” or a “syllable”, which cannot be considered equivalent. Moreover, we know from the history of SP that the concept of a LT was first borrowed from music and “extended” to other disciplines, i.e. that the multiplicity of signified of the sign “LT” is a voluntary construction[[7]](#footnote-7).

Although we illustrated the overloading mechanism with the sign “LT”, we can observe the same mechanism for other signs, for instance multi-disciplinary signs.

#### Bijection, surjection or injection between signified and disciplines?

Although we have shown the existence of several signified (several concepts) and their relationship with disciplines, one can wonder whether their relationship is a bijection (to a unique discipline corresponds a unique signified), surjection (each discipline has one or more signified) or injection (each discipline has either a unique signified, either no signified).

First, we can remark that there is a bijection in the case of the so called “multi-disciplinary signs” whose properties is exactly that each discipline has its own unique interpretation (signified) of the sign.

Then, we can also observe that some signs are very specific to a discipline (sometimes by construction), such that other discipline cannot interpret them. It is enough to conclude that in general, there is no surjection between signified and the set of disciplines.

Finally, on the question of injectivity, my hypothesis is that unlike oral languages, SP can be modeled as a regular language, hence a context-free language, such that the context (signs prior to the last executed sign) does not influence the meaning of the considered sign. We will see in a future section how this approach is successful in modeling the very basics of SP, while it is unclear whether all SP modes, rules and signs could indeed be represented by a regular language. In my hypothesis, there is indeed an injection between signified and disciplines, but I leave to experts in the domain the prospection of counter examples, such as context-sensitive examples that would demonstrate that a single sign can indeed have several signified for one discipline, that can be differentiated by the context instead.

### The difference between discipline (music, theater..) and technical apparatus, or “why we sometimes need to specify how to play a long tone, and sometimes not”

In the “Motivation” part, I differentiated the notion of “discipline” and “instrument”, that we could also call a “technical apparatus”. Indeed, one could wonder why the concept of a LT would be different in each discipline, but still the same for all instruments of the discipline.

In this part, I wil try to show you that the operability of a sign across disciplines is constructed “by SP” with learned (cultural) relations of analogy between signified whereas the operability of a sign across instruments of the same discipline is in general a construction external to SP that involves the human perceptual scheme.

DIFFERENCE D APPAREIL TECHNIQUE MAIS MEME PROROTYPE VS DIFF DE DISCIPLINE DONC DE PROTOTYPE (MEME SI L APPAREIL TECHNIQUE EST LE MEME).

### Speculative theory of differentiation of signified across instruments (technical apparatus) of the same discipline

Wrap-up: what is a long tone?

By taking the simple example of the LT, we will see how the categorical and prototypical perception scheme plays an important role in the mechanisms of SP.

What is a long tone?   
It would be tempting to define the long tone for each discipline by giving a set of characteristics that all long tone must have, for instance:

* In music, a constant pitch and volume over time
* In dance, a movement without accent…

We know however from the research on our perception scheme that the LT is not (and perhaps, cannot) be defined with such characteristics or traits. It shows us that the LT is not a musical element defined by a finite number of properties but is rather a category constructed by the human perception scheme[[8]](#footnote-10) around a “prototypical” element which exhibits the features that are usually said to define the concepts of a LT.  
In this framework, a content is perceived as “more” or “less” a LT, rather than either a LT or not.

### Sign overloading

I will try to show that that in general, the grouping relations linking those concepts together under a single sign are made possible by cultural schemes (analogies) and innate perception schemes of human mind. I will use the example of the sign LT to illustrate how these schemes allow for translating the sign LT into the concepts that are relevant for each discipline, instrument (or technical apparatus).

#### Innate perception schemes

##### How to perform a LT on percussions?

I often see questions arising about how to perform a sign for a specific instrument or discipline, for instance a percussionist asking how he should perform a long tone with drums.  
While the experienced performer will probably use different possible techniques intuitively (a fast roll, using brushes, playing on cymbals that have a long acoustic response), we can derive from this simple observation that the concept of the long tone doesn’t necessarily have a trivial interpretation and its realization may not be accessible for instruments like percussions.  
Let’s break down the conceptual operations that allow the percussionist to respond to a LT when its prototype is out the set of possibilities offered by his instrument.  
To achieve this, the percussionist would typically translate the concept of LT to another concept, the roll, easily accessible to the performer. The roll will allow him to approach the prototype of the LT by increasing the speed of the roll as much as he can.  
Another way to phrase this conceptually is to say that in the space of musical concepts, the prototype of the LT is the asymptotical, limit point of the concept of roll when its speed goes to infinity.

##### An implicit translation by our perception scheme

The important point illustrated here is that SP involves operations of translation from concepts whose prototype[[9]](#footnote-11) does not exist in a discipline to a concept of the discipline that relates to the prototype. In previous example, we saw that the percussionist could translate the sign LT to the concept of roll that is relevant to its instrument. But what is the implicit scheme allowing for this translation?

I interpreted the fast roll as an “approach” to the prototype of a LT in the space of musical concepts, suggesting that there is a metric that allow us to measure the distance between the roll at a certain speed, volume… and the prototype of a LT. Even though this topic is out of my field of expertise and of the scope of this text, we can remark that the acoustic response of a fast roll “approaches” the performer’s expectation of the acoustic response of a LT, such that the human perception scheme will associate the fast roll with the prototype of the LT.

##### Learning the translation

While the operations of translation are in general not obvious (hence the questions on the subject), I observed that most Soundpainters explicit what operations are valid to beginners in SP.  
  
IMITATION et mimetisme

SEul vs a plusieurs

#### Cultural analogies

## Personal observations during SP practice

My theory is that our categorical and prototypical perception scheme plays an important role on both how performers and soundpainters can interpret the artistic material that is being produced by the group, and how they will respond to the requests, i.e. what content the performer will produce.

I will try to support this theory by analyzing examples from my own experience as a performer, a soundpainter and discussions I had with other performers on the topic of the LT which appears as an easy example to observe some of the underlying schemes of perception and associations of musical concepts.

All 5 experienced soundpainters I have worked with teach the concept of a long tone by giving both characteristic traits and prototypical examples of a LT, showing “how it’s done” in each discipline. Their approach is usually to start defining orally the most characteristically trait of the concept, for instance “a fluid movement without accent” for dancers and then give illustrations using their body, with different speeds for each example. For musician, they would for instance mention that a LT is a sound with constant frequency as first definition of the concept and give examples by singing the prototype of a LT at different frequencies.

* From the examples given by the soundpainter, unexperienced performers are usually able to internalize the prototype of the concept of a LT for their discipline and produce examples of their own at different frequencies/speeds. I observed that the examples produced by unexperienced performers at first are very often close to the prototype.  
  But as performers become more experienced, I observed that they tend to increase the span of produced examples not only by broadening the distribution to the parameters that have been introduced to (volume, frequency, timbre…) but also by exploring different “dimensions” of the sound, such as vibrato, micro-distortions, extended techniques etc. In other words, they progressively “detach” from the prototypes of the LT concepts[[10]](#footnote-12) and their characteristic features by exploring more features of the content and reaching more extremal points. In other words, if we take the N-dimension space (one dimension per parameter of the LT) of all LTs accessible to a performer, whose boundaries are determined by the technical and imaginative limitations of the performer himself, the distribution of the production of a learning performer should first span a limited volume around the centroid of the space that can be considered the prototype of the concept of a LT. Then, the volume covered by the distribution and its variance would increase with experience and artistic research.  
  This expansion process was discussed especially in my experience with a year-lasting Soundpainting group, in which we had sessions dedicated to explore new generative processes and dimensions for LT and other very prototyped concepts. From the discussion, it was clear that our production of LT was largely prototypical and that extending the range of production required dedicated work and one explanation that I remember was that it takes a lot more cognitive load to produce a LT far from the prototype than a LT close to it. Therefore, under the constraints of immediate play, it was hard to propose something original. I conceived this training as a way to reduce the cognitive cost of the production of less-prototypical LT, therefore bringing more diversity to the responses of the performers.  
  During one week or shorter workshops that mixed both beginners and experienced performers in Soundpainting, I observed that experienced performers were responding to requests with a wider variety than beginners in Soundpainting as one could expect from previous observations, but also that performers with a greater technical level would also respond with a greater variety. My interpretation of this observation is that
  1. the cognitive load of the production of a content depends not only on the experience of the performer with SP, but also with his discipline
  2. the cognitive load is a key metric for understanding how far from a prototype a given performer can respond to a SP request

From my experience in Brazil, I can add to the latter the following remark:

* 1. the desire or willingness of a performer to respond to the request in a certain way is very cultural. In the previous discussion, I have interpreted the expansion of the variety of responses as a consequence of the decrease of its cognitive load for the performer, but it is important to remark that this expansion may not be observed at all if the performer himself is satisfied by a certain type of response, should it by prototypical or not. In fact, this expansion relies on the motivation of/relevancy for the performer to vary its responses for artistic reasons that depend on the context of the performance and the background of the performer, including his cultural background.[[11]](#footnote-13)
* Another consequence of the categorical nature of the concepts beneath SP signs is the inexistence of a clear frontier between the concepts themselves. For instance, one could argue that silence can be considered as an extremely low volume long tone, and purposefully respond to a request of LT with silence. My observation is that during learning phases, Soundpainters prefer that beginners show that they have understood the concepts by responding with prototypical examples instead of “extreme” examples.

*Goal: describe the concepts that SP deals with and propose a categorization of these concepts based on an interpretation of the mechanisms of SP (SP as a set of operations – transformations).*

## Structural concepts (the grammar of the language, that results in its language)

Those concepts are used to describe the structure of the language: its grammar. They are emergent from SP, in the sense that they have been formalized dozens of years after the creation of SP and refer to its particular grammar rather than objects that are also found in other contexts. I would call the latter ones “borrowed” concepts.

One should really think of these concepts as an equivalent of the “noun”, “verb”, “adjective” … that are used to describe oral languages that modern societies are familiar with in their oral languages.

It is important to note that sign languages (among which, SP) do **not** use the same concepts as modern oral languages to describe their grammar.

### Identifiers

### Content

### Modifiers (content parameters)

### Mode

All these structural concepts (grammar) are not culturally dependent.

DRAFT of ideas

* “Traditional” Soundpainting vs “Alternative” Soundpainting: a question a configuration (soundpainting as a language but not a configuration) means that it is not necessarily linked to the frontal relation between a composer (human) and an orchestra (of human performers). My use of the word “traditional” refers to the transmission and definition of SP by Walter, with his own use & configuration of Soundpainting performances. There is no “modern” Soundpainting but rather an expanding field of research and experimentations in how the language and its concepts can extend to very different contexts and configurations. Develop with Brasilian version of SP & baila baila.
  + The concepts depend on the configuration. For instance, a “note” in music is very different from the concept (or its absence!) of “note” in visual arts.
  + SP early concepts were built for music performances and translated later for other disciplines. (Q: why do we even use this word “discipline” in arts?)  
    -> Concepts inside SP are “stolen” from the configuration/context it deals with.
  + SP acts as a kind of translation from the concepts of different practices to signs
  + It is therefore important to understand that to my point of view, there is not a fixed set of concepts that would be incorporated by the SP language
  + but rather a fluid, dynamic and varying grammar

# Influence of context (from grammar to performance)

# Soundpainting recognition with Max/MSP

In this section, I will introduce my Max/MSP Soundpainting recognition tool.

First, we will motivate the creation of such a tool, capable of recognizing several SP signs but also of creating new ones, as a composer would do.

Then, we will explore the structure of this tool in a top-bottom approach, in contrast with constructivist model of SP that I presented in previous sections, by discussing the global structure of the program and then taking a deeper look into the features and key objects of each layer.

Moreover, we will investigate basic performance mechanisms to optimize the speed and accuracy of the system.

Finally, we will discuss the use of this tool for learning SP in connection with the theoretical aspects discussed in the previous parts. We will look at some implications of the tool for the formalization of SP and hopefully, the future of the language.

## A new configuration: motivations, goals, workflow & challenges

### Motivations

We have seen previously that SP is used in different configurations and their role as a super-structure of SP for the choosing the grammars of SP (modes), defining the roles of emitter/receiver (soundpainter/performer) and also the “default parameters” that define context-depend rules in the performance.

The use of a specific configuration is motivated by the results it aims to achieve and the qualities of the processes that it involves (creative processes, communication…).

There are several motivations for building and using a digital SP recognition tool.

* First, such a tool can be used to simulate a whole SP orchestra as a learning tool for individuals aside the collective approaches to learning SP. As for now, the tool only covers the basic structures of SP and would only be interesting for beginners in SP, who typically cannot rehearse with a group.
* Then for advanced or expert soundpainters, such a tool is rather interesting as a unique member of an orchestra for exploring new areas of composition, in particular the fields offered by artificial intelligence, machine learning and their complex generative processes.
* Finally, the tool could also be used as a controller with other digital elements that are already part of the performance: effect processors, amplification mechanisms, mixing devices, recording devices and many more. It is sometimes the case that such elements cannot be controlled real-time performance because of the complexity of their interface or controls, or that it requires performers to manipulate them. The interfacing possibilities offered by Max/MSP makes it an ideal choice for controlling these devices within the SP language directly.

### Goals

In the frame of the master thesis, I have chosen to focus on the use case of my tool as a learning tool.

In analogy with the constructivist model we discussed previously, we know that the learning tool must have at least the following features:

* A predefined dictionary of signs (that the user will be able to learn)
* A mechanism of sign creation
* Data management (loading a set of signs and saving new ones)
* A real-time sign classifier
* One or several modes (grammatical systems that define the rules of the language/communication)
* Audio/visual feedback in response to the soundpainter’s requests

The goal of my tool will be to implement these features, focusing on only one mode (the default SP mode) and implementing an audio feedback by simulating a small orchestra.

## The big picture: general description of the system

The was we have described the necessary features and the constructivist approach to SP is a direct inspiration for structuring the tool in terms of several independent layers, that are represented in Figure 1.

Each layer is composed of asynchronous external processes/objects (rectangles with sharp corners inside a layer) and internal processes/functionalities of the program in Max/MSP (rectangles with rounded corners). My work was to program the “internal” processes and functionalities as well as the information pipeline (arrows), whereas the “external” objects were not created by myself.[[12]](#footnote-14)

In the following parts, we will discuss each functionality and discuss the implementation details.

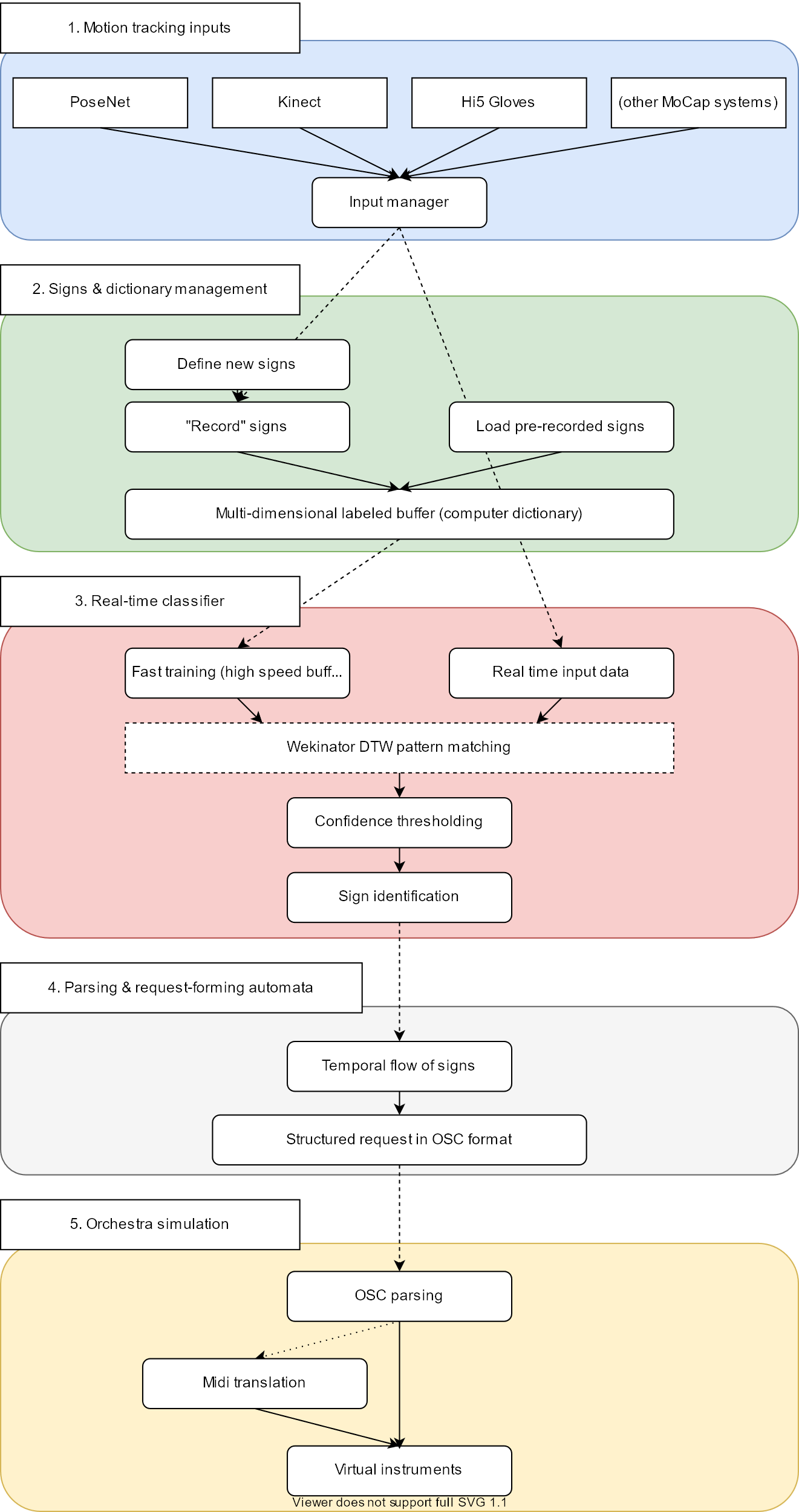


Figure 1 Summary of the recognition program structure

## Description of each layer

### Part 1: Motion tracking inputs

The role of the motion tracking layer is to compute a set of motion features from the movement of the user. There exist a variety of motion tracking systems with different technologies. In SP, there are some body parts such as the hands that are much more frequently to sign than others, therefore they require more precise tracking than the latter to classify amongst the signs.

My choice was to track the soundpainter motion at two different scales: full body and hands. This choice is motivated by my experience in SP and the set of technologies available that matches these different scales.

#### Full body tracking

##### Introduction to PoseNet

For the full body tracking, I chose to use primarly PoseNet[[13]](#footnote-15), a vision model that can be used to estimate the pose of a person in an image or video by estimating where key body joints are in 2D space. In my case, PoseNet uses the video input from the computer webcam.



Figure 2 Posenet skeleton tracking for "rest of the group"

PoseNet allows the user to choose different models and internal parameters that will affect its performance:

* The architecture of the model (MobileNet or ResNet)
* The input resolution of the video input
* The output stride of the model
* The depths of the convolution operations (for MobileNet only)
* The size of the model (ResNet only, only affects loading time)

With these settings, the user can adapt the model to its hardware to get the best performance.

##### PoseNet advantages

PoseNet has several advantages to its concurrent technologies:

* It takes it input from a webcam or any video input that can be recognized by the computer, so that
  + For many laptops with integrated webcam, there is no need of external hardware
  + It can be used with very common and cheap hardware in case the computer does not have its own webcam, making the costs typically very low
* It provides a direct feedback of its accuracy to the user by overlapping the skeleton joints with the video, allowing users to change settings according to how good they see the model performing
* It is an open-source project, led by giants (Google…) and supported by a vast community
* It is still under development and will probably continue to be improved over the years, so it has a much greater potential than hardware-dependent solutions that are getting obsolete very fast
* It integrates with Max (and other systems) very easily as it can be run in a little server

##### The main shortcoming of PoseNet: depth

The only major shortcoming of PoseNet with respect to other motion tracking systems is that it only operates in 2D and does not compute the depth of each joint.

As a workaround, I first built a simple calibration process which would allow to compute the depth of the torso of the user as well as its angle to the camera, before I realized that it was useless in my use case and would only bring noise in the data.

###### Depth in SP

My observation is that depth (z axis) is often not the most informative axis for recognizing SP gestures and even signs like “play” which uses the z dimension extensively can be recognized only by the movement of the body in 2D, from the point of view of the camera in front of the soundpainter.

However, capturing depth is important for one special sign in SP, which is often called “entering the box”. Its specialty is that this sign, consisting in putting one foot in an abstract “box” in front of the soundpainter, is used in many SP modes to significate “execute the request now”; whether the request has been defined previously (default mode) or is being signed while the soundpainter has “entered the box”. It is only meaningful when executed simultaneously to other signs, for instance “go gestures” in default mode or contents in launch mode. The opposite sign “exiting the box” also has the very specific meaning of “getting back to default mode”.

###### Recognition without depth: compromises and simplifications of SP grammar

Ideally, one would want to capture depth and abandon PoseNet for a better tracking method:

* Kinect systems can capture depth but have many other shortcomings in terms of performance, user experience and portability
* Motion capture suits are usually the most accurate and performant devices but their costs and and specificities make them unattractive for sharing the tool to the SP community
* OpenPose[[14]](#footnote-16) is the main realistic alternative to PoseNet at this moment; it supports 3D triangulation from multiple view (like two cameras orthogonal to the soundpainter) but could not be ported to Max without much hassle[[15]](#footnote-17)
* Some learnable triangulation methods are being developed recently[[16]](#footnote-18) yet far from portable to my project

PoseNet appears like the best compromise for this particular use case, its goals and under the constraints of this project.  
In theory, it would be sufficient to add a simple external hardware of software mechanism to know whether the soundpainter has “entered the box” or not to remove the greatest part of the problem. One could for instance think of using a numeric carpet of a simple tracker of relative feet distance on the z axis to achieve this.  
In fact, I chose to work primarily on the default mode as my use case. In this mode, the practical use of the “go gesture” “play” is always associated with “entering the box”[[17]](#footnote-19) while even if the use of other “go gestures” such as “Slowly enter” is not always with “entering the box”, a simplified version of the grammar can easily assume that “entering the box” is always performed with “go gestures”, without removing much of the SP performativity in this mode.

With this simplification in mind, a 2D tracking of the full body is enough to cover the default mode. But what about over modes, that also make use of the “box” in a different way?  
Interestingly, most other SP modes that I am aware of are using immediate requests rather than structured requests, and the soundpainter is always standing in the “box” when using these modes. In other words, signing “outside the box”, not in real-time, to make structured requests is particular to the default mode.[[18]](#footnote-20) By entering or exiting other modes (marked with their “enter” and “exit” signs), the program should be able to know whether requests must be considered immediate requests or structured, delayed ones[[19]](#footnote-21).

In conclusion, although more expensive or complex systems would allow for full 3D tracking of the body, PoseNet is suitable for building most features at the body scale relevant to SP and is the most adapted technology in the frame of this project.  
From the observation of correlations between missing depth features and known signs/features, we are able to make a small simplification of the SP grammar by assuming that the requests must be executed immediately when “go gestures” are used, and that in every mode but the default one, the requests are immediate. Under these assumptions, PoseNet is sufficient to build all necessary features to recognize SP signs at the body scale.

##### Building features

Meaningful features to feed the classifier with need to satisfy the following properties:

* The set of features should reduce to the lowest number of dimensions, while still being meaningful, in order to avoid the so-called “curse of dimensionality” problem. It can for instance be solved with a principal component analysis which in general provides the best set of features to a given classification problem; however, in our case, we want the features to always remain interpretable to the user, so that we cannot afford such a transformation. Instead, we should only keep the most significant joints in PoseNet output such as the wrists or elbows positions, discarding less significant ones (for SP) such as the nose or hips. However, if the user wants to build its own set of signs that rely heavily on these body parts, he would have to use such features.
* The features must be invariant to transformations that are not meaningful and do not correspond to any sign. In particular, the features must be translation- & rotation-invariants and independent of the dimensions of the body of the soundpainter.  
  In practice, this is achieved with PoseNet by taking the X & Y distance between each joint and the nose, which is considered a fixed point and then normalizing the X dimension with respect to the inter-shoulders distance and the Y dimension with respect to the distance between the middle of the hips and the nose.  
  The feature invariance will allow a sign to be recognized independently of the soundpainter’s
  + location and orientation (as long as the soundpainter faces the camera with a relatively small angle and does not get outside its field of view)
  + body dimensions (assuming that the general proportions of the human body remain constant)

With these constraints in mind and from my previous knowledge of SP signs, I could first order the body joints in PoseNet by their importance for SP signs recognition for basic signs in default mode:

* wrists
* elbows
* shoulders
* all other joints

In fact, I decided to only use wrists and elbows positions, which gave the best performance in my initial tests, reducing the feature space dimension to 8 instead of 34 (all 17 body joints X and Y values).[[20]](#footnote-22)

#### Hands tracking

THIS PART IS TO BE WRITTEN once the situation with the Hi5 gloves is clearer.

Mention HandPose and alternative gloves systems

#### Input manager

I initially created the ‘input manager’ to allow the user to select in the list of all possible inputs (PoseNet, Kinect, gloves…) those that he was using and to discard those that were unused. This way, the system would know how many features to expect and how to route the data flow correctly.

The user is now also able to add his own input to the system automatically, without manually creating additional routings all over the program. The necessary informations that the user must provide to use his own input are

* the name of the input inside the “Select inputs” panel
* the input data flow itself through a “send <input\_name>” object
* the “input size” (number of dimensions of the input) sent (only once) before the recording is launched[[21]](#footnote-23) through a “send <input\_name>\_size” object
* what classifiers should receive the input

### Part 2: Signs & dictionary management

We have seen that creating new signs is historically the core mechanism of SP. In my program, a sign is defined two properties:

* Its name
* Its category, in analogy with the syntaxic model of the default SP mode: WHO, WHAT, HOW, WHEN, OFF, NEUTRAL, LOGIC **(CHECK IF POTENTIOMETER/MAPPING instead of “HOW” is relevant when using with gloves)**. Other categories such as MODE are not implemented yet, given that the program only focuses on the default mode of SP.

These two properties are sufficient to allow the program to parse the sign, i.e. to construct a meaningful request from the temporal flow of signs.

Une image contenant capture d’écran, texte

Description générée automatiquementHowever, for a sign to be fully implemented (and have a concrete action), we will need two additional steps:

Figure 3 User interface for defining new signs. In the left rectangle, the user must type the signs he wants to in the format “category:sign\_name” without spaces or special characters. On the right panel, the user can see each corresponding sign in a dictionary form, ordered by category. TO REPLACE WITH CLEAN VERSION

* Recording training examples that are required by the classifier
* Programming the sign with the virtual instrument itself, so that it can be interpreted by the virtual instrument, just like a human performer must learn what a minimalism is before it can perform it

In the program, the recording process can be launched by the user right after a sign has been defined. The user can choose to either define one sign at the time and record one or several training examples for it, then saving the training data and adding another sign… or directly define a list of signs and recording all of them in the same session.

The recording session has the following form:

* 1. Initial preparation time of *I* seconds
  2. Each sign is recorded *N* times, in the following loop:
     + The recording is launched for *R* seconds
     + Break (preparation time for next recording) of *P* seconds

The user can change the values of *0<I, 0<N, 0<R<5[[22]](#footnote-24)* and 0<*P* according to his needs.

ILLUSTRATION FIGURE HERE

Each record is saved to a different buffer of the Multiple Buffer (MuBu) object and each active input data is saved into a different track. The user can navigate the recorded data in each buffer and track using the Multiple Buffer Interface (Imubu object).

ILLUSTRATION IMUBU

After the recording session, the data contained in a given track of a given buffer corresponds to a (N+1) x L matrix with N being the number of dimensions of the corresponding input and L being the number of steps in the recording sequence (sampling rate x duration of the sequence). The additional dimension corresponds to the time-tagging of the data. There are two main reasons for time-tagging the data:

* When several inputs are recorded at the same time, each has its own output rate; time-tagging the data in each track guarantees that during playback, the rate of each track is preserved
* Individual inputs can have varying output rates over time, for instance Posenet. Although we will see that the Dynamic Time Warping classification does “warp” the sequence in time and is therefore not sensitive to small variations of data rate, it is safe to assume that keeping the data timing in place always would better represent the original movements of the soundpainter

Wrapping up, inside the MuBu object, a sign is represented by labeled multi-dimensional buffers that contain the motion tracking data corresponding to each recorded example of the sign.

ILLUSTRATION LABELS ET DATA 2

Once the data has been recorded, the user is able to save the recorded buffers to .mubu files, by default in the ./data folder.

What would be ideal in my system would be to store the data in the following fashion:

./data  
 /track\_name (corresponds to the input name)  
 /sign\_label + unique\_id (.mubu or .txt)

The motivation for using this file and folder structure is that it best represents the data structure of the MuBu object itself and allows the user to clearly identify what the file corresponds to without looking at its metadata. The user would then be allowed to mix data from recording sessions that are inhomogeneous, i.e. with a different number of recorded inputs, by loading all the files that corresponds to the inputs he uses, even though they might come from very different sessions.

However, the MuBu object write and read mechanisms suffers from bugs[[23]](#footnote-25) that should be fixed by the developers in the near future (as of May 2020) and I had to implement a workaround before it gets fixed, by saving each buffer with all its tracks in a single file:

./data  
 /configuration\_#track\_name\_1\_#track\_name\_2…  
 /buffer\_name + unique\_id (.mubu, .txt)

This way, the buffer names are saved correctly, but the user is no longer able to mix data from different sets of inputs.

Loading the data is much simpler and can be achieved by a simple drag and drop of one or several data files in the dedicated zone of the program.

### Part 3: Classification

### Part 4: Grammar parsing

### Part 5: Orchestra simulation

## Performance

### PoseNet and Wekinator settings

We know that the DTW mechanism of Wekinator[[24]](#footnote-26) (the classifier that we will introduce in layer 3.) is by default downsampling its input to improve the DTW speed, such that the best compromise in performance and accuracy is to keep the number of FPS just below what Wekinator can handle without downsampling for a sequence of 2 seconds[[25]](#footnote-27). With Wekinator’s default settings (max sequence size = 10), the ideal number of FPS is 5[[26]](#footnote-28). On a fast computer, it would be worth to change Wekinator’s default max sequence length to 20-30 and run with around 10-15 FPS, which have proven to be more than enough for SP recognition or disable downsampling.

The qualitative improvement of the translation- & rotation-invariant transformation was also observed during the initial tests with those 8 features by moving in space and taking slightly different orientations to the camera. However, the improvement given by the normalization of the body joints with respect to body dimensions is still to be tested with more users.

### Threading

## Learning in SP and numerical tool

## Potential & future of the tool

### Emotion recognition

#### From PoseNet to building features for recognizing emotions in body gestures

Although we discussed the choice of meaningful features for recognizing SP signs within the default SP mode, I would like to discuss the motion descriptors library for Max/MSP “Modosc”[[27]](#footnote-29) and its potential use in other SP modes.

Let’s look at the SP mode “shapeline”. In this mode, all gestures and signs made by the soundpainter (except the “exit” sign of the mode) are interpreted in a figurative way by the performers, i.e. in an iconic or suggestive way rather than in a symbolic way. For instance, the soundpainter could use his facial expressions to convey emotional content or imitate the throwing of a virtual ball in the space and let the performers interpret (abstractly and freely) the dynamics of the scene.  
Interpretation in such a mode, as we have seen previously, involves particular cultural knowledges as well as knowledges about emotions, facial expressions… whether they are cultural or not.

The EyesWeb project[[28]](#footnote-30) proposes several “expressive cues” for analyzing body movement and gestures in relation with their emotional content and creating models of interaction between gestures and musical languages, in analogy to what the mode “shapeline” offers in SP. Some of these cues are features such as the contraction index, fluidity, curvature of a movement, jerk (first derivative of the acceleration), symmetry with respect to different points of the body, directivity etc.  
Although the authors of the this system are not directly demonstrating that these are the features the cognitive system uses for triggering emotions[[29]](#footnote-31) , the experiments and interactive performances they present suggest that these features do capture some sense of the emotions conveyed through gestures.

In Max, the “Modosc” library allows for computing of some of these cues. In future extensions of the project, it would be relevant to explore simple interpretations of emotional contents through gestures with music or visuals contents in the shapeline mode based on such features for PoseNet, even though more performant models would probably come from non-interpretable machine learning models in the near future.

#### FaceMesh and face emotions recognition

Tensorflow has just released in March 2020 the Facemesh package[[30]](#footnote-32) that infers approximate 3D facial surface geometry from an image or video stream and that can be ported to Max just as easily as PoseNet or Handpose models.

One use of this package for SP could be building a lightweight expression and emotion classifier that could also be used in SP modes such as the shapeline.

There are already several convincing attempts at recognizing facial emotions but Facemesh advantage is that it opens the way for fast, real-time emotion recognition from 3D mesh, hence independent of the user’s face color, dimensions, eyebrow shape, etc. Just like a normalized skeleton from PoseNet allows us to build a simple yet efficient model for recognizing SP signs without heavy training sets and models, the Facemesh could allow us to recognize emotions as facial signs… but also creating our own, new signs with the face.

### Topic B: …

### The future

# Conclusion

SIGN = something that stands for something else; signified & signifier  
3 != types of signs: symbols, icons and indices

No synonyms in SP

Are there homonyms? YESS

Meronymes… to check/discuss : sense relations

Sign overloading: what is an analogy?

Quantum-like theory of concepts to model human interaction (interference) in responding to SP signs? <https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00418/full>

Modes and defaults: attempt to define particular grammars inside a “universal” one?

Enquete statistique sur poids cognitif/temps de réponse/complexité de la réponse/débutant/experts

“categorization is the name of the cognition game and analogy is the mechanism that drives it all”.

Making analogy = raising the similar features of two mental things

Analogy is responsible for concept’s expansion.

Hierarchy of concepts? At least a complex form of structuring; Link with researchs in DH ontologies. Nice example: german way of constructing words

Utilization des signes motivée par la representation a priori du concept/résultat du soundpainter

Représentation mentale qui precede la réponse du performer; intéressant pour W (tout le monde?) si la réponse n’est pas telle qu’attendue

On ne peut s’adresser qu’à une discipline qu’on connait un peu

Immobilté du soundpainter dans la config ordi et config traditionnelle walter => aucune “deformation” du contexte de la performance artistique  
Notion d’espace

Distinction SP et performer: discussion de l’évolution, prospections de configuration possible?

Nouvelles configurations: le so

W a deja supprimé signe? Pk? Ex “race” supprimée du dict de l’académie  
Pas de synonyme en SP? Interpretation du synonyme dans les langues avec contexts: ??? nuance contextuelle (cf video )

Language SP manipulé par le sp (context free grammar) VS musical language (context sensitive) by performer

Context: importance au dela de la (context-free grammar): attentes du compositeur, configuration, par défaut…

Mode: préciser le sens en SP

“mode”, “forme”

Glossaire (definition des mots/concepts utilizes)

Biblio, table des illustrations.

4 semaines:

Facilité:

- partie technologique

* Brief history of SP

1. Eleanor H. Rosch. Natural Categories. Cognitive Psychology 4, 328-350 (1973), introducing the “prototype theory” [↑](#footnote-ref-1)
2. For a concise explanation of the ideas behind the prototype theory, check https://www.youtube.com/watch?v=mff\_sPnz\_gs [↑](#footnote-ref-2)
3. Ref M.R. class on pitch categorisation. Emphasis on the complexity of the categorisation process (non linear wrt similarity) [↑](#footnote-ref-3)
4. Note that the whole discussion of this section is only relevant to the signs used to signify contents. [↑](#footnote-ref-4)
5. In reference to the concept of overloading in programming languages. [↑](#footnote-ref-5)
6. There may be other cases in which a sign has several signified, even inside one discipline, but we won’t discuss this possibility. [↑](#footnote-ref-6)
7. We will discuss the motivations of this construction later. [↑](#footnote-ref-7)
8. it can also be manipulated by cultural schemes, but that is not the point here. [↑](#footnote-ref-10)
9. or the whole concept itself, but as we already saw, defining the frontier of a concept is not always possible. [↑](#footnote-ref-11)
10. I use plural here because I am assuming from the beginning that there is not a single concept of a LT for all disciplines but rather a set, a group, a of concepts, some relevant for only one discipline, some relevant for several disciplines, under the sign « LT ». In fact, I am following Maurice Bloch by thinking of signs and words as decoupled structures, where the sign LT can be thought as an alias for several underlying concepts (a line in painting, a fluid movement in dance…) : this is exactly what I am calling the overloading of signs. [↑](#footnote-ref-12)
11. In my case, from my experience in Europe, producing a large variety of contents was personnally exciting and part of the interest I and others share for the technique ; therefore it was an important consideration that may not be shared at all by other groups with different interests and motivations to use SP. [↑](#footnote-ref-13)
12. External objects that are “shipped in” with the program are distributed under GPLv3 license (or less restrictive). [↑](#footnote-ref-14)
13. With TensorFlow: https://github.com/tensorflow/tfjs-models/tree/master/posenet [↑](#footnote-ref-15)
14. https://github.com/CMU-Perceptual-Computing-Lab/openpose [↑](#footnote-ref-16)
15. The only realistic approach would have been to use the pytorch implementation of OpenPose (<https://github.com/Hzzone/pytorch-openpose>) and then try to run the python scripts in Max (Max does not interpret Python natively) with <https://github.com/grrrr/py>. It is very unlikely to work and could not be tested in the frame of my master thesis. [↑](#footnote-ref-17)
16. The most convincing project is https://saic-violet.github.io/learnable-triangulation/ [↑](#footnote-ref-18)
17. I had a theoretical discussion with Walter around the idea that in theory, one could use « play » without « entering the box » in requests such as « Strings LT Slowly enter Perc. LT Play WG After 30 seconds Off Play <enter the box> », which would mean that when the soundpainter « enters the box » :

    Strings slowly enter with a LT

    Percussions play immediately a LT

    Everyone stops after 30 seconds

    By not « entering the box » when signing the « Play » or « Slowly enter » gestures, it would allow everyone to schedule the 30 seconds count at the same time to perhaps stay synchronized with an external process. It would make sense to differenciate the timing properties of the « go gestures » (delay start by a certain amount) with « entering the box » which has for specific meaning « execute the request now » for certain configurations, such as when working with computers and time-sensitive systems. In practice, with human performers, the decorrelation between the two is either rare or never used within the default mode. [↑](#footnote-ref-19)
18. One could ask : « why then is it the case that the soundpainter always consider the virtual « box », even when the grammar of the mode implies that the request is always immediate ? I believe that it has to do with how SP has developed from the default mode to others and the pratical point of reference that the « box » is a common symbol for immediateness in all modes. [↑](#footnote-ref-20)
19. In practice, soundpainters tend to omit the « enter » and « exit » signs of modes that are not ambiguous, making things more difficult without depth tracking in case that « entering the box » is the only mark of the change of mode (for instance by using the launch mode directly with « entering/exiting the box » instead of signing Launch Mode and Tear). [↑](#footnote-ref-21)
20. I did not test the performance of this choice quantitatively, but only qualitatively by testing the change in performance that I could perceive with and without shoulders. It is obvious from the SP signs that I consider that aside shoulders, no other body joints could increase the performance of the recognition in real-time (curse of dimentionality problem). [↑](#footnote-ref-22)
21. The information flow at the opening of the patch follows a specific order. If implemented at startup, the triggering of the input size should be launched with a « loadpercent 91 » object to garantee that it is caught by the input manager. Otherwise, the number can be sent anytime before the recording is launched. [↑](#footnote-ref-23)
22. 5 seconds is the maximum sequence length that I allowed the recording buffer to store. Internally, it is a constraint from the MuBu object that stores the buffers and require a « maximum capacity » for each track, possibly for memory allocation issues. However, it is unrealistic that a sign does take more than 2 seconds to be executed. [↑](#footnote-ref-24)
23. https://discussion.forum.ircam.fr/t/mubu-write-to-file-bugs-suggested-improvement/21714/2 [↑](#footnote-ref-25)
24. DTW in Wekinator is based on the FastDTW implementation http://cs.fit.edu/~pkc/papers/tdm04.pdf [↑](#footnote-ref-26)
25. 2 seconds is the default value of the recording sequence for each sign. [↑](#footnote-ref-27)
26. We indeed have FPS\*sequence\_length = sequence size. [↑](#footnote-ref-28)
27. Modosc can be found at <https://github.com/motiondescriptors/modosc>, which relevant papers. [↑](#footnote-ref-29)
28. http://www.infomus.org/eyesweb\_eng.php [↑](#footnote-ref-30)
29. See for instance https://arxiv.org/pdf/1402.5047v1.pdf [↑](#footnote-ref-31)
30. https://blog.tensorflow.org/2020/03/face-and-hand-tracking-in-browser-with-mediapipe-and-tensorflowjs.html [↑](#footnote-ref-32)