Toward the Evaluation of Complex Hand Gestures for Expressive Voice Synthesis

Jonas Chatel-Goldman

Master in Cognitive Sciences IC²A

Université de Grenoble

Grenoble, France

MAGIC Lab

University of British Colombia

Vancouver, Canada

June 2010

Abstract

This thesis focuses on the gestural control and mapping of hand gestures in DiVA – a new interface for musical expression – with the aim of synthesizing audiovisual speech and song. The main purpose of the present work is to create a set of tools and guidelines for the analysis, evaluation and design of new gestural languages.

A full analysis & visualization system was developed, allowing for the interpretation and concise representation of complex gestural data by means of a real-time principal component analysis. The relevance of this application is illustrated through a detailed study of the language used in the VisualVoice project, and new means of improving target selection in gesture-to-voice mapping are suggested.

The second contribution of this thesis deals with the evaluation of hand poses and their associated transitions. We investigated Fitts' law – well-known model of speed-accuracy tradeoff – in two pointing task experiments, involving respectively simple movements of a single finger, and complex gestures of the entire hand. Results show a good fitting with the model, suggesting that Fitts' law can be successfully used to predict the difficulty of moving the hand from one pose to the other. The specific role of gesture bounds is also investigated, and shows a significant impact on movement time. These results suggest new ways to define gestural language and mappings relating glove controller variables to vocal synthesis inputs.

Abrégé

Dans ce mémoire, nous nous intéressons au contrôle gestuel et mapping des mouvements de la main dans DiVA – une nouvelle interface pour l'expression musicale – à des fins de synthèse sonore et visuelle de parole et de voix chantée. L'objectif principal est la création d'un panel d'outils et la mise en place de recommandations pour l'analyse, l'évaluation et le design de nouveaux langages gestuels.

Un système d'analyse et de visualisation complet a été développé. Il permet d'interpréter et de représenter des données gestuelles complexes de façon concise, via une analyse en composante principale effectuée en temps réel. La pertinence de ce logiciel est illustrée à travers une étude détaillée du langage utilisé au sein du projet VisualVoice, et de nouveaux moyens d'améliorer la sélection des cibles lors du mapping geste-parole sont suggérés.

Un second axe de recherche concerne l'évaluation des poses de la main et de leurs transitions associées. En particulier, la validité de la loi de Fitts – un modèle reconnu du compromis vitesse-précision – a été étudiée lors de deux expériences de pointage impliquant dans un cas de ne bouger qu'un seul doigt, dans l'autre des mouvements complexes de la main entière. Les résultats sont cohérent avec le modèle, ce qui suggère que la loi de Fitts' peut être adéquatement utilisée pour prédire la difficulté lors du déplacement de la main d'une position à l'autre. Le rôle particulier des effets de bord au sein des mouvements a également été exploré, et montre un effet significatif sur le temps d'acquisition. Ces résultats suggèrent de nouvelles méthodes pour la définition du langage gestuel et du mapping liant les données issues du gant aux paramètres de synthèse vocale.

Acknowledgments

First, I would like to thank Sidney Fels, my supervisor, who invited me in Canada and introduced me to this fascinating research topic which was previously unknown to me. Many thanks to my co-supervisor Nicolas d'Alessandro without whom this work could not have been completed. The advices and comments he provided me transcend the scope of this academic work. Many thanks to Pr. Pascal Perrier for his wonderful international network and the cheerful enlightenments he provided to me. Thanks to Robert Pritchard for having introduced me to several beautiful art-engineering works, and shared his wide hockey's keenness. Thanks to Lavana for her careful administrative support. Many thanks to Elena for her steady efforts in trying to improve my English, during our stirring ethico-cooko-musico discussions. Thanks to Nicolas (from Belgium), Nicolas (from France) and Elena for proofreading this thesis. Thanks to Nicolas (from France, 2) for the print services. Thanks to the people in the MAGIC lab, researchers and students, for their welcoming attitude. Finally, special thanks to my family for their understanding and inestimable support, and Débo for her unwavering love.

Contents

Chapter 1: Introduction

"Speech is rather a set of movements made audible
than a set of sounds produced by movements."
— Raymond H. Stetson

Design and evaluation of new interfaces for musical expression (NIME) is a flourishing field of study. The infinite possibilities brought by the evolution of computer music came along with the desire to obtain similar levels of control subtlety as those available in acoustic instruments [6]. Research in this field also provides a new entry point for fruitful investigations of the human motor and cognitive processes.

With the aim of producing real-time speaking/singing voice from hand gestures, the Digital Ventriloquized Actor (DiVA) offers a rich framework for inquiring a wide range of subjects from fine control of expressivity in voice production, to vocal tract models or skill acquisition. Specifically, investigation of the gestural control and mapping with sound synthesis gives insights about the generation and perception of body movements, as well as relations that tie gesture, speech and expressivity.

1.1. Visual Voice project: general overview

Context
The present research is part of the VisualVoice project, carried out at the UBC Media and Graphics Interdisciplinary Center[1]. It is based upon Glove-TalkII and GRASSP projects, developed since the early 1990s by Sidney Fels and Bob Pritchard [9, 30]. VisualVoice gathers students and researchers ranging across several departments (Computer Science, Electrical & Computer Engineering, Music and Linguistic) and explores interdisciplinary collaborations engaged through expression and multimodality.

Project
As the only actual wearable gesture-to-speech system, DiVA has been part of the late-breaking work on new musical interface design (NIME) since its very beginning. The system ultimately aims at *"synthesizing audiovisual speech and song, by means of an intermediate conversion of hand gestures to articulator parameters of a three-dimensional vocal tract model"* [35]. In addition, a visual face synthesis is computed, which is mapped to the produced voice. DiVA is primary intended to be used for music and theatre stage performances, as well as in the everyday community. From the artistic point of view, the ambition is to expand on the ability to create new means of vocal expression. DiVA enables singers and composers to explore different connections, nuances, and subtleties in creating a novel media space that integrates both the human and artificial voice/face expression.

DiVA gives the opportunity to study the human behavior in performing various types of evaluation of highly skilled performers. For several actors playing the same selected audio-visual piece, one can compare the produced voice and gestures under strict experimental constraints [34]. Also, recording and analyzing performer's learning process and evolution in skill acquisition provide valuable contribution to understanding expert performance and human-computer interaction (HCI) design.

[1] Please refer to the appendixes for a detailed view of the structure of the project

1.2. Motivations and direction for the current research

Actual research in the DiVA project focus on three main aspects:

- **3D articulatory voice synthesis**.
  The current system integrates a Holmes formant-based speech synthesizer based on work by Fels and Hinton [9, 32]. Through this approach, voice articulation is mainly represented as trajectories of few parameters. Although being convenient for a hardware implementation, this technique lacks of naturalness and expressivity. Therefore efforts are made toward a more up-to-date and realistic model, providing better voice quality and controllability. The final speech synthesis method will involve a real-time 3D biomechanical simulation modeling human vocal tract and upper airway anatomy [11].

- **Face synthesis**
  The reception of subtly expressive musical passages benefits greatly when the audience is able to see the singer's face [12]. Research is in progress to enhance the DiVA system with a synthetic face that supports such additional expressivity. Here the facial movements are coordinated with sound production directly, making good use of actual muscle-based, parametric, and kinematic control models [26].

- **Gestural control and mapping**
  Both input gestures and gesture-to-voice mapping play a strategic role in the subtle control of voice expression. The question of how to design a gestural language for voice production is not trivial, and has to deal with several issues such as hand motivity and multifinger synergies, symbolic representation of the gestural language or hand-based coarticulation[1]. At the same time, the mapping must provide an interface that easily fits to users' peculiarities, while enabling creativity and virtuosity.

1.3. Personal contribution on the gestural mapping

This thesis focuses on the third aspect of the DiVA project: gestural control and mapping. In the chain of information treatment from gesture to sound, both stages play a strategic role. First, gestures convey performer's vocal intent through a vocabulary composed of specific hand poses. Mapping then translates the hand positions and sends the necessary control parameters to the voice synthesis. Obtaining an expressive and natural voice flow thus greatly depends on the choice of hand poses. Furthermore, speech is not a simple concatenation of discrete phonemes, but a continuum involving *coarticulation*. Therefore, the dynamics underlying gesture (ie. transitions from one pose to another) have to be taken into consideration in the creation of a gestural language and in the mapping strategy.

The main purpose of this Master project is to create a set of tools and guidelines for the analysis, evaluation and design of new gestural languages.

As a primary analysis tool, **visualization** is important in enabling a straight interpretation of the data captured when moving the hand. Moreover, one can ask for a visual feedback to have a better mental representation of gestures during the instrument learning phase [3]. In this thesis, a full visualization system was developed, allowing for the concise representation of complex gestural data by means of a real-time Principal Component Analysis (PCA).

[1] Coarticulation refers to the fact that a phonological segment is not realized identically in every context, and shows considerable influence from neighboring segments (overlapping articulation). For a detailed revue, see [16].

The **evaluation** of gestural poses and their associated transitions is a complex question, involving related fields such as experimental psychology, physiology and human-computer interaction. In this context, it may be useful to benefit from the many existing studies on the design and evaluation of input devices for general interaction [37]. One major goal in these studies is the improvement of accuracy and/or time response in pointing tasks, following the relationship known as **Fitts' law** [22]. Using Fitts' law, one can quantify the difficulty to reach a defined target. Our intention is to apply this very convenient property to characterize the transitions between complex poses of the hand.

Fitts' law is inherently a 1D model, and these two last decades have seen intensive HCI research in modeling 2D and 3D pointing [1, 15, 21]. Extending Fitts' law to complex gestures - involving a higher dimensionality and high level target representation - is a challenge that has still to be overcome. The second contribution of this thesis is an investigation of multivariate pointing in light of the recent progress in the modeling of univariate pointing. By conducting Fitts' experiments with the glove used for DiVA as input controller, we explore the effects of borders, distance choice and previous learning on speed and accuracy. The results suggest new ways to define gestural language and mappings relating glove controller variables to synthesis inputs.

Chapter 2: Background

Gestural control of voice synthesis deals with many different and heterogeneous fields of research, such as design of new musical interfaces and controllers, speech and singing synthesis, human kinetics, etc. The purpose of this chapter is consequently not to propose a complete “State of the Art”, but to clarify the context, and concepts used from then on. We start with an overview of actual considerations on digital musical instrument in section 2.1. In section 2.2 we give a brief outline on the relations that tie gesture and speech, as well as a short state of the art of gesture-to-voice synthesis. A few notions of hand anatomy are presented in section 2.3. Finally, a detailed characteristic description of the DiVA 2.x system is given in section 2.4.

2.1. Speech and hand movements

Relations between gesture and voice production have been investigated in details these two last decades, in the light of strong evidences for a common neurological basis, gesture being *“at the cutting edge of early language development”* [28]. For instance, we have not yet discovered a culture in which speakers do not move their hands as they talk! [13]. This tied link and united performance play an important role in expressivity, as illustrated by chironomia, the art of using hand gestures for successful rhetoric and oratory [4]. In its ultimate purport, gesture can act as a complete substitute for speech in the purpose of communication, whenever oral communication is not yet possible. The best example is the sign language[1], which enables to fluidly express a speaker's thoughts using visually transmitted sign patterns. Like oral languages, elementary units – “phonemes” for voice, “cheremes” in the case of signed language – are organized into meaningful semantic units. Signs can either be *iconic*, delivering high-level semantic information based on symbols, or *arbitrary*, for intents of low-level spelling (see Figure 2.1).
In the context of VisualVoice, revisiting the production of speech by means of hand gestures offers a new perspective for exploring voice expressivity, in the directions of daily communication and musical purpose.

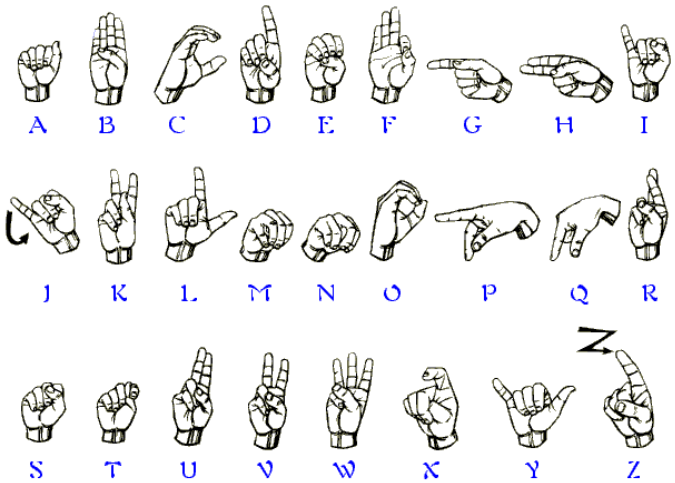

Fig. 2.1: Illustration of the dactylologic alphabet in the American Sign Language (ASL)

[1] For a full review on sign language, please refer to [33].

2.2. From gesture to voice

2.1.1. Digital musical instrument

Wanderley defines a digital musical instrument (DMI) as "*an instrument that contains a separate gestural interface from a sound generation unit, [both are] independent and related by mapping strategies*" [36]. This common way to represent a DMI is illustrated in figure 2.2, where three distinct units capture and shape the signal from gestures to produced sound.

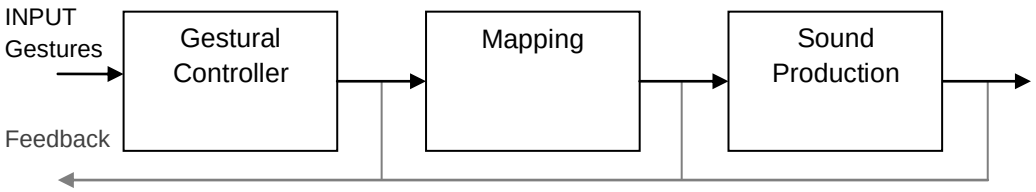

Fig. 2.2: Representation of a Digital Musical Instrument

Performer's gestures[1] are first acquired by the controller, either directly (sensors monitor player's actions) or not (actions are derived from the analysis of sound). Once gesture variables are available, they are related to the synthesis input variables using a particular mapping strategy, and sound is synthesized by the third unit. At different stages in this architecture, there can also be visual, auditory or tactile-kinesthetic feedbacks.

While in traditional acoustic instruments the gestural interface is also a part of the sound production unit, the dissociation of gestural control and sound generation in DMI enables to map any gestures or movements to any class of sounds [26]. This characteristic - offered by the important outgrowth of computer capacities - makes it possible to extrapolate the musical functionalities without any limit, but in the same time poses the question of how to design and perform this new kind of instrument.

2.1.2. Gesture-to-voice synthesis

Research on speech synthesis has enabled to attain acceptable intelligibility and naturalness, which paves the way toward finer control of prosodic nuances. In particular, efforts are made to be able to deal with subtle expressive variations rather than archetypal emotions (anger, fear, despair, etc.) [2, 20]. In this context, expressive speech synthesis and analysis, as well as gestures representation and instrument design are central points to be investigated.

Few works have been proposed on the gestural control of (singing) voice synthesis. In *Voicer* and *Calliphony*, a joystick and graphic tablet are used as controllers to produce the voice in realtime by means of glottal flow models [19, 20]. Performing with a pen gives a natural control on intonation, as it benefits from the skills in writing acquired since childhood. In this direction, theory of device embodiment gives guidelines for the design of highly expressive systems [9]. Within such framework, N.d'Alessandro used the *Luthery Model* for approaching the conception of a tablet-based instrument (RAMCESS), mixing prerecorded voice material and an interactive model of the glottal source [2]. Finally, in DiVA voice is produced using a glove as controller and "embodied" gestural representations of the vocal tract. A notable particularity of this project is the use of an articulatory model as speech synthesis method.

[1] For a detailed study on the notion of *gesture* in the context of human-computer interaction and in the musica domain, please refer to [6].

2.3. Anatomy of the hand

This is not our intention to explain in details every bone, muscle and tendon that compose the hand. However, a short summary of the main articulations is interesting, since their control is at the foundation of hand movements further acquired by the digital instrument.

The articulation of the human hand is constituted of three sets of different joints:

- Interphalangeal articulations are the hinge joints between the finger bones. We can distinguish the *proximal interphalangeal joints (PIP)* situated between the first (also called proximal) and second (intermediate) phalanges, and the *distal interphalangeal joints (DIP)* between the second and third (distal) phalanges. The only movements permitted are flexion and extension.
- *Metacarpophalangeal joints (MCP)* bind the phalanges and the metacarpals. The movements which occur in these joints are flexion, extension, adduction, abduction, and circumduction. Movements of abduction and adduction are very limited, and cannot be performed when the fingers are flexed.
- *Radiocarpal (wrist) joint* refers to the anatomical region surrounding parts of the forearm bones, parts of the five metacarpal bones and the series of joints between these bones. The movements permitted in the wrist are various and complex. They comprise marginal movements (abduction, movement towards the thumb or the little finger) and movements in the plane of the hand including flexion (tilting towards the palm) and extension (tilting towards the back of the hand).

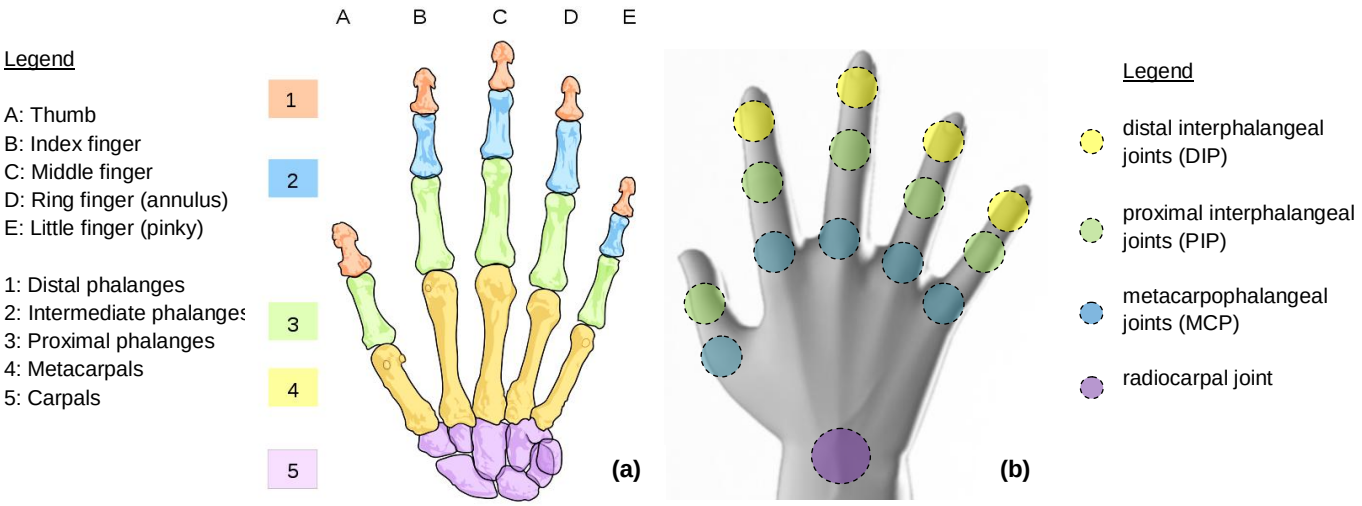

Fig. 2.3: Dorsal view of the human hand, (a) bones anatomy, (b) simple representation of the articulations

2.4. Gestural mapping in DiVA 2.x

2.4.1. Choice of the gestures

DIVA 2.x is a DMI that performs continuous hand gesture-to-voice mapping, using a number of hardware items to track a performer's hand movement and gestures. Differing from sign language, in this system speech is produced phonemically: each gesture corresponds to a single phoneme. There are 11 cardinal vowel sounds and 15 cardinal consonants mapped this way. The current hand poses were determined on the basis of a mental representation of the human vocal tract and articulators [9]. Consonants are produced by touching fingers to the thumb, much like how occlusive sound are naturally produced in the vocal tract when mobile organs (tongue, lip, lower mandible) are press against fixed articulators (hard palate, incisors) to stop the airflow. Vowels are produced by moving the hand in

the horizontal space, in a layout that is similar to the first and second formants of the vowel space (F1, F2).

2.4.2. Hardware devices

Performer's gestures are tracked by three hardware components:

- A wireless *CyberGlove*™, worn on the right hand, measures 18 angles of the hand articulations as shown in Figure 2.4. We will focus on this main controller along this thesis, as it has the central role of acquiring the complex gestures mapped to the phonemes.
- A *Polhemus Patriot tracker*™, worn on the right arm, measures 6 degrees of freedom of the performer's hand (X, Y, and Z position; yaw, pitch, and roll). It is principally used for moving into the vowel space (X, Y position) and pitch (Z position).
- A *TouchGlove*, essentially a modified keyboard, is worn on the performer's left hand. It has eight contact sensors that are activated by the performer pinching the appropriate pad with his/her left thumb. These are mapped to plosives, which proved to be too difficult to perform using continuous gestures due to their rapid release rates.
- Finally, the performer uses an *insole foot petal* to control the master volume of the sound produced by the system.

(a)

(b)

1	Thumb Rotation
2	Thumb MCP
3	Thumb PIP
4	Thumb Abduction
5	Index MCP
6	Index PIP
7	Middle MCP
8	Middle PIP
9	Middle-Index Abduction
10	Ring MCP
11	Ring PIP
12	Ring-Middle Abduction
13	Pinkie MCP
14	Pinkie PIP
15	Pinkie-Ring Abduction
16	Pinkie Rotation

Fig. 2.4: Picture of the CyberGlove™ with the sensors superimposed (a) and described in table (b) Except DIP joints, the angle values of almost each articulation are acquired, enabling for a precise and complete "snapshot" of hand gestures.

2.4.3. Software components

The system is intended to be used by a multitude of performers, each with their own unique hand shape and performance style. As such, an adaptive interface has been developed in DiVA 2.x, that allows each performer to create a set of proper "accents" (ie. configuration parameters). Three neural networks are used to implement this interface, playing the role of the mapping layer in a DMI representration. As described in [12], one Normalized Radial Basis Function (RBF) network is specialized to map the X and Y coordinates of the right hand to vowel formants while another maps right hand finger movements to consonant formants. The third network blends these two formant outputs together based on how much of a vowel or consonant shape the performer's hand is in. The centers of each RBF in each network are

set to the respond to a hand posture associated with a cardinal sound, and trained on multiple samples. Therefore the network responds to its primary regression intent, in approximating the complex hand poses space with a number of multivariate Gaussian functions (represented in Figure 2.5). Normalized response NR of each RBF is defined as:

$$R_i = \frac{\sum (M_i - I_i)^2}{\sigma_i^2} \qquad NR_j = \frac{R_j}{\sum R_j} \quad (1)$$

Where M and σ are the mean and standard deviation over all training samples, for each sensor parameter i, and I is the current real-time value for that same sensor.

We illustrate the selection process with a simple example in using only two sensors X and Y. In terms of selectivity, if the standard deviations for both the X and Y dimensions are very small, the width of the Gaussian function in both dimensions will be very narrow, and the performer will have to produce a gesture close to the phoneme's center for having favorable value returned by the corresponding hidden unit in the RBF network. If the X standard deviation is low, but the Y standard deviation is high, the performer will have to hit a very specific location in the X direction, but with a much greater range of favourable Y values.

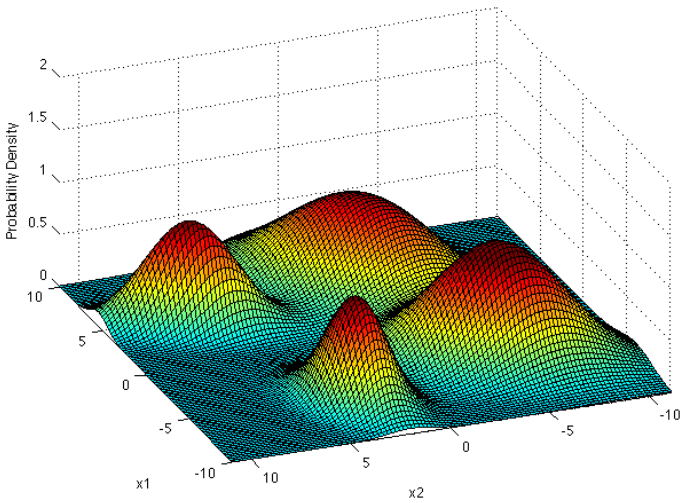

Fig. 2.5: Example of Radial Basis Function responses in two dimensions. Width and position of the normalized radial-basis functions are set with the standard deviations and means of the training samples.

2.4.4. Limitations

In the RBF network, the outputs of each hidden unit are normalized, so that they sum to 1, and their response comes to represent the probability that the performer is trying to produce the corresponding phoneme, with the closest units returning much higher values than the units that are farther away. As a result, the system interpolates for intermediate positions, performing a linear mapping between the closest RBF outputs. The drawback here is that such a linear interpolation leads to audible artifacts on the synthesized sound, since it approximates the physical mechanisms behind phonemes transitions, which are not linear by nature.

Inputs of the neural network are the raw data issued by the different controllers, CyberGlove included. One particularity in the hand motivity is the fingers' synergies, ie. their mechanical coupling. Feeding the RBFN with isolated sensor values does not take into account these coupled degrees of freedom, although they are important for a complete interpretation of the gesture data. This issue will be discussed in chapter 3.

Chapter 3: Visualization and analysis by means of real-time PCA

3.1. Introduction

Providing tools that enable a straight interpretation of the data captured when moving the hand is one of the primary aims of this master project. The use of a very precise sensor with high dimensionality as input controller (eg. the CyberGlove) is essential to acquire the subtle hand movements - which are further mapped to the voice parameters - and convey performer's intention for an expressive voice. However, the resulting complexity imposes a serious drawback in the way the data can be visualized and interpreted. Consequently there is a need for reducing the data dimensionality, in order to be able to properly visualize and interpret the hand movements.

The dependencies between the various sensors are another issue to be considered. Due to physiological but also technical considerations, there is a substantial amount of information redundancy in the glove data. This redundancy is to be considered as a correlation that holds some of the relations tying the fingers' kinetic. To distinguish the part of data that hold "important information" (this notion is examined in section 3.3.7) is a primordial step in the design of a mapping strategy.

These issues can be addressed by a **real-time data reduction**, in projecting the immediate sensor values into a new space to extract signals that best describe the main characteristics of the original motion capture data set. We emphasise on the "real-time" condition for the analysis, reduction and visualization of the resulting signals. Giving the researcher and performer an instantaneous visual feedback is a highly desirable property, since it allows for an intuitive understanding of the dynamic of the hand movements. The real-time property for gesture mapping is also extremely important to avoid any latency, since gestural mapping is a part of the digital instrument.

Principal Component Analysis (PCA) is used in this project, as it provides an objective and common method to reduce the dimensionality and to evaluate the variances and correlation within a given data set.

In section 3.2 we present in detail the method used to compute PCA in this project, as well as terms and notions referred to in the sequel. In section 3.3 we recall the main assumptions and limitations underlying PCA. How it is applied in a set of tools enabling the analysis of gestures in DiVA is described in section 3.4. Section 3.5 clarifies the central role of training set. An in-depth analysis of current DiVA gestural language is given in section 3.6. Finally, the notions of sensor variance and distances choice are discussed in sections 3.7 and 3.8.

3.2. Background in PCA

PCA is a statistical technique commonly used as a tool in exploratory data analysis. It is alternatively named the discrete Karhunen-Loève transform, the Hotelling transform or the proper orthogonal decomposition (POD) depending of the field of application. A detailed description of the mathematics behind PCA can be found in [8, 31]. The idea is to combine information that demonstrates high covariance within the data set, in finding the orthogonal axes that maximize the inertia of the data cloud, as illustrated in Figure 3.1. PCA is a two-step algorithm that includes the decomposition process and the reconstruction process[1].

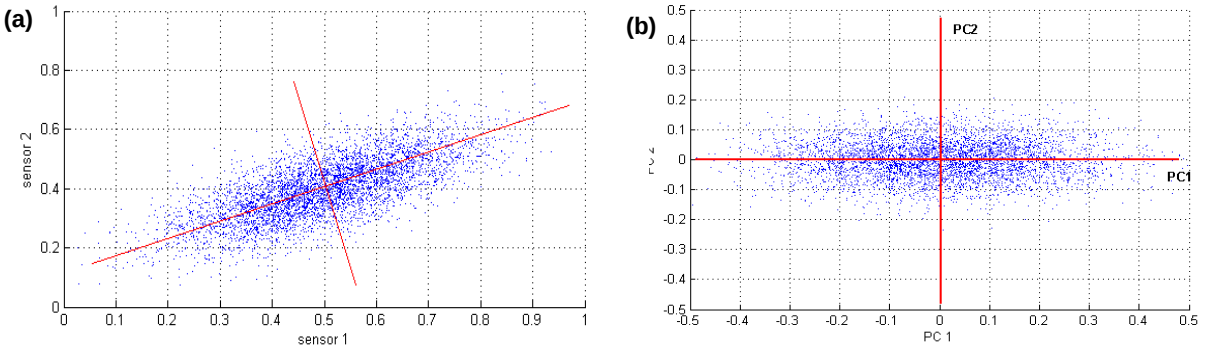

Fig. 3.1: Example of a PCA process on a bivariate Gaussian distribution. First the orthogonal axis that minimize the sum square error within the data set are found (a), then the data set is projected into this new base (principal components) (b).

The covariance matrix of input signals is first decomposed in terms of its eigenvectors and eigenvalues. Once the eigensystem is obtained, the original data set is projected into the new basis maximizing the covariance, and reconstructed according to its principal components (PC). These components are ordered so that those with the highest eigenvalues are presented first. PCA is a non-destructive process preserving the global information; data reduction is performed with the extraction of a lower dimension data set, obtained by selecting the components that seem to best explain the behaviour of the original one.

3.3. Properties and limitations

When applying PCA, one must keep in mind its intrinsic properties and underlying assumptions:

- PCA is **non-parametric** (no prior knowledge is incorporated in the process), **bijective** (one-to-one correspondence between the initial and reconstructed sets) and **independent of any hypothesis about data probability distribution**.
- PCA is theoretically the **optimal linear scheme**, in terms of **least mean square error**, for the projection of data to a new coordinate system such that the greatest variance along the input data set corresponds to the first principal component, the second greatest variance to the second principal component, and so on.
- PCA only finds the independent axes of the data under **Gaussian** assumption. It comes from the primary motivation of the method, which is to decorrelate the data set, ie. remove second-order dependencies.
- PCA only finds the independent axes of the data under an assumption on **linearity**. The initial data set is assumed to be a linear combination of a certain basis.

[1] A detailed description of PCA computation is given in the appendixes.

3.4. Applying PCA to the analysis of gestures in DiVA

3.4.1. Overview

Within the DiVA frameworks, we developed a full system for the real-time analysis of hand gestures. Important decisions taken for the development and programming issues are discussed in section 4.4.2. A sum up of code structure and data flow is given in section 4.4.3. The application, which implements the principal components analysis and visualization with a user-friendly and didactic interface, is described in details in section 4.4.4.

How glove data is processed follows the layout given by the two-step PCA algorithm:

- *Record and processing – decomposition step*

A record session is necessary to provide the **training data set** that serves as a basis for the analysis. Every input sample generated during a sequence of hand movements is saved in memory for this purpose. Isolated gestural targets are also recorded during this session, which allows for a later analysis and representation of these complex hand poses. The essential role of the training set and precise record protocols are discussed in section 5.1.
After pre-processing of the recorded data, the eigensystem is computed from the training set in order to obtain the associated projection vectors.

- *Projection and visualization - reconstruction step*

This phase corresponds to the core of the real-time analysis. It simply consists in projecting the current input data into the new space defined by the training set, without any modification of the eigensystem. We then provide a complete visualization for the intuitive representation of principal components. Gesture trajectories are interpreted trough the movements of faddish spheres in a cube, representing the projected hand data into the PCA space.

3.4.2. Practical choices

The application is intended for use in both the present study and future research within the DiVA project. Thus, development choices are made following constraints such as modularity, reusability and inter-operability. In particular, it must enable an easy migration from one plate-form to another, supporting any upcoming strategic decision on the running system.

Programming language is C++, since it is the language chosen for the DiVA software. As “middle-level” and object oriented language, it allows for both good efficiency and modularity. Due to its popularity, a large number of libraries are available, which is particularly convenient.

We only used free, open-source and cross-platform environments and libraries in the development of this application:

- *NetBeans* is used as Integrated Development Environment.
- *Eigen* is used as template library for linear algebra. It is versatile, fast and robust.
- *openFrameworks* library is used for the user interface and 3D display. It is especially designed to assist the creative process by providing a simple and intuitive framework for experimentation.

We also used the OSC (Open Sound Control) protocol for the communication between the input controller, the application and other DiVA software. OSC's advantages include interoperability, accuracy, flexibility, and enhanced organization and documentation.

3.4.3. Code structure and data flow

We take full advantage of C++ as an object-oriented language using classes, inheritance, overloading, etc. Any modification applied to data is monitored carefully; the information can be checked at any moment in the process sequence. Data are recorded in text files, allowing for a deftly use and analysis.

Detailed view of the class organization, information path and processing, and file format are given in the appendixes.

3.4.4. Software presentation

Recording

A full part of the software is dedicated to the record of hand movements and gesture targets. The user is asked to realize a specific sequence of gestures, which are automatically recorded. Each target is registered a certain number of times, which leads to the creation of one cloud of glove data per target, enabling for a further analysis of variance and mean calculation.

The record procedure is the following:
1. A picture representing a given hand pose is displayed.
2. The user moves the hand to “reach” the target, and press the backspace button for target selection (at this point, reaching decision can only be left to the user’s appreciation). The current glove samples are recorded and added to the target data cloud.
3. A new task begins with a new target, using the previous one as initial position.

The choice of transitions to be executed is important, in the sense that it completely defines the gesture space taken into consideration as training data set. We use a particular pathfinding algorithm to randomly generate an optimal sequence of gesture poses, so that each transition is only parsed a given number of times (Euler path). This point is discussed in detail in section 3.5.

Fig. 3.5 Display of hand pose during the recording session

Visualizing

The main aim of this application is to enable a straight interpretation of the hand movements. In this context, efforts have been made to clarify as much as possible the real-time representation of gestures. Our elaborated visualization tool includes the following features:

- Hand movements are represented by means of faddish spheres moving inside a transparent cube (figure 3.5). Gesture trajectories are made easy to visualize with a parameterizable “tail” reminiscence. The cube can be rotated and scaled ad libitum. Visualization can be paused at any moment to capture and analyze the hand movements.

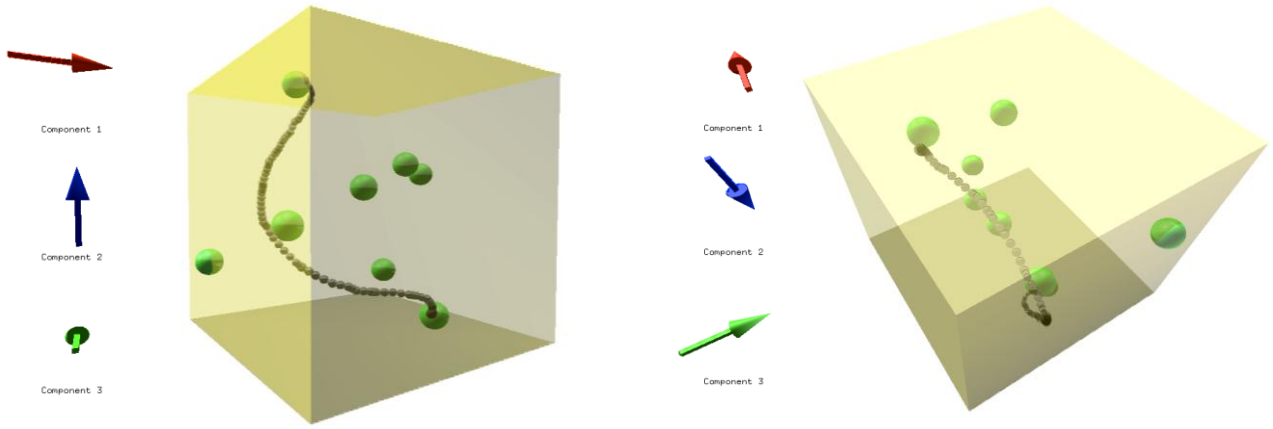

Fig. 3.6: Gestures are visualized trough the apparent motion of spheres (in black) inside a cube. Recorded targets are also projected into the PCA space and displayed (in green). Here axes are the three first principal components.

- Projecting the movements in 3D enables to represent three successive principal components (PC) simultaneously. All dimensions are equally scaled for display relatively to the maximum dynamic in training base among every component. The axes can be browsed easily, shifting from PCs to PCs, which is convenient for their interpretation (figure 3.6).

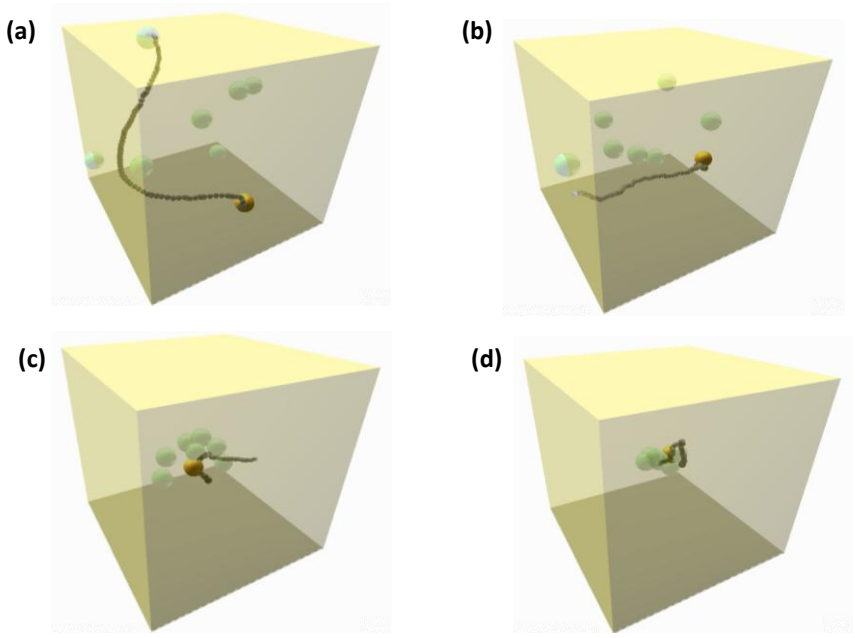

Fig. 3.7: The same movement can be seen along different CPs, in order to obtain a full representation of the projected data. (a) $CP_{1,2,3}$ (b) $CP_{2,3,4}$ (c) $CP_{3,4,5}$ (d) $CP_{4,5,6}$.

- Means of each recorded hand poses are represented into the cube. These latter can be browsed and recognized with a simple mark up and the display of their corresponding picture. The entire data cloud of each target can also be visualized to have a complete view on the recorded data and its dispersion (figure 3.8).

Fig. 3.8: Observation of the targets and their data cloud in the projected space. Top right picture shows the hand pose represented by the highlighted cloud (in orange). Transparent spheres are the recorded positions; solid ones correspond to clouds' means.

- Finally, a collision management system is developed. It enables to delve the concept of distance in the gestural space (discussed in section 3.5.X), and is also used for the Fitts' law experiment involving the whole hand (see chapter 4).

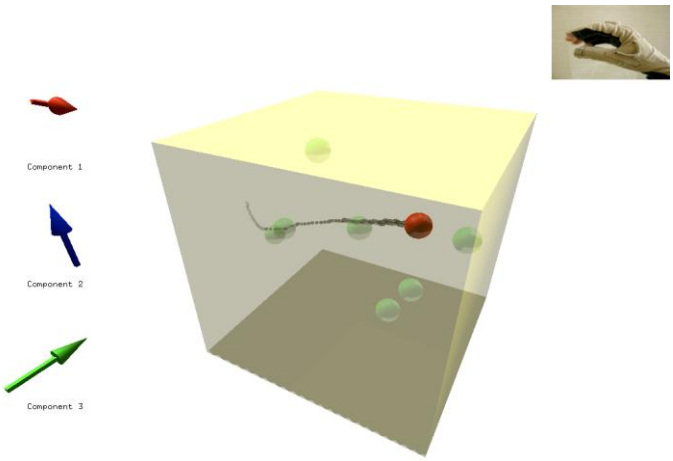

Fig. 3.9: A visual feedback shows when targets are reached (in red)

3.5. Central role of the training set

New glove samples are considered as supplementary points during the visualization, ie. do not cut into the eigensystem. The decision of keeping the training base unchanged after the record session is based upon the following arguments:

- By recording under strict constraints the movements that constitute the training set, one can precisely define the gesture space in which input data is projected. This allows for the investigation of new gesture languages, corresponding then to specific subspaces in the space of all possible hand movements. Any movement that does not belong to the defined gesture set is projected outside the cube. By doing so, the limits of the language become evident, which offers a convenient entry point for the exploration of new hand poses.

- It is essential that the movements done outside the record session do not impact the projection itself, which would be the case if new data was taken into consideration in the eigensystem. Indeed, any visualization or experimentation doesn't make sense if the space and experimental conditions constantly change without control. By clearly dissociating the training phase (deconstruction step) and the visualization phase (reconstruction step), the analysis keeps its objective aspect. In this case, only the personal appreciation that triggers the target selection brings an unavoidable subjective layer.

However, creating the base data set is not trivial. Since every sample acquired during the record session cut into the base data set, any hand movement performed at this moment has to be carefully directed and controlled. The following considerations give guidelines for setting up a judicious record protocol:

- The projection space is influenced by the number of time specific positions in the initial space are recorded. For example acquiring 10 times position A and 1 times position B in a record session will not make the same results than the opposite situation. Therefore the initial set of movements has to be equally balanced among the different hand poses.

- Data is continuously acquired in, but also between the defined hand positions. Therefore the transitions between these gestures should be taken into consideration in the construction of the new gesture space. The ideal case is thus to perform every possible transition the same number of time, in order to obtain a uniform space.

As introduced previously, our record protocol consists in presenting a determined sequence of targets to be performed by the user. The optimal sequence is equivalent to a path that parses each transition a given number of times, not more, not less. This problem corresponds to the famous Seven Bridges of Königsberg problem solved in 1736 by the mathematician Euler [5]. More specifically, in graph theory our precise case corresponds to a strongly connected, directed graph. Latter properties verify the necessary conditions for the existence of an Eulerian path between the hand poses [38]. We can thus construct a random Eulerian path out of this graph by adapting commonly used Fleury's algorithm (cf. appendixes).

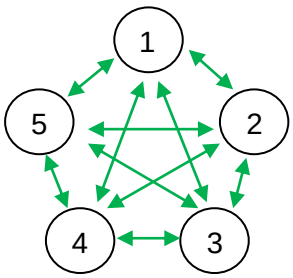

Fig. 3.10: illustration of a strongly connected and directed graph. An Euler path would go from *edge* to *edge* (eg. hand poses) visiting each oriented *vertex* (eg. gesture transition) exactly once.

3.6. Analysis of gestures

Based on the present framework, we distinguish three plans that will be investigated in the analysis and interpretation of gestures:

1. Analysis of movements. Real-time projection and visualization of hand movements gives interesting insights on finger synergies and allows for a functional description of gestures.
2. Analysis of gestural language. Record and projection of hand poses enable to better understand the space occupation of these targets within the gestural language, but also the space occupation of the entire language itself in the universe of all possible hand movements. Analysis of the distribution of targets' cloud (ie. dispersion of record points) gives indications on acquisition of gestures and motor accuracy.
3. Analysis of collisions and target pointing. Describing the transition between targets, and modeling their difficulty is an important step for the design of new gestural languages. This complex issue is investigated in chapter 4.

Problem of comparison and record of a pseudo-random training set

As the PCA projection relies entirely on the initial data set, comparisons are not possible between gestural languages when using different training sets. Although analysis of space occupation *inside* a given language is still possible, it is necessary to have a common base for different languages, ie. to project several hand poses with the same eigensystem.

From this perspective, an ideal basis would be the one that completely represents the universe of all possible hand movements, enabling to visualize the *absolute* distribution of any projected gestural language in the entire hand motor space.

We make an attempt in this direction in recording a large set of pseudo-random hand movements, being fully conscious of the inherent lack of objectivity of such procedure. In particular, we emphasize on trying to visit every possible border in term of hand movements, in order to extend the limits of the gestural space to its maximum (this notion is developed in chapter 4).

Experimental conditions

The following investigations are based on data provided by pilot studies. They were carried out on 3 research fellows, between 22 and 27 years of age, having either normal vision or wearing corrected lenses, and right-hand dominant.
We focus on the analysis of pseudo-random training sets and on the gestural language already in use in DiVA. We recall that the aim of the current work is not to inquire for new hand signs, but to provide tools for future designs of gestures.

Analysis of the pseudo-random training set

We recorded 10K samples of pseudo-random movements as training set, corresponding roughly to 10 minutes of glove data recording. The instructions given to the pilot subjects were to try to visit every possible hand gesture, with an emphasis on maximal flexion/extension. They were also suggested to concentrate on each finger separately, and then on the entire hand. Record was stopped after subjects produced the 10K samples. PCA was finally computed on this data set.

The first 6 components provide enough information to explain 90% of the variance within the data set, as demonstrated in Figure 3.9a. Looking closer to eigenvectors for this specific PCA enable to explain more precisely the contribution of the different sensors to resulting principal components. Figure 3.9c gives an illustration of the eigenvectors matrix through a color scale, enabling for a fast interpretation of the coefficients (related values table is given in annexes). PC_1 appears to be related to the joints second

from fingers tip (sensors 6, 8, 11, 14), PC_2 to the joints where the fingers meet the palm (sensors 5, 7, 10, 13), PC_4 to the thumb movements (sensors 1, 2, 3). Last PCs (PC_{12}→PC_{16}) are almost only contributed by punctual sensors dedicated to abduction movements (respectively sensors 9, 15, 2, 13, 4). It indicates that these sensors acquire fairly uncorrelated movements, which in the same time do not explain an important variance. Real-time visualization of hand motions corroborates this analysis of sensors' contributions to principal components.

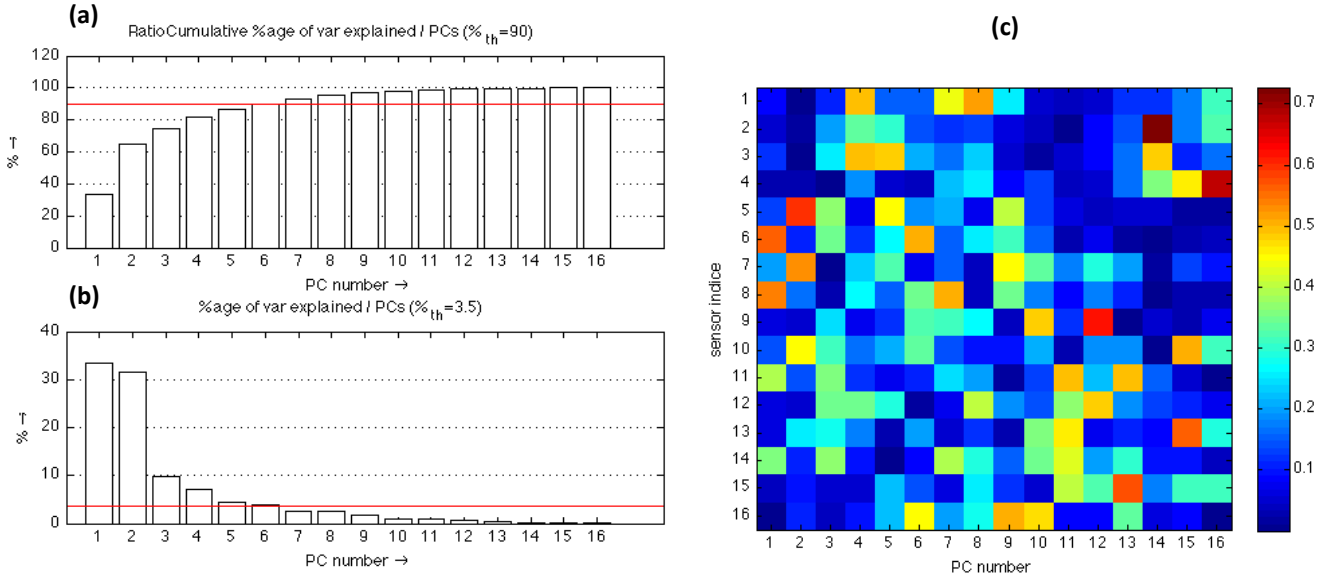

Fig. 3.11: Inertia accumulation **(a)** and fraction **(b)** for pilot subject 1's pseudo-random database (10K samples). **(c)** Illustration of eigenvectors (absolute values). High coefficients contribute greatly to the resulting principal components.

Assuming that our data set offers a significant representation of the entire range of hand movements, these results suggest that when moving the hand without specific constraints, most of the variance is explained by the fingers phalangeal and carpal articulations ($PC_{1,2}$). This has to be taken into consideration in the design of specific hand positions, especially considering the issue of focus when performing specific gestures, as detailed in next section.

Analysis of the language used in DiVA 2.x

The record protocol here is the one previously described in section 3.4.4. Every possible transition between gesture poses is visited, which means that each target is recorded (Nb_{target}-1) times. DiVA language being constituted of 15 hand poses (described in chapter 2), for each pilot subject we recorded 15*14=210 points in the target data set. Record of every glove samples received during the session gives the resulting training data set. Its average number of sample is 49592 over 3 pilot subjects.

Once again, 90% of the variance is explained by the first 6 PCs (Figure 3.10a). Inertia is more distributed here than with the pseudo-random training set, in particular the gap between PC_2 and PC_3 is less important, as shown in Figure 3.10b. A possible reason is that DiVA language uses a small part of the dynamic of phalangeal and carpal articulations, the gestures being concentrated around the same hand positions in the aim of representing the vocal apparatus. More specifically, in the DiVA language oppositions of the thumb with the index and middle fingers, and movements of the ring have an important role in the articulation of synthesized voice [9]. This fact meets the large contribution of sensors 6, 8 and 11 (index, middle and ring second joint) respectively in PC_1, PC_2 and PC_3 as shown in Figure 3.10c.

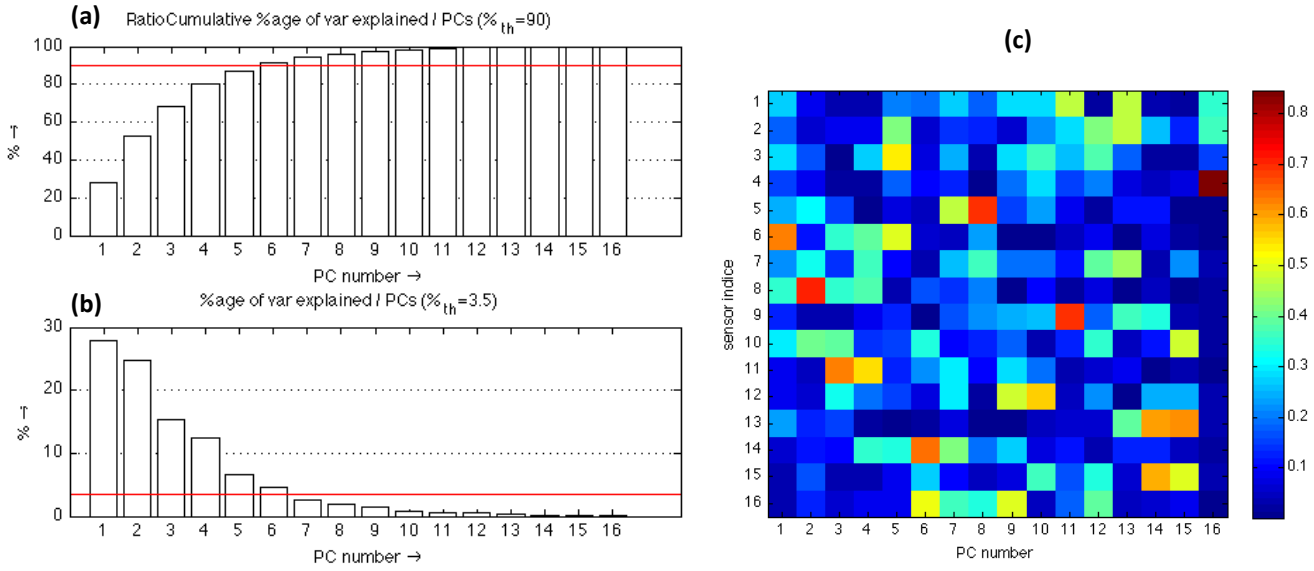

Fig. 3.12:Inertia accumulation **(a)**, inertia fraction **(b)** and illustration of eigenvectors (absolute values) **(c)** provided by pilot subject 2's training database for DiVA2 language (55K samples).

Space distribution of gesture poses in the DiVA language is made visible by visualizing the target positions after projection[1], as depicted in Figure 3.11. It brings to light the non-uniform repartition of hand poses in the space build upon the transitions from one gesture to another. In particular, we observe the presence of small clusters, corresponding to close gestures in terms of Euclidean distance. It is consistent with how the DiVA language is designed, ie. as subtle variations around few key positions, which here can be seen as the groups into the cube. Recorded target positions are superimposed on their means in Figure 3.11b. Visualization of targets' cloud underlines an important overlapping of the recorded hand poses inside each cluster, although in this case visual spheres radius has been arbitrary chosen.

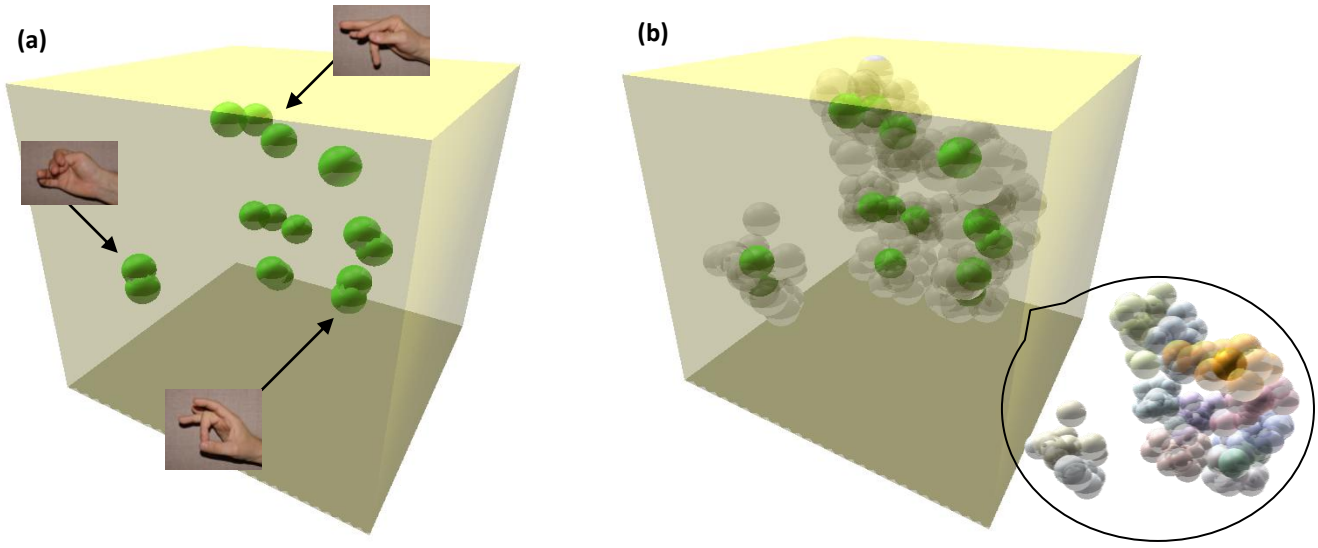

Fig. 3.13: Visualization of DiVA 2.x language (three principal components) for pilot subject 2. Solid green spheres show target means **(a)** and transparent spheres represent targets' recorded positions (one color per target in the insert) **(b)**. Visualization of hand poses distribution underlines an unequal space occupation, with the presence of several groups (arrows and pictures).

[1] Please refer to the appendices for a detailed view of each association with hand poses in the projected space.

Projecting the same set of targets onto a pseudo-random training set also recorded by the subject enables us to represent DiVA 2.x language in an approximation of the entire hand motor space (Figure 3.14). It makes clear that this language doesn't use the full range of possible hand movements, but instead is confined to a small area of the gesture space.

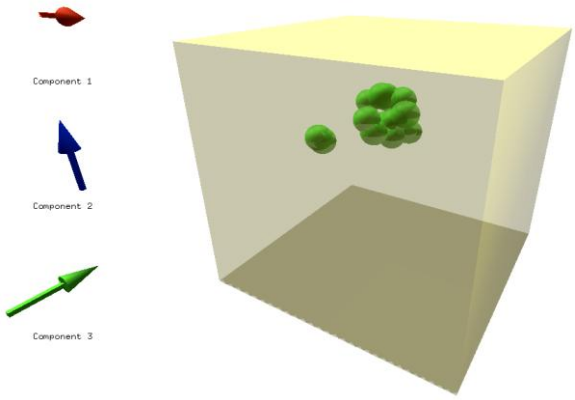

Fig. 3.14: Visualization of DiVA 2.x language (three principal components) projected on the basis built upon the pseudo-random training set, for pilot subject 2. The cube represents here the limits of possible hand movements. DiVA 2.x hand poses are concentrated in a small part of the available space.

Through a detailed analysis of the language used in DiVA 2.x, we have identified some of its main characteristics: importance of index, middle and ring fingers, and variation around hand key poses. We have also observed possible drawbacks in its design, namely a reduced and unequal space occupation. The initial intent in the VisualVoice project was to map these close gestures to sounds that are also perceptively close, enabling for a subtle gestural control of the voice. However, such proximity makes it very difficult for the system to properly dissociate two different hand poses. Furthermore, it increases the risk of collision on the gestural trajectories, leading to undesirable audible artifacts.

We suggest two possible, non-exclusive solutions for solving this issue:

- To change the gestural language, emphasizing on a more uniform repartition.
- To change the distance used in the space where targets are represented. This point is detailed in section 3.8.

3.7. Notion of variance in the context of gesture-to-voice mapping

Within a given hand pose, every articulation and every finger doesn't have the same importance. Indeed when performing a gesture, one focuses on specific parts of the hand for several reasons:

- *Attention (cognitive process)*

It is well known that attentional focus affects motor performance [39]. In order to deal effectively with the control of specific hand muscles, one has to concentrate on them while withdrawing the others. This "limitation" is inherent to human allocation of processes resources.

- *Language design*

There is an explicit focus on small parts of possible movements and positions when designing new gestures for a voice synthesis purpose. For example, the vocal tract analogy in DiVA language often leads to concentrate on subtle movements of fingers in contact with the thumb, taking less care to other parts of the possible gestures.

Therefore, it is particularly relevant to consider focusing on specific regions of the hand in the gesture-to-voice mapping, as it enable to “zoom” on important parts of the gesture while filtering non-important information. Until now, this has not been taken into account in DiVA where the same importance is given to every sensor, whichever the hand poses. As a result, although each gesture only involves a few fingers, the performer has to remember and reproduce the positions of all five fingers for each gesture, which implies a number of disadvantages. Firstly, it places a greater burden on the performer’s memory, slowing down the gestural learning process and increasing the frequency of errors. Additionally, requiring the performer to move all of the fingers into right position for each pose greatly reduces fluidity and decreases sound quality. Over various instances of a single gesture, the fingers not involved in that gesture may be coming from a number of different positions, depending on what the previous gesture was. In consequence, it becomes extremely difficult for the performer to produce exactly the intended sound, and the intelligibility of the speech produced decreases a considerable amount.

Is there a way to automatically extract signals that best describe the main characteristics of each gesture? As a data analysis and reduction tool, can this issue be solved with PCA?

As described in section 3.3, PCA makes the assumption that the main information in the data set is explained by high variances. In order to understand if the variance can be a good indicator of the importance of a sensor in the gesture, we now distinguish different parts of a hand pose:

- Fingers which are important in the gesture. These one are supposed to be realized accurately, since they are the focus of attention. Their relative sensors are thus supposed to have a small variance.
- The other fingers, less important in the gesture. As one does not concentrate on them, relative sensors are supposed to be variable and have a high variance. However, non-important fingers can also be static, even if we do not concentrate on it.

As a consequence, it seems inappropriate to use directly the variance of each sensor as a descriptor of its importance within a given hand pose. For addressing the issue of dynamic focus on specific part of the gestures in future works, we suggest the incorporation of prior knowledge on the gestural language in the mapping strategy.

Nonetheless, target selection can still be improved in taking into account the global variance along the entire training data set, as explained in the next section.

3.8. Notion of distance and target selection

This section analyzes the importance of distance choice for target selection.

In PCA, the entire projection phase is just a linear combination of the data set with a matrix of eigenvectors. If the new basis is not a subset of the eigenvectors (ie. no compression and data loss) then the distances are preserved with the original space, as illustrated in Figure 3.15. It is clearly visible that keeping the same scale, both data distribution and target shape remain identical. Therefore, target selection (ie. subspace closed by target contour) is obviously the same in the – physical – space of sensors and after projection in the space of principal components.

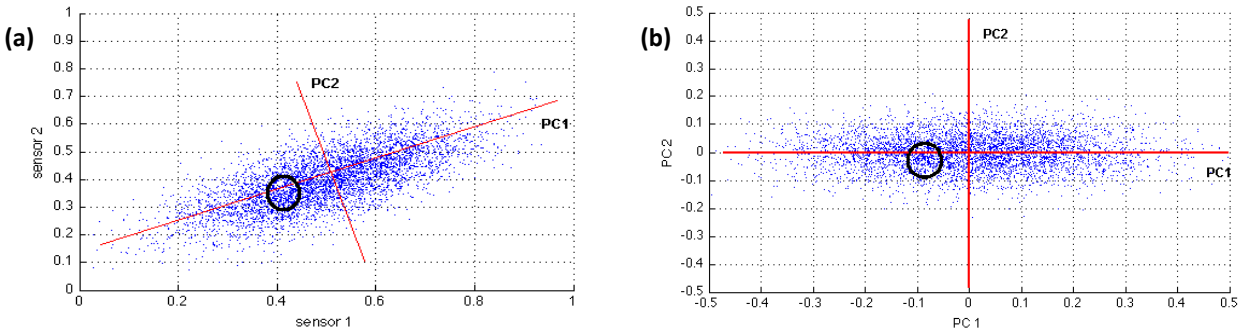

Fig. 3.15: Example of target selection with a bivariate Gaussian distribution: (a) in the sensor space (b) after projection in PCA space using l2 norm with unitary weights. If no compression is applied on data, distances are not affected by the linear decomposition/reconstruction.

We propose to benefit from the fact that PCA decomposes and sorts the dimensions by decreasing inertia, to adapt target form to the variance present in the data set. The assumption here is that the principal components with larger variance correspond to interesting dynamics and lower ones correspond to noise. Accordingly, we decide to weight the PCA space with its eigenvalues. If the first PCs explain a considerable part of variance, such weighting is almost equivalent to reduce and compress the data set, but has the advantage to not require any choice in the number of PCs to cut. Weighting with eigenvalues leads to compress the last dimensions (Figure 3.16a), which for target selection is equivalent to only distort target's contour, effectively adapting to the variance within the data set (Figure 3.16b).

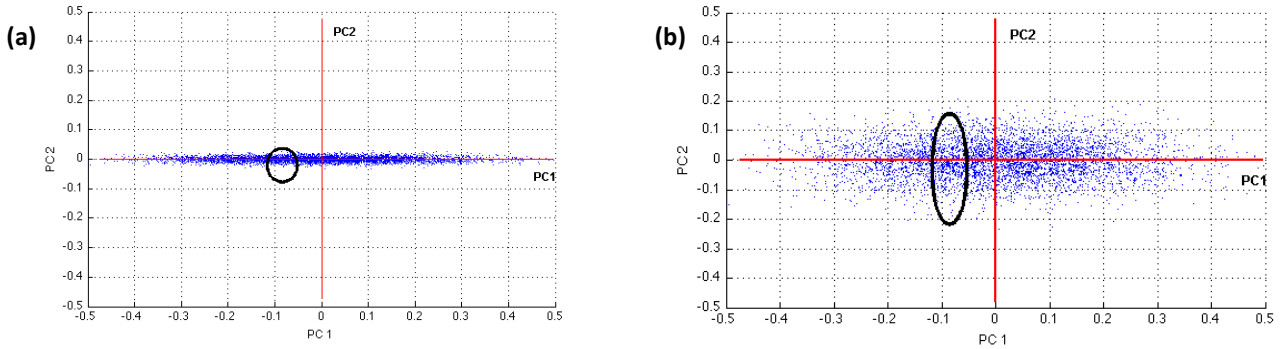

Fig. 3.16: Example of target selection: (a) in PCA space using l2 norm weighted with eigenvalues (b) equivalent target using l2 norm with unitary weights. Keeping the same scale, weighting is equivalent to compress the data set or distort the shape of target according to the weight coefficients (eg. eigenvalues).

Empirical pre-studies have shown a great improvement in the ease to reach and point at targets, when weighting with eigenvalues. Such distance choice will be investigated in details in chapter 4.

3.9. PCA analysis: conclusion

In this chapter, we described the most important elements of the analysis and visualization system in terms of its user interface, data processing and data manipulation. Our application is primary intended to facilitate the design of new gestural languages, enabling for an easy interpretation of the characteristics of hand poses within a given set of gestures. We have illustrated this purpose with a non-exhaustive analysis of the language used in DiVA 2.x, highlighting possible drawbacks in the choice of hand poses. Finally, we discussed two important aspects of glove data interpretation and manipulation – the notions of variance and distance choice – and suggested new means of improving target selection in gesture-to-voice mapping.

Chapter 4: Applying Fitts' law to complex hand gestures

4.1. Overview and interest within the DiVA framework

The design of a new musical instrument is commonly treated with an idiosyncratic approach, on a case-by-case basis [2]. In the specific context of DiVA, defining the input gestural language - which conveys performer's vocal intent - is an intrinsic part of the instrument creation. The current gestural vocabulary has been made up following the same empirical principle. Although such workflow emphasizes on intuition, it lacks of objective methodologies for the evaluation of gestures.

In 2001, Wanderley and Al. [37] proposed a first attempt toward the application of HCI methodologies to the musical domain. However, these researches focused on the evaluation and design of input devices for musical expression (ie. controllers only). With the same approach, we intend to benefit from the substantial HCI literature in evaluating input gestural language with a generalized procedure, allowing for the objective characterization of any new gesture as input control.

We make the general assumption that hand poses can be seen as targets that the performer tries to reach when producing the synthesized voice. This assumption shows its limits as soon as we consider the voice to be emitted as a continuum. Indeed, by nature coarticulation leads to *not* reach the targets, in a strategy of saving motor energy or increasing dexterity. Nonetheless, we assume that a path is interpolated between target "windows" of various sizes, for both vocal and gestural articulation [18]. Such approximation enables the use of existing models of human behavior for target acquisitions and pointing tasks, such as Fitts' law. This model coming from the HCI research offers a strong background for investigating the elements that shape the ease/difficulty of moving between complex poses of the hand, as well as the role of expert skill acquisition and mental representation in the performance of these musical gestures.

After a short presentation of Fitts' law in section 4.2, we clarify our motivations and assumptions in section 4.3. In section 4.4 we describe the context of the experiments carried out in these studies, which are explained in details in sections 4.5 for tasks involving one finger only, and in section 4.6 for tasks involving the whole hand. Finally, directions for future studies are given in section 4.7.

4.2. Background in Fitt's law

Fitts' law is one of the most reliable quantitative models of the human motor behavior. It was first introduced on the basis of ideas from information theory (Fitts, 1954). It predicts that the movement time T to acquire a target actually depends on its width W and its distance D according to the relation (4.1):

$$T = a + b.log_2\left(\frac{D}{W} + 1\right) \qquad (4.1)$$

Where a (sec) and b (sec/bit) are constants reflecting the efficiency of the pointing system. Commonly, Fitts' law is also written as T = a + b x ID, where ID stands for the index of difficulty.

Fitts' law is inherently a 1D model: in its origins the target width was considered along the movement direction, leaving the target height practically at infinity. These two last decades have seen intensive HCI research in modeling 2D pointing [1, 17, 21]. More recently, a very thorough study conducted by Grossman and Balakrishnan [15] investigated pointing at trivariate targets in a 3D environment. Researches in this field are mainly about understanding of the factors that underlie the pointing task (movement amplitude, approach angle, target height, width and depth), or other factors such as the device resistance and the influence of muscle groups [40]. To our knowledge, no previous work focused on Fitts' law validity under a learning constraint.

It is noticeable that due to the usual applications of the model (user interfaces and/or input device design) common HCI studies narrow on the acquisition of graphical targets. Another approach concerns research in the fields of motor behavior and kinesiology, supporting Fitts' law over a wide range of movements and muscles groups [29]. However, to date we are not aware of any work that considers targets as specific hand poses.

Finally, it is interesting to point out that despite Fitts' law is a robust and widely used model, the underlying mechanisms of the speed-accuracy trade-off remain a mystery [24].

4.3. Direction of the current study

This work is first motivated by the need to extend Fitts' law to complex 3D targets, defined here as specific hand postures. In particular, we want to investigate three main assumptions on the effects of borders, distance choice and training on target acquisition performances. Ultimately, the objective is to build a solid foundation for the design of new gestural languages.

Nature and effects of borders

By nature, flexions and extensions of fingers are limited in their range by the motor possibilities peculiar to each individual. When trying to apply Fitts' law to the hand movements directly, one must examine the physical characteristics of specific positions within the motor dynamic. In terms of effort and accuracy, what shapes the gestural space along the entire movement range? Is this space uniform, or are there disparities and particular constraints? More precisely, what are the nature and properties of "borders" in the hand poses? Once again, we propose to take inspiration from HCI works to help in investigating these issues.

Fitts' model implicitly applies for targets within the dynamic range of the input controller (including human possible movements), where one can potentially exceed the target area in the course of movement execution. However, targets on the extremities of movement dynamic are also commonly taken into consideration with the assumption of infinite width. Indeed, movement is then stuck to the border, which leads to remove the target's width constraint. This convenient property is widely applied in graphical user interfaces (GUI) where the most used buttons are placed at the edges and corners of the computer display, thus being particularly easy to acquire since the pointer remains at the screen edge regardless of how much further the mouse is moved.

Intuitively, in the case of hand movements, borders can be seen as the maximal flexion, extension, adduction, abduction, and circumduction of the fingers. In addition, we suggest that any contact between fingers should also be considered as dynamically creating a border. In this case, fingers' movements are blocked, which indeed corresponds to an "edge" situation.

Our intention is to investigate precisely the nature of the borders in the case of hand gestures, and their effects on movement time for pointing tasks.

Distance choice

We have seen in chapter 3 that the choice of a distance which best describes the gestural space has a great importance in the representation and mapping of hand poses. In fact, finding new appropriate distances is a central question in actual research on Fitts' law, especially when inquiring for new model candidates in multivariate pointing [1, 15]. However, there is a major difference in the way the distances can be investigated, between "regular" studies on bi- or tri-variate targets and the present gestural space exploration using a glove as input controller. In the first common case, pointing tasks involve targets that are well-defined, in both 2D and 3D space, and set as independent variables. These conditions enable to explore distance choices *post-hoc*, which is very convenient for the investigation and comparison of different models. On the other hand, our study deals with complex gestural targets, proper to each individual (dependant variables) and using high dimensional data. The characteristics of these targets thus really depend on the distance choice, and consequently it has a direct impact on the selection criteria during the pointing tasks. In this context, it is required to set the distance choice on for all before the experiment.

Our interest in defining and evaluating appropriate distances is double:

- To find a model comparable to the original Fitts' law that would best model the time to acquire a target in the complex gestural space.
- To obtain the best results for target acquisition, in terms of movement time and difficulty.

Assumptions on mental representation

Our main hypothesis is that the formulation and learning of a symbolic gestural language (high level target representation) changes the distances employed in the speed-accuracy trade-off. As part of a probable constant of the motor system in interaction with perceptive system, we suppose that Fitts' law remains valid for modeling the time to reach a "symbolic" and high dimensional target, but only considering a shift from a physical to a "mental" representation of the target. In this framework, both target distances and target widths are to be seen as "mental representations" which would shape the task difficulty.

This assumption meets Moore and Fels' research about human/device interaction, with concepts like intimacy or embodiement [10, 25]. Moore stated: "*the control intimacy determines […] the psychophysiological capabilities of a practiced performer*". Intimacy is to be seen as a deep level of integration and communication with a device (eg. a musician with his/her instrument), and is critical in the perspective of subtle expression of ideas and emotions. Many factors influence the degree of intimacy and the rate at which intimacy grows, such as learning, training and high level representation of control and mapping [10]. In the same way as a skilled guitarist adapts to play a musical piece within timing constraints whichever are finger positions, we argue that training for increasing dexterity in the execution of target gestures, with emphasis on high level target representation, should increase intimacy and lead to a minimization of mental distances. Intuitively, this assumption appears especially meaningful in the performance of complex movements, where a simple mental representation of the gesture pose enables to reach it with more ease.

Such conjecture is yet hard to assess, due to the difficulty to evaluate and access "mental" distances, intimacy and gesture representation. The aim of the present study is not to bring forthright and definitive answers on these issues, but guidelines for further investigations. Especially, we will focus here on the validity of Fitts' law on simple gestures when the accent is put on target embodiment through the learning of symbolic gestures, and express possible directions to study these concepts on more complex gestures.

4.4. Experimentation

4.4.1. Manipulations

We decide to tackle these issues from two extreme perspectives in terms of gesture complexity. The study therefore implies two main phases involving respectively one finger (**1F**) and the whole hand (**H**). The interest of studies involving two fingers (2F) is discussed later.

Each phase consists in a specific Fitts' law experiment with pointing tasks, depending on the assumption to be assessed. Figure (4.1) summarizes the possible manipulations.

Experimentation / Hypothesis	One Finger (1F)	Two Fingers (2F)	Entire Hand (H)
Borders (a)	✓	X	✓
Distance (b)	-	X	✓
Representation (c)	✓	X	…

Fig 4.1: Outline of the experimental manipulations

By running Fitts' experiments on one finger (1F) we intend first to verify that Fitts' law well applies using the CyberGlove as input device. This is made easier by the experimental conditions, which match here the basic Fitts' model with only one degree-of-freedom (dof). To our knowledge, the use of the entire hand (H) for defining the targets in a Fitts' experiment has never been investigated before. Both phases (1F and H) allow for studying the nature and effects of borders (a), which leads us to carry out two distinct experiments (**1F/a**, **H/a**).

Considering the distance choice doesn't make sense with only 1 dof (1F). In future research we intend to investigate this issue with the whole hand in (**H/b**). Finally, we also intend to delve our assumption on mental representations in the simplest case only (**1F/c**). Verifying this last hypothesis on more complex movements is an awkward issue, which cannot be dealt properly before validating the applicability of Fitts' law to the entire hand. *Due to time and space constraints, (**1F/c**) and (**H/b**) experiments will not be approached in this thesis.*

4.4.2. Participants

Subjects are healthy adults, between 22 and 35 years of age, having either normal vision or wearing corrected lenses, and right-hand dominant. Left-hand dominant subjects cannot be included into this experiment, due to a material lack of glove controller for the left hand.

At the time this thesis is written, experiments are still on progress. Therefore only one female and two male volunteers participated to the experiment.

4.4.3. Apparatus

Testing is conducted on MacBook computer with a 13'' LED monitor set to 1280*800 resolution, running Mac OS X (Intel Core 2 Duo 2.4Ghz processor, 4GB RAM). All participants use the same CyberGlove input controller. Software is authored in C++ using Eigen and Openframeworks libraries. It presents trials to participants while logging their hand activities in text files. The software runs full-screen, with a white background color. All other applications and nonessential services are disabled.

4.5. Investigating Fitts' law with one finger: experiment (1F/a)

We make the arbitrary decision to choose the index as the single finger used for the (1F) experiments, focusing more specifically on the second articulation from finger's tip (proximal interphalangeal joint PIP). This choice is motivated first by the configuration of sensors in the CyberGlove, with the presence of a sensor for this specific joint. Secondly, empirical investigations tend to indicate that this articulation:

- Offers the largest range in terms of movement dynamic, with a good accuracy.
- Enable for well-defined borders (maxima in terms of flexion/extension), since it has low flexibility. In comparison, intercarpal articulations are flexible and depend on the movement of the other fingers, which lead to unstable and shifting borders.

4.5.1. Goals

The intention here is to verify whether Fitts' law can be applied to the modeling of speed-accuracy trade-off with one single finger (1F), and in which conditions. In particular, the effects of bounds (extremes of the finger flexion/extension) on movement time are investigated. The empirical constants of the model are determined for our specific input device (CyberGlove). It is a first step in investigating Fitts' law with the entire hand movements.

4.5.2. Procedure and design

First, participants are asked to move the finger on its full range in order to perform the calibration, as shown in figure 4.4. It basically consists in the measure of the maximum range $\Delta = Pos_{max} - Pos_{min}$

After this calibration step, participants perform a conventional sequence of Fitt's reciprocal pointing tasks. Since the range of possible finger movements is unique for each participant, we normalize both the target size and the target distance with Δ. The independent variables are the target size (W: *0.05Δ, 0.1Δ, 0.2Δ*), distance between targets (D: *0.3Δ, 0.55Δ, 0.75Δ*), nature of the target (Border: b, $\bar{b}$) and the movement direction (Dir: *up, down*). A fully crossed design results in a total of 36 combinations of W, D, Border and Dir.

According to the 1-D Fitts model (Eq. 4.1) these conditions comprise 9 distinct IDs ranging from 1.585 to 4.087 bits: {1.585, 2.00, 2.222, 2.662, 2.700, 2.807, 3.170, 3.585, 4.087}

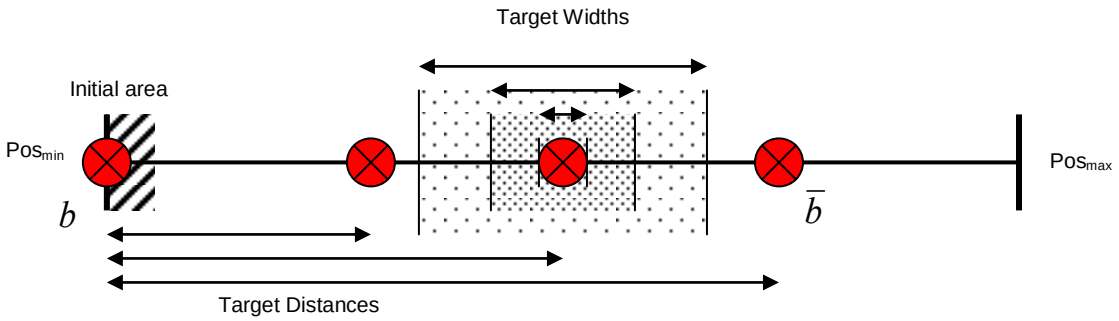

Fig 4.2: Design schematic representation

Note on the calibration

Using Δ for both target distances and width enable to have the same relative amplitudes between participants. However, it also implies a trade-off between finger dynamic and precision (W increases/decreases with Δ), which is not necessarily true.

Note on the choice of independent variables

In order to also investigate the influence of movement direction on the performances, we take the maximal and minimal positions as initial target. It allows us to run the target acquisition in the two opposite direction, in a reciprocal manner.

Note on the reciprocity

Each possible departure area has its center in even Pos_{Min} *or* Pos_{Max}. This implies that every reciprocal task ends with a target on the border, since the participant has to return from the specified target to one of the extremities. This way, there are as many trials with the b condition as trials with the $\bar{b}$ condition.

The experiment includes two sessions: a practice session, to allow participants to get used to the tasks and conditions, and a data-collection session, wherein participants test the 36 different D-W-Border-Orig combinations in a random order. Within each condition, participants perform 8 trials for a total of 36 × 8 = **288 target acquisitions.**

On the screen, the finger position is represented by a vertical cursor (yellow stroke) that moves into a rectangular shape according to the movements of the finger. After each pattern of 48 acquisitions (6 times in total), the participant is encouraged to take a pause if needed. At this time, experimentation completion is also shown.

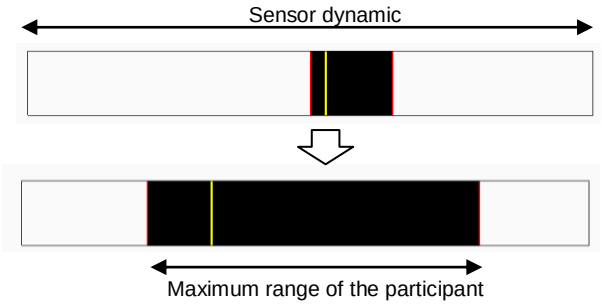

Fig 4.3: Sensor calibration

The participant is asked to reach the target as fast and accurate as possible, in order to not introduce a particular bias toward speed or accuracy. The different steps of a target acquisition are explained in detail in Figure 4.4.

Note on target selection:

Participants move the cursor with movements of their right hand, and select the targets by pushing a button with their left hand. In opposition to “classic” Fitts’ tasks (eg. mouse or stylus “point & click”), the use of both hands to perform the selection may introduce a bias due to the required mental overload. However, we cannot come across any simplest method for the target selection.

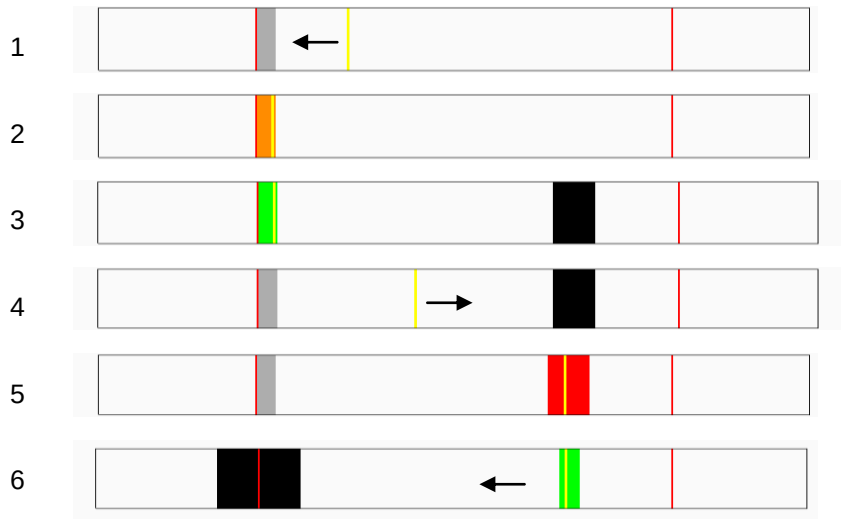

Fig. 4.4: Example of a target acquisition. **(1)** The initial position is displayed in grey. **(2)** Moving the cursor in the initial area changes its color to orange and triggers a random countdown.**(3)** At the end of the countdown, the color swaps in green, the target to reach is displayed in black and sound is played indicating that the initial area can be leaved. **(4)** As the participant moves the cursor out from the initial target, the movement time is incremented until the task end. **(5)** Task end occurs when the participant push the spacebar key to select the target. A visual/audio feedback informs the participant that the target was hit (target turns red) or missed (target turns red + buzzing sound). **(6)** Initial area and Fitts' target swaps, enabling the participant to perform the reciprocal pointing task

4.5.3. Results

The dependant variables are movement time (MT) _defined as the time between leaving the departure initial area and selecting the target in a trial set, and error rate _defined as the average number of errors per trial. Errors occurs when participants clicks when the cursor is outside the target.

Outliers are defined based on MT: any data point further than 1 standard deviation away from its condition's mean is removed. Errors are also removed. A total of 6.5% of the data were removed as outliers.

Analysis of data

We present here some preliminary results, as only three persons participated to the experiment for the moment. Analysis of variance is thus not very relevant, but gives an idea of the main tendencies.

The independent variables D ($F_{2,17}$=9.3 p<.005) and W ($F_{2,17}$=31.8 p<.0001) all have a significant effect on the movement time MT. Recall that for each pointing task, we tested movements in both directions (Dir: down/up, equivalent to flexion/extension). Analysis of variance shows that Dir does not have a significant effect on MT ($F_{1,35}$=0.21 p=.647).

Of particular interest is the effect of the border condition B on MT, which shows to be significant ($F_{1,35}$=64.8 p<.0001). We therefore split the MT data according to the two B conditions, and fit it to Fitts' model (Eq. 4.1) using a least-squares method. Results are shown in Figure 4.5. The R^2 values for the regression are important (>.85): it indicates a good fitting with the model.

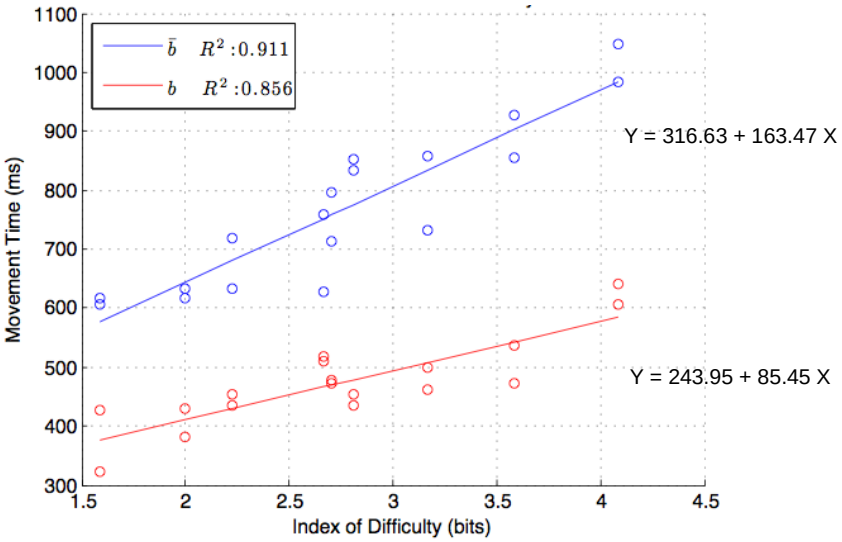

Fig 4.5: Movement time in function of difficulty index for subject 1

These results tend to indicate that Fitts' law validity does not depend on the border condition. In addition, targets in the border are reached significantly faster, and are less affected by target index of difficulty. Therefore, using the CyberGlove as input gesture controller does not obstruct Fitts behavioral model, and paves the way toward investigations of complex hand poses.

4.6. Investigating Fitts' law with the entire hand: experiment (H/a)

4.6.1. Goals

Our intent is to verify whether Fitts' law can be applied to the modeling of speed-accuracy trade-off in the context of complex gestures involving the entire hand. Especially, the effects of gesture borders on performance and model fitting will be investigated.

4.6.2. Design and analysis of hand poses

Dealing with complex hand poses in a Fitts' experiment is not trivial, and involves choosing appropriately the positions to be considered as targets. A major drawback here is that targets cannot be considered as independent variables anymore. The hand poses, peculiar to each participant, are recorded at the beginning of the experiment following the procedure described in chapter 3. Such record phase can be considered as a calibration step, each participant having his/her own hand shape and interpretation of the gestures.

The choice of gestures used as targets is guided by the following considerations:

- Hand poses must be as clear as possible, in order to minimize the cognitive load necessary to visualize the gestures before their performance, and to avoid differences of interpretation between non-specialist participants. Here the experimental design and especially how the gestures are presented to the participants play a central role for their correct interpretation.

- Targets must be distributed uniformly, and along the entire gestural space since we want to reduce the influence of possible (unknown) disparities in this same space.
- Targets must be chosen so that a maximum variance is explained by the three first principal components to enable a rather complete representation in three dimensions.
- Finally, in order to investigate the effects of borders, ultimately a given hand pose has even to entirely respect the border condition or not at all; positions at the crossing (ie. parts of the same gesture reach a bound, other parts don't) must be avoided. We recall that our definition of borders is the maximal flexion, extension, adduction, abduction, and circumduction of the fingers, as well as any contact between them.

We address these issues by choosing 6 specific hand poses, shown in Figure 4.6. Three first positions respond to the border condition; here participants are asked to flex/extend the hand as much as possible. Three last positions respond to the non-border condition; they require being as supple and relaxed as possible in order to not reach any border.

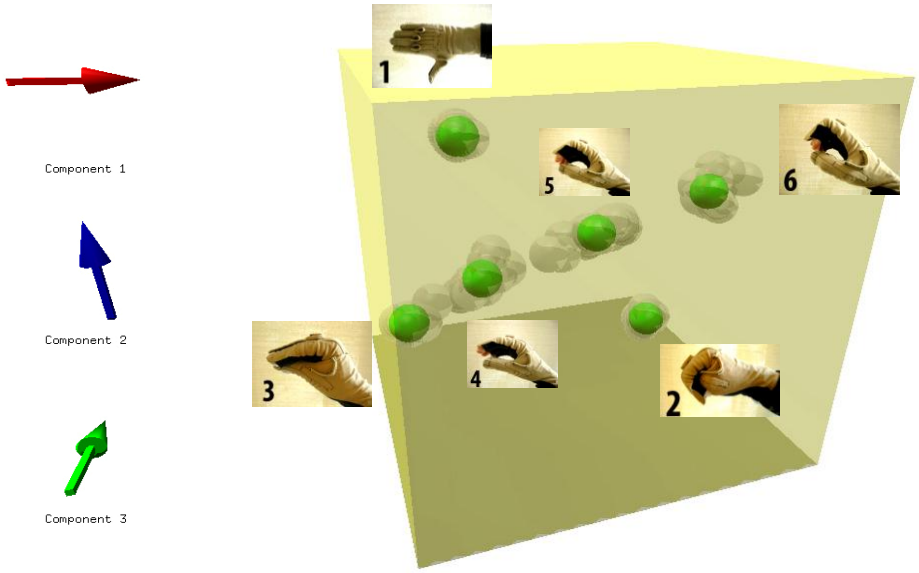

Fig. 4.6: Visualization of the gestural language designed for the experiment in $PC_{1,2,3}$ for subject 2 (in grey the recorded points, in green means of each cloud). Data is projected on the basis built upon movements acquired during the initial record phase, and weighted with eigenvalues; each target is recorded 15 times. Visualization shows a uniform repartition of hand poses in the space, and also highlights disparities in the dispersion of record points between the targets.

A short analysis of the data set (Figure 4.7a) indicates that more than 95% of the variance is explained by the three first axes, enabling the software to display a complete representation of the targets in the cube. This fact is due to the range of movements chosen, principally moving the fingers closer/further from the palm (positions 3, 4, 5, 6). These motions involve flexions/extensions of PIP and MCP joints that are almost only explained respectively by PC_1 (sensors 6, 8, 11, 14) and PC_2 (sensors 5, 7, 10, 13), as illustrated in Figure 4.7b. In a lesser extent, PC_3 explains the movements of the thumb (sensors 1, 2, 4) chosen to reach two extreme borders between the hand poses 1 and 2.

It is interesting to notice that the dispersion of the record points is not constant and shows a great influence of the border condition. It is illustrated in Figure 4.6 and 4.7c were trajectories and targets on borders are concentrated on small area, whereas intermediate positions show a greater dispersion. This point will be discussed in section 3.3.4.

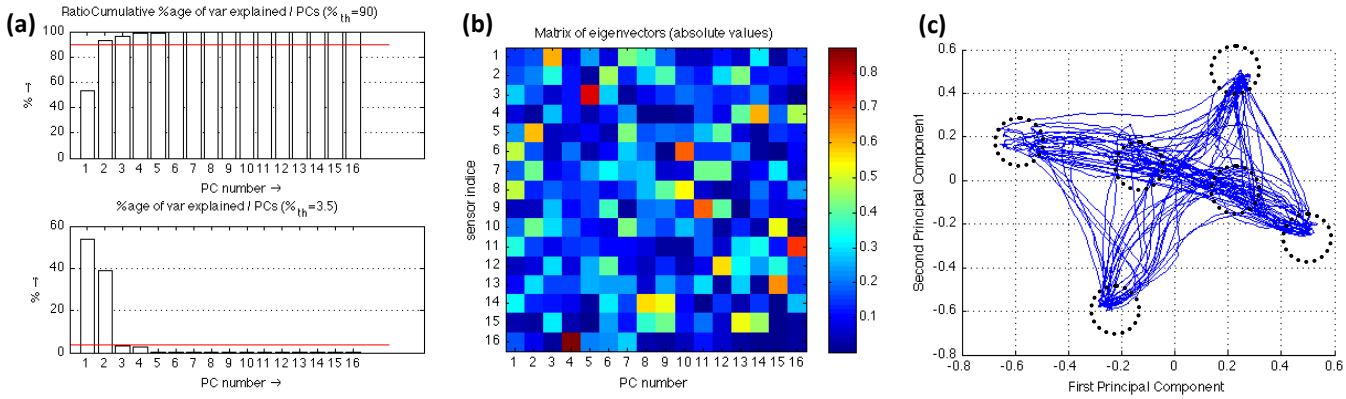

Fig. 4.7: Inertia accumulation and fraction **(a)** and illustration of eigenvectors (absolute values) **(b)** for subject 2's database (≈11K samples). **(c)** Representation of the gesture trajectories in PC_1 and PC_2 during the record session. It is clearly visible that each possible transition between targets (black dashed circles) was visited.

4.6.3. Procedure and design

This study involves two phases: first record of hand poses, then the Fitts' experiment itself.

Before starting the first session, the experimenter distributes instructions[1] explaining the different hand poses. Participants are asked to reproduce them in order to verify their good interpretation, and additional precisions are given if necessary. The record is then realized through an authored application, following the protocol described in section 3.4.4. Each of the 6 targets is acquired 15 times, which leads to record each possible transition 3 times, for a total of 90 selections by the participant.

Means of recorded targets are then displayed into the cube. This representation is shortly explained to the participant, who is asked to reach each target until she/he feels comfortable. The space is weighted with eigenvalues to take into account variance within the data set, which greatly improves the ease to reach targets. Every distance (target radius, distance between targets) thus depends on the recorded data set.

Thereafter the Fitts' law experiment begins, with a sequence of pointing tasks using the recorded targets. In order to balance the target acquisitions equally, each possible transition is visited, giving the following number of distance (D) values: 0.5 x N_{target} x (N_{target}-1) = 15. Distance values depend on the recorded targets, and cannot be set as independent variables. The latter are the target radius (R: 18, 25, 30 "angle" units) and nature of the target (Border:$b, \bar{b}$). There are a total of 15x3x2=90 combinations of D, R and Border that are tested by participants in random order. Within each condition, participants perform 4 trials**, for a total of 360 target acquisitions**.

Different steps of a target acquisition are summarized in Figure 4.8. Participants are asked to be as fast and accurate as possible in reaching the targets. After each pattern of 72 acquisitions (5 times in total), the participant is encouraged to take a pause if needed. At this time, experiment completion is also shown.

[1] Instructions shown to participants are given in appendixes.

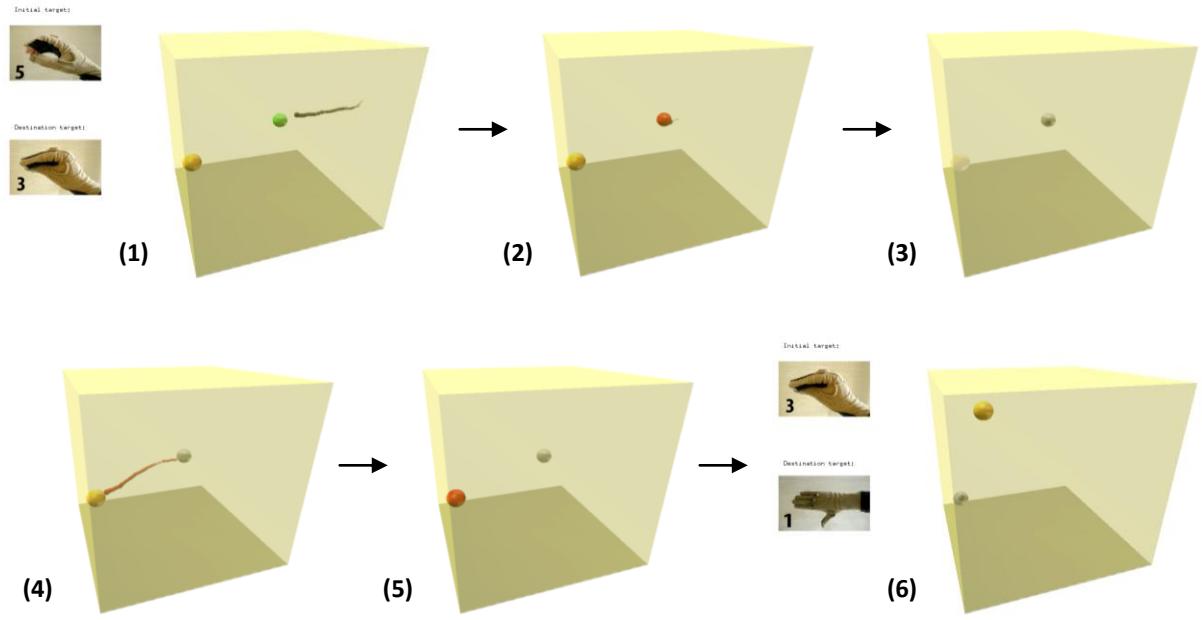

Fig. 4.8: Example of target acquisition. **(1)** Pictures of initial (A) and destination (B) target are displayed, as well as their respective spheres (A: green, B: orange). **(2)** Moving in A turns it red and play a specific sound. It also triggers a random countdown, at the end of which A is displayed in grey, while B flashes on and off **(3)**. Participant moves toward B **(4)** and reaches it **(5)**. Pressing the spacebar triggers next task, keeping B as initial target with a new destination target **(6)**. If the target is missed, a buzzing sound is emitted.

4.6.4. Results

The dependant variables are movement time (MT), error rate and final position (at the moment of selection). Outliers are defined on the same basis as experiment 1F/a. Consequently, a total of 8.2% of the data were removed as outliers.

Analysis of data

Effects of the border condition B on MT are still significant ($F_{1,89}$=139.5 p<.0001). As for the previous experiment (1F/a) we split the data according to the two B conditions. The model used to fit the data is a simple adaptation of Fitts' law (Eq. 4.1), where we replace the target width with the hyperspheres' radius R (Eq. 4.2):

$$T = a + b.log_2\left(\frac{D}{R} + 1\right) \qquad (4.2)$$

The linear regression is realized using a least-squares method. Results are shown in Figure 4.9, and show that using directly the original Fitts' law leads to an important spread of the data. Thus we obtain small R^2 values for the regression (< 0.7), which indicates that better models should be found for applying Fitts' law to highly dimensional data. Considering the overall behavior, we found the same tendencies as in (1F/a): Fitts' law validity doesn't seem to be compromised when using the entire hand, neither to depend on the border condition. Targets in the border are reached significantly faster, and are less affected by target index of difficulty.

These results are consistent with the idea that the gestural space of the hand is not uniform, but shows disparities within its motor dynamic. Our intuitive assumption on the particular ease of movements toward borders is supported by the fact that participants reach them faster and more precisely than

intermediate positions. We believe that this is due to both mechanical constraints, ie. finger movements cannot exceed their dynamic, and “mental” constraints, ie. intermediate and continuous positions are harder to visualize and learn than extreme positions and discrete contacts between fingers.

Once again, such conclusions have to be considered cautiously before carrying out more experiments.

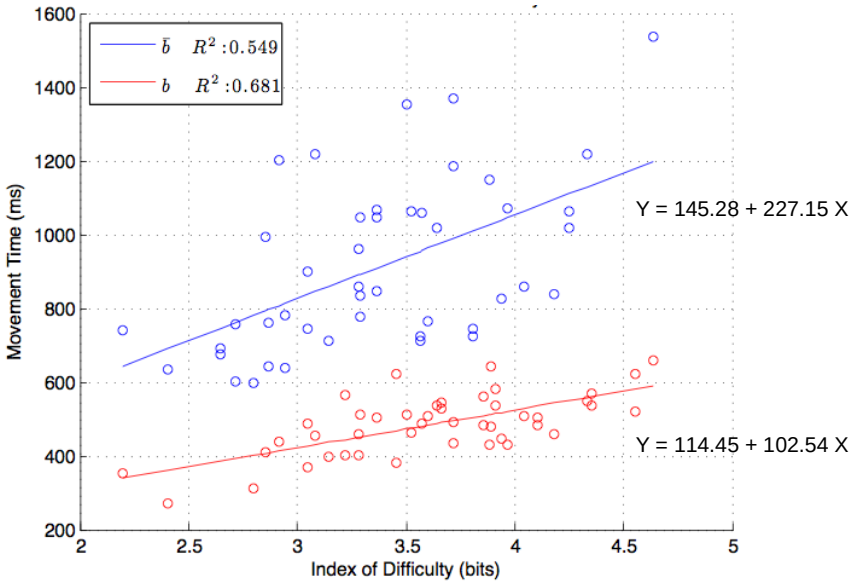

Fig 4.9: Movement time in function of difficulty index for subject 1.

4.7. Toward extensive studies

In order to better understand the mechanisms underlying motor control of hand movements, deeper studies have to be conducted on finger synergies, distances choices, impact of embodiment and skills acquisition.

One judicious way to better understand finger synergies and coupled degrees of freedom is to consider the simplest case of gestural targets defined by two fingers only. This paradigm is at the junction between the experiments presented in this thesis, involving one finger and the whole hand. With two fingers, the complexity in terms of mental representation emerges, which enables to inquire the effects of learning and training (leading to target embodiment) on movement time and accuracy. This latter question will be investigated a near future, under strict experimental constraints on one finger only[1].

Finally, distance choice has to be looked at in detail with the purpose of building new models that better describe the speed-accuracy tradeoff in complex hand gestures.

[1] Considerations on the design of these experiments are presented in the appendixes.

Chapter 5: Conclusion

In this thesis, we focused on the gestural control and mapping of hand gestures in DiVA – a new interface for musical expression – with the aim of synthesizing audiovisual speech and song.

A full analysis & visualization system was developed, allowing for the interpretation and concise representation of complex gestural data by means of a real-time principal component analysis. The relevance of this set of tools was illustrated through a detailed study of the language used in DiVA 2.x. It highlighted possible drawbacks in the choice of hand poses: a reduced and unequal gestural space occupation with an important overlapping between targets. New means of improving target selection in gesture-to-voice mapping were suggested, namely to change the distances employed while emphasizing on uniform distribution of the hand poses. Finally, as a guideline to facilitate the design of new gestural languages, we propose to use our system in an iterative creation process, with a design scheme similar to the prototyping spiral for a human-computer interface (HCI) device [2].

The second contribution of this thesis is the evaluation of hand poses and their associated transitions. Benefiting from research in HCI on the modeling of target pointing, we investigated Fitts' law – well-known model of speed-accuracy tradeoff – in two experiments involving target acquisition tasks with the precise controller used in DiVA for acquiring hand gestures (CyberGlove).

In the first study, participants were asked to move only one articulation (index PIP). Results showed a good fitting with the model, suggesting that Fitt's law can be successfully investigated using the CyberGlove as input controller. We then explored target pointing tasks involving movements of the entire hand, participant being asked to reach complex hand poses. Using the application developed in the first part of this thesis, hand poses were recorded and visualized as targets after a real-time data reduction. Results of this second study tend to indicate that movement time can still be predicted with the target distance and size, despite an important spread of the data. Subsequent research could inquire better models for applying Fitts' law to complex hand gestures.

The specific role of borders – defined as movement bounds and contacts between fingers – was also investigated. In both studies, results tend to indicate that Fitts' law validity is not compromised by the border condition, which has however a significant impact on movement time. Border targets are reached significantly faster than intermediate positions, and are less affected by the task difficulty. This effect of borders should be taken into consideration when designing new gestures for driving the speech synthesis in DiVA, as it has a direct impact on the ease for producing the mapped phonemes. We believe that border positions foster gestures embodiment as they are particularly salient in the sense of their ease to be memorized and performed. In this context, future studies within the DiVA framework should inquire the effects of learning and training on gesture performance, with the ambition to push further the design and potential expressivity of new gestural languages.

References

[1] J. Accot and S. Zhai, “Refining Fitts' law models for bivariate pointing”, *ACM CHI*. p. 193-200, 2003.

[2] N. d’Alessandro, “RAMCESS: Realtime and Accurate Musical Control of Expression in Voice Synthesis”, PhD Thesis, Université de Mons, 2009.

[3] D. Arfib, J. M. Couturier, L. Kessous, and V. Verfaille, “Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces”, *Organized Sound*, vol. 7, no. 2, pp. 127–144, 2002.

[4] G. Austin, “Chironomia, or a Treatise on Rhetorical Delivery”, London: 1806. Ed. Mary Margaret Robb and Lester Thonssen. Carbondale, IL: Southern Illinois UP, 1966.

[5] N. Biggs, E. Lloyd and R. Wilson, “Graph Theory 1736-1936”, Clarendon Press, Oxford, 1976

[6] C. Cadoz and M. Wanderley, "Gesture-Music" in *M. Wanderley and M. Battier, eds. Trends in Gestural Control of Music*, IRCAM – Centre Pompidou, 2000.

[7] F. Craik and R. Lockhart, “Levels of processing: A framework for memory research”. *Journal of Verbal Learning and Verbal Behavior 11*, 671-684, 1972.

[8] C. Duby, S. Robin, “Analyse en composantes principales“, Institut National Agronomique Paris – Grignon, 2006.

[9] S. Fels and G. Hinton, “Glove-TalkII: A neural network interface which maps gestures to parallel formant speech synthesizer controls”*, IEEE Trans on Neural Networks*, Vol 9, No. 1, pp. 205-212, 1998.

[10] S. Fels, “Intimacy and Embodiment: Implications for Art and Technology” in *Proc. ACM Workshops on Multimedia*, pp. 13–16, 2000.

[11] S. Fels, F. Vogt, K. van den Doel, J. Lloyd, O. Guenter, “Towards realizing an extensible, portable 3D articulatory speech synthesizer”. *International Workshop on Auditory Visual Speech Processing*, pp. 119–124, 2005.

[12] S. Fels, B. Pritchard, A. Lenters, “ForTouch: A Wearable Digital Ventriloquized Actor” in *Proceedings of the New Interfaces for Musical Expression (NIME09)*, pages 274-275, 2009.

[13] P. Feyereisen, J-D Lannoy, “Gesture and Speech: Psychological Investigations”, Cambridge University Press, Cambridge, 1991.

[14] W. Gray, W. Fu, “Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head”. *Cognitive Science 28*, 359-382, 2004.

[15] T. Grossman, R. Balakrishnan, “Pointing at trivariate targets in 3d environments”, *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 447–454, 2004.

[16] J. Hardcastle and N. Hewlett, “Coarticulation: Theory, Data and Techniques”, Cambridge: Cambridge University Press, 1999.

[17] E. Hoffmann and I. Sheikh, “Effect of varying target height in a Fitts movement task”, *Ergonomics*. 37(6), p. 1071-1088, 1994.

[18] P. Keating, "The window model of coarticulation: articulatory evidence", in *Papers in Laboratory Phonology I* , ed. J. Kingston & M. Beckman, Cambridge University Press, pp. 451-470, 1990.

[19] L. Kessous, “Gestural control of singing voice, a musical instrument”, in Proceedings of the 2004 Conference on Sound and Music Computing (SMC’04), IRCAM, Paris, France, 2004

[20] S. Le Beux, C. d'Alessandro, A. Rilliard, B. Doval, “Calliphony: A system for real-time gestural modification of intonation and rhythm”*,* Speech Prosody, Chicago, 2010.

[21] S. MacKenzie and W. Buxton, “Extending Fitts' law to two-dimensional tasks”, *ACM CHI*. p. 219-226, 1992.

[22] S. MacKenzie, “Fitts' law as a research and design tool in human-computer interaction”, in *Human-Computer Interaction*, volume 7, pp. 91–139, 1992.

[23] S. MacKenzie, “Movement time prediction in human-computer interfaces” in R. M. Baecker, W. A. S. Buxton, J. Grudin, & S. Greenberg (Eds.), 2nd ed., pp. 483-493, 1995.

[24] S. MacKenzie and R. W. Soukoreff, “An informatic rationale for the speed-accuracy tradeoff” *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* – pp. 2890-2896, 2009.

[25] F. Moore, “The dysfunctions of midi”, *Computer Music Journal*, 12(1):19-28, 1988.

[26] A. Mulder, “Virtual Musical Instruments: Accessing the Sound Synthesis Universe as a Performer”, in *Proceedings of the First Brazilian Symposium on Computer Music*, 1994.

[27] M. Nazari, P. Perrier, M. Chabanas & Y.Payan, “Simulation of dynamic orofacial movements using a constitutive law varying with muscle activation”, *Computer Methods in Biomechanics & Biomedical Engineering*, 2010.

[28] Ş. Özçalışkan and S. Goldin-Meadow, “Gesture is at the cutting edge of early language development”, *Cognition*, *96* (3), B101-B113, 2005.

[29] R. Plamondon, A. Alimi, “Speed/accuracy trade-offs in target-directed movements”, *Behavioral and Brain Sciences,* 20, 279–349, 1997.

[30] B. Pritchard, S. Fels, “GRASSP: gesturally-realized audio, speech and song performance”, *NIME06*, 272-276, 2006.

[31] J. Ramsay and B. Silverman, “Functional Data Analysis”, New York, Springer-Verlag, 2005.

[32] J. Rye and J. N. Holmes, “A Versatile Software Parallel-Formant Speech Synthesizer,” *Joint Speech Research Unit Report*, no. 1016, 1982.

[33] W. Sandler and D. Lillo-Martin, “Sign Language and Linguistic Universals*”.* Cambridge, UK: Cambridge University Press, 2006.

[34] A. Savard, "When Gestures are Perceived through Sounds: A Framework for Sonification of Musicians’ Ancillary Gestures", M.A. thesis, McGill University, 2008.

[35] VisualVoice, http://www.magic.ubc.ca/artisynth/pmwiki.php?n=VisualVoice.HomePage.

[36] M. Wanderley, “Gestural Control of Music”, *Proceedings of the International Workshop on Human Supervision and Control in Engineering and Music*, Kassel, Germany, pp.101-130, 2000.

[37] M. Wanderley, N. Orio, and N. Schnell, “Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI,” *Computer Music Journal*, vol. 26, no. 3, pp. 62–76, 2002.

[38] R. Wilson, R. Balakrishnan and G. Sethuraman, “Graph theory and its Applications”, in Proceedings of a Conference at Anna University, Chennai, India, Narosa Publ., pp 188, 2004.

[39] G. Wulf, “Attention and motor skill learning”, Champaign, IL: Human Kinetics, 2007.

[40] S. Zhai, “Human performance in six-dof input control”, PhD Thesis, University of Toronto, 1995.

[41] S. Zhai, J. Kong and X. Ren, “Speed--accuracy tradeoff in Fitts' law tasks - on the equivalency of actual and nominal pointing precision”, *International Journal of Human Computer Studies*, 823-856, 2004.