When Gestures are Perceived through Sounds: A Framework for Sonification of Musicians' Ancillary Gestures

Alexandre Savard

Music Technology Area
Schulich School of Music
McGill University
Montreal, Canada

February 2009

A thesis submitted to McGill University in partial fulfillment of the requirements of a degree of Master of Arts in Music Technology.

2009/02/03

Your file Votre référence
ISBN: 978-0-494-66989-1
Our file Notre référence
ISBN: 978-0-494-66989-1

Abstract

This thesis presents a multimodal sonification system that combines video with sound synthesis generated from motion capture data. Such a system allows for a fast and efficient exploration of musicians' ancillary gestural data, for which sonification complements conventional videos by stressing certain details which could escape one's attention if not displayed using an appropriate representation. The main objective of this project is to provide a research tool designed for people that are not necessarily familiar with signal processing or computer sciences. This tool is capable of easily generating meaningful sonifications thanks to dedicated mapping strategies. On the one hand, the dimensionality reduction of data obtained from motion capture systems such as the Vicon is fundamental as it may exceed 350 signals describing gestures. For that reason, a Principal Component Analysis is used to objectively reduce the number of signals to a subset that conveys the most significant gesture information in terms of signal variance. On the other hand, movement data presents high variability depending on the subjects: additional control parameters for sound synthesis are offered to restrain the sonification to the significant gestures, easily perceivable visually in terms of speed and path distance. Then, signal conditioning techniques are proposed to adapt the control signals to sound synthesis parameter requirements or to allow for emphasizing certain gesture characteristics that one finds important. All those data treatments are performed in realtime within one unique environment, minimizing data manipulation and facilitating efficient sonification designs. Realtime process also allows for an instantaneous system reset to parameter changes and process selection so that the user can easily and interactively manipulate data, design and adjust sonifications strategies.

Acknowledgments

First, I would like to thank Marcelo Wanderley, my supervisor, who introduced me to this fascinating research topic which was previously unknown to me. His enthusiasm has been incredibly encouraging during these years. Many thanks to my co-supervisor Vincent Verfaille without whom this work could not have been completed. The advices and comments he provided me with transcend the scope of this academic work. His tremendous MATLAB and LATEX expertise has been immensely helpful throughout every aspect of this project. Thanks to Vincent, Marcelo, and Bertrand Scherrer for proofreading this thesis and especially to Rebecca Barnstaple and Alexia Moyer for their wonderful help regarding the final editions. Thanks to the people in Music Technology, professors and students, for their welcoming attitude. Finally, special thanks to my family for their understanding and inestimable support.

Abrégé

Ce mémoire présente un système multi-modal de sonification combinant la vidéo conventionnelle à des sons de synthèse générés à partir des données provenant de systèmes de capture de mouvements. Ce système permet une exploration rapide et efficace des données de gestes auxiliaires de musiciens, pour lesquels la sonification complémente la vidéo en mettant l'accent sur certains détails qui peuvent au premier coup d'oeil passer inaperçus. L'objectif principal de ce projet est d'offrir un outil de recherche à des utilisateurs qui ne sont pas nécessairement familiers avec le traitement du signal ou encore les sciences informatiques. Cet outil permet d'obtenir aisément des sonifications significatives, grâce à des stratégies de mapping dédiées. D'une part, la réduction de dimensionalité des données recueillies à l'aide de systèmes de capture de mouvement tel le Vicon est fondamentale car elles peuvent excéder 350 signaux décrivant les gestes. Ainsi, l'utilisation d'une Analyse en Composantes Principales permet de réduire objectivement le nombre de ces signaux à un sous-ensemble contenant l'information la plus pertinente en termes de variance. D'autre part, les données de mouvements présentent une grande variabilité entre les sujets: des paramètres additionnels de contrôle de la synthèse sonore sont donc offerts dans le but de restreindre la sonification aux gestes significatifs, décelables visuellement en terme de vitesse et de distance parcourue. Ensuite, des techniques de conditionnement du signal sont proposées pour adapter les signaux de contrôle aux valeurs requises pour les paramètres de la synthèse sonore, ou encore afin de permettre à l'utilisateur de mettre l'emphase sur l'information qu'il juge importante. Toutes ces traitements des données s'effectuent en temps-réel à l'intérieur d'un seul environnement, ce qui minimise la manipulation des données et facilite la conception d'une sonification efficace. Le temps-réel permet aussi une réponse immédiate du système aux changements de paramètres et selection de traitements, si bien que l'utilisateur peut aisément manipuler les données de manière interactive, élaborer et ajuster des stratégies de sonifications.

Contents

List of Figures

Chapter 1

Introduction

Gestural research, which explores human body biomechanical capabilities and communication potential, is a flourishing field of study. Specifically, the study of movement during a musical performance provides a strong basis for investigating gestures' spatial and temporal organization. For several instrumentalists performing the same musical excerpt of standard repertoire, one can compare performers' gestures under strict experimental constraints.

Ancillary gestures refer to movements not directly related to sound production. These gestures are part of the performance as they contribute to the overall information communicated by the performer to the audience. Analyzing ancillary gestures, which convey this hard-to-quantify information, provides insight about generation and perception of body movements.

1.1 Motivation

When using sonification to study ancillary gestures, researchers are especially interested in investigating both the interaction between different types of gesture and the periodic gestural patterns. Scanning large video footage while looking for relevant excerpts is time consuming. Increasing the playback rate of videos is one solution to make this preliminary investigation process faster.

As human capacity to integrate speed-up information contained in videos is limited, cognition overload may occur and important details can be missed. One way to emphasize visual information is to use supplementary auditory feedback as gestures occur in the video. By using multi-modal representation, information is more easily integrated as attention can

be focussed on specific details while listening to the global information stream.

Furthermore, gestures' periodic patterns may occur over long periods or be masked by less significant gestures. Using the combination of visual and auditory displays, speed-up data is likely to have higher potential to reveal periodic patterns than video only.

Finally, sonification can be used during realtime motion capture sessions to provides performer with direct feedback concerning their gestures. It then becomes a pedagogical tool, as performers can investigate their stage presence through sonification of their gestures.

1.2 Project Overview

Interest in sonification, and gesture sonification specifically, has recently increased dramatically. Researchers interested in using sonification to obtain a prior awareness concerning gestural data currently need to develop their own sonification tools as there is no consistent environment specifically adapted for gestural research. The researcher is then required to be familiar with digital audio signal processing and computer science, that can explain why sonification is not yet a widely used gestural data exploration technique. In this thesis, a full sonification system was developed allowing for the multi-modal representation of gestural data.

A previous project at the IDMIL[1] proposed a sonification strategy for musicians' ancillary gestures [1]. The sonification was achieved using two different programming environments: MATLAB (for offline data manipulation and pre-processing), and MAX/Msp (for audio processing) in realtime. The sonification was subject-specific without any control parameters allowing modification of the resulting sound synthesis. The first goal of this thesis is to implement both data processing and audio generating algorithms in real-time to allow for instantaneous feedback of parameter modifications, which is also called interactive sonification.

Modern motion capture systems consistently describe body movements by generating several signals (possibly over 350) to describe positions and orientations of every part of the body. This large quantity of data requires some prior expertise concerning human movements in order to find the most relevant information to be sonified. As one intention of gesture sonification is to simplify prior data analysis, this expert-knowledge required to

[1]Input Devices and Music Interaction Laboratory, McGill University

design an adequate system contravenes its efficiency; the second goal of this thesis was then to reduce data manipulation in sonification design. Principal Component Analysis was used to objectively perform a reduction of the amount of data.

Since different performers may not emphasize the same gestures in terms of velocity or path distances, signal conditioning is required to adapt the control signals to the sound synthesis parameters. Important information may be masked by less important gestures performed during a motion capture session. This situation specifically occurs when one is interested in comparing several performers, some performing less explicit gestures than others. As a third contribution of this thesis, non-linear warping curves are used to adapt signals within ranges where they can be compared. Warping techniques are also used to emphasize certain gesture characteristics, making them easily perceivable.

1.3 Thesis Overview

Chapter 2 describes fundamentals concerning sonification. It discusses four works stressing different aspects related to gesture sonification. In chapter 3, details are presented concerning the gestural data processing and sound synthesis algorithms implemented in our sonification system. These algorithms were introduced in a previous project and have been implemented as realtime data processing techniques in this project. Chapter 4 describes the parameters used to restrain sonification to significant information that actually describes gestures. The data reduction model is presented in chapter 5, which is at the heart of the system's ability to efficiently explore the data. Finally, Chapter 6 presents different signal conditioning techniques to arbitrary emphasize important information and adapt signal to sound synthesis parameter requirements. Chapter 7 presents the final sonification system developed during this thesis. Chapter 8 brings the overall conclusion and a discussion concerning a number of improvements to the sonification system that are suggested as future works.

Chapter 2

Background

The field of sonification is clearly an interdisciplinary research area merging theoretical and technical knowledge from various areas of expertise. It builds upon an extensive foundation of psychological research in sound perception or cognition and contributes to our understandings of information acquisition. Another aspect of research in auditory displays relates to the design of tools and sound production algorithms proper to diverse practices of sonification. Finally, a last area in this research field concerns sonification design itself and its applications to different situations and contexts. This chapter briefly covers key concepts associated with sonification and its application to performers' ancillary gestures.

2.1 Issues in sonification

Limits of representation

The amount of data in our computational world is rapidly increasing. Available media technologies are now powerful enough to manage impressively large data sets allowing analysis of problems and phenomena involving several computational dimensions. Climate and atmospheric phenomena are, for example, are currently described in terms of temperature, atmospheric pressure, and humidity level, distributed through extremely vast volumes of gaseous fluids [2]. Scientists using computers to inspect such a large amounts of computationally processed results are now facing the problem of integrating data information in order to extract meaning.

To support analysis and provide insight into the data set, visual representations using

tools such as graphs or videos are traditionally used. Unfortunately, visual representations alone is inadequate to describe complexity within data sets, especially high dimensional ones. Relevant information may hide under several data layers; information that would bypass the understanding of the researcher if not displayed using an appropriate tool. A major issue in data representation is to efficiently manage the problem of cognition overload [3]. Large quantities of data may quickly become perceptually confusing for the researcher due to a natural limitation in our abilities to understand multiple streams of information. Nevertheless, visualization is still an active research field in data representation, but alternate approaches may also be examined.

An alternate representation

Research in sonification can be considered analogous to research in data visualization. The idea behind sonification is to complement (or simply replace) visual representation of data by auditory cues, generated from the data itself, which provides information about its content. Sonification can be defined as "the transformation of data relations into perceived relations in an acoustic signal for the purpose of facilitating communication or interpretation" [4]. In other words, sonification is the art of making data audible, at the same time providing an additional means for its exploration as a complement to other perceptual channels such as vision. In this present context, gesture data may then be assimilated by a human analyst using visual displays, auditory displays, or combinations of both.

These alternate representations of data became possible with rapid development in the last decades of techniques that allow for effective computer generated sounds. Perhaps the best known example of sonification is the Geiger-counter, used to detect invisible radiation levels. In more recent sonification history, the "quantum whistle" [5] refers to another very successful example. The physicists involved in this project claimed that sonification proved itself to be very efficient in clarifying evidence of hard-to-visualize quantum oscillations using a traditional oscilloscope. Once sonified, the periodic behavior of the experimental results became clearer.

2.1.1 Psychoacoustics and multimodal perception

Foundation in psychology of sound perception

Characteristics of the hearing channel may suggest answers to many relevant questions concerning scientific displays. There exists extensive research literature regarding intensity, frequency, and temporal discrimination of sounds [6]. The ability of the human ear to differentiate within and between several sound features is of fundamental importance in sonification. The ability of the auditory system to distinguish close frequencies (3 – 5 cents) or close abrupt temporal changes (50 ms) have been well documented. Even if non-objectively quantifiable, timbre can play an important role as it introduces several ways to characterize auditory stimuli such as brightness or noisiness. Spatial localization is another aspect of the auditory system that can be exploited for the purpose of sonification.

Furthermore, the auditory channel demonstrates astonishing abilities to integrate several streams of information. Integration refers to the aptitude to make the distinction between different layers of non-similar sounds. Polyphonic music is probably one of the richest examples of superimposed melodic streams. In psychoacoustics, two main types of auditory integration are distinguished: grouping (vertical and spectral integration) and streaming (horizontal and temporal integration) [7]. One must understand the characteristics of sound perception in order to make a clear relation between the data and the appropriate sound features. However, despite flourishing literature in this area, there is still no objective method for determining the best strategy to translate data relations into sounds [4]. Bregman proposes that sonification should follow the rules derived from the theory of "Perceptual Streaming" [8] in order to create auditory scenes.

Foundation in multimodal integration

As previously stated, sonification is also used in the field of multimodal representations once combined to visualization. There is strong evidence of interaction between different perceptual channels [9]. A well known example demonstrating this last issue is the McGurk effect [10]. Wrong combinations of consonant speech sounds and videos showing lip movements of different consonants lead the listener to understand inaccurate consonant sounds. For instance, hearing the sound "da" while looking at lips pronouncing "ka" misleads the observer/listener to understand "ga".

Following this idea of perceptual interference between different modalities, it has been

argued that the amount of visual information that can be simultaneously processed can be increased by using properly designed multimodal representation [11]. Numerous psychological studies in multimodal perceptual integration refer to several criteria summarized in [12]. Among the most important are: temporal stimulus coincidence, spatial stimulus contiguity, or similar stimulus duration and intensity. Multimodal displays benefit from this interaction between perceptual channels to enhance human perceptual efficiency. Sound may convey additional information not necessarily easily observable, such as force or spin. An inherent problem in visual displays such as numerical values or graphs is that they require a large amount of attention from the user. Sonification may be used to free cognitive processes required to perform visual analysis.

2.1.2 Sonification tools

Sound and meaning

Sonification tools are designed in order to create sounds that are intentionally meaningful [13], [11]. The choice of appropriate sounds to depict specific attributes related to data is influenced by the sound features to be exploited. In this context, sounds are considered as vectors that convey information through the hearing channel, such as words or music. However, a huge distinction remains between these two means of communication and sonification, related to the fact that arbitrary rules in both music and speech ensure a proper exchange of information, which is not necessarily the case in sonification [13]. Non-speech sounds would be preferred, as they do not convey idiosyncratic information such as vocabulary or tonal harmony. In sonification, sounds acquire their meaning by identifying and modifying specific features that exploit the previously cited characteristics of human hearing.

Sonification for data exploration and monitoring

So far, a particular attention has been paid to psychological issues related to sonification. Another important aspect of sonification is the actual design of the sonification tools, which must take into account several details. These aspects are ruled by psychoacoustic principles that were previously enumerated; other aspects relate to the system's internal structure. These criteria are determined by the intended use of the system. There are two major approaches for sonification tool design: exploration and monitoring [3]. Any sonification

tool, however, deals with two fundamental system components: data-related processing and the actual display which includes various techniques to generate sound. The current project introduces a framework for monitoring musicians' ancillary gestures which is flexible enough to be adapted to different individuals' specificity.

Interaction in auditory displays

Considering the last assertion, another aspect of sonification system design stresses its capacity to control the sonification process. Recent trends focus on the use of dynamic human interaction to explore data sets while they are being transformed into sound [14] [15]. Here, sonification enters the realm of human-computer interaction. Manipulating the data in realtime improves the potential for exploration [3]. Realtime gives the user instantaneous auditory feedback on his/her actions. This is a highly desirable property, since it allows quick and intuitive learning of overall sonification reactions to parameter changes. An instantaneous response to any parameter modification provides indirect but useful information about the data, which may leads user to further improvements of the sonification design. This is an important reason for a system to process information and produce the sonification in realtime. The interaction can occur on different levels of the sonification process, ranging from data normalization and scaling to the optimization of sound synthesis parameters [16]. Choosing the adequate parameter set becomes an integral part of the sonification process [17].

2.2 Gesture sonification

The concept of gesture covers a substantial area, combining expertise in different fields. It includes not only physiological aspects but issues that refer to communication and meaning. Musicians' ancillary gestures specifically refer to performers' body movements that are not directly involved in instrumental sound production. Although our understanding of gesture is growing, analysis of human motion remains challenging due to its high dimensionality and variability for both intra-subject and inter-subject considerations. An advantage in using sonification is that it may help highlight hidden information that would be hardly detectable otherwise. For instance, looking exclusively at fixed angle videos (i.e. videos showing the same side of the performer) augment the risk of missing important movement details.

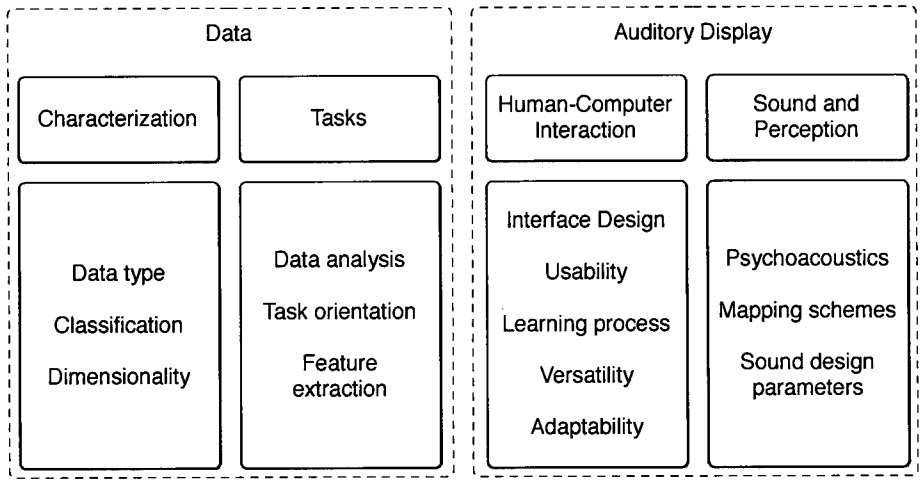


Fig. 2.1 Interaction parameters of a model proposed in [18] and applied to sonification of gestures.

Previous research tends to demonstrate a strong correlation between the performers' ancillary gestures and the musical score [19] [20]. The same instrumentalist will tend to perform similar ancillary gestures for the same musical excerpt, remaining coherent over time. This correlation suggests a direct relation between some gestures and the performed music. They are an implicit part of the learning process. Observing different performers, similar movements are, at some level, even noticed. These similarities are due to several factors: material/physiological (respiration, fingering), structural (rhythm, melodic contour), or interpretative (performer's mental model) [21]. Performers exhibit a tendency to group these gestures according to structural elements of the musical piece. Musicians refer to this as phrasing. The movements of the performer often coincide with rhythmic structures suggested by the music and gesture cluster frequently occurs at the end of the musical phrase.

2.2.1 Gestural control of sound synthesis and sonification

Arbitrariness of mapping

The frontier between gesture sonification and music-oriented computer generated sound, controlled by gesture, is definitely narrow. Both use the output provided by gestural

acquisition systems to generate auditory events that are somehow related to the gestures that produced them. However, these two disciplines strongly differ in the significance carried by their respective sound events. While in music the composer' intention is subtle, sometimes even hidden, sonification, prefers an explicit rendition of a gesture in order to make the association between this gesture and its corresponding sound feature as obvious as possible.

Digital Musical Instruments consist of controllers that interface human actions with computational music related processing [22]. They are principally characterized by the dichotomy between the input interface and the sound generation system (if compared to traditional musical instruments). This is the same for sonification, where data obtained from the motion capture system is an input that obviously does not produce any sound by itself. Mapping of gesture information to computer-generated sound receives increasing attention in the field of musical input devices. Digital musical instrument designers investigate conscientiously various mapping schemes and their related potentials[23].

[24] proposes a bidirectional mapping model between a *related-to-gesture* perceptual space and a *related-to-sound* perceptual space. In a musical context, artistic aspiration can give rise to specific requirements in terms of mapping. It is not the case in gesture sonification where most of the prerequisites relate to *related-to-gesture* characteristics. Ideally, the choice of sound synthesis should be made according to the type of gesture to be depicted by the sonification. However, interference due to auditory perception characteristics may occur. Among these constraints, an obvious one is related to streaming of simultaneous sound. The fundamental frequencies must overlap carefully using strategies like different panning angles or different timbres so that the different sounds are still distinguishable.

Motion capture data mapped to music

A motion capture system can be adapted to be used as a musical performance instrument. There have been various attempts to establish a methodology for control of parametric sound synthesis through motion capture systems. One of these attempts[1] describes a system based on motion capture, designed to control a musical soundtrack and computer generated animations rendered from the gestural data of a dancer [25] [26].

[1]The attempt considered is definitely not sonification. It is however included here since the authors faced several problems that are common to both approaches.

The authors used a Vicon motion capture system with a model built on 30 marker positions describing the overall movements of dancers. As a prior assumption, the raw data obtained may contain a high information potential regarding the expressiveness of the dancers. However, they stipulate that a process is needed in order to significantly reduce the large amount of data (900 events/second, 30 markers at 30 Hz) without loosing salient information.

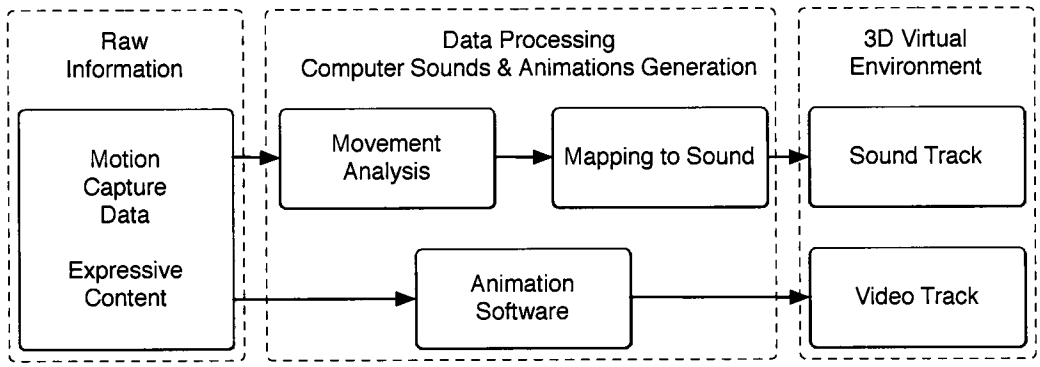


Fig. 2.2 Motion capture data mapped to music and animation in order to translate dance movements into auditory and visual computer generated events (from [25]).

They propose two contrasting methods to extract information from the data frames. The first one is based on segmenting the data frames so each segment contains one significant gesture. Abrupt changes in marker positions result in peaks in the acceleration that are then used as temporal marks for each segment. A number of parameters can be derived: the overall time of a given segment (i.e. time to perform this gesture), the total distance, the average velocity, or the ratio between the total distance and the direct distance (i.e. evaluation of the curvature).

The second method involves pattern recognition and principal component analysis (PCA). In this method, movements are analyzed and classified according to their similarity to previously classified sets. Gestures are considered as a sequence of states. The position of each marker is described according to the body's center of mass. PCA is performed to derive the most important features of the data set. Movements are recognized by computing their euclidean distance with a training set previously processed.

2.2.2 General framework

Sonification using EMG sensors

Another study looks at a realtime sonification toolkit developed to aid the analysis of a general data set [27]. The authors aim to build a multi-disciplinary framework for data sonification experiments. As a first application of the toolkit, two specific data sets are considered: helicopter flight recording and physiotherapy data provided by EMG sensors. The latter is of particular interest here since EMG sensors provide information about muscular activity[2], and thus, about human body movements.

The Interactive Sonification Toolkit has two fundamental functions derived from the types of interaction discussed in section 2.1.2. The first allows the user to manage the data streams while the second enables the conversion of the streams into sound. Once the data are mapped with sounds, it is possible for the user to interact with the sonification in realtime.

The user interface consists of two main parts: the data scaling page and the interactive sonification page. The data is streamed into several "sonification channels" that allow for several simultaneous sonifications to be defined. Within the data scaling page, the user can determine the input data, the normalization parameters, and specifications related to pitch shifting or stretching of the data. The second part of the interface is the "sonification page" that provides a selection of sound generation techniques as well as control over different synthesis related parameters.

While optimized for EMG sensor data, this sonification does not provide any information on the gestures. However, the architecture of the system is well suited to movement data exploration. It is reportedly user-friendly and allows a very specific data processing strategy as well as a more flexible approach in a perspective of data exploration.

Framework for sonification of motion capture data

A completely different work describes three software solutions to sonify human motion as a framework [29]. The authors claim that it results in a very fast prototyping environment for sonification. They use the Vicon motion capture system to track movements. The software part is principally composed of Chuck used to interface both Marsyas and STK, two object

[2]A previous attempt to sonify muscular activity through EMG sensor data was presented in [28].

oriented programming libraries related to sound analysis and generation. They discuss about sonification using their framework for data of different fields including traditional performance on musical instruments, performer acting out emotion using their body, and data from individuals having impairments in sensory-motor coordination.

Sonification using their framework is a three part process: data collection with the Vicon system, data import into the proper synthesis language, and finally the sonification algorithm selection. Marsyas is a collection of feature extraction techniques for signal processing that are suitable for marker movement. It also makes possible implementation of machine learning and classification algorithms. STK (the Synthesis ToolKit) has been developed to provide an extended framework of algorithms involved in sound synthesis. The last component of the framework, Chuck, is a concurrent programming environment that incorporates precise control over time and thus, realtime reliability. Different modules manage independent processes simultaneously. Different parts of the data can be added during runtime in a synchronized way. It gives the user the opportunity to modify sonification at every moment.

In order to provide a tool as flexible as possible in terms of sonification design, the system adopts an open structure (i.e. every algorithm or component can be interfaced). As a drawback, it lacks a consistent visual interface. Chuck is mostly a command line interface. Users need extensive knowledge about signal processing and computer programming to be able to achieve advanced results. Furthermore, Chuck is an alternate audio specific programming language that requires a learning period even for proficient programmers. All these requirements added to the prerequisites about sonification in general and physiological knowledge make this system unsuitable for users that are neither experts in audio signal processing or computer sciences.

2.2.3 Sonification based on expert knowledge

Sonification of sport movements

In the area of human gesture analysis, sports are a very rich and diversified source of investigation. They provide a number of situations where very specific gesture combinations have to be executed in order to achieve a precise task. A recent research on sonification of sport movements [12] [30] presents relevant arguments to strengthen the reasons in favor of the usage of multimodal representation in research related to gesture, combining videos and

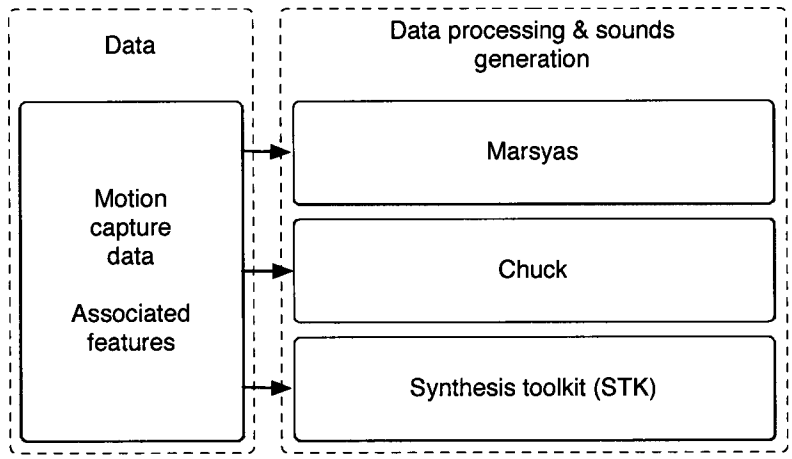


Fig. 2.3 Architecture of the framework for sonification of motion capture data (from [29]).

sounds for scientific purposes. An extensive foundation of knowledge about the potential for auditory information to enhance visual perception in a research context has not yet been achieved and this paper proposes an innovative evaluation method in this direction.

The experiment consist of videos of different subjects performing jumps. The videos are augmented by computer-generated auditory stimuli that are controlled by parameters derived from characteristics inherent to the jumps. More precisely, a force plate that generates electric current proportional to the applied pressure is placed under the subject performing jumps. It provides a summary of the overall muscular tensions involved in this action and thus a relevant gesutre feature to be sonified. Starting from the hypothesis that sounds are strongly linked to kinetic events, two experiments are elaborated to validate the role of sonification in multimodal displays.

They provide the result of a two part experiment that evaluates the perception and reproduction of sport movements (i.e. jumps) using visual data enhanced by sonification. They first asked subjects to evaluate the difference between two consecutive jumps under three conditions: visual treatment exclusively, audio treatment exclusively, and visual-audio treatment together. In a follow-up, the subjects were required to reproduce the height of jumps using (separately) visual information and visual-audio enhanced information. Participants had better results in evaluating audio-visual enhanced excerpts than visual-only

ones.

The justification of the experimental design is based on the merging of three areas of research in the field of human perception that include previously discussed multisensory integration and perceptual dissociation. In an ecological approach [31] [32], sound perception is regulated by prior knowledge of physical mechanisms that actually produce "natural" sounds. The perceived sound is mentally associated with its source and categorized according to this criteria. Ideally, in order to reach a high level of accuracy in multimodal representation, the computer-generated sounds have to be apparently caused by the stimuli. The gesture feature extraction is of primary importance in achieving accurate sonification using this approach.

What is not covered in this experiment are specification such as: distance from the ground or knee angle during jump preparation. More elaborate sounds could have lead to more accurate evaluation as well as consideration of what features best contribute to provide information in an evaluation or reproduction task.

Sonification of musicians' ancillary gestures

A work presented in [1] aims at developing techniques which allows for non-realtime sonification of clarinetists' movements. It continues the previous research investigated in that the sonification design is based on expert knowledge. Sonification is used to enhance the listener's perception with the intent of furthering our detailed understanding of performers movements. A prior familiarity with gesture characteristics greatly influenced the sonification approach.

Once primary awareness of gesture related features is established, sonification can bring some support to the analysis of specific details and interactions. The relevant movements that occur during a clarinet performance are described in [20], [33]. Body curvature, bending of knees, weight transfer, or circular movement of the clarinet bell are the most salient gestures considered. The sounds are chosen to suit the characteristics of their related gesture features. Sounds are carefully parametrized so that when listening to the result, one acquires specific details about the temporal evolution of position, orientation, and velocity of a given gesture.

Rather than focusing on low-level features such as absolute location or distance between markers, the gesture descriptions involve parametric calculations that are more physiolog-

ically relevant: polynomial curvature, circle radius, approximation of the center of mass. At the other end of the sonification process, the sound synthesis algorithms the authors stress render very precise descriptions of the gestures revealing the general evolution of the movements. They characterize each gesture in a way that allows for several distinctive streams (i.e. several gestures) to be sonified simultaneously, and then compared.

The underlying structure of the system is a four part architecture including several data processing layers which range over a variety of data transformation from data acquisition to sound synthesis. Intermediate processing includes parameter computation throughout mathematical algorithms and parameter scaling. A more detailed description of all the techniques used to adapt the data and convert it into sound is presented in the next chapter. More generally, each gesture considered is treated using the exact same succession of algorithms.

Over-specificity is definitely less appropriate for data exploration. It brings difficulties related to calibration of the system. Such a precise sonification design may not suit every performer. Some gesture evaluation techniques may provide better results if applied to certain performers rather than others.

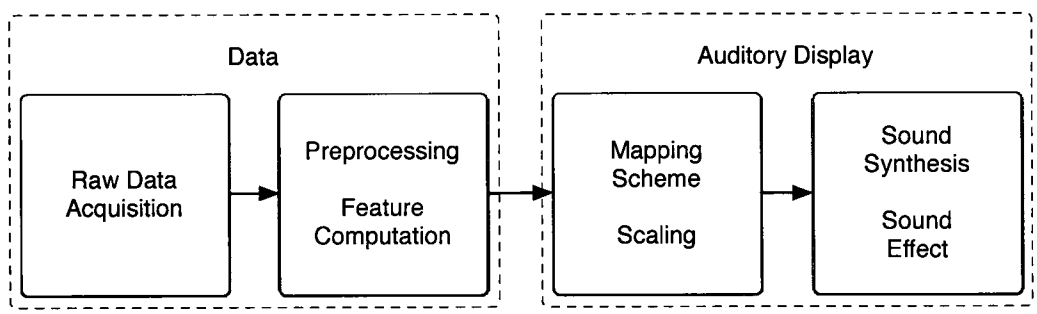


Fig. 2.4 General structure of the sonification system presented in [1].

Conclusion

This chapter presented fundamentals concerning sonification. As a multidisciplinary research field, sonification merges together several areas of study: psychological research in sound perception, sound cognition, and information acquisition. New sonification tools

design also requires knowledge in human-computer interaction and signal processing. Previous successful attempts in gesture sonification (i.e. sport movements, musicians' ancillary gestures, sonification framework) suggested that multi-modal representation is promising in regard to information integration.

Chapter 3

Revisiting previous work

Before going deeper into gesture analysis and gesture sonification, it is important to clarify the idea of "gesture features", what actually describes the gestures. This chapter takes a closer look at a recent research project on clarinetist's gesture sonification presented in [1] at the IDMIL. It proposes several gesture evaluation techniques (body curvature, knee bending, clarinet circular movement, and weight transfer) and various sound mapping strategies. Each gesture extraction feature and sound generation algorithm will be briefly discussed. The methods previously designed as non-realtime preprocesses in MATLAB from the previous project are implemented in realtime in the present work.

The two projects mainly differ in the data set they exploit. The previous project was based on data collected using an Optotrak 3020 three IR-camera system [34] with seven marker positions. Due to a limited number of body markers (see Figure 3.1), the different methods of approximating gesture were strongly influenced by the information available. The Vicon motion capture system used in the current project allows for a better description of the overall body by using a significantly larger number of markers (i.e. 38 markers) than the Optotrak. The sonification system designed in this project is compatible with both motion capture systems since the Optotrak's marker placement is a subset of the more complete Vicon's one. A detailed characteristic description of the Vicon system is presented in section 4.1.1.

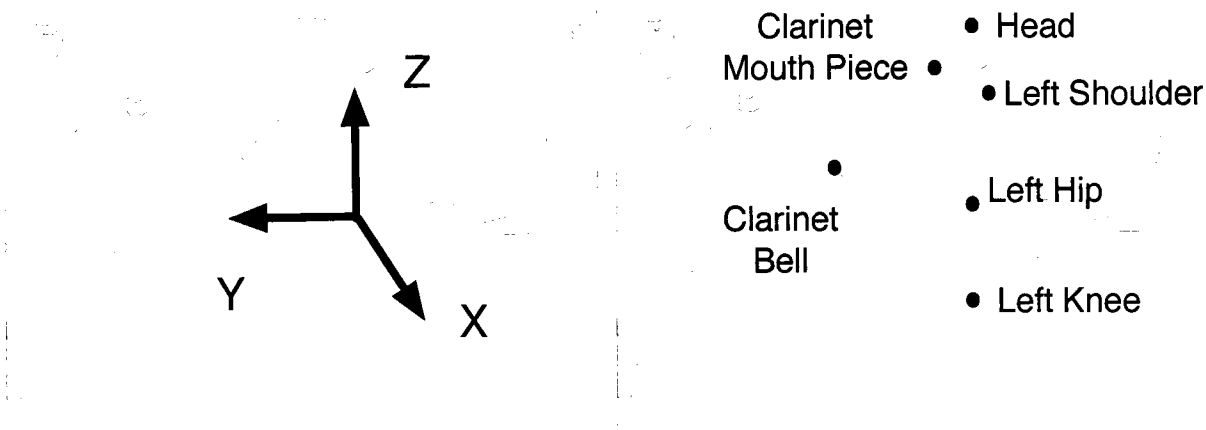


Fig. 3.1 Marker placement of the Optotrak system. The axis are defined so that the x axis belongs to the horizontal plane in reference to the floor and points to the left while the y axis points to the front in the same plane. The z axis is oriented toward the ceiling

3.1 Gesture Features

3.1.1 Body curvature

The human spine is composed of several articulations, each slightly contributing to the overall body curvature. This gesture is consequently evaluated in relation to several markers (i.e. head, back, hip). The curvature is not uniformly distributed among the articulations contributing to the gesture; integrating all the information in one unique variable is thus problematic. Two paths have been explored in the previous project and are described here: one fits a circle over the data and an other uses polynomial regression.

Circle fitting

This method consists in evaluating the radius of a circle which maps the position of three markers. As the markers are brought into a straight line, the radius of the circle tends to a large, even infinite value. In opposition to this, a curved body fits a smaller circle. Given the equation of a circle of radius r and centered on (h, k):

$$(x - h)^2 + (y - h)^2 = r^2 \tag{3.1}$$

the coordinates of the three markers must satisfy

$$(y_1 - h)^2 + (z_1 - k)^2 = r^2 \tag{3.2}$$
$$(y_2 - h)^2 + (z_2 - k)^2 = r^2 \tag{3.3}$$
$$(y_3 - h)^2 + (z_3 - k)^2 = r^2. \tag{3.4}$$

Isolating k and h:

$$k = \frac{(y_1 - y_3)u - (y_1 - y_2)v}{(y_1 - y_3)(z_1 - z_2) - (y_1 - y_2)(z_1 - z_3)} \tag{3.5}$$
$$h = \frac{(z_1 - z_3)u - (z_1 - z_2)v}{(z_1 - z_3)(y_1 - y_2) - (z_1 - z_2)(y_1 - y_3)} \tag{3.6}$$

where

$$u = \frac{-(y_2^2 - y_1^2) - (z_2^2 - z_1^2)}{2} \tag{3.7}$$

and

$$v = \frac{-(y_3^2 - y_1^2) - (z_3^2 - z_1^2)}{2}. \tag{3.8}$$

It is then easy to find the radius of the resulting circle in the following way:

$$r = \sqrt{(y_1 - h)^2 + (z_1 - k)^2}. \tag{3.9}$$

Polynomial evaluation of curvature

An alternate approach to evaluating body curvature is to consider the curvature of a polynomial correlating to the spatial location of the three markers. The mathematical curve is defined using polynomial regression. Given any points $(x_i, y_i = f(x_i)), i = 1, 2, ..., m$, the coefficients $p_j, j = 0, 1, 2...n$, of a polynomial $P_n(x)$ of any order $n \in N$ that best fit the points' coordinate in a least-square sense are evaluated. Since for body curvature only three positions ($m = 3$) are considered, a polynomial of the second order ($n = 2$) is sufficient:

$$P_2(x) = p_0 + p_1 x + p_2 x^2. \tag{3.10}$$

The residual R is defined as the sum of the squared vertical differences between the actual coordinates (x_i, y_i) and the polynomial predicted ones $(x_i, P(x_i))$:

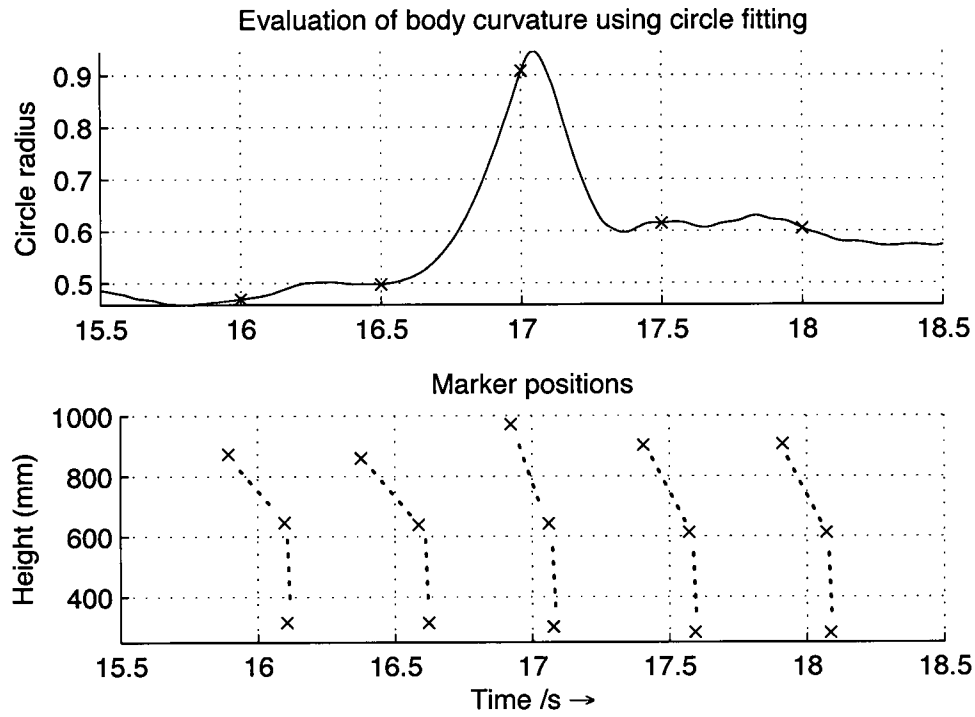


Fig. 3.2 Circle fitting curvature evaluation technique. The top figure shows the resulting signal of a change in the body curvature over a two second period evaluated via circle fitting technique. The bottom figure presents the corresponding marker position in the YZ plane with a step increment of 250 msec.

$$R^2 \equiv \sum_{i=1}^{n} [y_i - (p_0 + p_1 x_i + p_2 x_i^2)]^2. \tag{3.11}$$

In order to get the best approximation of the polynomial, the residual sum must be minimized. Taking the partial derivative according to each coefficient and finding the local minimum leads to:

$$\frac{\partial}{\partial p_0} R^2 = -2 \sum_{i=1}^{n} [y_i - (p_0 + p_1 x_i + p_2 x_i^2)] = 0 \tag{3.12}$$

$$\frac{\partial}{\partial p_1} R^2 = -2 \sum_{i=1}^{n} [y_i - (p_0 + p_1 x_i + p_2 x_i^2)] x_i = 0 \tag{3.13}$$

$$\frac{\partial}{\partial p_2} R^2 = -2 \sum_{i=1}^{n} [y_i - (p_0 + p_1 x_i + p_2 x_i^2)] x_i^2 = 0. \tag{3.14}$$

We then obtain the following system of equations:

$$\begin{cases} np_0 + p_1 \sum_{i=1}^{n} x_i + p_2 \sum_{i=1}^{n} x^2 = \sum_{i=1}^{n} y_i \\ p_0 \sum_{i=1}^{n} x_i + p_1 \sum_{i=1}^{n} x_i^2 + p_2 \sum_{i=1}^{n} x^3 = \sum_{i=1}^{n} x_i y_i \\ p_0 \sum_{i=1}^{n} x_i^2 + p_1 \sum_{i=1}^{n} x_i^3 + p_2 \sum_{i=1}^{n} x^4 = \sum_{i=1}^{n} x_i^2 y_i. \end{cases} \tag{3.15}$$

Using the matrix form that is easier to handle:

$$\sum_{i=1}^{n} \begin{bmatrix} 1 & x_i & x_i^2 \\ x_i & x_i^2 & x_i^3 \\ x_i^2 & x_i^3 & x_i^4 \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} y_i \\ x_i y_i \\ x_i^2 y_i \end{bmatrix}. \tag{3.16}$$

From this system:

$$\mathbf{A}\overrightarrow{P} = \overrightarrow{B}, \tag{3.17}$$

the coefficient vector is given by

$$\overrightarrow{P} = \mathbf{A}^{-1}\overrightarrow{B}. \tag{3.18}$$

Once the three polynomial coefficients p_i are found, the degree of curvature k at an arbitrary position x can be easily obtained using:

$$k = \frac{\frac{d^2}{dx^2}P_2(x)}{(1 + \frac{d}{dx}P_2(x)^2)^{3/2}}, \tag{3.19}$$

where

$$\frac{d}{dx}P_2(x) = 2p_2x + p_1 \tag{3.20}$$

and

$$\frac{d^2}{dx^2}P_2(x) = 2p_2. \tag{3.21}$$

The curvature evaluation of the resulting polynomial(s) is arbitrarily performed at the middle marker location. The fact that the polynomial is continuously defined between markers is particularly convenient since it is possible to change the location of the curvature evaluation. However, matrix inversion is not always possible and may restrict the usage of this method.

Evaluation

The algorithms described in the previous project (i.e. circular and polynomial fitting) are efficient curvature evaluation techniques that still make sense when applied to Vicon's data sets. Circle fitting demonstrates a strong infinity asymptotic behavior (see Figure 3.3) when the three markers lie on a straight line, implying a circle radius which extends to infinity. The polynomial fitting method follows a more linear behavior pattern and provides as a parameter the location of the curvature evaluation. Both algorithms are implemented as realtime processes but polynomial regression is preferable to circular fitting principally because of its linear behavior. No serious problems occurred concerning matrix inversion even if this issue should be improved in future developments of the sonification system.

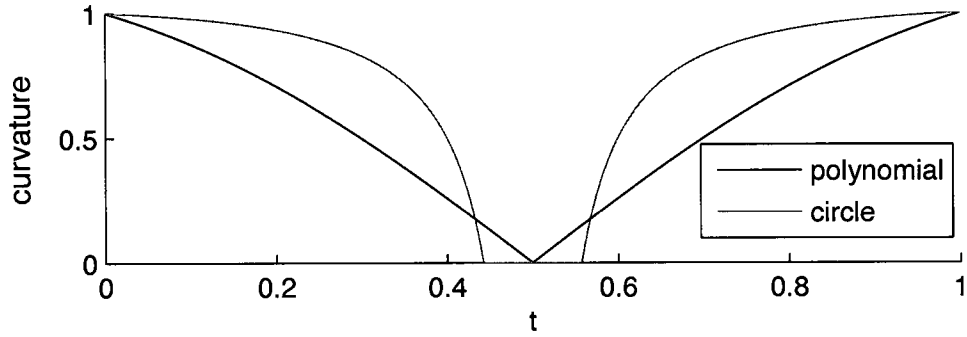


Fig. 3.3 The circle fitting technique has an asymptotic behavior. As the parameter t augment, the circle fitting technique tend to infinity. The middle position corresponds to the aligned marker position.

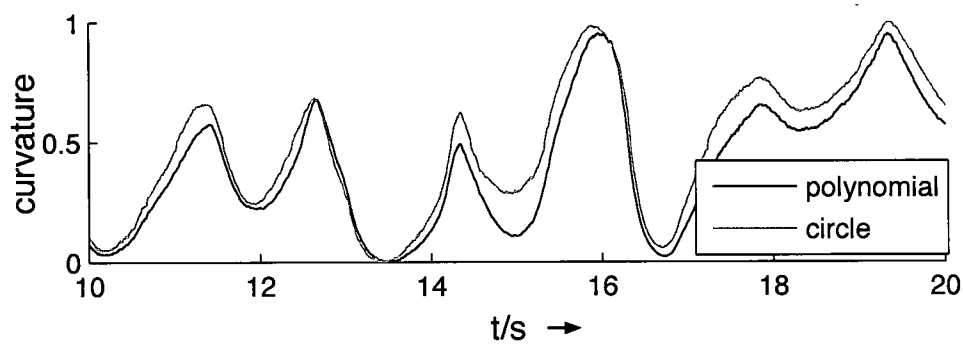


Fig. 3.4 Comparison between polynomial regression and circle fitting method.

3.1.2 Knee Bending

The Optotrak marker set does not provide any information about either ankle position. Knee bending must be approximated using knee and hip positions projected on the transversal axis. Their distance d on this axis is proportional to the knee angle (see Figures 3.5, 3.6). When standing up straight, the two positions should mostly coincide once projected on the y axis. The main problem related to this approximation is the fact that it is not robust to the overall body orientation of the performer. When rotating the body while not moving the feet, changes in the difference between the respective positions may be interpreted as a bend of the knee. Slight variations in the approximation are also attributed to different orientations of the hips.

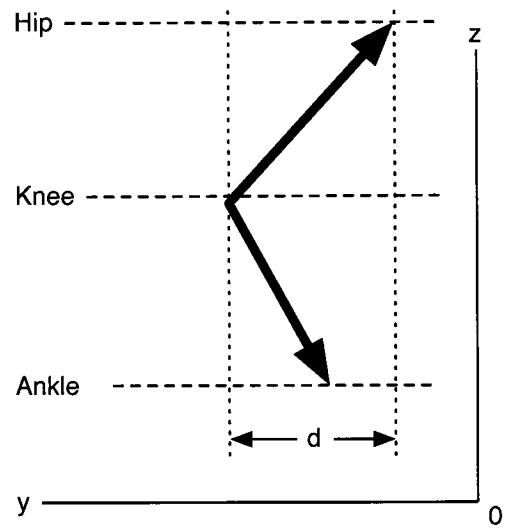


Fig. 3.5 The distance between knee and hip as an approximation of the knee bending.

3.1.3 Weight Transfer

A convenient way to describe weight transfer is to evaluate the performer's center of mass defined as the weighted mean position of all markers. In the project using the Optotrak data set, the center of mass was evaluated using the mean position of the head, left shoulder, and left hip markers. Figure 3.7 shows the close relationship over time between this

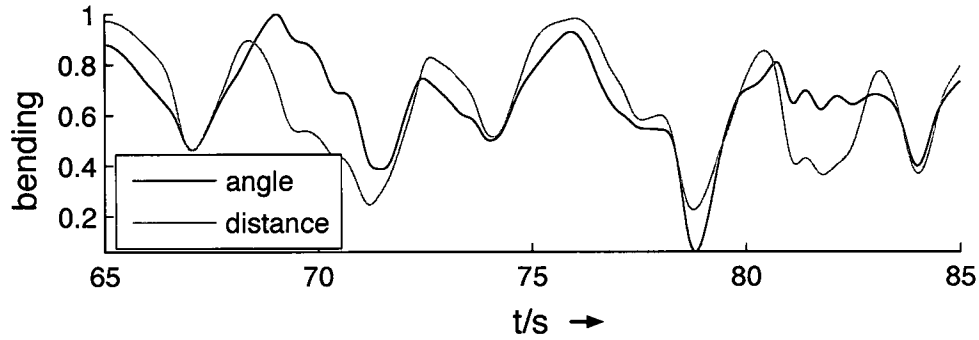


Fig. 3.6 Evaluation of the knee bending using relative distance and knee angle.

approximation of the center of mass and its evaluation provided by the Vicon.

It is interesting to note that different instrumentalists may present different weight transfer patterns in relation with their respective instrument. For example, Figure 3.8 shows samples of weight transfer patterns projected on the floor of three different performers: two clarinetists and one violist. Looking at violist's patterns, the weight transfers produce straight lines that mostly maintain the same orientation of the instrument. In opposition to this, the clarinetists' weight transfers are not specifically oriented and occur in various directions.

3.1.4 Circular movement of the clarinet bell

Clarinetists clearly perform circular motion with their instrument bell. Figure 3.9 is an example of several circular motions performed by the same subjects during different time intervals in the xy plane.

While performing a perfect circle over a period T corresponding to a number N of samples, the mean position $\overline{M}(t)$ of the past N value $M(t) = (x(t), y(t))$ should appear exactly at the center of the circle:

$$\overline{M}(t) = \left(\frac{1}{N} \sum_{i=t-N+1}^{t} x(i), \frac{1}{N} \sum_{i=t-N+1}^{t} y(i) \right). \tag{3.22}$$

From this result, it is possible to evaluate two main parameters:

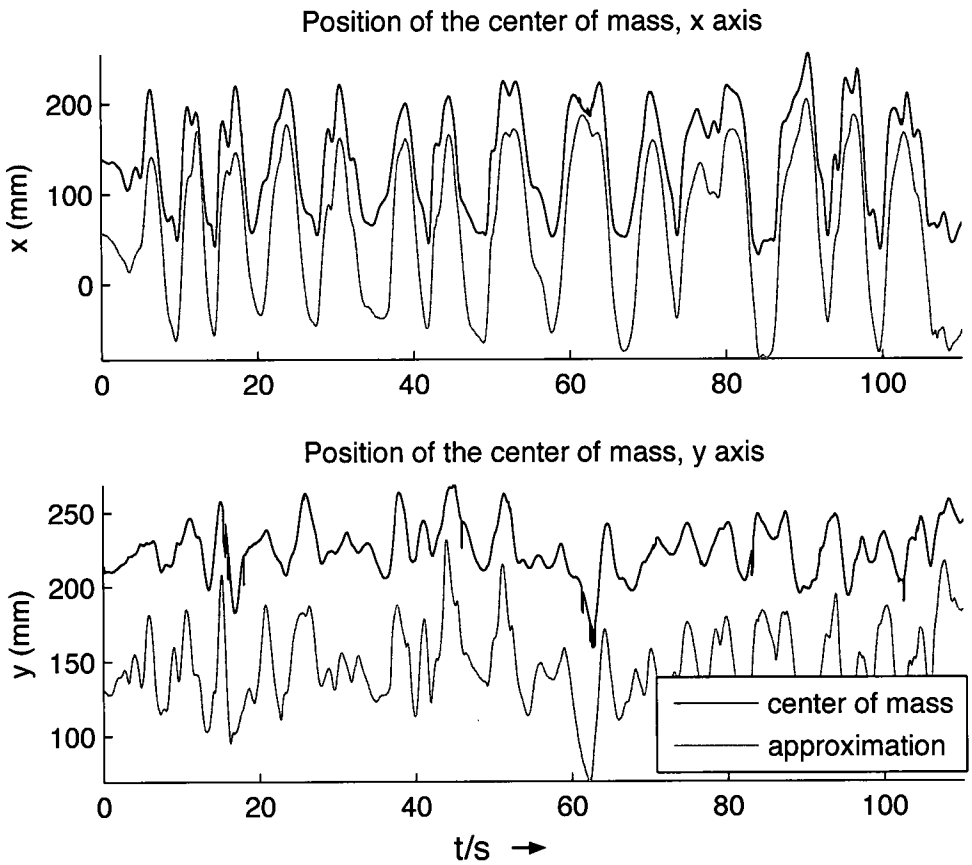


Fig. 3.7 Comparison of center of mass provided by the Vicon and its approximation using the Optotrak.

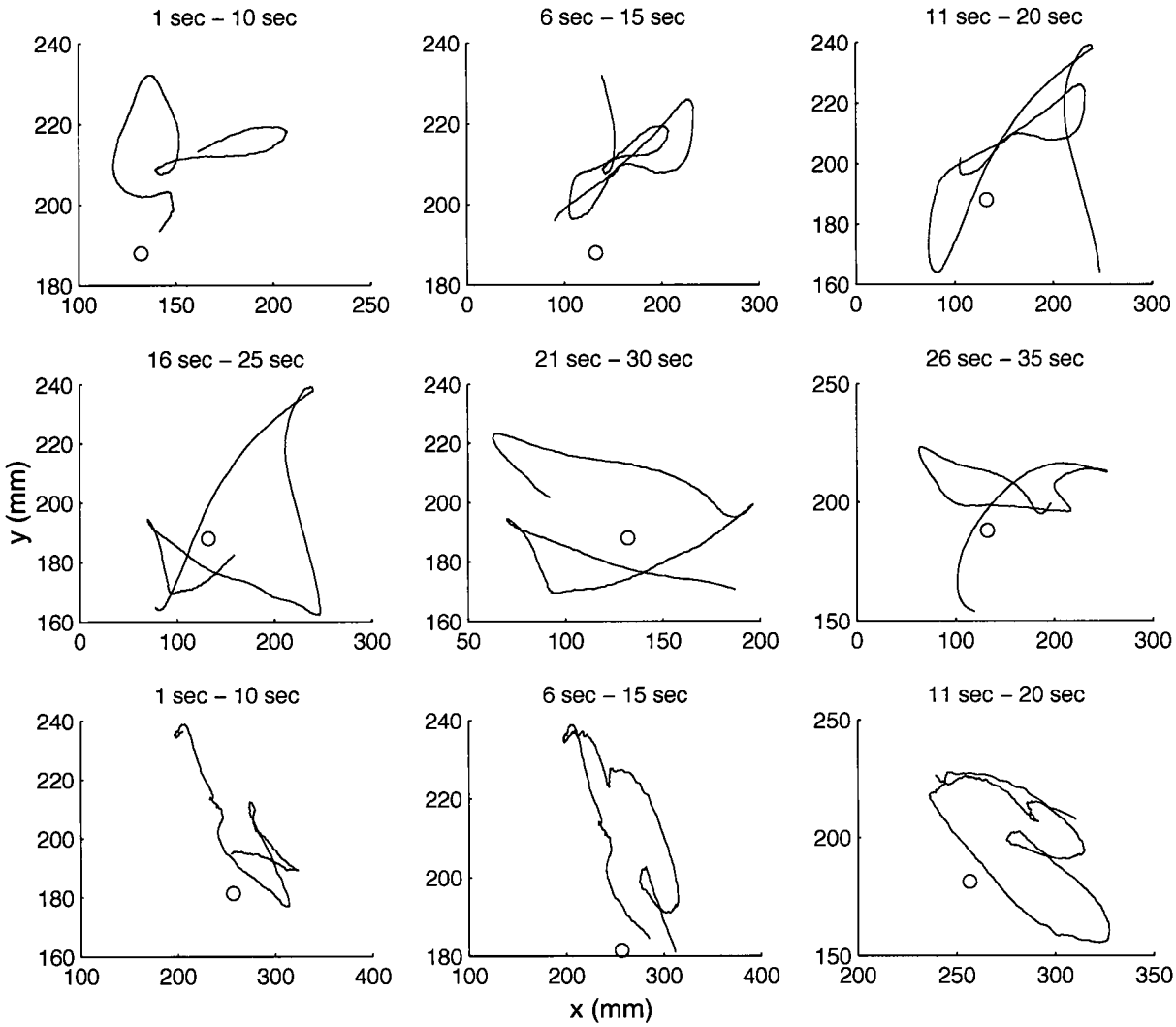


Fig. 3.8 Trajectories of center of mass projected in the xy plane (floor). The circles mark the starting position at the beginning of the performance.

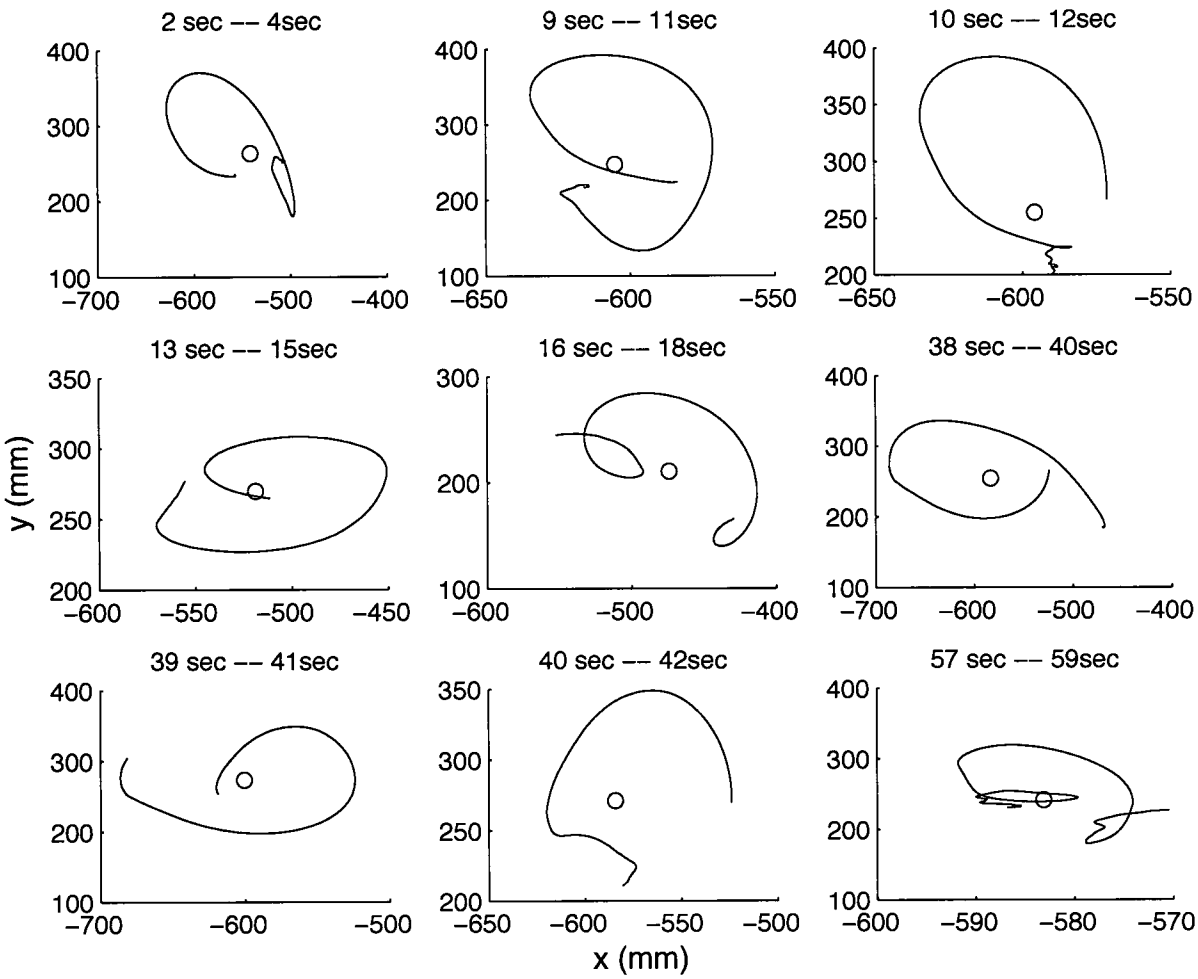


Fig. 3.9 Examples of circular movements around a mean value (circle) in the xy plane.

- the distance d between the current position and the mean position

$$d_N(M(t), \overline{M}(t)) = \sqrt{(M_x(t) - \overline{M_x}(t))^2 + (M_y(t) - \overline{M_y}(t))^2}, \tag{3.23}$$

- the angle between the $\overrightarrow{0y}$ axis and the vector of the current position in the mean reference position

$$\theta_N(t) = i_1\frac{\pi}{2} + i_2\frac{\pi}{2} + \tan^{-1}\left(\frac{M_x(t) - \overline{M_x}(t)}{M_y(t) - \overline{M_y}(t)}\right). \tag{3.24}$$

The algorithm also detects circle-like movements such as arcs and ellipses. It must be implemented in a relative position according to the head since it could otherwise be very reactive to fast swiping movements. Since it uses the mean value over a fixed number N of delayed samples, it is efficient for circular gestures of a specific temporal length and circle radius. Too slow or too small circles result in signals of low amplitude. Despite this inconvenience, there are several relevant features that can be used to describe the gestures:

- distance between mean and current positions,
- angular displacement,
- position velocity,
- angular velocity,
- angular acceleration.

3.2 Computer-generated sounds

In this section, the auditory tools implemented within the system are discussed outside of considerations related to the entire field of sound semantics or gesture-to-sound mapping. It will rather briefly introduce sound synthesis algorithms implemented in this project and describe more specifically what parameters they offer in terms of sound feature control. An comprehensive discussion about sound synthesis and their use in sonification is presented in [35].

3.2.1 Sound synthesis

Additive synthesis

Additive synthesis is a direct application of Fourier's theorem, which stipulates that any complex waveform can be locally represented as a linear combination of sinusoidal elements with coefficient A_k and phase σ_k for every frequency f_k:

$$x(t) = \sum_{k=0}^{\inf} A_k \sin(f_k t + \sigma_k). \tag{3.25}$$

The idea is that one can use a sine generator or a combination of several sine generators of different frequencies to generate complex periodic waveforms. Several oscillators offer control of the fundamental frequency of the sound as well as the relative frequencies and amplitudes of every frequency component, which would be more related to timbre. In theory, a very wide range of possible sounds can be generated or reproduced. However, in practice, only few oscillators are commonly used since it is still a computationally consuming technique.

In a special case of additive synthesis, a technique widely used is the beating effect resulting from adding two components of close frequencies. Given a system of two oscillating components having the same amplitude A:

$$x_1 = A\cos(\omega_1 t), \tag{3.26}$$

$$x_2 = A\cos(\omega_2 t), \tag{3.27}$$

$$x = x_1 + x_2 = 2A\cos\left(\frac{\omega_1 - \omega_2}{2}t\right)\cos\left(\frac{\omega_1 + \omega_2}{2}t\right). \tag{3.28}$$

Where $\frac{\omega_1-\omega_2}{2} \leq 20Hz$, $2A\cos\left(\frac{\omega_1-\omega_2}{2}t\right)$ is an oscillating amplitude modulation envelope producing a beating effect. This effect is convenient to convey information since it provides an accurate temporal appreciation.

Another special technique is the glissando version of Shepard tones based on an implementation of Jean-Claude Risset [36]. Using phase-delayed modulation envelopes applied on overlapping oscillators whose frequencies are set according to the Shepard scale [37] creates an auditory illusion of a continuously ascending or descending fundamental frequency.

The speed of the ascending or descending frequency becomes an informative parameter.

Frequency modulation synthesis

Frequency modulation synthesis [38] is a very well known synthesis technique which produces a complex waveform by modulating the "carrier" frequency ω_c or the phase ϕ_c by another periodic oscillator. Phase modulation is described as:

$$y(t) = A\cos(2\pi\omega_c + I\cos(\omega_m t + \phi_m)). \tag{3.29}$$

Given the carrier frequency ω_c, the modulation frequency ω_m can be found using:

$$\omega_m = h \cdot \omega_c. \tag{3.30}$$

where h is the harmonic ratio given as a parameter. Increasing I (the modulation index) results in increasing the number of side bands, which is directly linked with brightness. Changes in harmonicity ratio may require a bit more training to be completely distinguished from changes in brightness but can still be used as an informative parameter for sonification. It is then possible to describe different aspects of a complex gesture using three control parameters: carrier frequency ω_c, modulation index I, and harmonicity ratio h.

3.2.2 Sound effects

Panning

Panning is mainly used to represent spatial location. It essentially creates a virtual sound source positioning. Dynamic changes in the panning angle can be used to represent displacement and may be particularly efficient in enhancing moving visual stimuli. Binaural amplitude difference is a widely used panning technique that provides efficient results in the case of left and right stereo panning. Horizontal panning (left – right) is completely defined given only one parameter, θ, the panning angle that ranges from 0 to π radians. To preserve the original sound intensity regardless of its virtual source location, the respective amplitude for each stereo channel is evaluated using the formulas:

$$A_R = \left| \frac{\sqrt{2}}{2}(\sin\theta + \cos\theta) \right|, \tag{3.31}$$

$$A_L = \left| \frac{\sqrt{2}}{2}(\sin\theta - \cos\theta) \right|. \tag{3.32}$$

Intensity preservation is verified as:

$$A_R^2 + A_L^2 = 1. \tag{3.33}$$

Filtering

Filters are linear time-invariant systems used to manipulate the amplitude and phase of the spectral contents of sound. They can be used to augment or reduce specific frequency components from rich spectrum input sound signals. White noise is probably the best to achive this requirement but other type of noise (pink, brown...) can be used if one is interested in enlarging the range of timbre quality. A very simple but versatile filter is the biquadratic filter that allows for several types of filtering envelopes to be defined using four coefficients, namely: low-pass, high-pass, band-pass, and band-rejection.

$$y[n] = a_0 x[n] + a_1 x[n-1] + a_2 x[n-2] - b_1 y[n-1] - b_2 y[n-2]. \tag{3.34}$$

This flexibility is explained by the fact that it is defined using two poles and two zeros that can be parametrized as one can see on the z-transform of the transfer function:

$$H(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}}. \tag{3.35}$$

Using filters, the determining parameter in regards to perceived effect is cut-off frequency in the case of low-pass/high-pass filters, and central frequency in the case of the band-pass/low-pass filter. The filter quality (Q) may also be a relevant parameter, specifically in relation to band-pass filtering. It can be used to emphasize the filtering effect.

Conclusion

This chapter described various gestural feature extraction algorithms as well as sound synthesis techniques implemented in the sonification system. The data processing and sound synthesis were achieved in Max/MSP, a programming environment design for realtime audio processing. It allows for the users to perform changes in sonifications without any

required compilation, which facilitates sonification design.

Chapter 4

Gesture data: acquisition and processing

Chapter 3 introduced gesture feature extraction algorithms developing perspectives based on expert-knowledge. As explained in section 2.2.3, marker positions alone are insufficient to describe complex interactions in the body. The feature extraction processes as proposed in the last chapter produce continuous signals following a sonification approach said to be data driven.

In a multi-streaming context where several sonifications are displayed to the user, it is convenient (if not fundamental) to sonify only portions of data that contain relevant gestural information. One needs to filter out the information that is not considered to be a gesture. This chapter presents the data provided by the Vicon motion capture system and describes briefly the various gesture features derived using the biomechanical model. Finally it discusses the question of gesture definition and more especially which control parameters are relevant to filter out undesired information in realtime.

4.1 Data provided by motion capture systems

4.1.1 Motion capture system

Motion capturing consists in the accurate tracking and recording of markers fixed at strategic positions over the subjects's body. For this project, different gesture data sets have been recorded using a Vicon 460 [39] with six infra-red cameras. This motion capture system

tracks passive markers without labeling information. Marker identification is performed afterwards during the computation of the biomechanical model through relative position analysis. Unfortunately, this process introduces delays and may have an effect on the perspective of realtime tracking and sonification. Passive markers free subjects from any power settings or wires that could hamper their gestures. The markers have to be fixed as close to the bone as possible since the structural information is relayed out by the skeleton. Furthermore, the elasticity of the skin could introduce undesirable variability into the results. All these considerations are taken into account to provide a three-dimensional representation (x, y, z) of each marker position evolving over time, for instance at a frame rate of 100 Hz.

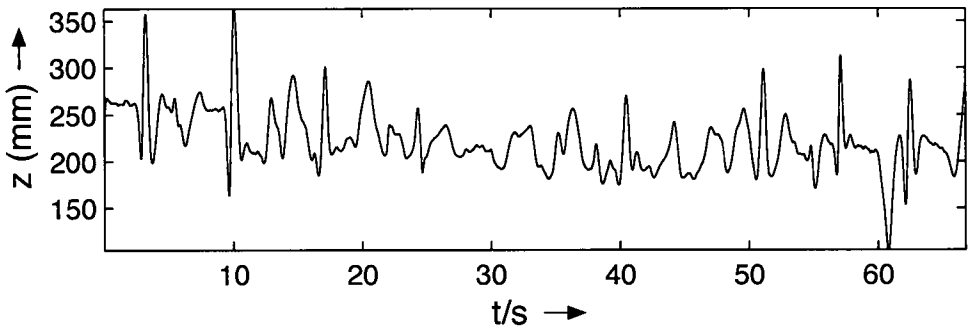


Fig. 4.1 Position of the clarinet bell.

4.1.2 Marker placement models

The markers placement as well as the camera positions depend on the research perspective. There exist several widely used marker placement protocols. Each one is associated with a slightly different biomechanical model that includes measurements of relevant articulation angles and joint rotations. For this project, the plug-in-gait model has been used. It provides 38 marker positions that globally describe the configuration of the body posture. Two additional markers were attached to the instrument at both extremities, the mouthpiece and the bell, in order to track the motions of the instrument. The overall marker set provides a more precise description of the performer's body movements than the earlier Optotrak system, wich was unfortunately not able to process the same amount of markers. The amount of markers available as well as their configuration obviously influence the

choice of gesture evaluation techniques. The main purpose of marker placement protocols is to enable the comparison between different data sets.

There is no marker located on the performers' fingers when using the plug-in-gait protocol. Tracking the complex motion of the hands, wrists, and fingers would have required a huge amount of markers concentrated in this area as well as an alternate camera positioning. Even if very interesting, it is beyond the scope of this work to apply sonification on gestures directly related to sound production. It would introduce several sound semantic issues, including the possibility that the sonification would interfere considerably with the actual sound produced by the instrument. Even if the instrumental sound would not be audible, confusion in the interpretation of the sonification could arise from the similarity between the gestures, especially ones that produce or modify instrumental sounds [40], and sounds actually produced by sonification.

4.1.3 Motion capture sessions

The data to be sonified were collected during several motion capturing sessions conducted in previous projects at the IDMIL. The performers were asked to play an excerpt of Brahms' *Sonata for clarinet op. 120 n° 1.* It is a piece of standard repertoire that presents an average degree of difficulty, maximizing the number of potential participants. Several recordings of the performance were carried out asking the performers to play with different conditions (i.e. natural, upright, sitting down). All subjects are advanced instrumentalists studying at universities. Additionally, this musical excerpt is also part of the viola repertoire, which makes possible the comparison of performances on different instruments.

A previous discussion highlighted the close connection between performers' ancillary gestures and the performed music. Figure 4.2 shows the relation between two distinct performances of the same musical excerpt by the same subject. It was recorded using the Optotrak motion capture system and the performer was asked to play in various manners, namely: immobile, standard, and expressive.

4.1.4 Conventional visualization

In addition to the motion capturing system, a conventional video camera is used to record each performance synchronously. Both recordings reveal distinct but complementary information. For instance, even if video provides less precise gestural tracking, it gives a good

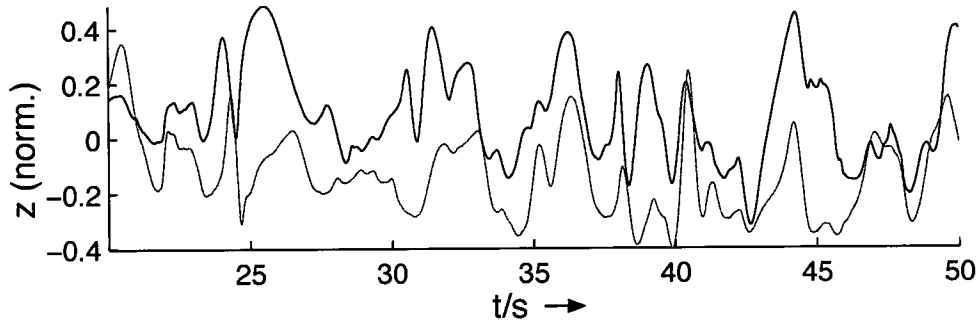


Fig. 4.2 Comparison between two different levels of expressiveness performed by subject 1: expressive (dark) and standard (thin). The two performances demonstrate similar behaviours.

overview of the whole performance (see Figure 4.3). This global comprehension is essential to interpret local details in their context. On the other hand, data from motion capture systems are objective measurements of the performance that can be mathematically analyzed. However, they are not easy to handle or visualize, especially when considered as signal frames. It actually reinforces the need to develop new representations or simply augment existing ones in order to enhance the information integration.

4.2 Data processing: biomechanical model

Relevant information concerning gestures may not be directly accessible exclusively from marker position; algorithms are required to extract it. The sonification of gesture can be enhanced by considering other descriptions such as joint angles of articulated body structures, which was the case in [1]. The feature extraction algorithms combine several local positions into a global description, which leads to relevant control parameter signals for sound synthesis. A realtime implementation potentially requires control of the selection of both markers and feature extraction algorithms to be applied.

4.2.1 Evaluation of joint angles

Over 65 joint angle evaluations are embedded in the plug-in-gait biomechanical model and describe physiologically meaningful articulations such as knees, hips, or neck. Each is described in terms of Euler angles that decompose the orientation of rigid bodies according

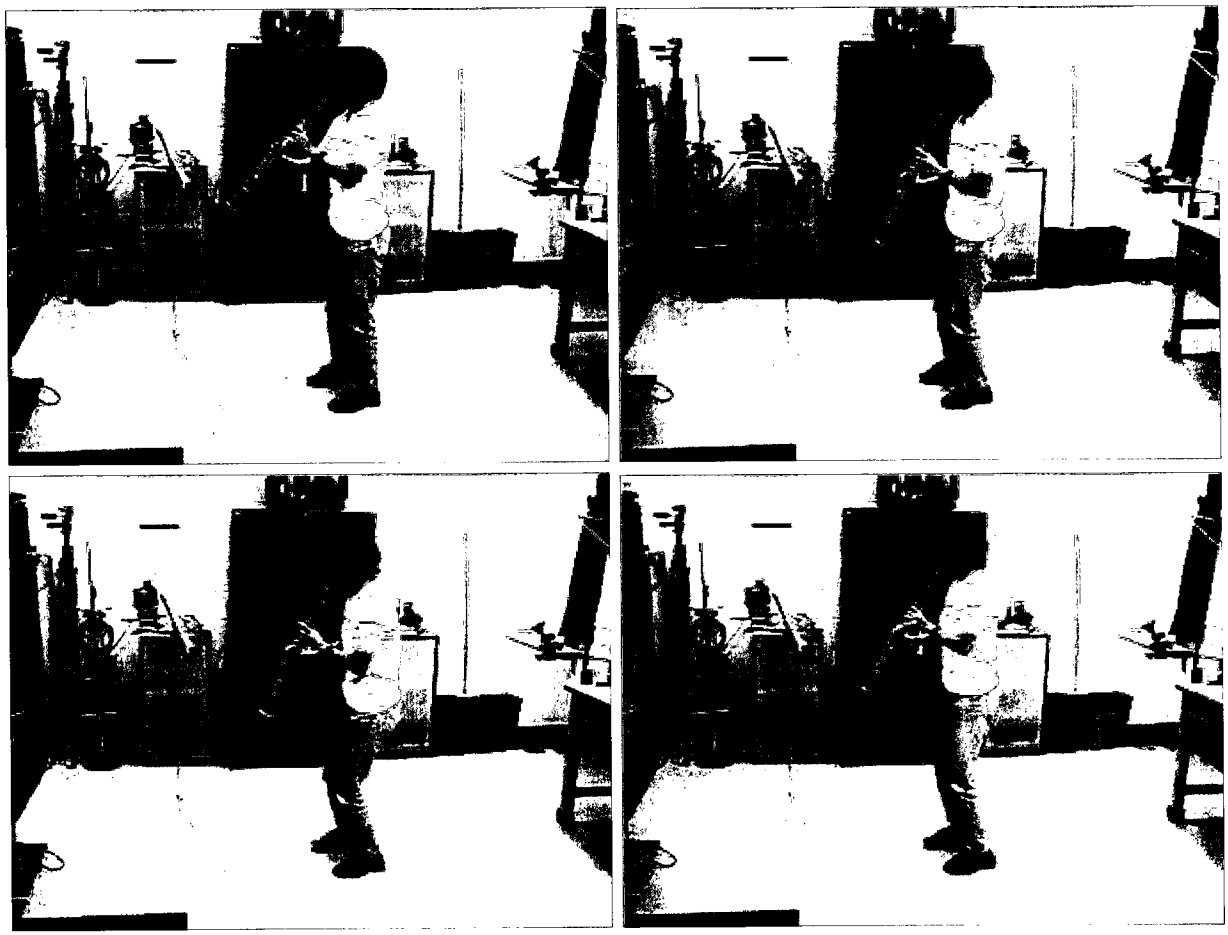


Fig. 4.3 Several postures resulting from different gestures performed by subject 1 (captured using the Optotrak system).

to three rotational parameters (α, β, γ) within a fixed reference (x, y, z). The reference is determined at the specific marker position where the angles are to be evaluated and conserves the same absolute orientation as that provided by the motion capture system.

There are additional angles, even if not physiologically relevant, that can contribute to the overall understanding of the performers' movements, especially angles involving the instrument itself. The orientation of the instrument is established by considering two positions and one reference; in the case of the clarinet, corresponding to the instrument's bell and mouthpiece. This last marker is designated as the reference for a vector pointing in direction of the instrument bell. Once the vector is normalized, the rotation parameters are easily derived from inverse sine and cosine transformations. They result in a robust evaluation of the instrument orientation relative to the head and body movements of the performer.

Returning to the evaluation of the body curvature, angles at precise locations along the neck and the spine (i.e. C7 and T10) can be used to refine the evaluation of the curvature (see Figures 4.4, 4.5). The first one is mainly related to head movements while the second concerns the whole upper body. The information they provide is obviously correlated but still shows subtle distinctive behaviors. The performer could move the head only or otherwise proceed with movements involving the whole upper body without any specific motion of the head.

4.2.2 Relative positions and changes of reference

An approach to extract information about a marker or a group of markers is to evaluate their relative distances. This is similar to a change of reference where the position of a given marker is expressed in relation to another. Omitting one or more axes results in a projection over the remaining axis. The gesture detection can suffer from the interference of other movements, for instance, the circular movement of the clarinet bell affected by fast changes in the body curvature. One solution is to isolate movements of interest within a locally-defined coordinate system. It is easily defined by the difference between a marker position $(p_x(t), p_y(t), p_z(t))$ and a reference $(r_x(t), r_y(t), r_z(t))$, not necessarily fixed, as:

$$d(t) = (p_x(t), p_y(t), p_z(t)) - (r_x(t), r_y(t), r_z(t)).$$

The data expressed in the new reference could conflict with the visual representation of

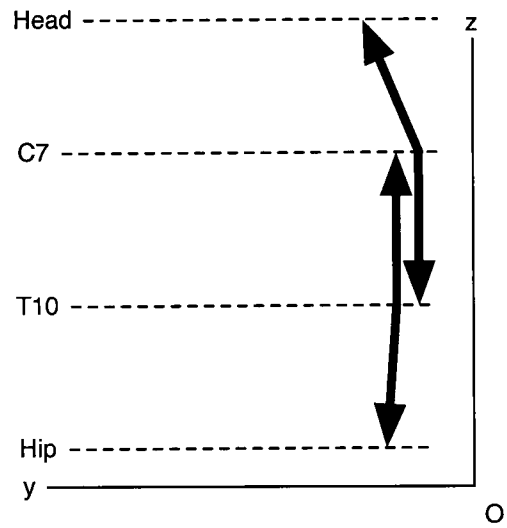


Fig. 4.4 Angles evaluated at the neck (C7) and the back (T10).

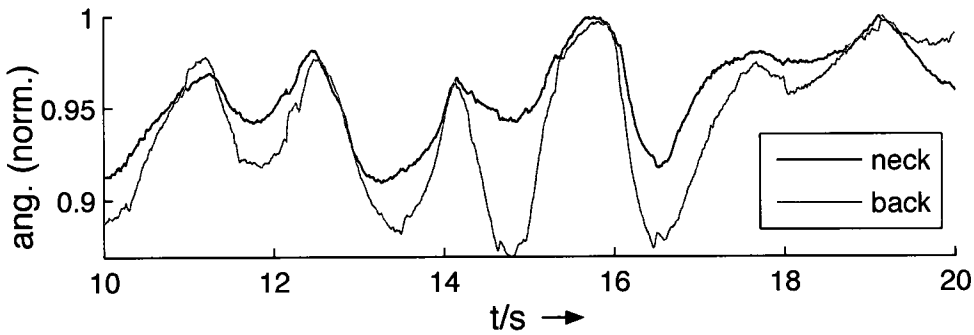


Fig. 4.5 Comparison between neck (C7) and back angles (T10).

the original data. As an example, one could be interested in marker positions relative to the center of mass. Using this reference, the feet, which are not suppose to move, would acquire exactly the opposite motion as the center of mass in the original reference. Conversely, the hips will appear to be motionless. Looking at the video of the performance, someone would hardly connect sounds generated by the sonification in such a reference. The changes of reference have to be performed carefully and should be restrained to close relations within the same body region. It is mainly performed to isolate specific body parts from global movements such as the center of mass position changes.

4.3 Gesture definition in the context of sonification

In this project, gestures are considered in terms of movement and velocity, in opposition to contexts where gestures are used to communicate using signs, which would require elaborate algorithms to achieve gestural pattern recognition. Despite this fact, information of idiosyncratic nature is still conveyed by gestures as certain levels of expressiveness are distinguished among different performances and subjects. A performance said to be *expressive* is characterized by fast gestures with abrupt transients. Some other types of performance tend to have slower and continuous gestures without evident interruption. The different processes applied on data must not alter this information. Considering these aspects, gestures are mainly defined in terms of velocity and path distances.

4.3.1 Velocity

Kinematics is a branch of mechanics that is concerned with motion without any reference to the generating forces. In kinematics, velocity is proportional to the amount of motion. It represents a quantification of an amount of space $f(x+h) - f(x)$ covered over a certain amount of time h, and is compute as:

$$\frac{df(t)}{dt} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{4.1}$$

It can be considered that a gesture occurs each time the following condition is reached:

$$\frac{d}{dt} g(x(t)) \neq 0 \tag{4.2}$$

for any gesture extraction algorithm g as defined in chapter 3.

Differentiation

The gesture signals considered are discrete signals sampled at a rate of 100 Hz, (with period $T = 0.01$ seconds). The velocity is estimated using the backward finite difference approximation. It is a causal system that involves simply evaluating the difference between the current sample $x[n]$ and the previous one $x[n-1]$ (see Figure 4.6). No future samples are considered in the approximation of the finite difference, a causality property that suits the realtime condition required in this project.

$$\frac{d}{dt}x(t) \approx \frac{x[n] - x[n-1]}{T} \tag{4.3}$$

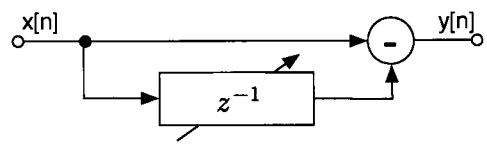


Fig. 4.6 Chart diagram of finite backward difference filter.

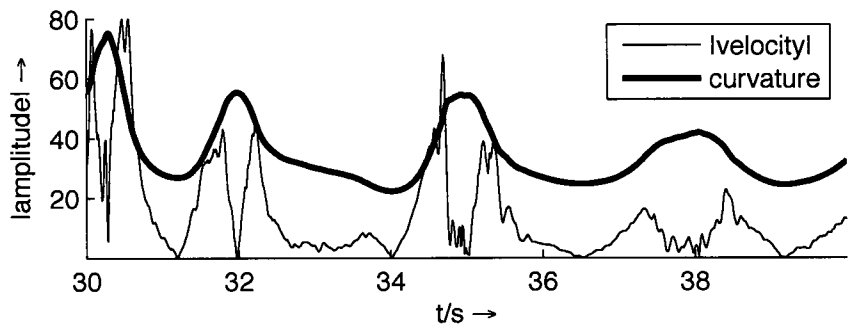


Fig. 4.7 Evaluation of the body curvature's velocity using a finite backward difference.

The finite backward difference is adapted to musicians' ancillary gestures in the sense that it preserves the high frequencies (representing fast gestures) and attenuates low frequencies (representing slow gestures). Finite backward difference is a high-pass filter that

allows for high frequency to pass through it. The magnitude spectrum $S(e^{i\omega T})$ is given by

$$\left|S(e^{i\omega T})\right| = \left|\frac{2}{T}\sin\left(\frac{\omega T}{2}\right)\right| \tag{4.4}$$

with radian frequency $\omega = 2\pi f$ (see Figure 4.8). Once processed by a finite backward difference, the signal's amplitude is proportional to the velocity of the gesture. Linked to the amplitude, this parameter attenuates slow gesture sound's intensity.

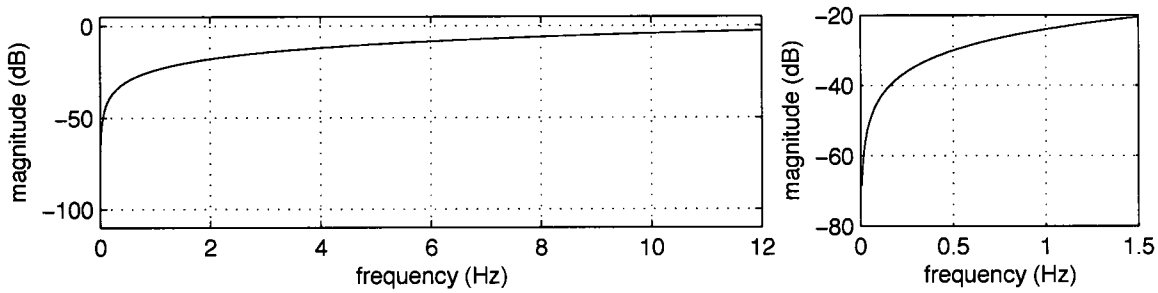


Fig. 4.8 Magnitude response of the finite backward difference (left) and zoom on the first 1.5 Hz (right).

A differentiator filter is elaborated to strongly attenuate frequencies below 1 Hz using the Parks-McClellan optimal FIR filter design algorithm. Figure 4.9 shows the 10th order differentiator filter's magnitude response. The resulting filtered signal (Figure 4.10) is very rich in high frequencies, which mainly contribute to represent high velocity. It demonstrates that the information of interest (one that can be used as an amplitude envelope) lay above 1 Hz for ancillary gestures.

Filtering

Gestures are usually performed at very low velocity corresponding to low frequencies (0 to 5 Hz). High frequency components present into the signal are mainly due to noise generated from marker tracking algorithms, skin and clothes elasticity, or haptic noise. Since backward difference is a high-pass filter, it tends to amplify unwanted high frequencies. Looking to Figure 4.7, a same gesture starting and ending at the same specific position is characterized by two distinctive velocity curves corresponding to forward and backward trajectories. Taking the absolute value of the difference also adds high frequency components into the

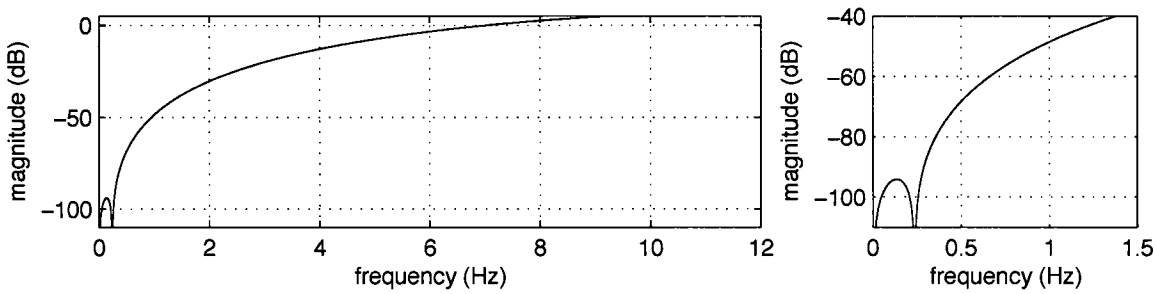


Fig. 4.9 Magnitude response of a differentiator especially designed to attenuate frequencies below 1 Hz.

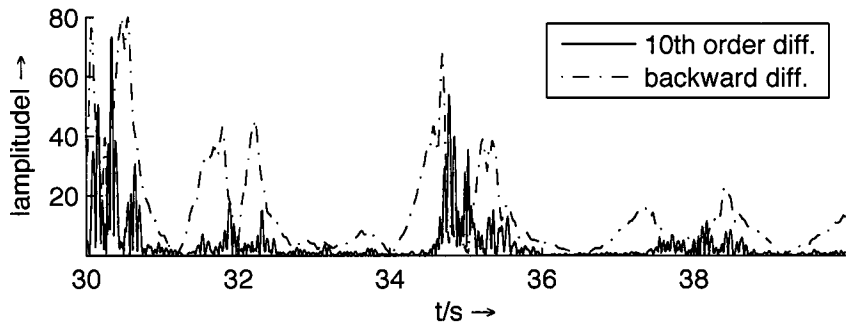


Fig. 4.10 Filtered velocity signal where frequencies below 1 Hz have been attenuated.

velocity signal.

Gesture velocity signals are low-pass filtered using the Parks-McClellan optimal FIR filter design algorithm using cut-off frequency respectively set at 2,4,6, and 8 Hz. A 80 dB stopband attenuation and a 10 Hz transition lead to a 25th order filter that maximize the frequency magnitude response. A cutoff frequency of 2 Hz is usually too low and pertinent information concerning gestures is filtered. Cutoff frequencies of 4 and 6 Hz are more appropriate as the original velocity curves are slightly smoothed. Removing high frequency components due to absolute value tends to unify the two distinctive velocity curves (back and forth movements) which is convenient once the velocity signal is applied as an amplitude envelope for the gesture's sonification.

4.3.2 Path distance

Path distance refers to the total amount of space between starting and ending positions of a gesture. There is ambiguity when movements are fast enough to generate sonifications but does not realize a consistent path distance, for instance, when the performer's body parts are slightly waving around mean positions. Once sonified, they still produce continuous sound. Can movements be considered as gestures even if they are hardly visually perceptible? Gestures of this type may be undesirable as they may not convey clear information about the performer's intentions or shadow relevant gestures.

Ideally, gestures' starting points and ending points would be detected in order to determine if the distance between them is large enough to sonify it. However, for a realtime application, this procedure is not possible since the sonification would occur once a complete gesture is performed. Figure 4.11 shows several examples of situations where very explicit gestures are performed along with less explicit ones. Explicit gestures are characterized by a salient curve in the velocity signal while gestures of low velocity are usually less significant.

One can easily argue that to be considered as a gesture, a movement must exceed an arbitrary path distance threshold α:

$$\left| \frac{d}{dt} g(x_i(t), y_i(t), z_i(t)) \right| \geq \alpha, \tag{4.5}$$

where g refers to the trajectory in space of a given marker. The threshold is set depending on

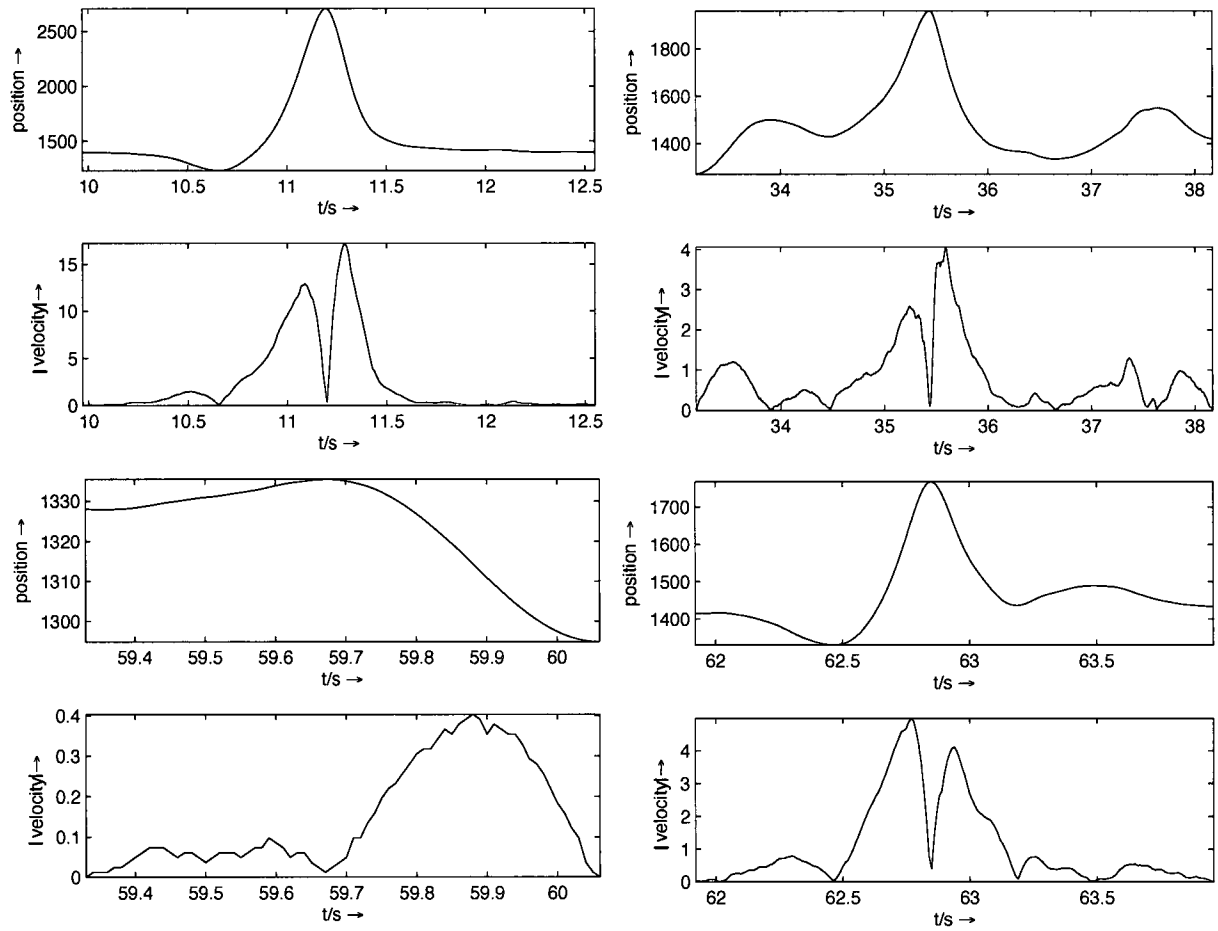


Fig. 4.11 Position (top) and corresponding velocity (bottom) signals of four different gestures.

the context, the type of gesture, or the user's intentions. For instance, if several gestures are to be listened simultaneously, one would be interested in listening only to explicit gestures using a high threshold.

Truncation

Truncation[1] is performed on velocity signals in order to suppress less importants gestures in terms of path distances. Figure 4.12, 4.13 show the amount in percentage of gesture velocity signals that have been truncated according to a threshold that ranges from 0 (no truncation) to 1. As the threshold augments, the percentage of movement truncated augment non-linearly. The slope of the curves are abrupt for low threshold values, which implies that slight changes of threshold values may have great consequences on the truncation of gestures and consequently on the resulting sonification.

Even if each gesture is normalized according to itself, several gestures demonstrate different behavior to the truncation. These differences are due to two different considerations:

- gesture-dependent,
- subject-dependent.

Gesture dependent

As shown in Figure 4.12, different gestures (or feature extraction algorithms) tend to have different responses to truncation. Some of these differences are due to biomechanical considerations, for instance, weight transfer seems to be generally slower than the knee bending with least abrupt transients.

On the other hand, the extraction algorithms influence the velocity evaluation of the gestures as well. Circular movements are detected using the mean average position over a temporal window around half a second long. It leads to a less reactive velocity estimation if compared to instantaneous posture evaluation such as polynomial regression or euclidian distance measurement.

Subject dependent

Figure 4.13 shows how different subjects present different curves for various threshold

[1]Truncation is a signal warping technique and its implementation is described in chapter 6.

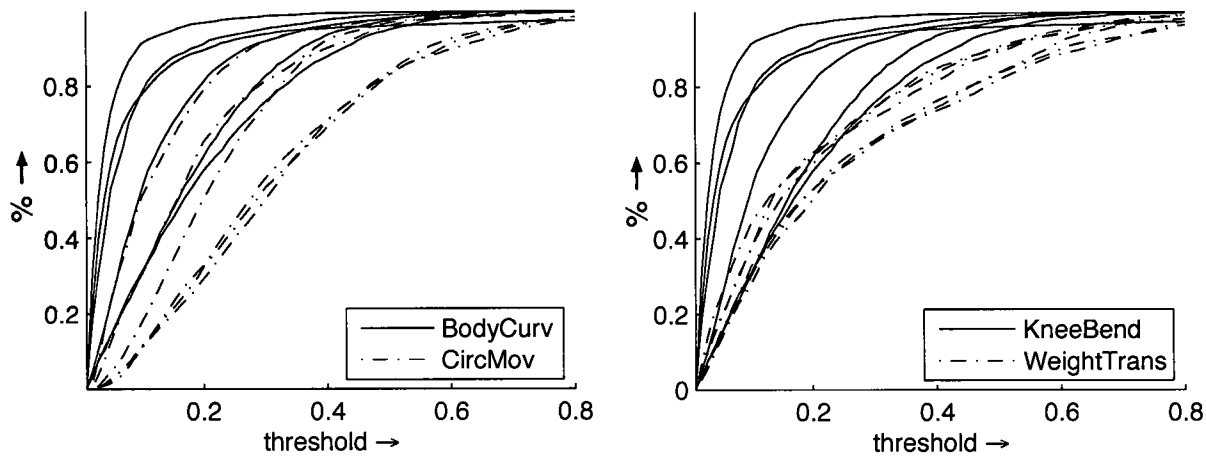


Fig. 4.12 Truncated signals in % for several gestures: body curvature vs circular movement (left), and knee bending vs weight transfer (right). As the truncation threshold augments, types of gesture behave differently.

values. Some subject perform faster gestures than others or emphasize certain gestures compared to others. For example, considering knee bending on Figure 4.13, a threshold set at 0.1 truncates from 35 to 60 % of the gesture signal to be sonified. While a threshold set at 0.05 still ranges from 25 to 45 % of truncation depending on the subjects considered.

Truncation effect

Unfortunately, a high threshold tends to delay the attack of the sound in comparison to the actual gesture's transient. Looking at Figure 4.14, a threshold that exceeds 0.1 (normalized scale) may produce perceptible delays larger than 100 ms for medium velocity gestures (normalized velocity around 0.5). The amount of delay becomes larger as the considered gesture is slow. Such delays are definitely not suitable for gesture sonification as they may confuse the user's ability to associate a specific sound to its corresponding gesture. Figure 4.14 demonstrates that the fast gestures are less affected by the truncation than the slow ones. For a similar threshold value, the delay between the actual gesture (the visual stimuli) and the sonification is less important (below 50 ms).

In an ideal situation, assuming a constant acceleration a, the amount of delay d given

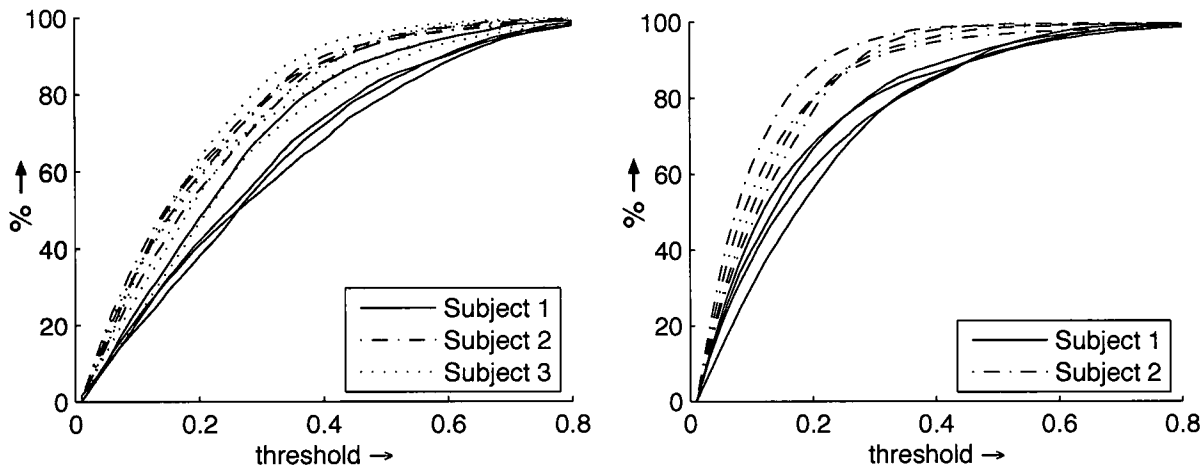


Fig. 4.13 Effects of the truncation threshold in percentage for several subjects: body curvature (left) and knee bending (right).

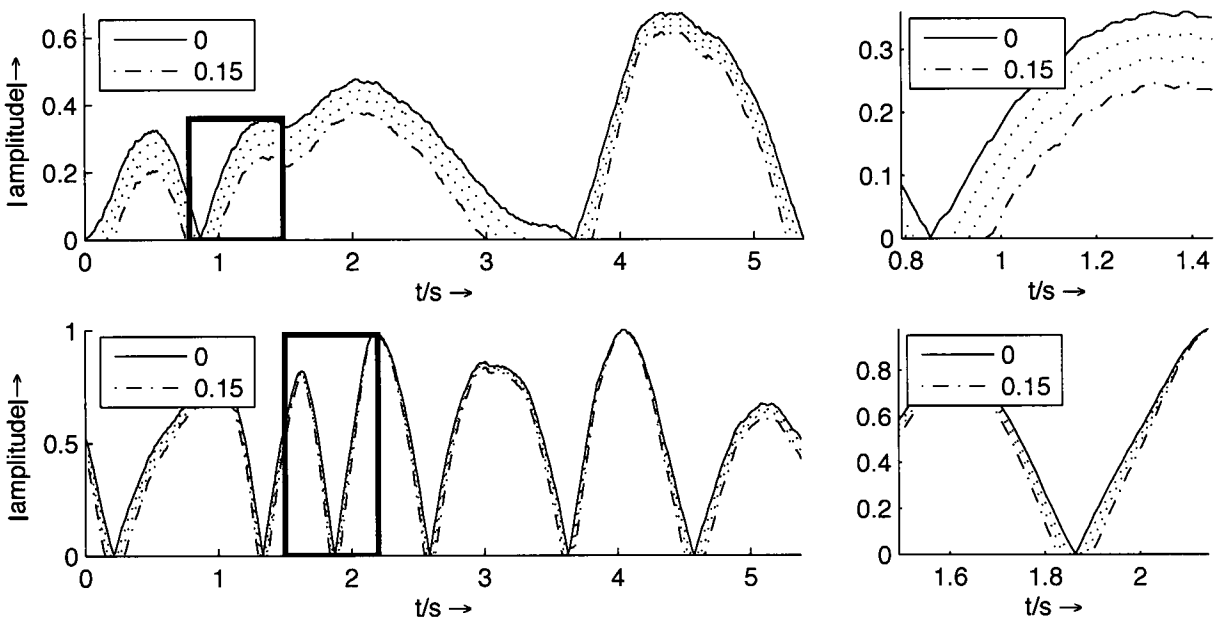


Fig. 4.14 Effects of truncation on gesture of medium (top) and fast (bottom) velocity. A high threshold delay medium gestures' sonification by temporal value greater than 50 ms.

a velocity threshold v_{thresh} is determined by:

$$d = \frac{v_{thresh}}{a} \tag{4.6}$$

However, gesture acceleration is definitely not constant nor analytic (see Figure 4.15) and devising a compromise between truncation effects and the amount of truncated gesture signal cannot be achieved analytically.

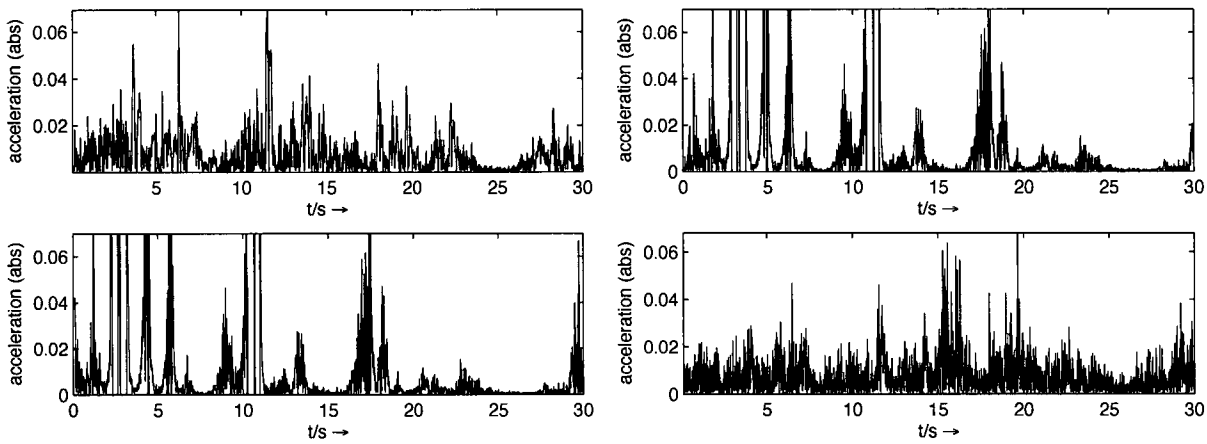


Fig. 4.15 Acceleration signals for circular movements of the clarinet bell, body curvature, knee bending, and weight transfer respectively.

Gestures such as the body curvature and the knee bending are fast and are evaluated using reactive algorithm. In practice, efficient truncation thresholds for these gestures range from 0.03 to 0.08. On the other hand, slow gestures or less reactive algorithms require a higher threshold between 0.05 and 0.15, which is the case for the circular movements of the clarinet bell or the weight transfer. Chapter 6 presents other solutions involving more elaborated signal warping techniques than truncation are presented in. These alternative solutions provide better results than truncation especially for situations where several performers having different gestural behaviors are compared.

Conclusion

As movement data presents high variability depending on the subjects, several control parameters are introduced throughout this chapter to restrain the sonification to significant gestures only. These parameters are mostly defined in terms of gesture's velocity and

gesture's path distance. On the one hand, a gesture must be executed fast enough to be visually consistent. On the other hand, a gesture must achieve a sizable displacement in order to contrast with slight waving movements around average positions.

Chapter 5

Data reduction

Data reduction intends to provide a convenient analysis tool that allows for a fast exploration of data. Due to physiological but also technical considerations, there is a substantial amount of information redundancy in the plug-in-gait biomechanical model. Since the Vicon system uses passive markers, information redundancy is essential in order to avoid confusion between symmetrical body parts during the identification process. Only 32 of the 38 markers provide relevant structural and postural information while the other 6 are used to facilitate the recognition task among similar body parts. Considering the amount of data provided by the motion capture system, the various features computed in the biomechanical model, or simply the extra features that can be possibly extracted, it may become inconvenient for the user to manage such a large number of possibilities in a fast and reliable way.

There are gesture features that obviously provide an important contribution to the overall understanding of movement organization such as center of mass. However, several less obvious correlations in the data cannot be completely described with a unique feature, for example: bending the knee to conserve balance while curving the body forward. On the other hand, some gestures are mainly consequences of other distinct ones. The circular movements of the instrument bell are mainly the results of several less noticeable articulations (i.e. elbows, wrists, shoulders). Gesture features such as those previously described are therefore reduced descriptions that together combine several sources of information. Questions remain regarding the selection process that can be applied to this collection of features. Are there any subsets that are always relevant regardless of the individual?

Principal component analysis is used in this project as it provides an objective method to evaluate the correlation within a given data set and thus, indicates the most important features characterizing this data set.

5.1 Principal component analysis

Principal component analysis (PCA) is a technique used in statistics to simplify a high-dimensional data set into a set of lower dimension while preserving the main information present in the original set. It is alternatively named the discrete Karhunen-Loève transform, the Hotelling transform or the proper orthogonal decomposition (POD) depending on the field of application. A detailed description of the mathematics behind PCA can be found in [41], [42], or [43]. The idea is to combine information that demonstrates high covariance within the data set. PCA is a two-step algorithm that includes the decomposition process and the reconstruction process.

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, in order to be reconstructed according to its principal components. These components are ordered so that those with the highest eigenvalues are presented first. PCA is a non-destructive process preserving the global information; a lower dimension data set is obtained by looking at the components that seem to best explain the behavior of the original one (see Figure 5.1).

Given a data set x_n, each vectors x_n are centered (which means their mean value is subtracted) and normalized in order to bring them into a common comparable range. The variance of the projection on a vector u_1 is defined as:

$$\frac{1}{N}\sum_{n=1}^{N}(u_1^T x_n - u_1^T \bar{x})^2 = u_1^T S u_1^T, \tag{5.1}$$

where the mean $\bar{x}$ and covariance matrix S are respectively given by:

$$\bar{x} = \frac{1}{N} = \sum_{n=1}^{N} x_n \tag{5.2}$$

and

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T. \tag{5.3}$$

The projected variance $u_1^T S u_1$ is maximized with respect to u_1 using a Lagrange multiplier:

$$\frac{\delta}{\delta u_1}(u_1^T S u_1 + \lambda_1(1 - u_1^T u_1)) = 0 \tag{5.4}$$

leading to the stationary point

$$S u_1 = \lambda_1 u_1, \tag{5.5}$$

which implies:

$$u_1^T S u_1 = \lambda_1. \tag{5.6}$$

It means that choosing u_1 equal to the eigenvector having the largest eigenvalue λ_1 maximizes the variance for the first principal component.

The covariance matrix of the data matrix X is given by:

$$S = \frac{1}{N}X^T X. \tag{5.7}$$

It can be decomposed in terms of its eigenvector v_i and eigenvalue λ_i:

$$S = \frac{1}{N}X^T X v_i = \lambda_i v_i. \tag{5.8}$$

The first principal component $\tilde{x}_1$ is computed using the eigenvector v_1 which have the larger eigenvalue λ_1.

$$\tilde{x}_1 = v_1^T X. \tag{5.9}$$

Additional principal components are defined using the remaining eignevalues (ordered by decreasing values) and their related eigenvectors. Finally, for each principal component, the related explanation rate τ_i is derived from the eigenvalue:

$$\tau_i = 100\frac{\lambda_i}{\sum_{j=1}^{N}\lambda_j}. \tag{5.10}$$

There are essentially two questions that can be answered with PCA. What signals have redundant information that can be summarized? What signals best characterize a data set

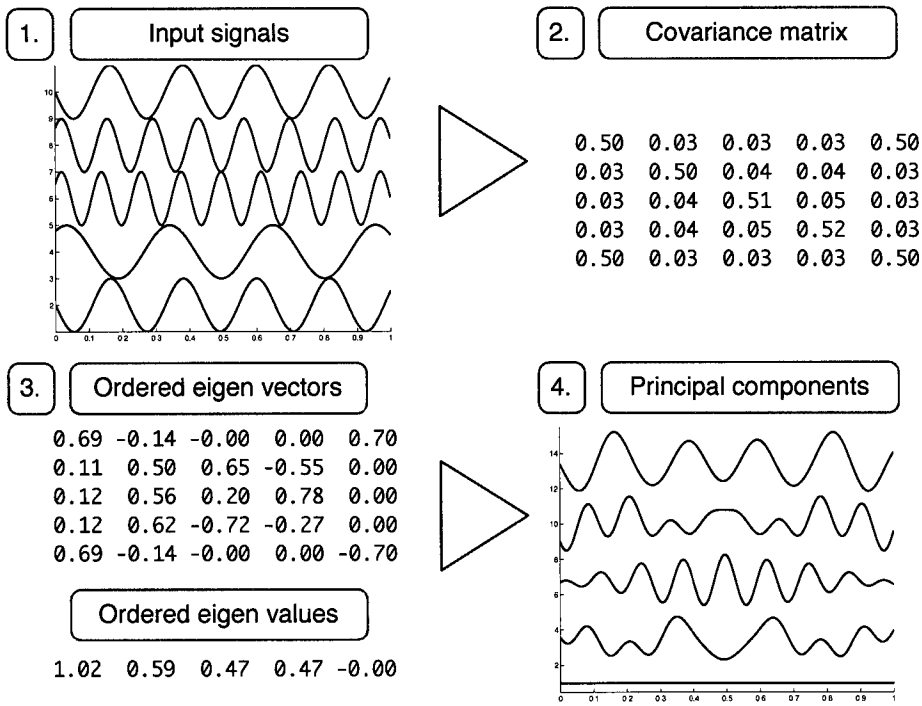


Fig. 5.1 Principal component analysis: decomposition and reconstruction processes.

in terms of linear combination? The first one concerns the purpose of data reduction. The second, leads to an interesting avenue for automatic feature extraction. PCA is extremely informative but has to be handled carefully; the principal components may not most of the time be interpreted in terms of physical parameters. Investigations of periodic gestural patterns using PCA were presented in recent publications and consider the specific cases of hula hooping [44], juggling [45], and violists/clarinetists' center of mass cite[Delphine].

5.1.1 Center of mass

When applied to a data set that includes all the position markers of a clarinet performer, the first three principal components are clearly associated to the motion of the center of mass along the three main axis. From this perspective (i.e. considering all position markers), the correlation between them is very strong when weight transfers are performed. It is sufficient to describe 85 % to 90 % of the markers' movement, depending on the performer.

The subsequent principal components (that represent almost 15 % of the markers' movements) are necessarily related to the gesture features even if their respective eigenvalues are low. Compared to weight transfers that involve the whole body, there is less data available for specific body parts such as head, arms or legs. It reduces the amount of correlation for these specific parts. An alternate approach needs to make the principal components related to local parts of the body appear in the first instances. There is no guarantee that a principal component related to a specific feature will get the same eigenvalue every time and thus, will appear at the same rank.

One solution is to consider only markers and features that belong to a specific body part. The whole marker set can be roughly separated in 4 different subparts: the head, the upper trunk, the lower trunk, and both legs. Correlations within a given subgroup of markers for each body part are thus improved. In the following, to get a better idea of the behavior of specific parts of the body, PCA is applied to selected groups of markers/features in order to identify the signals that share similar behaviors with the first principal components (and so contribute the most to them).

5.1.2 Head position and orientation

The movements of the head are described by the plug-in-gait biomechanical model using four three-dimensional coordinates at different extremities of the cranium (LFHD, RFHD,

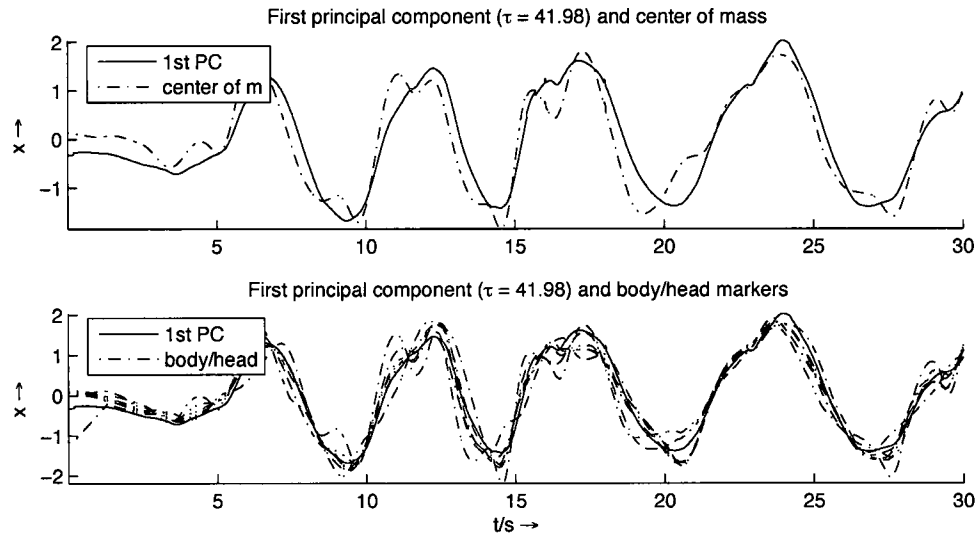


Fig. 5.2 1st principal component x-projection of subject 4. The PCA is performed on the complete data set, which provides a first principal component strongly similar to the center of mass and body/head movements.

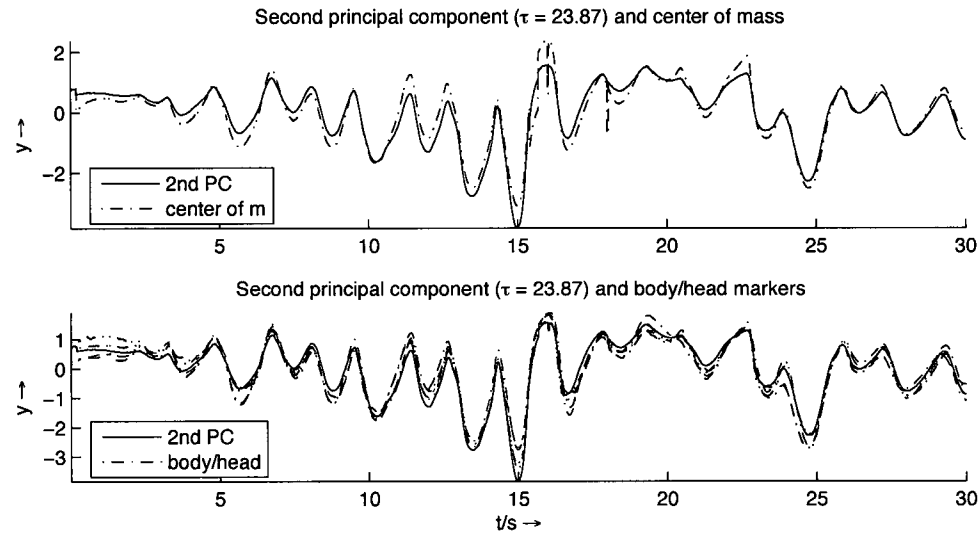


Fig. 5.3 1st principal component y-projection of subject 4. The PCA is performed on the complete data set, which provides a second principal component strongly similar to the center of mass and body/head movements.

LBHD, RBHD) and two sets of angles describing respectively the head orientation and the neck articulation. Figure 5.4 shows how the four positions behave similarly. Differences result from changes in head orientation. Therefore, it is possible to reduce (without an important loss of information) the number of signals from 12 coordinates and 6 angles (head and neck) to 3 coordinates and 3 angles describing its orientation.

The position of the head is simply derived by taking the mean average position of the 4 markers. Figure 5.5 demonstrate the close relationship between the position, the orientation and the principal component of the head over the x axis. Even if the center of mass is included in the data set processed by PCA (in order to include the information of the whole body), it is interesting to note that it does not contribute to the overall understanding of this subset's motion. Similar results are obtained for the y and z axis. However, the correlation between angle and position is less obvious for the latter, as seen in Figure 5.6. Figure 5.7 demonstrates that principal components obtained from subjects performing few slight movements (subject 2 and 3) present a certain amount of variability if compared to ones obtained from subjects performing very explicit movements (subject 4), which follow exactly the corresponding gesture feature.

5.1.3 Body positions and orientations

The trunk data can be divided into two distinct groups of four markers demonstrating distinctive behavior: upper and lower trunk. The upper body set is composed of the sternum (STRN), clavicle (CLAV), 7th cervical vertebrae (C7), and 10th thoracic vertebrae (T10) markers while four markers set at the extremities of the pelvis describe the lower body (LASI, RASI, LPSI, RPSI). The additional information provided by the model concerns the angles of the spine and pelvis. Applying PCA to the overall data set extracts principal components that respectively fit the spine and pelvis angles as demonstrated in Figure 5.8. Therefore, these two angles are good summaries of the movements for this part of the body. The biomechanical model provides only very partial information about the thorax and the abdomen, and is therefore inadequate to detect breathing. Both upper and lower parts of the body are considered to be unarticulated. Similarly to the head subset, the two position subsets can be reduced to a unique mean position combined with an orientation. From this description, extra features can be easily ascertained from the difference between angles, for instance: the overall body orientation in the horizontal $x - y$ plane, the torso torsion, or

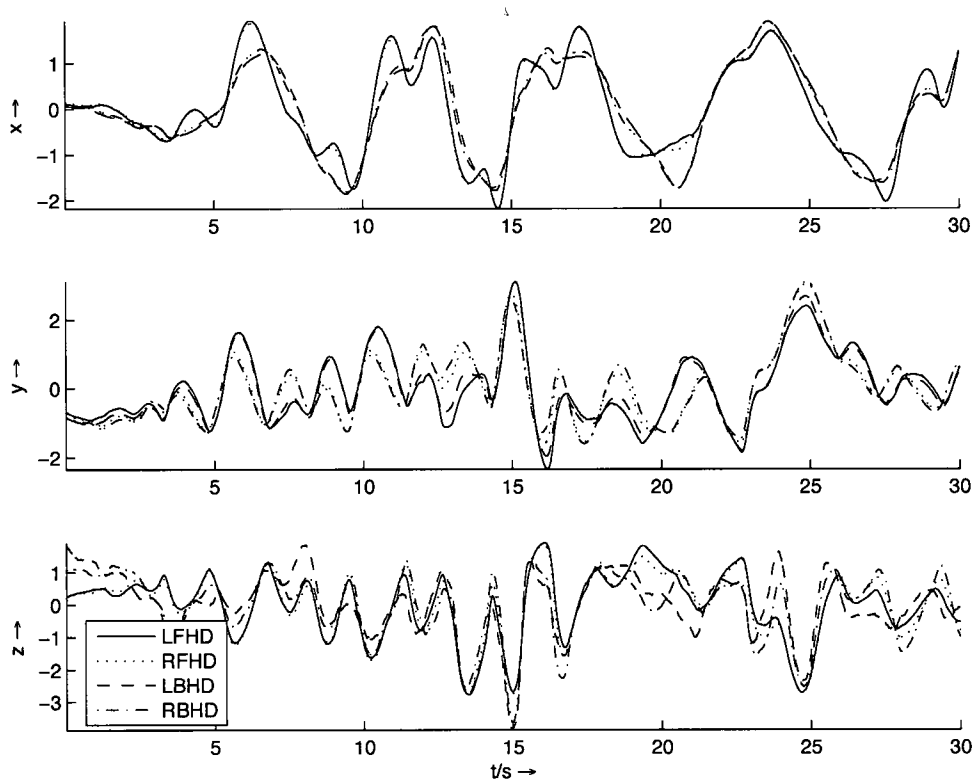


Fig. 5.4 Comparison between head marker positions of subject 4.

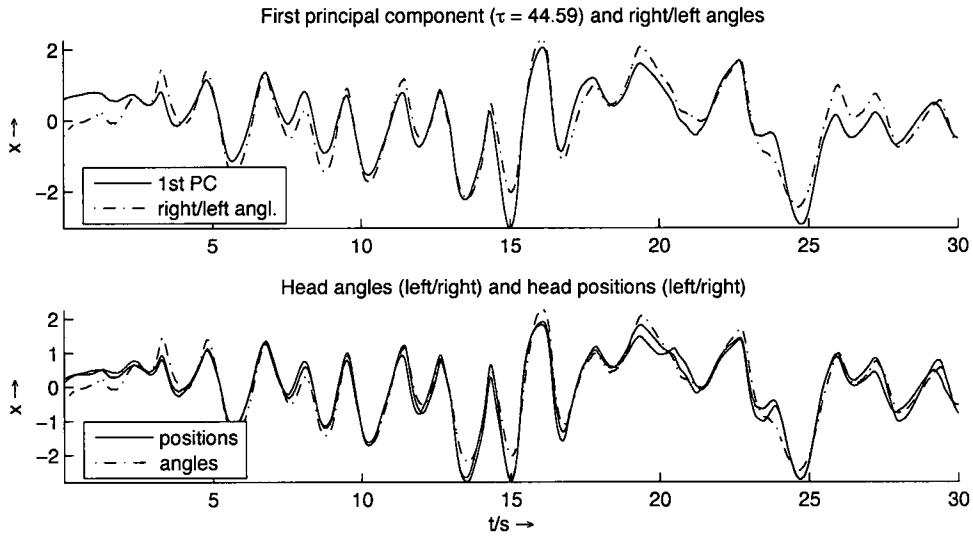


Fig. 5.5 1st principal component x-projection of subject 4 compared to head orientation and position (PCA performed on the head subset).

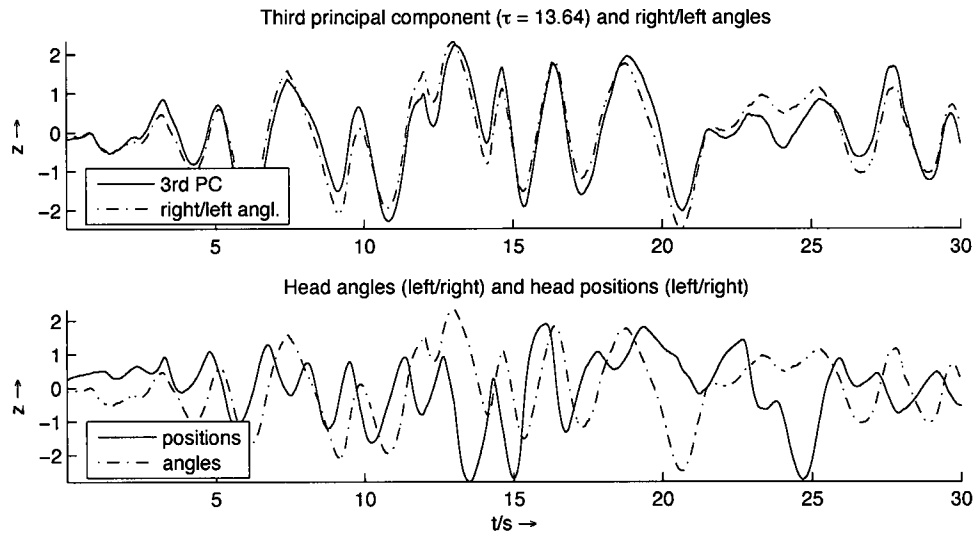


Fig. 5.6 3rd principal component z-projection of subject 4 compared to head orientation and position (PCA performed on the head subset).

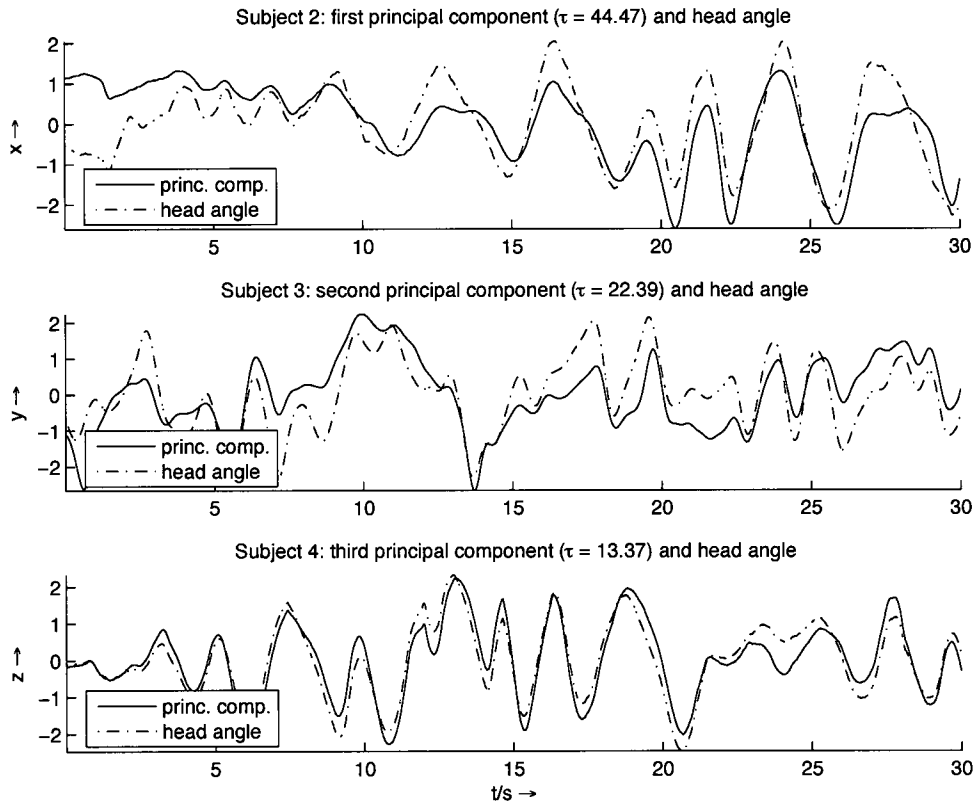


Fig. 5.7 Principal components for subjects 2, 3, and 4 (PCA performed on the head subset).

the relative orientation of the head with body.

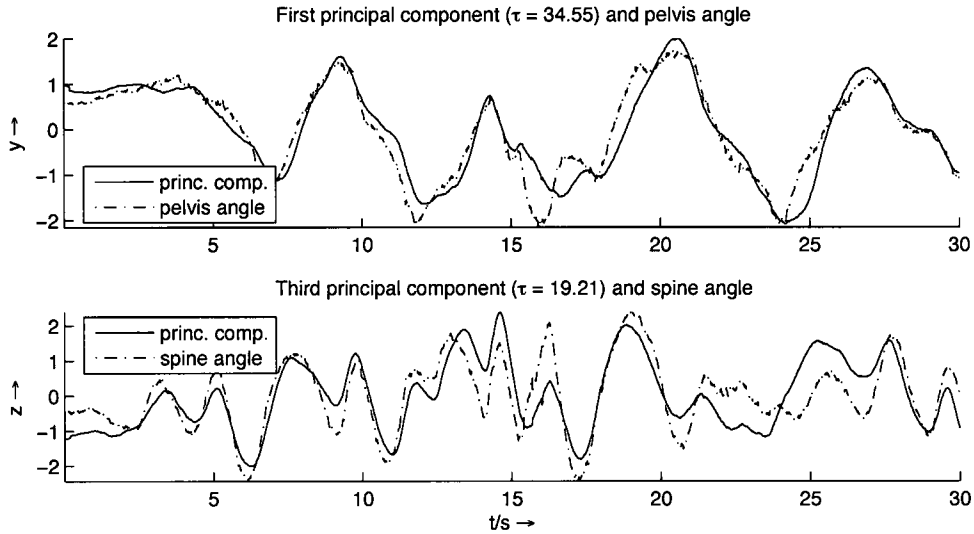


Fig. 5.8 Principal components of subject 4 (PCA performed on the combination of upper trunk and lower trunk subsets).

5.1.4 Leg positions and angles

Leg movements involve a greater number of joint articulations than other parts of the body considered so far. Furthermore, they are relatively independent and may perform distinct movements. Three angles describe the movement of each leg: the ankle, the knee, and the hip. The model provides several information positions such as ankles (LANK, RANK), knees (LKNE, RKNE), tibia (LTIB, RTIB), and thigh (LTHI, RTHI).

When PCA is applied to the leg positions/features data set, the first principal component is associated with left-to-right weight transfer. More interestingly, the subsequent components are related to knee movements. Even if the PCA is performed on the entire leg subset (i.e. combining the two legs), Figure 5.9 shows that pertinent information about both knees can be extracted. For x and z angles, the principal components behave the same way the mean average would. They are more reactive to changes even if the movements alternate from one leg to the other. A closer look at Figure 5.9, especially the z axis, reveals a smoothing of the resulting principal components. This benefit is mainly derived from the fact that the information's presence in both angles and positions. With

PCA, the noisy components are averaged during linear combination; this smoothing is a very desirable property. Applying PCA to each leg separately improves the results, the principal component following the movements of one knee.

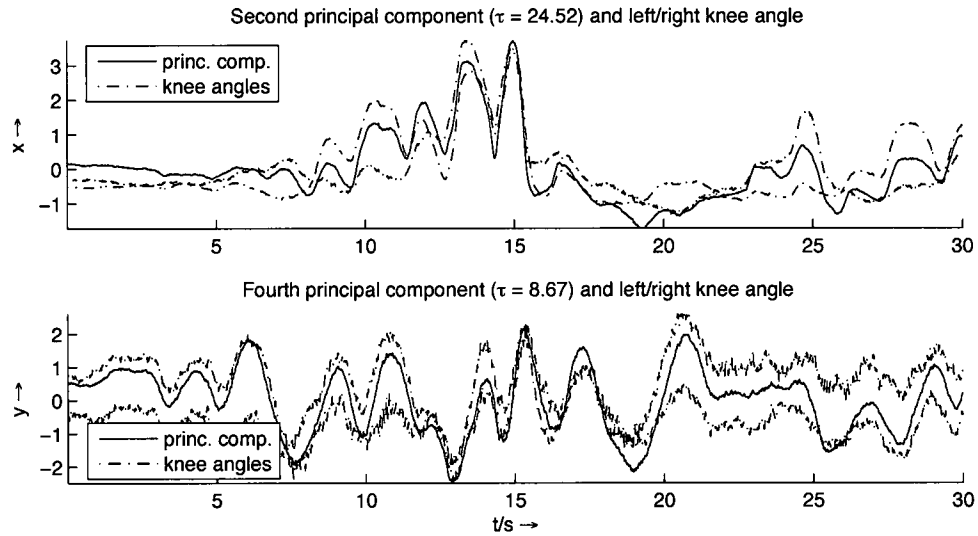


Fig. 5.9 Principal components of subject 5 (PCA performed on a combination of left and right leg subsets).

5.2 Reduced model

According to the previous investigation of markers/features subgroups, the reduced model consists of the following positions:

- head mean position,
- C7, T10,
- pelvis mean position,
- left and right knees,
- left and right wrists,

and angles:

- head orientation,
- spine angle,
- pelvis orientation,
- left and right knee angles,

corresponding to a reduction from 165 signals to 33 signals. An identical set of angle has been selected for both model to reinforce information related to specific articulations. Performing PCA on the complete set of angles results in several complex eigenvalues in the very last components. This is due to the fact that information conveyed by these components is redundant with information conveyed by other components. This information can be discarded as it does not convey any additional significance. There is no such problem using the proposed subset of angles, which provide exclusively new information.

Figures 5.10, 5.11 shows two performer's sets of principal components. The components of the complete marker set are displayed on the left. The first 5 or 6 reduced model's principal components (displayed on the right) strongly resemble ones of the complete data set for every subject and performance. Furthermore, they appear at the same position according to their respective eigenvalue. In general, the set of the first 20 principal components of the complete data set includes the first 15 principal components of the reduced model. However, subsequent components may not appear at the same position and they have to be reordered in order to compare them.

As already assessed, the three first components correspond to the center of mass. However, changes in the position of the center of mass are produced by several other gestures. As an example, movements of the center of mass along the z axis (2nd component in Figures 5.10, 5.11) are related to body curvature (5th component). Looking closer to eigenvectors for this specific PCA in Figure 5.12 (which represents contributions of different signals to resulting principal components), the signals that contribute the most to the 5th principal component are the positions of the head, T10, C7, and pelvis combined with angles of the head and spine along the y axis, which refer to body curvature. Even if the 2nd and 5th components appear very similar, the information summarized by these components is not exactly the same: one refers to the center of mass, the other to body curvature.

The first 6 components provide enough information to explain more than 95 % of the variance within both data sets as demonstrated in Figure 5.13, 5.14. The clarinet bell's

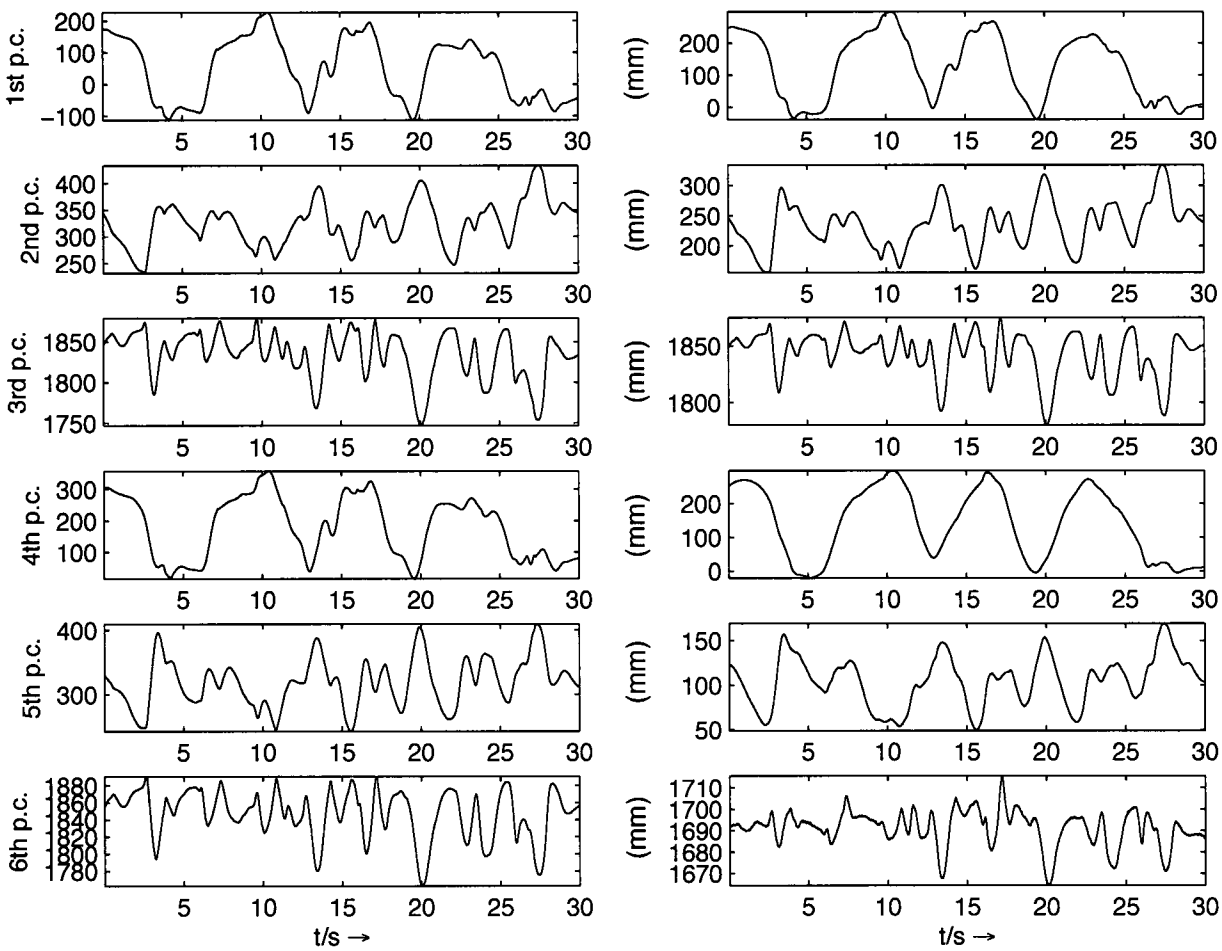


Fig. 5.10 PCA performed on both complete model (left) and reduced one (right) for subject 4.

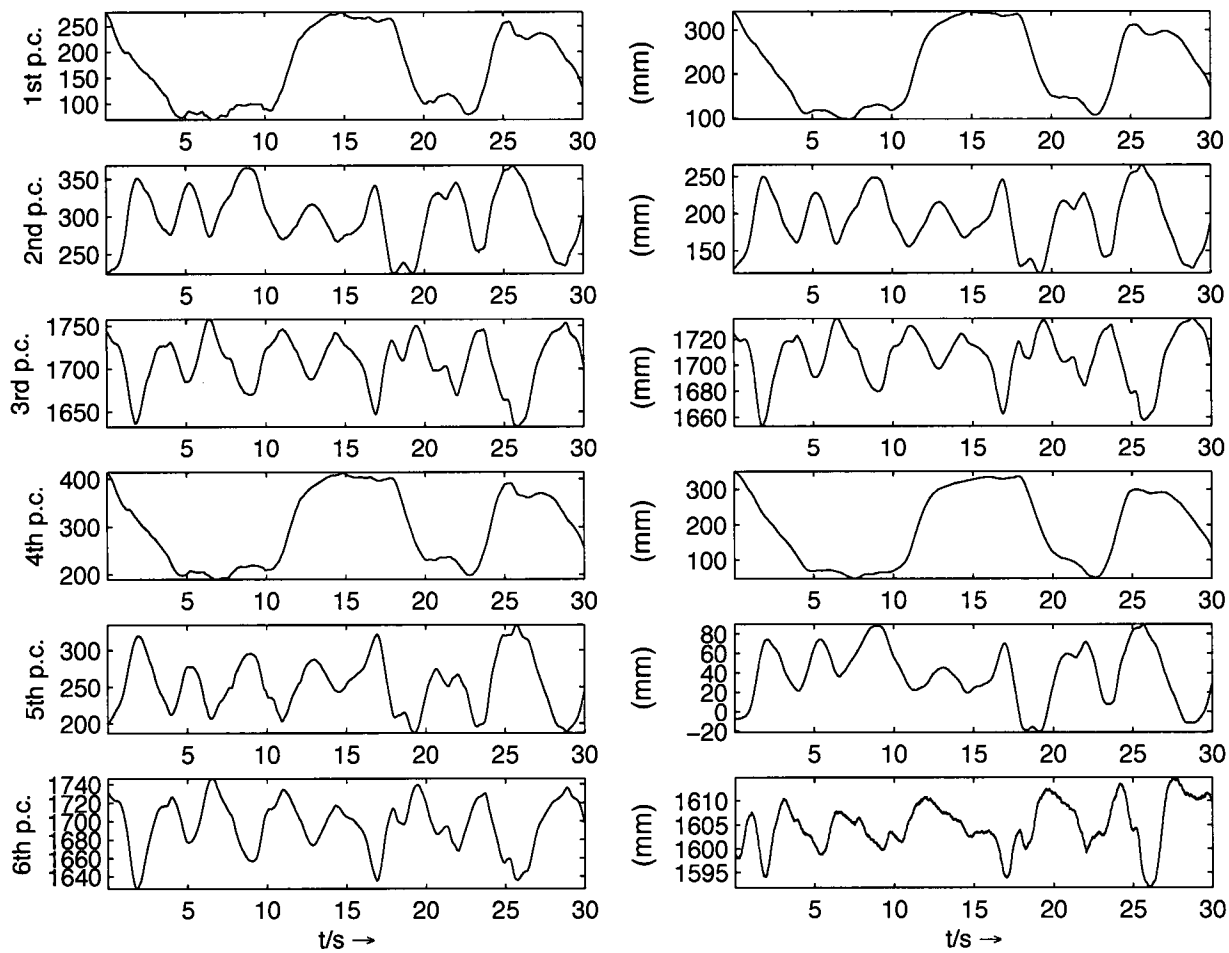


Fig. 5.11 PCA performed on both complete model (left) and reduced one (right) for subject 6.

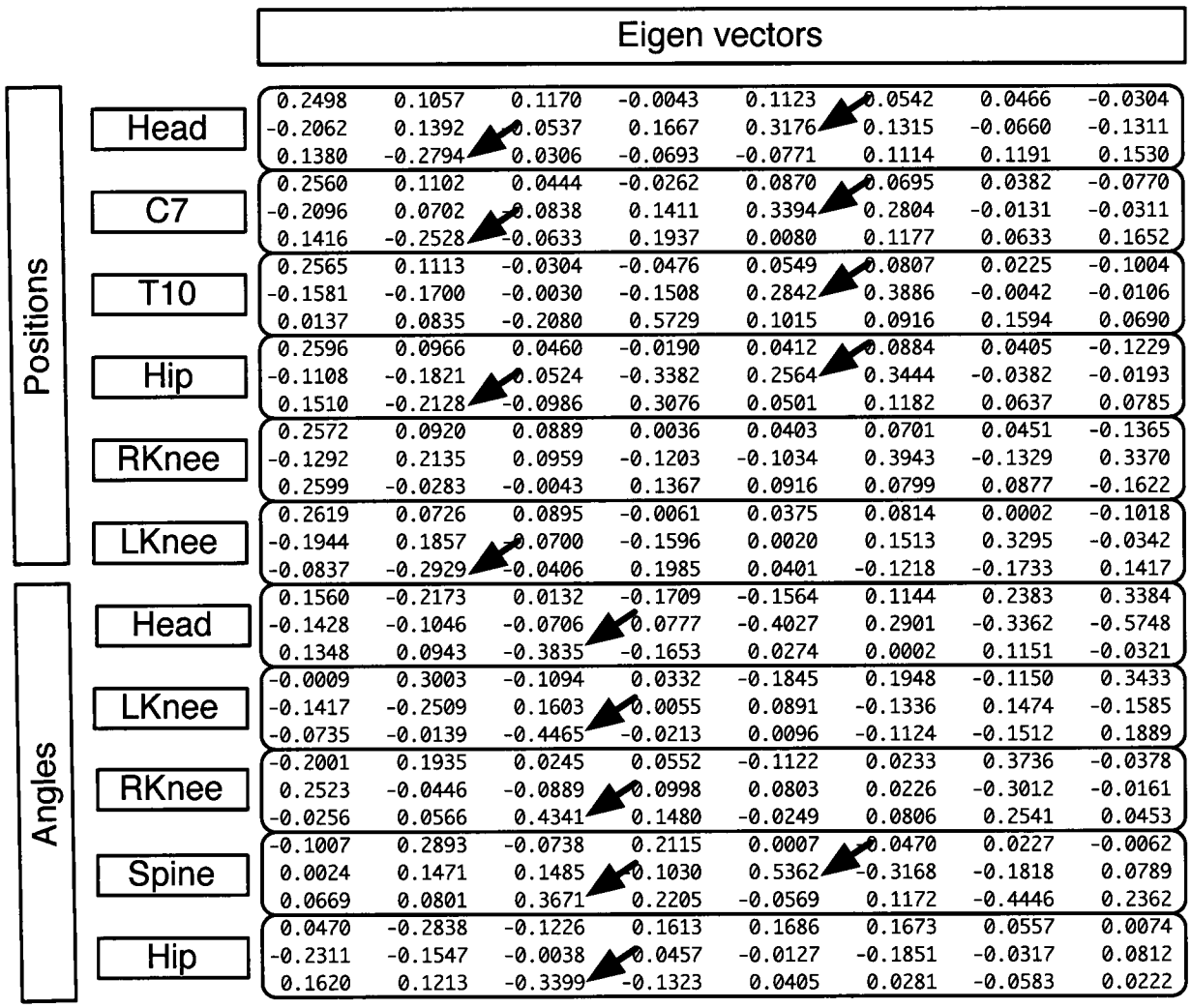


Fig. 5.12 Example of eigenvectors. The arrows point to high coefficients in terms of linear combination once the original data set is projected into the new basis.

circular movement, present in the complete model through signals describing the hands, is missing from this group of principal components. Markers showing this movement perform with relative independence with the rest of performer's body. Information related to circular movement thus appears in components of little importance despite the fact that it is an important gesture to consider. It is thus possible to claim that the reduced model is a good summary of the original data set and provides relevant information about the performer's ancillary gestures.

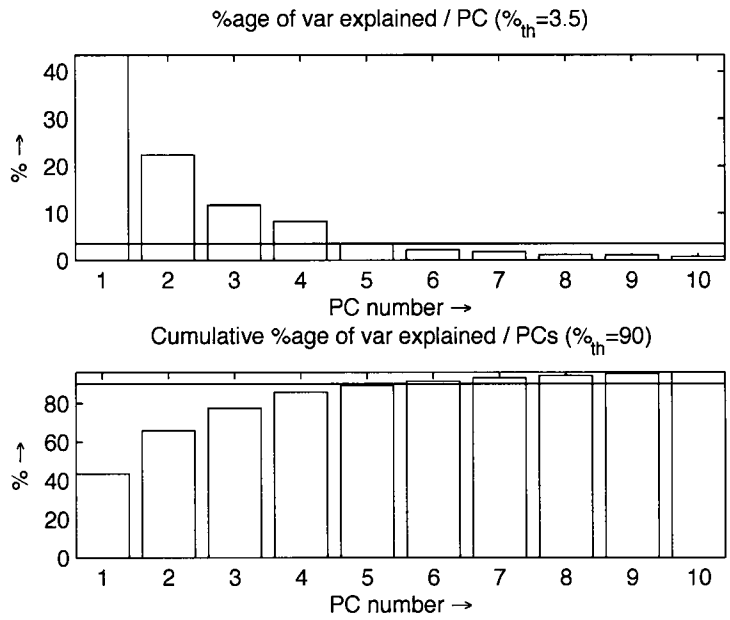


Fig. 5.13 Percentage of explanation for subject 4's complete marker model.

Even if principal components of both models are very similar, their respective ranges may differ from one model to the other. This means that in the complete model, this information is shared by several markers/angles while in the reduced model, this information is preserved and localized into very specific parts (markers/angles) of the body.

5.3 Realtime usage of PCA

The results presented in sections 5.1 and 5.2 demonstrate that PCA can be used to extract the main information within a data set. The interpretation of eigenvectors in terms of

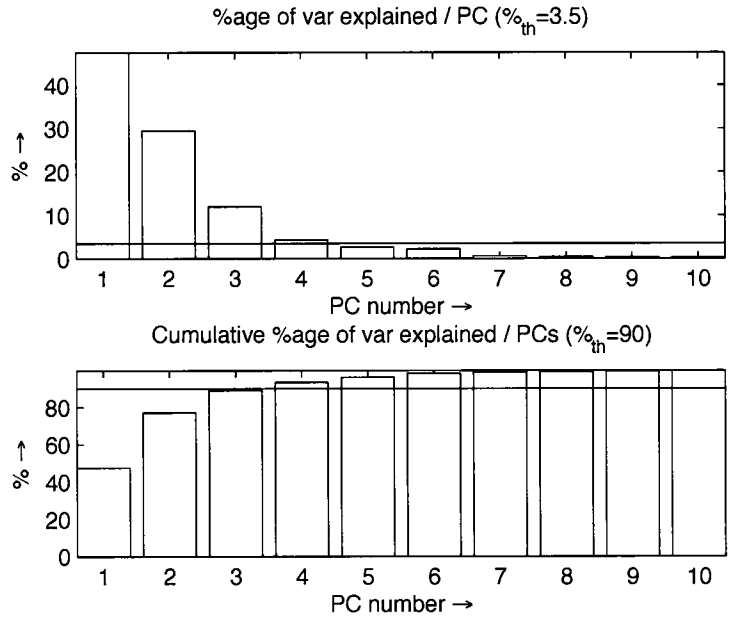


Fig. 5.14 Percentage of explanation for subject 6's complete marker model.

weighting coefficients leads to a realtime usage of PCA. One can thus listen to principal components instead of specific gesture features. A PCA is performed offline on the sample data set during a calibration process. It is assumed that the sample performance of the subject considered depicts his overall gestural behavior. The resulting eigensystem is then loaded into the sonification system. The realtime data stream is projected onto this static eigensystem.

For a signal stream of n signals $s(t) = \{s_1(t), s_2(t), ...s_n(t)\}$, given a set of eigenvectors $v_i = \{a_{i,1}, a_{i,2}, ..., a_{i,n}\}$ corresponding each one to a specific principal component, $i = 1, 2, ...n$, the feature extraction is simply computed with the following linear combination:

$$s'_i(t) = s(t) \cdot v_i = \sum_{k=0}^{n} v_{i,k} s_k(t).$$

One important benefit of this approach concerns the fact that no subjective appreciation of the gesture to be sonnified is required. For instance, when a subject demonstrates a preference to perform weight transfers along the y axis instead of x, this movement should appear in the very first principal component. This characteristic is particularly convenient

as a first exploratory technique. Even with no prior knowledge, a coherent sonification can be rapidly designed. Figures 5.15, 5.16, 5.17 show the differences between the principal components obtained from the projection into the data set's eigensystem and the projection of the data into a calibration eigensystem sampled from the same subject.

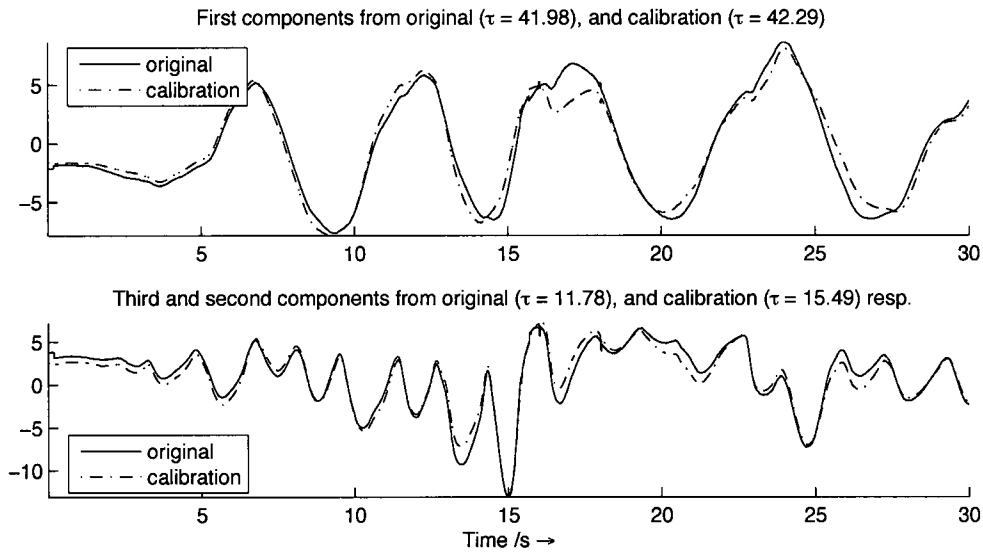


Fig. 5.15 Subject 4's principal components generated using original and calibration eigensystems of the entire data set.

This technique can be limited in that it is difficult to predict gesture features related to a given principal component. Identification of the features has to be achieved empirically and may become clouded by alterations of the original information. Among these signal alterations, the most obvious is slight changes in slope. Some curves are more pronounced than others, and this may have effects on the signal scale. In general, subjects that perform clear and explicit gestures see their principal components matching their related feature accurately (refer to Figure 5.7).

Additionally, data has to be centered and reduced, what leads to an eventual loss of relative differences between the signals. Another alteration not depicted in the figures concerns the sign of principal components that can be inverted without any physical reason. Positive or negative components may arise during the eigen decomposition process and the user needs to keep in mind that both must be systematically tested when listening to principal components while looking to the video.

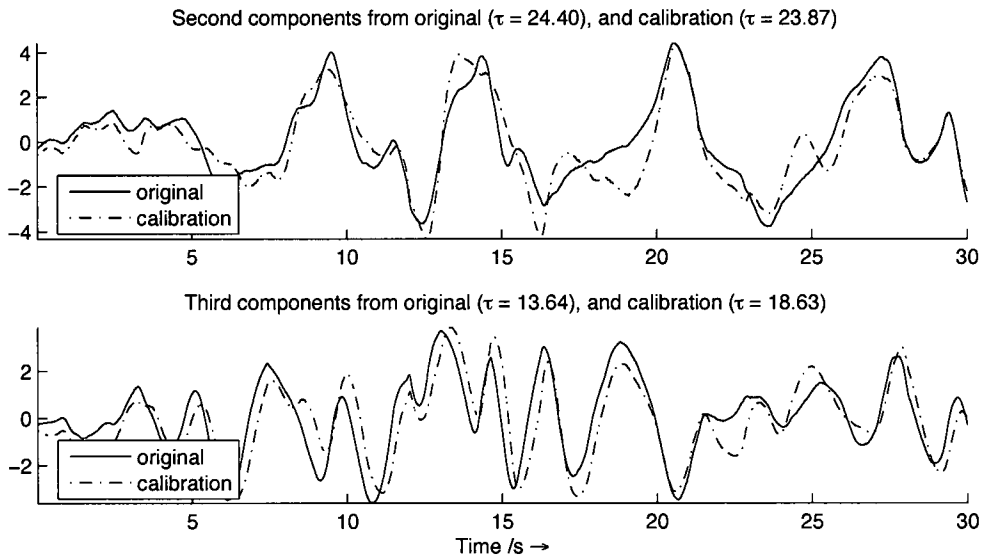


Fig. 5.16 Subject 4's principal components generated using original and calibration eigensystems of the head subset.

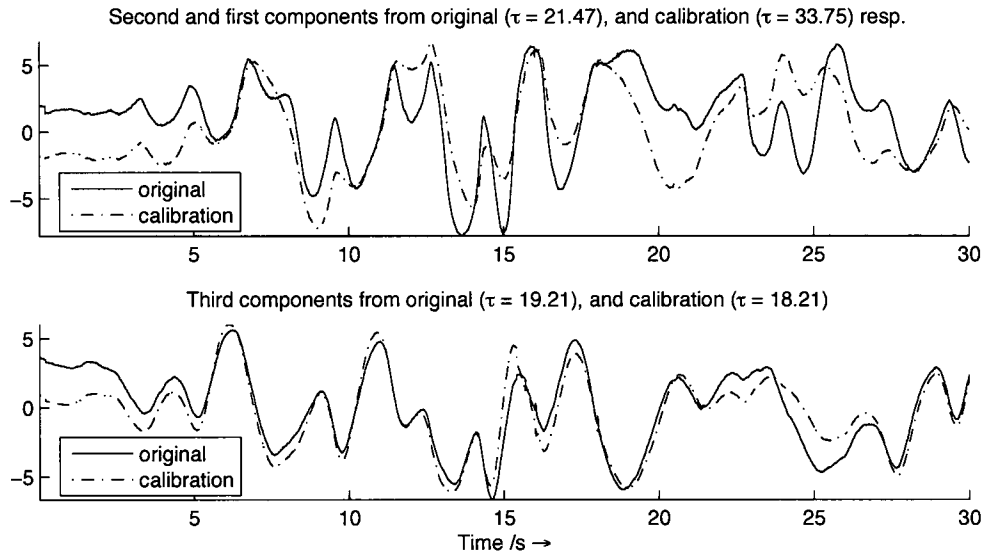


Fig. 5.17 Subject 4's principal components generated using original and calibration eigensystems of the body subset.

Conclusion

This chapter introduced a data reduction process to extract signals that best describe the main characteristics of the original motion capture data set. Through PCA, the combination of signals having high covariance leads to a subset of signals which preserve the overall information contained into the complete model. From the perspective of sonification, the reduced model is simple and convenient to handle. Features selection and mapping onto sound synthesis parameters is then facilitated.

Chapter 6

Mapping and gesture sonification

Chapter 4 presented techniques used to acquire information about gestures. The resulting gesture feature signals are seen as vectors conveying the gesture information to the sound synthesis algorithms. This chapter concerns components inherent to signal conditioning that allow for the coherent transformation of the gesture feature from a geometric representation to a sonic representation. The position signals, the gesture features, or the principal components, will be simply referred to as gesture features. Several sound synthesis techniques have already been presented in chapter 3. Rather than introduce new sound synthesis techniques, the discussion will stress the processes used to condition the data for sound synthesis requirements.

6.1 Mapping in the context of sonification

6.1.1 General principles

Mapping essentially consists of all the processes that play a role in linking the control signals to control parameters of either sound synthesis or sound effect algorithms. General discussions of mapping consider different types of control signals that regroup features extracted from gestures, sounds [1], or even music compositional parameters. The present work is more concerned with mapping gestural variables to sound synthesis control parameters, a subject of such interest in developing new controllers for musical expression. In the context of sonification, mapping is an essential operation central to data exploration. A different

[1]For instance in the context of adaptive audio effects [46], [47].

mapping produce different data enhancements the same way it modify the possibilities and playability of digital musical instruments as originally discussed in [48].

Mapping schemes presented in the literature [23] [24] classify different mapping strategies using several criteria: simple/complex, explicit/implicit, and static/dynamic. Simple mapping refers to a one-to-one linking: each control parameter corresponds with a unique control signal. This strategy is most appropriate for sonification as it avoids confusion between the different control signals that could arise from many-to-one or many-to-many schemes. Explicit mappings are preferred for sonification over implicit or "black box" ones. These are intended to be clearly understood by the user. The use of explicit mapping offers better visibility as to what is computed. Finally, for the purpose of precision, accuracy, and repeatability, that static mapping strategies are better than dynamic ones.

Mapping does not only concern matching control signals with parameters as it also regroups various signal conditioning processes. Several features extracted from a gesture data set cannot be directly heard through sound synthesis. They require intermediate processes also present in musical contexts. These processes have a direct impact on the nature of information conveyed by auditory events and thus, are fundamental for accurate sonification design. The mapping scheme implemented is depicted in Figure 6.1. *Related-to-gesture* and *Related-to-sound* parameters as defined in [24] are at the extremities while special attention to three intermediate layers: filtering, normalization, and warping due to their impact on the informational content. Since the different types of intermediate processes depend on the feature to be mapped, the discussion will first summarize the general mapping principles borrowed from the field of digital musical instrument design before getting deeper into the signal conditioning.

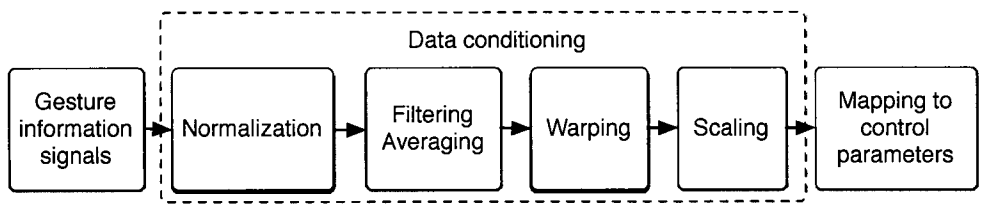


Fig. 6.1 Mapping scheme in a data sonification context.

6.1.2 Mapping of control signals to sound parameters

As suggested in [1], gesture velocity, or more exactly the gesture feature derivative, should be linked to the sound amplitude. This physical attribute of sound is strongly related to the perception of intensity. It follows an ecological approach [31] [32] to the relation between sounds and kinetic events in gesture multi-modal representation [30]. Loud sounds are produced by powerful vibration carrying a lot of energy and are somehow related to high velocity. By contrast, absence of motion should result in no sound at all, which is coherent with the notion of derivative used to evaluate the velocity of gestures.

It has already been discussed that complex mapping is not appropriate for sonification. Reason further to those already expressed in section 2.2.1 is that complex mapping cannot be learned instantaneously [23]. The mapping must remain simple in order to shorten the user learning period, which occurs on two different levels. The first one concerns the actual learning of the sound properties that convey information. The second one consists in assimilating the user interface.

One-to-one and one-to-many mapping strategies are convenient in implementing sonification systems to solve problems related to the management of control signals and control parameters. Each gesture feature is paired with its derivative. The user can choose to apply the derivative of the gesture feature to the sound's amplitude, or simply listen to the signal continuously. This leads to a slightly different interpretation of the sonification. Listening to the position of the center of mass without modulating the amplitude provides information about the absolute position of the overall body in the absolute space; once modulated, the information refers to the actual gesture (i.e. weight transfer).

It appears that one-to-many mapping strategies also enhance a given gesture with multiple sound features, which makes sense in the context of sonification. As an example, to emphasize weight transfers, the left/right displacement is mapped to both the panning angle and the LFO frequency. With some practice, the exact position of the centre of mass can be accurately estimated by combining low oscillation beating effect and the panning angle.

6.2 Normalization

There are three fundamental reasons for using normalization. The first one concerns adapting signals to the requirements of control parameter ranges; to be performed efficiently, scaling requires signals to be normalized. In section 6.3, techniques to slightly modify the behaviour of control signals in order to enhance some of their implicit characteristics will be discussed; these warping techniques require the signals to be normalized. Finally, to compare data from different subjects, normalization can be used to scale the control signals relative by to each other.

Two different types of normalization of the control signals are proposed in the sonification system. Unipolar normalization results in signals that are exclusively positive $[0, 1]$ and is primarily intended for control parameters such as frequency or amplitude that require positive values. The normalized signals $x_{norm}(t)$ are obtained with:

$$x_{norm}(t) = \frac{x(t) - x_{min}}{x_{max} - x_{min}} \tag{6.1}$$

where $x(t)$ is the input signal, while x_{max} and x_{min} are defined respectively as its maximum and minimum value.

Bipolar Normalization is used when the desired signal needs to be rendered between $[-1, 1]$. It is mainly used to preserve the sign and simply achieved by dividing the input signal by its maximum absolute value.

$$x_{norm}(t) = \frac{x(t)}{\max(|x(t)|)} \tag{6.2}$$

Normalization is determined by the type of gesture represented and the control parameters for which the control signal is used. For example, frequencies or amplitudes must be defined in terms of positive values. Panning angles, however, make sense if define as negative values corresponding to "left" and positive values corresponding to "right". Other synthesis parameters such as the ascending or descending fundamental frequency of the infinite glissandi are also defined in terms of positive and negative values respectively.

It is important to note that normalization cannot be performed in realtime without specifying arbitrary maxima and minima. This is a major drawback, requiring the system to be calibrated. Calibration can be achieved by defining a range-of-motion in which

the normalized signals are expected to be constrained. Different normalization limits will lead to different normalization strategies. There are several subjective considerations that restrict the complete and successful automatization of the normalization process.

Inter-gesture normalization

Several gesture feature extraction algorithms will produce several different ranges of information. While the angles range from 0 to 2π, the relative distance in millimeters may produce values greater than 1000. Normalization is required if someone is interested in comparing gestures that are not of the same type. Each control signal is individually normalized so that their respective maxima are the same. The relative difference between gestures is lost and thus, the level of expressiveness is not conserved by this normalization process. It allows for many hard-to-see details concerning the gestural patterns of gestures of different types or different levels of expressiveness to be revealed without perceptual bias due to dissimilar signal amplitudes. On the other hand, it may give prominence to useless information (e.g. noise in the feet position measurement).

Inter-performance normalization

Given a selection of gesture features, both the comparison between different subjects or the comparison of different performances of a same subject require that, for each gesture type, normalization be performed according to a maximum for all subjects. For each gesture, control signals keep their relative differences every subject. This allows for comparison of the performances and their relative gestures' velocity, since the relative amplitude for each gesture of the same type is conserved.

As discussed in section 3.1.2, evaluation of knee bending with the Optotrak system was not robust to changes in hip orientation. Two performances can most likely present significant difference in hip orientation resulting in an offset differences as shown in Figure 6.2. These gesture features must be centered according to their respective mean value in order to compare them.

The histograms of Figure 6.3 present several common situations that occur when investigating gestures. A problematic aspect occurs when a subject occasionally performs wide gestures that are not representative of the overall performance. Such wide gestures will narrow the *range-of-interest* (top left). Intervals of gesture are not necessarily the same as

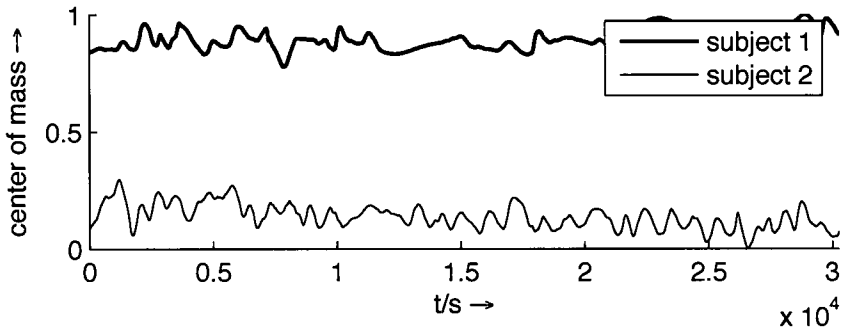


Fig. 6.2 Wrong normalization with offset of two knee bends affected by different orientation of the body.

performer's posture change (top right). One subject may perform very small gestures in comparison with another (center right) or the gestures may be performed most of the in a specific region of the *range-of-motion* (center left). Looking at the center of mass, one subject could systematically perform the weight transfers in the same direction (bottom left).

6.3 Signal warping

Once properly normalized according to a strategy described in section 5.2, control signals are ready to be scaled and directly linked to sound synthesis parameters. Additionally, warping techniques can be applied to the data in order to modify the behaviour of the sonification which in turn enhances/attenuates desired/undesired information. These are optional processes, useful to clarify the sonification as they reinforce certain inherent gestural characteristics. The following discussion is inspired from [49] and adapted for sonification of gestures.

The following are examples of situations where a modification of the behaviour of the control signal would enhance the resulting sonification:

- truncate the data in order to filter out undesired information,
- attenuate very slow gestural information that has been amplified or, as opposed to this, increase significant information that has been attenuated by a comparative normalization strategy,

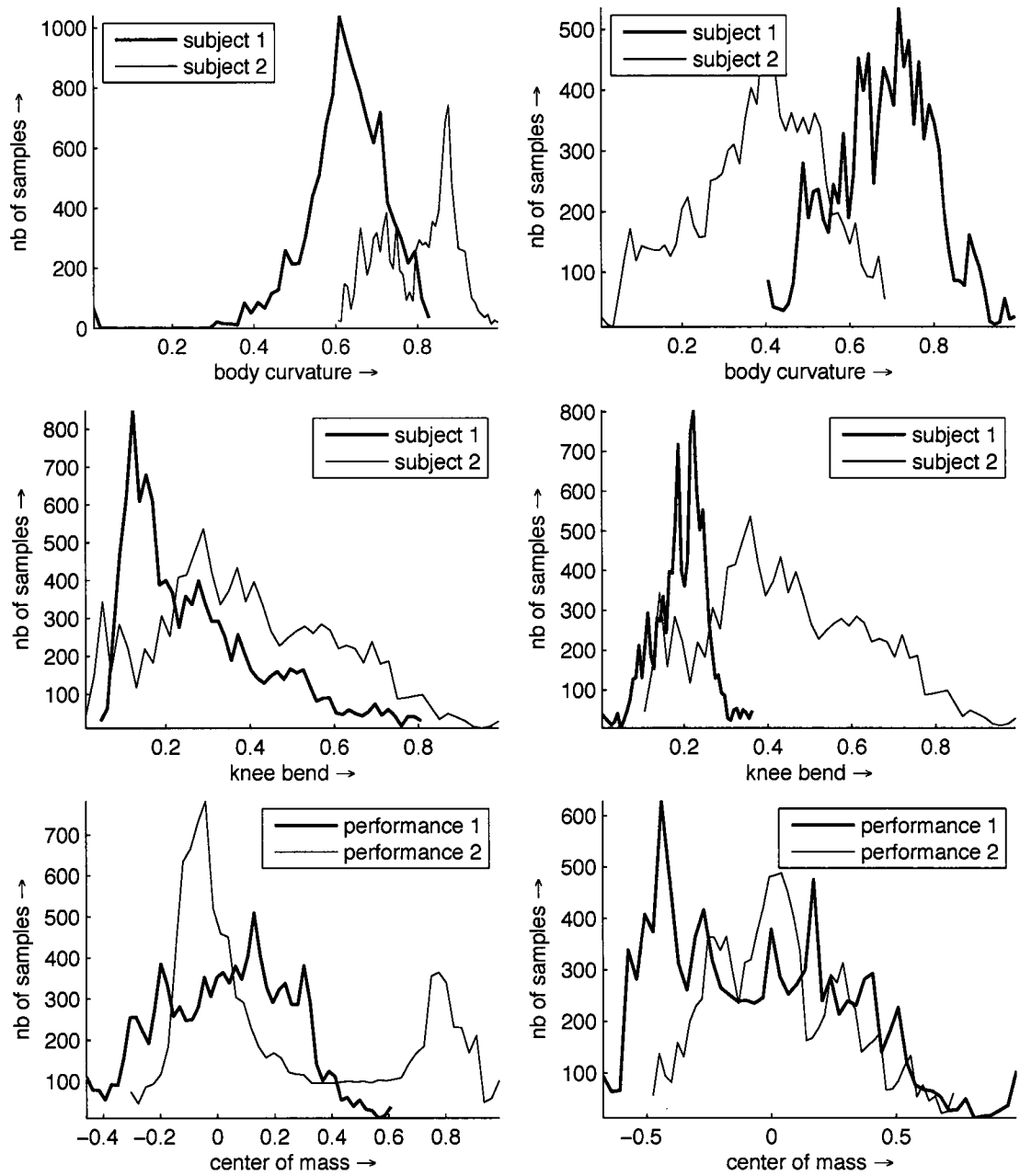


Fig. 6.3 Several histograms of different gestures.

- enhance variation within a signal to emphasize different characteristic positions,
- warp the signal in order to exploit the full range of a sound synthesis parameter.

The input signals $x[t] \in [0, 1]$ are modified using a transfer function $H(x[t])$ stored in a lookup table.

$$y[t] = H(x[t]) \tag{6.3}$$

Signal warping functions are chosen according to the physical behaviour they model into the signals [47]. Warping techniques must offer also significant parameters to the user in order to precisely quantify the modification applied to the signals. These deterministic operations present several advantages: precision of the quantification, repeatability, and realtime processing.

6.3.1 Application to amplitude

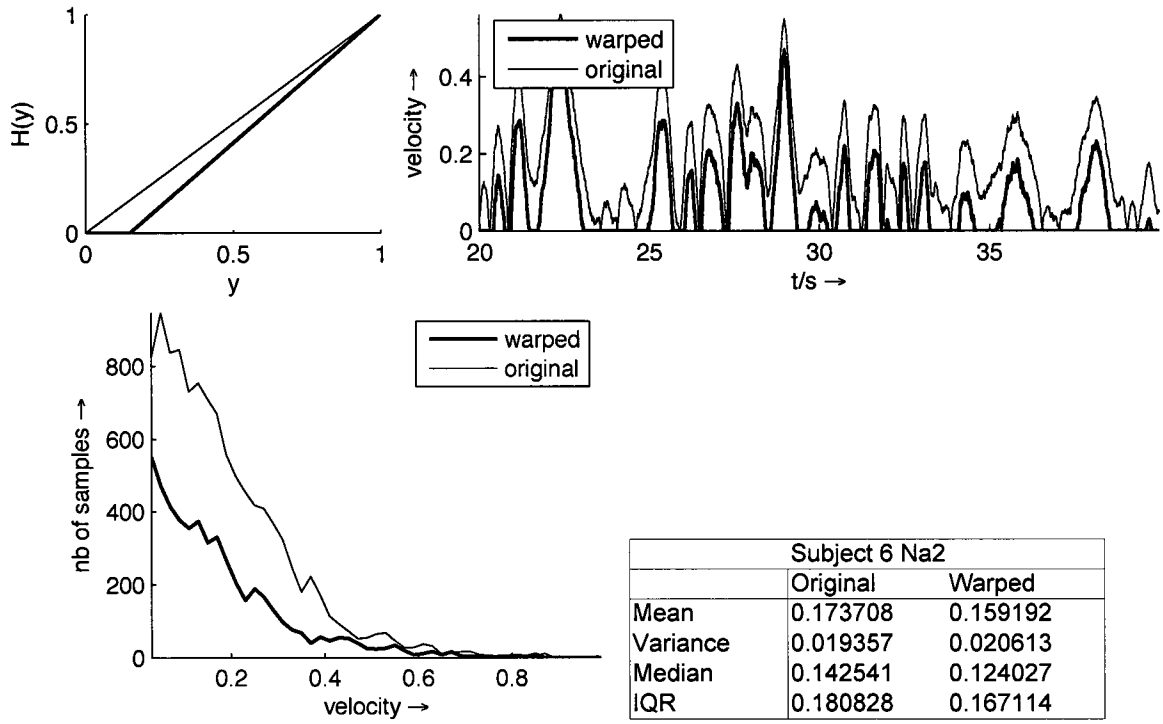
Truncation

In chapter 4, the evaluation of the gesture's velocity was introduced with a discussion on filtering out what is arbitrarily considered not to be gestural. Truncation can be applied to other parameters, but its application to amplitude constitute a good example of where this conditionning technique is required. Every value of an input signal $x[t]$ that is below a certain threshold is set to zero. In order to conserve the original range $[0, 1]$, the truncated signal must be stretched out.

As depicted in Figure 6.4, even if stretching the signal compensates for the reduction in amplitude and that the mean value[2] of the histogram is approximately the same, the result is that the gestures of low amplitude signals are significantly reduced in comparison to the high amplitude ones. It could become confusing and even impossible to detect gestures of low amplitude when several sonifications of different gestures are streamed simultaneously. This situation becomes problematic especially when two performances of different expressiveness are compared as the reduction could affect the performance that already presents low amplitude signals.

$$x_{trunc} = \frac{i_1 t_m + i_2 t_M + i_3 x[t]}{t_M - t_m} \tag{6.4}$$

[2]The statistics are computed without considering the 0 value as it indicates an absence of gesture.



	Subject 6 Na2	
	Original	Warped
Mean	0.173708	0.159192
Variance	0.019357	0.020613
Median	0.142541	0.124027
IQR	0.180828	0.167114

Fig. 6.4 Effects of a truncation on the body curvature's absolute velocity, which would bias the truncated signal (transfer function, warped signal, histogram).

where

$$i_1 = 0 \text{ if } x(t) > t_m, i_1 = 1 \text{ if } x(t) \leq t_m$$

$$i_2 = 0 \text{ if } x(t) < t_M, i_1 = 1 \text{ if } x(t) \geq t_M$$

$$i_3 = 1 \text{ if } t_m < x(t) < t_M, i_3 = 0 \text{ elsewhere.}$$

Compression

The use of compressor is common in audio processing. It reduces the dynamic range of an audio signal, especially ones those demonstrating very salient attack transients such as drum or voice. The attack transient may be high compared to the sustain and release part of the signal. The compressor makes sure that the audio signals will not clip. On the other hand, it balances the whole signal by attenuating the high amplitude so that low ones can be amplified. Alternatively, it can be used as an audio effect to literally "compress" the audio signal.

In the context of gesture conditioning, a slight variation of the original audio compressor has been implemented since, as a requirement, the output signal must remain between 0 and 1. Two slopes, $S_1 \geq 1$ and $S_2 \leq 1$ (gain and compressor) that respectively amplify and reduce the signals of low and high amplitude levels are defined given two parameters: the compression threshold m and the compression ratio α.

$$S_1 = \frac{n}{m}, \qquad S_2 = \frac{1-n}{1-m} \tag{6.5}$$

$$S_1 = \alpha S_2, \qquad \alpha \geq 1 \tag{6.6}$$

Under these restrictions the inflection point (m, n) is given by:

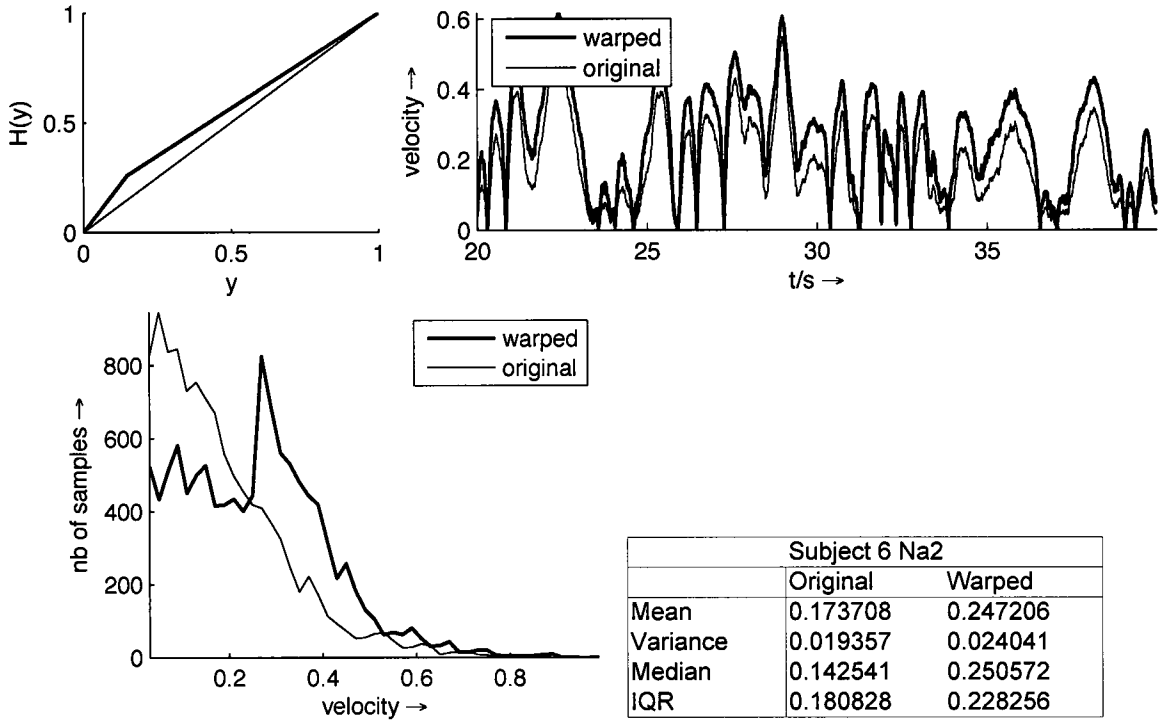
$$n = \frac{\alpha m}{(1 + (\alpha - 1)m)} \tag{6.7}$$

The compression is given by:

$$x_{comp}[t] = \begin{cases} S_1 \cdot x[t] & x \leq p \\ S_2 \cdot (x[t] - m) + n & x > p \end{cases} \tag{6.8}$$

The compression of control signals is convenient, if not required, for performances that

present some occasionally large gestures. It attenuates the relatively large gestures that could hide smaller ones. However, the technique described below is not uniformly applied to the entire signal due to the two different slopes used. A distortion, especially at the inflection point, can occur that slightly changes the behaviour of the signal, see Figure 6.5. The distortion may affect the perception of subtle changes in the gesture.



	Subject 6 Na2	
	Original	Warped
Mean	0.173708	0.247206
Variance	0.019357	0.024041
Median	0.142541	0.250572
IQR	0.180828	0.228256

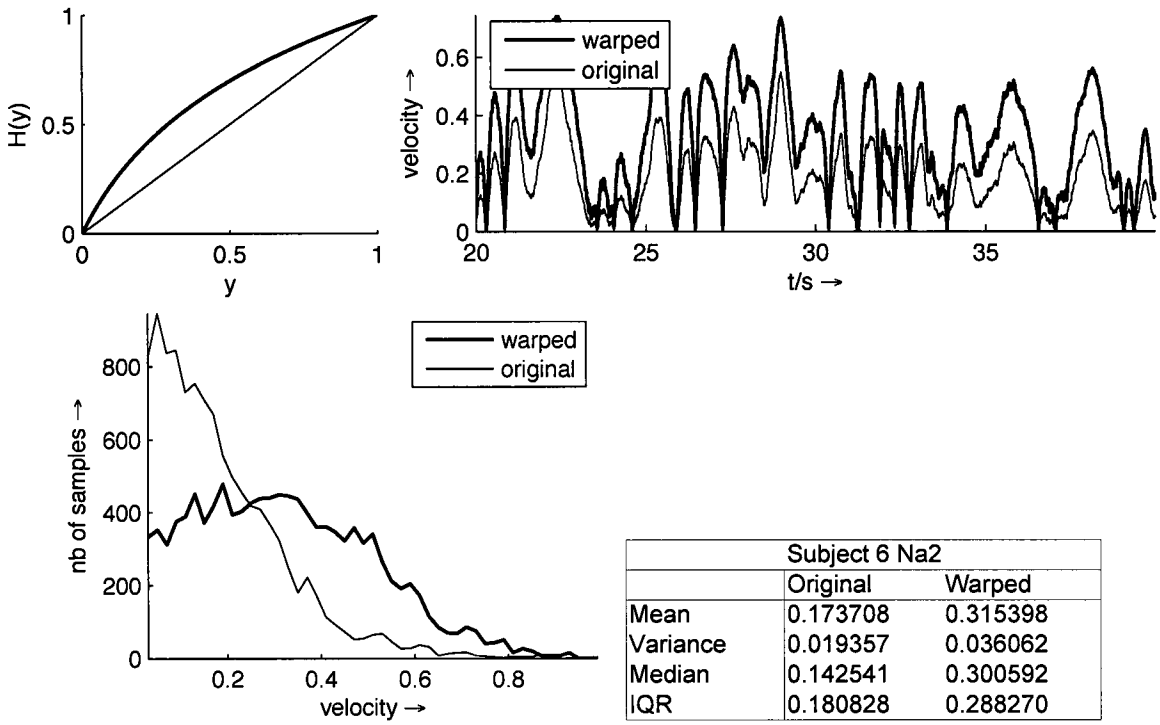
Fig. 6.5 Effects of a compression on the body curvature's absolute velocity (transfer function, warped signal, histogram).

Logarithmic warping

To avoid the effect of an inflection point, logarithmic scales are used to compress the high amplitude levels of the signals. The effect of the compression increases as the amplitude gets higher. It preserves the subtle changes of the signal since the slope of the transfer function is uniformly monotonic and growing. Using this transfer function one must find the portion of the logarithmic curve that fits the compression needed. To do so, the signals are scaled according to a factor α along the curve. As the parameter α increases, the

compression becomes stronger. To avoid the negative part of the logarithmic curve, the input signal $x[t]$ must be translated by an offset of 1. Looking at the histogram, the mean value is now nearer toward the center of the parameter's range.

$$x_{log}[t] = \frac{\log(\alpha x[t] + 1)}{\log(\alpha + 1)} \tag{6.9}$$



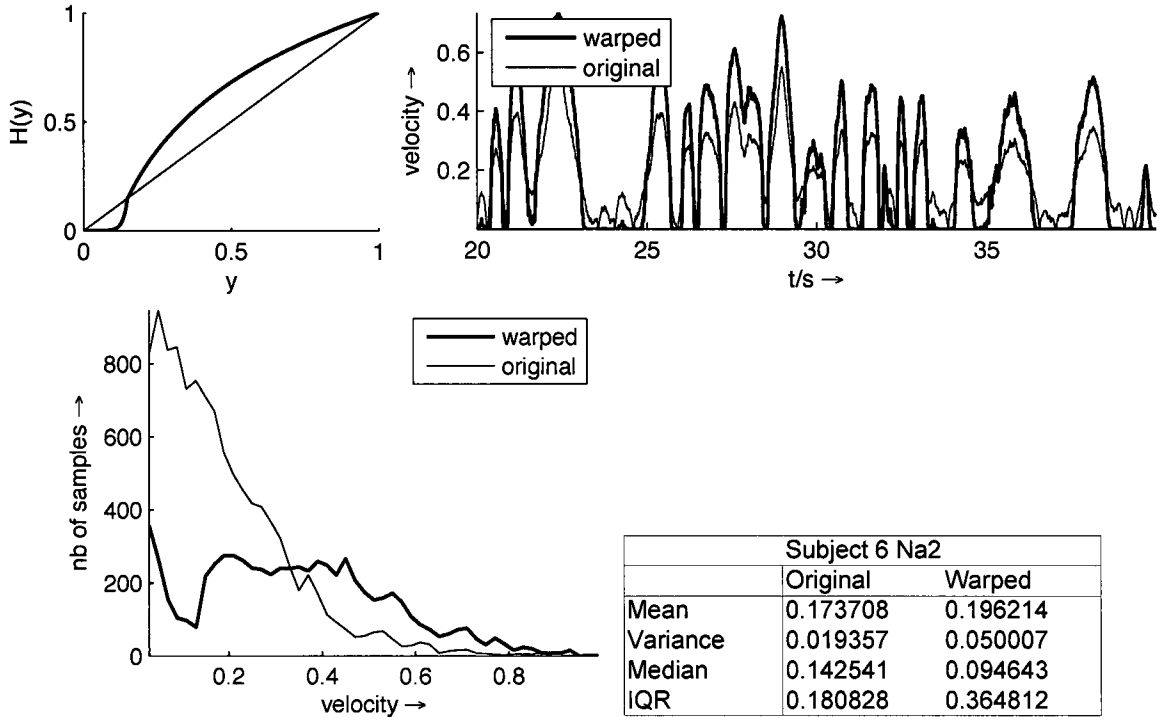
	Subject 6 Na2	
	Original	Warped
Mean	0.173708	0.315398
Variance	0.019357	0.036062
Median	0.142541	0.300592
IQR	0.180828	0.288270

Fig. 6.6 Effects of a logarithmic transfer function on the body curvature's absolute velocity (transfer function, warped signal, histogram).

The previous warping technique tends to amplify very low signal values and is desirable when significant information has been attenuated during the normalization process. However, there are situations (very low amplitude gesture or noise introduced in the data processing) where signals must be attenuated. For this reason, logarithmic warping needs to be extended using an alternate function specifically designed for this purpose, one that will compress the noise or the low amplitude signals. The exponential function $f(x) = e^x$, defines a curve that fits this criteria. Under a certain threshold $p \in [0, 1]$, the exponential curve is used while the logarithmic one is used for the other part as depicted in Figure 6.7,

6.8. Similar to the process applied to the logarithmic curve, a scaling factor β is applied to strengthen the effect.

$$x_{exp-log}[t] = \begin{cases} p\frac{exp(\beta x[t])}{exp(\beta p)} & x \leq p \\ p + (1-p)\frac{\log(\alpha x[t]+1)}{\log(\alpha(1-p)+1)} & x > p \end{cases} \tag{6.10}$$



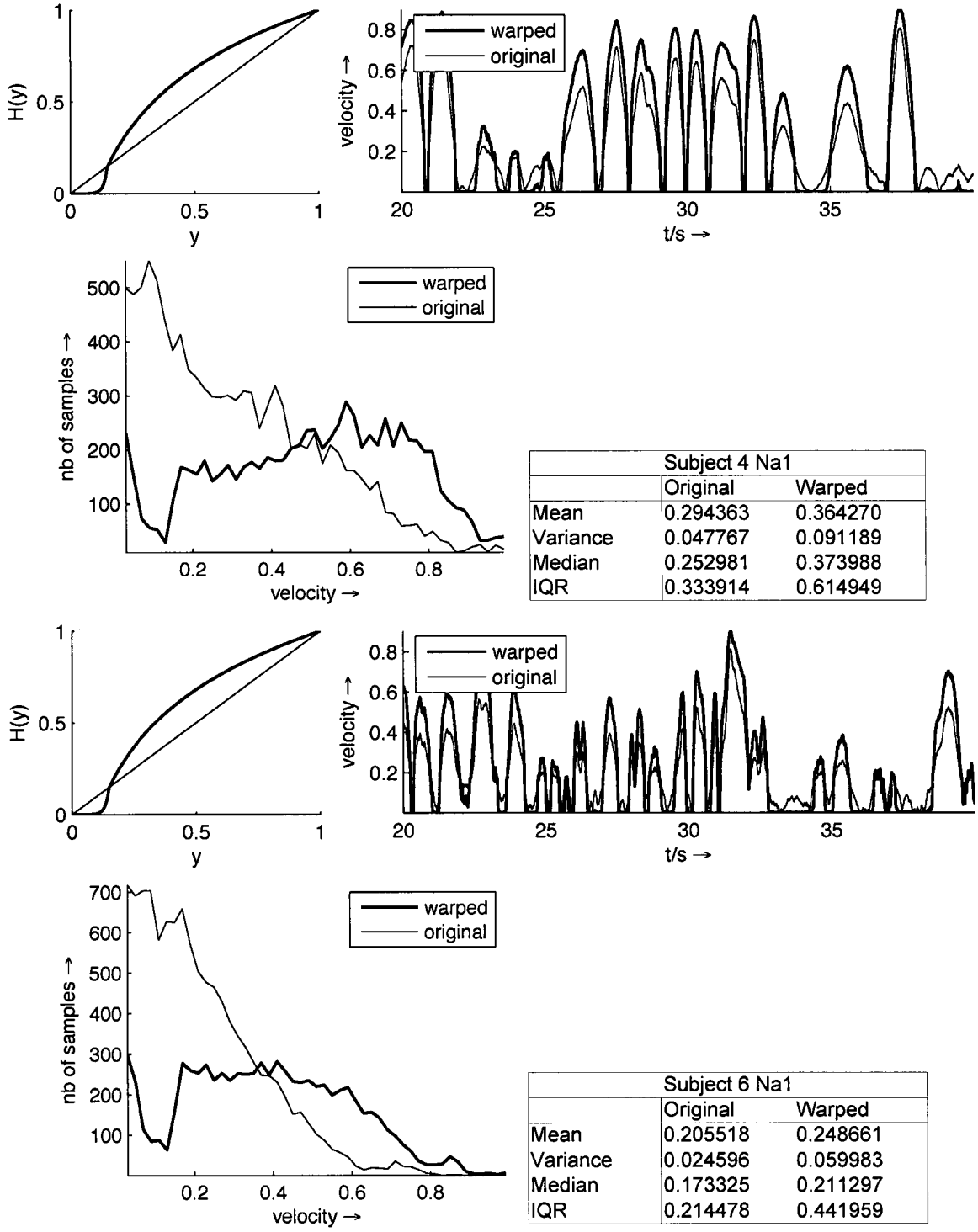
	Subject 6 Na2	
	Original	Warped
Mean	0.173708	0.196214
Variance	0.019357	0.050007
Median	0.142541	0.094643
IQR	0.180828	0.364812

Fig. 6.7 Effects of concatenated exponential and logarithmic transfer function on the body curvature's absolute velocity (transfer function, warped signal, histogram).

6.3.2 Application to alternate features

Sinusoidal warping

The warping techniques described so far share the same general impact on the signals as they reduce the low values. Alternate techniques can be introduced to achieve different purposes. For instance, sinusoidal-like transfer functions have interesting properties as they emphasize a signal's specific regions that can then be easily distinguished by the



Subject 4 Na1		
	Original	Warped
Mean	0.294363	0.364270
Variance	0.047767	0.091189
Median	0.252981	0.373988
IQR	0.333914	0.614949

Subject 6 Na1		
	Original	Warped
Mean	0.205518	0.248661
Variance	0.024596	0.059983
Median	0.173325	0.211297
IQR	0.214478	0.441959

Fig. 6.8 Effects of concatenated exponential and logarithmic transfer functions on the body curvature's absolute velocity of two levels of expressiveness: expressive (top) and standard (bottom).

listener. Using these types of functions, the signals will tend to converge more quickly to their maxima, minima, and other arbitrary defined positions. The main purpose of these functions is to enhance the difference between specific arbitrary positions. As an example, it appears to be particularly useful to distinguish the two extreme left and right positions of the weight transfer. The modification is induced to the input signal using the following formulae[3]:

$$x_{sin}[t] = \frac{1 + \sin(\pi(x[t] - 0.5))}{2}. \tag{6.11}$$

An extension of the simple sinusoidal transfer function is the concatenation of two or more sinusoidal curves as shown in Figure 6.9. It shows the effect on the signal of an additional inflection point set at 0.25. It introduces an additional level where the values will be attracted. Returning to body weight transfers, it is then possible to emphasize an additional position, namely the centre, in order to clearly identify this performer's posture.

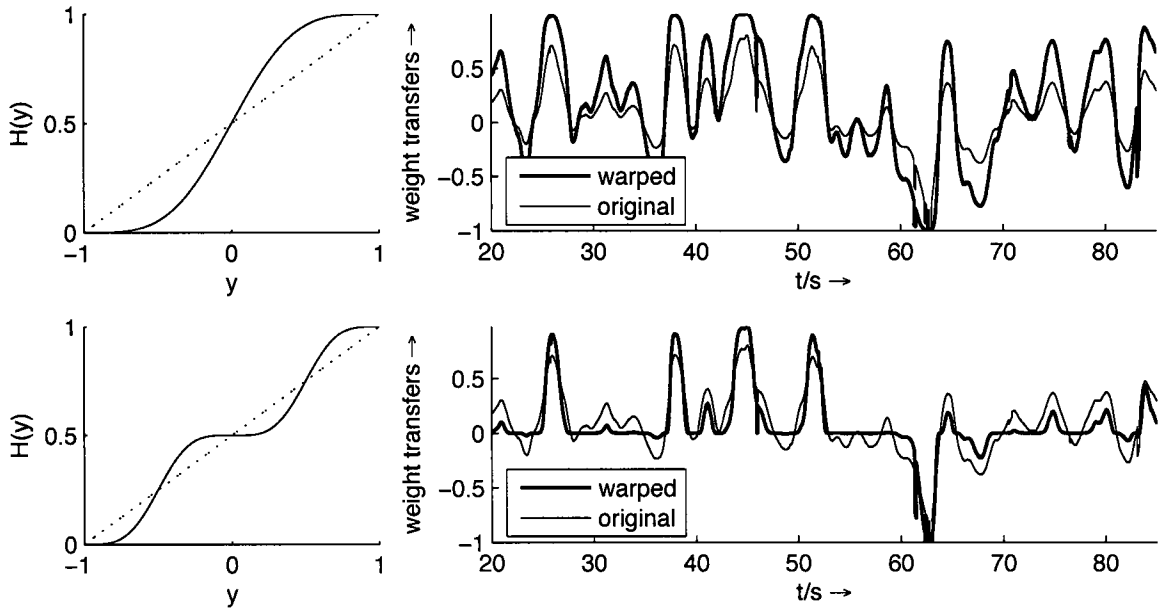


Fig. 6.9 Effects of two sinusoidal-like transfer functions on the left/right weight transfers.

[3]To reinforce the sinusoidal curve, sin(sin()) is used instead of sin().

Limitation of warping

Warping can be used to enhance characteristics already present in the control signals and make them more easily detected, especially, in a context where several streams of information are to be listened to simultaneously. However, these techniques should not be overused as it may create behaviour that did not initially exist. For this reason, a parameter that controls the amount of warping effect applied to the input signal is also provided (see Figure 6.10).

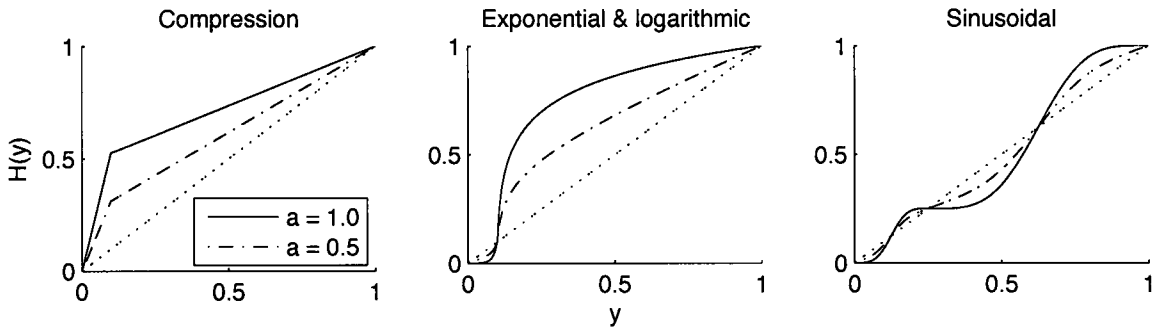


Fig. 6.10 Different transfer functions with scaling factors.

6.4 Scaling

As introduced in section 5.2, control signals may not be the range of sound synthesis control parameters. Control signals' maxima and minima may exceed numerical limits for which the sound synthesis produce efficient results. Otherwise, in the eventuality that the ranges are suitable for the parameters, they are also required to be compatible with the characteristic of the auditory system. Too narrow ranges could fail to illustrate a gesture meaningfully or worse, cause a given gesture feature to be audible. Scaling is simply processed by multiplying the normalized warped signal $x[t]$ with a scaling factor a and adding an offset b:

$$y[t] = a \cdot x[t] + b. \tag{6.12}$$

Conclusion

In this chapter, the discussion stressed the signal conditioning of the data to be sonified. Warping is useful to adapt several signals to ranges where they can be compared or to respectively emphasize or attenuate desirable or undesirable information. The techniques described in this chapter are tools offered to the user to assist a successful sonification of gesture in a fast and reliable way. They are essential to ensure that the sonification is designed in a short period of time, enhancing the experience of both user and participant in a context where a subject would be invited to experience sonification in realtime.

Chapter 7

Sonification system

This chapter presents the most important elements concerning the implementation and usage of the gesture sonification system developed during this thesis using the Max/MSP realtime signal processing environment. The sonification system does not require advanced knowledge in numerical signal processing, computer programming, or sonification, and can be easily calibrated to individual subjects. The programming structure benefits from Max/MSP environment supporting encapsulation, which is convenient for further extension of the system.

7.1 Sonification desktop

The sonification desktop is the main interface from which users design and manipulate gesture sonification (see Figure 7.1). Usual navigation controls (start, stop, timer) are provided to manipulate the data. A switch allows the system to be set in calibration or sonification mode (see section 7.4). The various data and sound processing techniques are regrouped into several sonification channels (see section 7.3). Several menus allow for the selection of data, processing algorithms, sound synthesis, and calibration preferences. Given selected data or sound processes, specific patches ("subinterface") can be open to modify parameters related to these processes (see Figure 7.2). Sonifications as well as the control signals that generated them can be saved as PCM audio files (wav format). Recorded sonification can be reloaded as well.

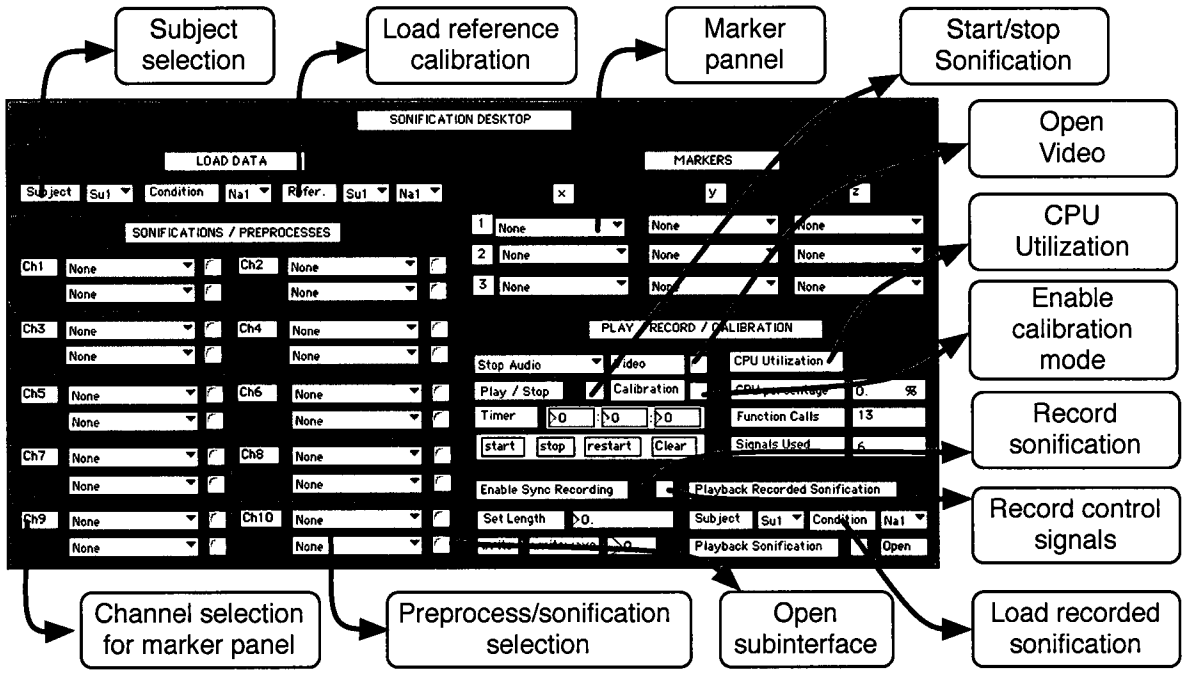


Fig. 7.1 The sonification desktop.

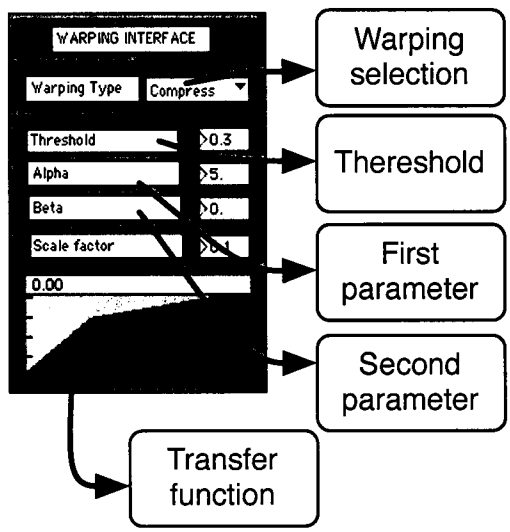


Fig. 7.2 Example of "subinterface" (warping process interface).

7.2 Data formating

The sonification system is designed to support realtime data processing. Max/MSP is an audio synthesis programming environment which puts priority on audio processing. To ensure synchronization between video and sound, the system processes both gesture features and sound synthesis at the audio sampling rate. Data are imported into Max/MSP as wav audio files that are generated from the Vicon's c3d ASCII file format using MATLAB. A 100 Hz wav file is generated for every marker position (x,y,z) and every features is computed by the plug-in-gait biomechanical model. The sfplay~ object loads the data and computes the up-sampling at MSP's audio sampling rate without any corruption of the data content.

The amount of computation is reduced using the poly~ object. Specifying "down 4" as an argument, the internal processes are performed at a sampling rate of 11025 Hz instead of 44100 Hz. Up-sampling and down-sampling the signals do not affect the data content. Furthermore, unused DSP chain subparts are muted using the poly~ object, which ceases the DSP calculations.

7.3 Sonification channel

The sonification system's architecture is inspired from audio mixing consoles. It is divided in several sonification channels regrouping different classes of processes:

- raw data selection and data loading,
- data processing and feature extraction,
- signal warping,
- sound synthesis.

Figure 7.3 presents a detailed diagram concerning the relationship between classes of processes within one channel. Channels allow for the design of specific sonifications given one or several input signals. This architecture is convenient for the design of multiple sonifications in parallel, one for each gesture to be considered. The output sonification of each channel is sent to the sonification mixer (see Figure 7.4) which gives users the possibility to balance different sonifications in order a global appreciation of the overall sonification or, as opposed to this, to stress specific gestures.

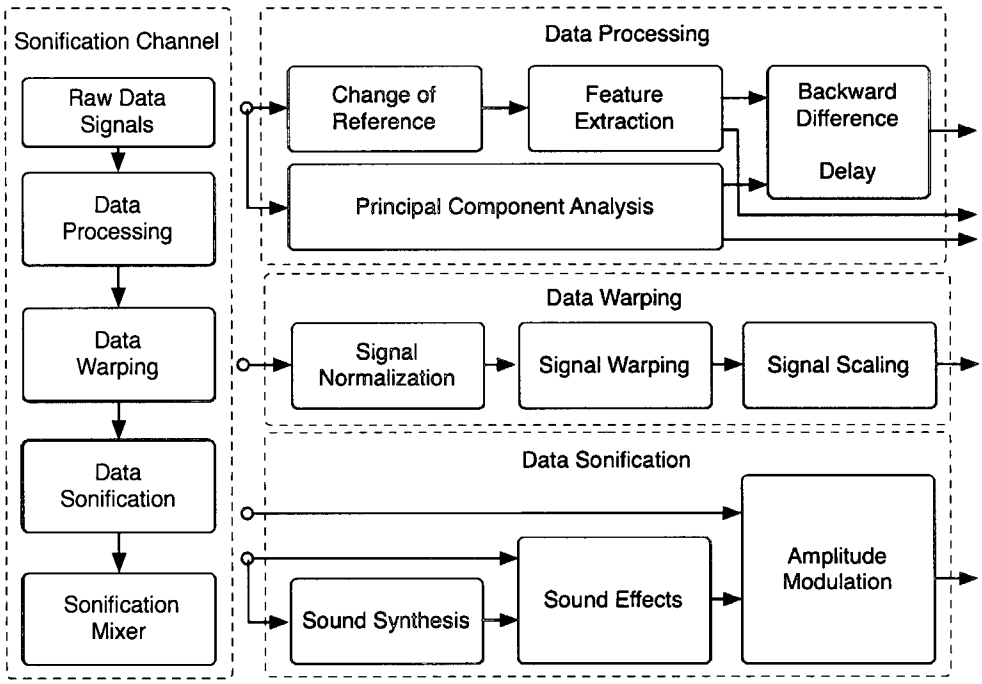


Fig. 7.3 Sonification channel.

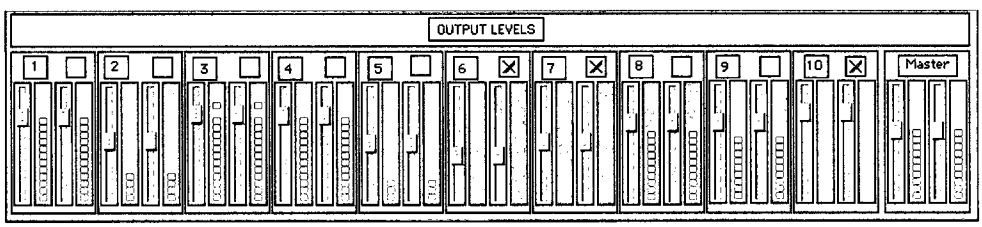


Fig. 7.4 Sonification Mixer.

Extensions can be easily incorporated to existing sonifications using additional channels. The modularity of the different algorithms allows for most of the gesture features to be mapped to any type of sound synthesis. Restrictions occur when gesture feature algorithms extract more control signals than the sound synthesis can manage. In these cases, only the most significant gesture features or sound synthesis parameters are linked together. The modular aspect of the system allows for a programmer to easily add his/her own data processes or sound synthesis algorithm without significant efforts in terms of integration into the system.

7.4 System calibration

Achieving a relevant sonification requires the system to be calibrated for a specific performer. The calibration may also depend on the comparative strategy to be adopted. Scaling and other warping techniques cannot be efficiently performed without normalization. This is only possible with knowledge concerning the interval in which each gesture feature is defined. This range is referred to as a "range of motion". The concept of "range of motion" is also reliable to a group of subjects in order to compare their relative positions or velocities.

7.4.1 Realtime calibration

To calibrate the system, one needs to specify the appropriate strategy (see section 6.2) and then process each of the motion capture data sets. In calibration mode, collections containing pre-normalized gesture features' maxima and minima are automatically updated for every motion capture processed. The calibration can be performed on data from one subject only, on different performances of the same subject, or on a group of subjects. These collections are saved for further usage. For several recorded motion capture sessions, the system's calibration can be performed rapidly by increasing the data playback rate. The extrema are used to normalize the data once the system is set back in sonification mode. This calibration procedure is convenient for realtime processing especially when a subject would be asked to perform a "range of motion" to calibrate the system in order to perform a live sonification.

7.4.2 Offline calibration

As the Max/MSP environment can hardly support a computational extensive process such as PCA, this part of the calibration must be performed offline in another computing environment. Successively to data formating from c3d file format to wav file format, PCA is performed on relevant subgroups of marker positions and features. The eigensystem are then loaded in the sonification system and applied as coefficients for linear combinations of signals as a specific subject is selected. The analysis is performed in MATLAB and the output is formated according to Max/MSP's "collection" object requirements. Once a specific body part is selected, users may listen to the chosen principal component in realtime.

7.5 Data processing and sonification design

7.5.1 Data processing

Data selection is an important component of the sonification system allowing the user to choose which raw data is to be driven into sonification channels. A reduced model of the most relevant body features as proposed in section 5.2 is chosen as a default selection of relevant gesture features to be sonified:

- center of mass,
- head position and orientation,
- upper trunk (shoulder, sternum) average position and orientation,
- lower trunk (hips) average position and orientation,
- overall body curvature,
- left / right knee angle,
- instrument circular movements.

For more specific uses, one can also select the original marker position and biomechanical features computed by the plug-in-gait biomechanical model. Additional gesture features can be defined using distance between markers or defining new angles. Once a sonification

is designed, the system allows for rapid switches from one subject to another (conserving the settings and parameters) for comparison perspectives.

In an explorative perspective, one can be interested in hearing to the principal components for a specific subgroup of markers. Here are listed the most relevant subgroup of marker positions and features on which the PCA is performed:

- head,
- upper body,
- lower body,
- Left / right legs.

The ordered eigenvectors that specify the coefficients for the linear combination are imported in the sonification system. A channel for which a specific PCA has been selected is fed with all the necessary signals to generate the principal components for a given subgroup of markers. The user can specify which principal component he/she is interested in.

7.5.2 Speed-up sonification

One application of sonification is to highlight gestural patterns occurring over long periods of time during a performance. These patterns may become more easily perceived as the multi-modal representation playback is sped-up. As an additional benefit, the required amount of time for listening and visualizing several motion capture sessions is thus shortened. One advantage of the sonification techniques described in this project is that increasing sonification playback rate does not affect its pitch. Even if increasing sonification playback rate shortens sound length and affect events succeeding rates, sounds are still recognizable as timbre and pitch remain the same. Users can change the playback rate using a scalar factor (1 implies normal playback):

- 2 plays sonification/video at twice the speed,
- 0.5 plays sonification at half the speed,
- 0 stops the sonification/video.

Conclusion

This chapter presented the most important elements of the sonification system in terms of its user interface, data processing, and data manipulation. Presenting briefly its programming structure, it discussed certain aspects of the system that facilitate data sonification and exploration. Among these aspects are the organization of the different processes into one unique structure: the sonification channel. Simple calibration process and capability to rapidly switch between different subjects are characteristics which make the system convenient to use for gesture sonification.

Chapter 8

Conclusion

8.1 Contribution

The main contribution of this work is the creation of a musicians' ancillary gesture sonification system. The system can be used in two different ways:

- to process and sonify data directly acquired from motion capture systems in realtime using the same computer environment,
- to sonify principal components summarizing the movement of specific body parts from an exploratory perspective.

The system does not require the user to have specific knowledge in programming nor signal processing. It is mainly designed for researchers in physiology-related fields, interested in rapid exploration of motion capture data sets.

To facilitate the achievement of initial results, the reduced model is proposed as a preliminary data subset. For users investigating specific details, it is also possible to access the entire data set. Through PCA, it is possible to describe movements of specific body parts using selected features and determine which of these features are best suited to the reduced model. Applying PCA to both data sets and comparing results for the original extended one to the reduced model, it is clear that the main information is not affected by the proposed data reduction.

The user control parameters are simple and intuitive, which makes the sonification design straightforward. Normalization is automatically performed under specified inter-gestures or inter-subjects strategies. Signal warping techniques using transfer functions

adapt the control signals to control parameter requirements, which is necessary when comparing different performers' sonifications, especially when one or several subjects perform more explicit gestures than others. Non-linear curves are useful as one can map the different control signals applying different scales. Signal warping can be used to strengthen signals' behavior and make evident details that were previously non-obvious.

The system can be easily extended by users who have basic knowledge in programming and signal processing, as alternate gesture feature extraction algorithms or sound synthesis techniques can be easily added to the system.

8.2 Limitations and future work

The sonification strategy defined in this project may be improved in order to increase user's ability to perceive gestures. As it is not natural to intuitively associate non-physically generated sounds to gestures, the relation between a gesture and its sonification must be obvious and unambiguous. For example, sounds for back and forth performer's movements are mostly the same, although subtle changes in gesture's direction are still perceptible (ascending or descending pitch, left-to-right in opposition to right-to-left panning). These subtle changes could be exaggerated mapping the sign of the derivative to parameters that strongly modify sound's timbre. Emphasizing them, users' ability to perceive movement through sounds would be increased.

The investigation using PCA demonstrates that combining the correlated information leads to signals that adequately summarize groups of positions/features. This investigation can be considered a first step in the achievement of a system that would objectively determine the most important information describing performers' body gestures.

Using PCA, it is not possible to automatically identify the exact gestural feature related to a specific principal component. Furthermore, it demonstrates satisfactory results only when performers execute very explicit gestures (large spatial distance). Otherwise, the same analysis applied to subjects performing less explicit gestures provides results that are not clear enough to arrive at reasonable conclusions.

There are several alternatives in functional data analysis that could be investigated in order to achieve more accurate comparison between signals. Instead of correlation, absolute sample-per-sample distance between smoothed signals provides more robust results concerning subjects who perform less explicit gestures. As an alternate process, differential

data analysis takes into consideration signals' variation rate over time, an important aspect that is difficult to obtain with PCA.

The first attempts to design sonifications which conserve relative differences within a group of subjects have been only partially successful. One issue concerns the fact that subjects performing subtle and less explicit gestures produce sonifications that are less obvious to discriminate as their signals are considerably attenuated once normalization according to subjects performing wide and explicit gestures is performed. This is one reason why warping techniques were introduced in chapter 6. As knowledge concerning an objective relation between sound and visual perception of gestures in sonification is almost inexistent, further statistical analysis must be performed to strengthen knowledge about perception of ancillary gestures such as in [50], especially in terms of velocity and path distances in order to formulate efficient warping curves for sonification.

Finally, sonification's efficiency to represent gestures should be investigated in psychological studies. Several aspects are potentially relevant concerning :

- subjects' ability to associate videos with respective sonifications,
- gesture-to-sound mapping efficiency,
- characteristic types that, once sonified, best describe their related gestures,
- discrimination capability of a subject regarding slight variations of control parameters (PCA, exact gesture feature).

In this sense, the sonification system proposed here is ready to generate efficient sonification for the possible studies described above. Depending on the investigation, new motion capture sessions may also be required. As musicians tend to perform several gestures simultaneously, it is difficult, in currently available sessions, to find excerpts which contain one isolated gesture for every type of gesture considered. Musician should be asked to separately perform different gestures under several velocity conditions so that study involving sonified gestures discrimination can be easily prepared.

References

[1] V. Verfaille, O. Quek, and M. M. Wanderley, "Sonification of musicians ancillary gestures," in Proc. Int. Conf. on Auditory Display, (London, UK), pp. 194–7, 2006.

[2] A. Polli, "Atmospherics/weather works: a spatialized meteorological data sonification project," Leonardo, vol. 38, no. 1, pp. 31–36, 2005.

[3] G. Kramer, "An introduction to auditory display," in Auditory display: sonification, audification and auditory interfaces (A. Wesley, ed.), vol. 18, (Reading, MA, USA), pp. 1–77, 1994.

[4] S. Barrass and G. Kramer, "Using sonification," Multimedia Systems, vol. 7, pp. 23–31, June 1999.

[5] S. D. S. Pereverzev, A. Loshak and J. Davis, "Quantum oscillations between two weakly coupled reservoirs of superfluid He-3," Nature, pp. 449–51, July 1997.

[6] B. C. J. Moore, An introduction to the psychology of hearing. San Diego, CA, USA: Academic Press, 1997.

[7] S. M. Williams, "Perceptual principles in sound grouping," Auditory Display: Sonification, Audification, and Auditory Interfaces, pp. 746–748, 1994.

[8] A. S. Bregman, Auditory scene analysis: the perceptual organisation of sound. Cambridge, MA, USA: MIT Press, 1990.

[9] J. Vroomen and B. Gelder, "Sound enhances visual perception cross-modal effects of auditory organization on vision," Journal of Experimental Psychology: Human Perception and Performance, vol. 26, pp. 1583–90, Octobre 2000.

[10] H. McGurk and T. McDonald, "Hearing lips and seeing voices," Nature, vol. 264, no. 5, pp. 746–748, 1976.

[11] C. Scaletti and A. Craig, "Using sound to extract meaning from complex data," in Extracting meaning from complex data: processing, display, interaction (S. of Photo-Optical Instr. Eng., ed.), (Bellingham, WA, USA), pp. 147–53, 1990.

[12] A. O. Effenberg, "Using sonification to enhance perception and reproduction accuracy of human movement patterns," in Proc. Int. Workshop on Interactive Sonification, (Bielefeld, Germany), pp. 1–5, 2004.

[13] T. Hermann and H. Ritter, "Sound and meaning in auditory data display," in Proc. IEEE, vol. 92, pp. 730–41, 2004.

[14] A. Hunt and T. Hermann, "The importance of interaction in sonification," in Proc. Int. Conf. on Auditory Display, (Sydney, Australia), 2004. http://www.icad.org/websiteV2.0/Conferences/ICAD2004/papers/hunt_hermann.pdf.

[15] T. Hermann and A. Hunt, "An introduction to interactive sonification," in Multimedia, IEEE, vol. 12, pp. 20–24, 2005.

[16] S. Pauletto and A. Hunt, "Interactive sonification in two domains: helicopter flight analysis and physiotherapy movement analysis," in Proc. Int. Workshop on Interactive Sonification, (Bielefeld, Germany), January 2004. http://www.icad.org/websiteV2.0/Conferences/ICAD2004/papers/pauletto_hunt.pdf.

[17] A. Hunt and T. Hermann, "The importance of interaction in sonification," in Proc. Int. Workshop on Interactive Sonification, (Bielefeld, Germany), 2004.

[18] S. Saue, "A model for interaction in exploratory sonification displays," Proc. Int. Conf. Auditory Display, 2000. http://www.icad.org/websiteV2.0/Conferences/ICAD2000/ICAD2000.html.

[19] M. M. Wanderley, "Non-obvious performer gestures in instrumental music," Gesture-Based Communication in Human-Computer Interaction, pp. 37–48, 1999.

[20] M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch, "The musical significance of clarinetists' ancillary gestures: an exploration of the field," Journal of New Music Research, vol. 34, no. 1, pp. 97–113, 2005.

[21] M. M. Wanderley, "Quantitative analysis of non-obvious performer gestures," Gesture and Sign Language in Human-Computer Interaction, pp. 241–253, 2002.

[22] M. Wanderley and P. Depalle, "Gestural control of sound synthesis," Proc. of the IEEE, vol. 92, no. 4, pp. 632–644, 2004.

[23] A. Hunt, M. Wanderley, and R. Kirk, "Towards a model for instrumental mapping in expert musical interaction," in Proc. Int. Computer Music Conf., (Göteborg, Sweden), pp. 209–12, 2000.

[24] D. Arfib, J. M. Couturier, L. Kessous, and V. Verfaille, "Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces," Organised Sound, vol. 7, no. 2, pp. 127–44, 2002.

[25] F. Bevilacqua, L. Naugle, and I. Valverde, "Virtual dance and music environment using motion capture," in Proc. of the IEEE - Multimedia Technology and Applications Conference, (Irvine, CA, USA), 2001.

[26] F. Bevilacqua, J. Ridenour, and D. Cuccia, "3D motion capture data: motion analysis and mapping to music," in Proc. Workshop/Symposium on Sensing and Input for Media-centric Systems, (Santa-Barbara, CA, USA), 2002.

[27] S. Pauletto and A. Hunt, "A toolkit for interactive sonification," in Proc. Int. Conf. on Auditory Display, 2004. http://www.icad.org/websiteV2.0/Conferences/ICAD2004/papers/pauletto_hunt.pdf.

[28] W. Fitch and G. Kramer, "Sonifying their body electric: superiority of an auditory over a visual display in a complex multivariate system," in Auditory display: sonification, audification and auditory interfaces, vol. 18, (Reading, MA, USA), pp. 307–27, Addison Wesley, 1994.

[29] A. Kapur, G. Tzanetakis, N. Virji-Babul, G. Wang, and P. Cook, "A framework for sonification of Vicon motion capture data," in Proc. Int. Conf. on Digital Audio Effects, (Madrid, Spain), pp. 47–52, 2005.

[30] A. O. Effenberg, "Movement sonification: Effects on perception and action," IEEE Multimedia, vol. 12(2), pp. 53–9, April 2005.

[31] W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception," Ecological Psychology, vol. 5, no. 1, pp. 1–29, 1993.

[32] W. Gaver, "How in the world do we hear? exploration in ecological acoustics," Ecological Psychology, vol. 5, no. 4, pp. 285–313, 1993.

[33] M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch, "Expressive movements of clarinetists: quantification and musical consideration," Tech. Rep., MT2004-IDIM01, IDMIL, McGill, 2004.

[34] NDI, Optotrak, 2007. http://www.ndigital.com/certus.php.

[35] C. Scaletti, "Sound synthesis algorithms for auditory data representations," in Auditory display: sonification, audification and auditory interfaces (A. Wesley, ed.), vol. 18, (Reading, MA, USA), pp. 223–52, 1994.

[36] J. C. Risset, "Pitch control and pitch paradoxes demonstrated with computer-synthesized sounds," The Journal of the Acoustical Society of America, vol. 46, no. A, p. 88, 1969.

[37] R. Shepard, "Circularity in judgments of relative pitch," Journal of the Acoustical Society of America, vol. 36, no. 12, pp. 2346–53, 1964.

[38] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," Computer Music Journal, vol. 1, no. 2, pp. 46–54, 1977.

[39] Vicon Motion Systems, 2007. `http://www.vicon.com`.

[40] C. Cadoz and M. M. Wanderley, "Gesture–music," tech. rep., Ircam, 2000.

[41] J. Ramsay and B. Silverman, Functional Data Analysis. New York, NY, USA: 2nd ed. Springer, 2005.

[42] A. Daffertshofer, C. J. Lamoth, O. G. Meijer, and P. J. Beek, "PCA in studying coordination and variability: a tutorial," Clinical Biomechanics, vol. 19, pp. 415–428, May 2004.

[43] C. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[44] R. Balasubramaniam and M. Turvey, "Coordination modes in the multisegmental dynamics of hula hooping," Biological Cybernetics, vol. 90, no. 3, pp. 176–190, 2004.

[45] A. A. Post, A. Daffertshofer, and P. J. Beek, "Principal components in three-ball cascade juggling," Biological Cybernetics, vol. 82, pp. 143–152, February 2004.

[46] V. Verfaille and D. Arfib, "ADAFx: Adaptive digital audio effects," in Proc. of the Workshop on Digital Audio Effects, Limerick, pp. 10–4, 2001.

[47] V. Verfaille, M. M. Wanderley, and P. Depalle, "Mapping strategies for gestural control of adaptive digital audio effects," Journal of New Music Research, vol. 35, pp. 71–93, March 2006.

[48] A. Hunt, M. M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," in Proc. Int. Conf. on New Interfaces for Musical Expression, (Dublin, Ireland), 2002. `http://www.nime.org/2002/proceedings/paper/hunt.pdf`.

[49] V. Verfaille, Effets audionumériques adaptatifs: théorie, mise en oeuvre et usage en création musicale numérique. PhD thesis, Université Aix-Marseille II, 2003.

[50] B. W. Vines, Seeing Music: Integrating Vision and Hearing in the Perception of Musical Performances. PhD thesis, McGill University, 2005.