========= Introduction ==========

We present the most important elements concerning the implementation and usage of a gesture sonification system developed using both MATLAB and the Max/MSP realtime signal processing environment. Based on a prior attempt using the optotrak motion capture system, a new attempt extended the previous tool to perform sonification on a gesture data set acquired using a more recent Vicon motion capture system. The different between both data sets dimensionality was a main challenge: the data set describing musician gestures grew from 8 3-dimensional position sensor providing nevertheless significant information on one side, to more than 600 hundreds signals once extended by a biomechanic model. Methods investigated to reduce this amount of data will be discussed below.

========== System overview ==========

To produce a significant signification using the system developed during this project requires a two-step procedure that involves preprocessing of the gesture data set by a set of routine written in MATLAB to obtain meaningfull signals that can be easily loaded in the realtime audio processing engine.

Max/MSP is an audio synthesis graphical programming environment that is designed to process audio data in realtime. It puts priority on audio processing, above data and event processing. To ensure synchronization between video and sound, the system processes both gesture features and sound synthesis during audio processing callbacks in MSP. 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. Data up-sampling at MSP's higher sampling rate (11025 Hz) is performed at runtime while reading the data.

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.

The system 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. 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 a global appreciation of the overall sonification or, as opposed to this, to stress specific gestures.

======== Data reduction ==========

Due to the large amount of data and human auditory system's limited ability to process multichannel information, strategies have to be developed in order to extract relevant information from the data set, which information would then be used to drive signification. Principal component analysis were selected as a dimension reduction tool: the input signals are combined in order to maximize variance among the data set the gesture numerical representation is fully preserved throughout the transformation.

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. The idea is to combine information that demonstrates high covariance within the data set in a two-step algorithm that includes the eigenvalue decomposition process and the linear coimbination reconstruction process.

When applied to a data set made of 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. Unfortunately, centering the full data set does not improve our insight of the data set informational content. Doing so made the leg moving instead of the torso as the performers were asked to stay their feet fixed to the ground. The subsequent principal components (that represent almost 15 % of the markers' movements) necessarily relates to gesture features even if their respective eigenvalues are low.

In order to emphasize this information, PCA is performed on local regions of the body such as arms, legs, or torso. In addition, data set were augmented using derived gesture feature such as articulation angles to reinforce information related to specific articulations. 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. 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. Several signals can be discarded as they represent redundant information that does not convey any additional significance on their own.

======== Mapping of control signals to sound parameters ========

The frontier between gesture sonification and music-oriented computer generated sound, controlled by gesture, is definitely narrow. 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. In other words, complex mapping, mapping cannot be learned instantaneously [23], is not appropriate for sonification. Simple mapping shorten the user learning period to assimilate the actual sound properties that convey information but also the user interface of the signification system itself. It is important to ante that in the context of signification, the informational content of signals must not be modified by the processing techniques. On the other hand, applied correctly, processing can help to enhance characteristic that could have remain hidden otherwise.

As suggested in [1], gesture velocity, or more exactly the gesture feature derivative, is 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.

One-to-one and one-to-many mapping strategies are convenient in implementing sonification systems and solve several problems related to control signals and control parameters management. 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 more conveniently to the actual gesture (i.e. weight transfer).

It appears that one-to-many mapping strategy from one gesture feature signal to multiple sound features enhance a given gesture, 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 estimated by evaluating accurately low oscillation beating effect and the panning angle.

---------- Normalization ----------

Normalization were used for three fundamental reasons. 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. It is important to note that signals maxima and minima must be know to achieve good normalization. There is, at least, two different normalization strategy applied in the system.

Inter-gesture normalization

Several gesture feature extraction algorithms will produce several different ranges of in- formation. 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 nor- malized 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 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).

---------- Signal warping ----------

Once properly normalized according to one of the strategy above, control signals are scaled and linked to sound synthesis parameters. Scaling is mandatory as most of the sound synthesis techniques require the control parameter to respect a certain range. 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 but 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 nor- malization strategy,
- 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 $y[t] = H(x[t])$.

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 value2 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.