Quantitative Evaluation of Percussive Gestures by Ranking Trainees versus Teacher

Lei Chen, Sylvie Gibet, Pierre-François Marteau, Fabrice Marandola, Marcelo M Wanderley

Quantitative Evaluation of Percussive Gestures by Ranking Trainees versus Teacher

Lei Chen
IRISA
Campus of Tohannic
University Bretagne Sud
56000 Vannes, France
lei.chen@univ-ubs.fr

Sylvie Gibet
IRISA
Campus of Tohannic
University Bretagne Sud
56000 Vannes, France
sylvie.gibet@irisa.fr

Pierre-François Marteau
IRISA
Campus of Tohannic
University Bretagne Sud
56000 Vannes, France
marteau@irisa.fr

Fabrice Marandola
Sorbonne-Universités/MNHN
CIRMMT, McGill University
Montréal, Canada
fabrice.marandola@mcgill.ca

Marcelo M. Wanderley
IDMIL
CIRMMT, McGill University
Montréal, Canada
marcelo.wanderley@mcgill.ca

ABSTRACT

In this paper we characterize timpani gestures by temporal kinematic features, containing most information responsible for the sound-producing actions. In order to evaluate the feature sets, a classification approach is conducted under three main attack categories ($legato$, $accent$ and $vertical\ accent$) and sub-categories (dynamics, striking position). Two studies are carried out: intra-subject and inter-subjects classification. Results are presented in terms of a quantitative ranking of students, using professional gestures as training set, and their gestures as test set.

Author Keywords

percussive gestures, classification, expressive variations, evaluation.

ACM Classification Keywords

H.5.5. Information Interfaces and Presentation (e.g., HCI): Sound and Music Computing: Signal analysis, synthesis, and processing

INTRODUCTION

While performing a music instrument, musicians establish a rich, well-structured, expressive interaction with the instrument. In order to master such gestural interaction and to control the fine-tuning of the sound-producing gestures, many training years are necessary. In such a learning process, a strong coupling between sound effects and control gestures is established. Instrumental gestures are progressively refined so that the produced sounds satisfy the desired goals. This process is usually guided by a professional teacher who gives the key elements to improve the regularity and the quality of the performance. In most of instrumental playing, the study of pedagogy around musical training has led to a gestural categorization that follows the gesture-sound relationships.

In previous work a wide range of musical gestures has been studied through the analysis of the gestural signals that are responsible for the sound production, but not so much concern percussive gestures [10], [1], [11]. Yet these gestures are interesting because they are short striking gestures which can be largely varied along the musical scores, depending on the musical intention. For percussive gestures, different playing modes are usually defined according to the following axes/dimensions: (i) Attack; (ii) Dynamics; (iii) Striking position. Attack refers to the initial stage of sound envelope, which is directly related to sound quality, especially in percussion. In music, expressive terms such as legato, tenuto, staccato, etc. correspond to different types of attacks and indicate various ways of shaping a sound. Dynamics (sometimes called intensity) corresponds to the loudness of the produced sound. The striking position indicates the location where the stroke is played. It is interesting to note that these dimensions may characterize gesture or sound descriptors, or both of them, which is due to the strong gesture-sound coupling that drives percussive gestures. We will also consider this categorization along the above axes and study the effects of distinct expressive variations on the performance.

In this paper we focus on timpani gestures. These gestures are characterized by large kinematic and expressive variations, with all the upper-body articulations of the percussionist strongly involved during gesture executions. Furthermore, the whole kinematics of the timpani gesture embeds not only the sound-producing gesture, but also the preparing and retracting gestures that are supposedly used to ease the performance while reducing the produced effort.

MOCO'16, July 05–07, 2016, Thessaloniki, GA, Greece
 ISBN 978-1-4503-4307-7/16/07...$15.00
DOI: **http://dx.doi.org/10.1145/2948910.2948934**

Prior knowledge in musicology and percussion teaching has led to the identification of different sets of gestural attributes in articulation and phrasing for musical performance on timpani. Cook [6] proposes the following attributes: (i) manipulating the height of the stroke, which naturally results in grip alterations in order to change the stroke height while keeping dynamic balance; (ii) manipulating the grip, which involves varying finger contact and grip pressure on the stick and results in changes of the stiffness of the system made by the combination of the mallet and the hand of the player; (iii) altering the place where the strokes hit the surface, which involves modifying the vibrations of the playing area; (iv) manipulating the lift of the stroke by changing the arm, wrist and finger movements (configuration and biomechanical state), thus altering the stroke's velocity and time of contact of the head of the mallet with the head of the timpani. F. Marandola proposes to consider: (i) the tightness of the grip depending on the context. This tension has a significant impact on the *fullness* of the sound, which can be related to the richness of the overall timbre. Hence not enough grip results in less fundamental, and too much grip results in stiffer attack, and greater risk of sound distortion, as well as a reduced range of expressivity; (ii) the depth of the stroke, which is directly linked to the contact time of the stick with the timpani?s skin

Inspired by these works, we propose to describe percussive gestures by kinematic temporal trajectories that best characterize these gestural attributes from motion captured data. Figure 1 shows a phase-diagram of the velocity versus position of the right hand mallet 3D-trajectory when a performer plays repeated $legato$, $accent$ and $vertical\ accent$ attack beats. We can observe the different patterns achieved when executing different types of repeated attacks.

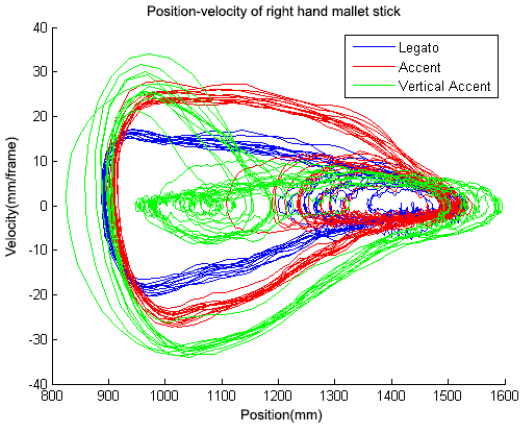

Figure 1. Velocity versus position for 3 variations of attack: $Legato$**,** $Accent$**,** $Vertical\ Accent$ **(data from FM).**

The first purpose of our study is to find out a minimum subset of features that effectively describe timpani gestures, both spatially and temporally. However, to capture a complete human motion, many body segment joints need to be considered, and to describe the angular values or 3D positions of the joints, many attributes are needed accordingly. It is therefore necessary to reduce the dataset by finding spatial as well as temporal reduced representations, while preserving the expressive qualities of the movements responsible for the variations along the percussive performance. The second purpose of our study is to evaluate different performers on the basis of their kinematic similarities with a professional musician. In particular different students' performance are compared to their professor's ones, resulting in a ranking between them according to their capability to reproduce the teachers gestures. A study on percussive gestures is conducted and validated through a classification approach under three main attack categories ($legato$, $accent$ and $vertical\ accent$) and sub-categories (dynamics, striking position), the first one for intra-subject, the second one for inter-subjects classification.

RELATED WORK

Previous research has focused on the analysis of gestures for particular instrumental gestures. Most of the work is directed toward the design of new musical and interactive interfaces. Especially regarding percussion-related systems, an important research direction is the development of devices to track performer gestures for controlling sound synthesis processes [12]. Despite the availability of various devices, the most accurate system for tracking percussive gestures remains camera tracking systems [14]. These systems offer an effective method for capturing, analyzing and virtually reconstructing the whole body of a performer. Hence, using motion captured (MoCap) data, it becomes possible to identify the gesture profiles and the gesture characteristics that are responsible for the sound production.

One of the issues addressed by the analysis of gestures is to better apprehend the underlying gestural processes involved in the control of sound, and the way it is related to motion data recorded during real performances. A traditional approach consists in identifying from MoCap data the set of features that best characterize the movement. A state of the art of computable descriptors of human motion is described in [13], presenting both low-level descriptors that compute quantities directly from raw motion data, and higher level descriptors that qualify the meaning, style or expressiveness of $3D$ motion capture data.

Dahl [11] thoroughly studied the striking of drum performance, and tried to establish a correspondence between movement parameters and sound production. The preparatory movement, the rebound and the timing of a stroke were respectively analyzed to study their control over the sound properties. In [8], percussive movement and timing strategies used by professional percussionists were observed. Despite large differences between the preparatory movement of different subjects, results showed a larger preparation height for the accented strokes. Further work [9] on striking velocity and timing were conducted with several percussion players performing accented strokes at different dynamic levels, tempi and on different striking surfaces attached to a force plate. A consistent individual movement pattern was maintained for the different players under different playing conditions. Another result showed that an increasing height and striking velocity led to an increasing dynamic level, and to a lengthening of the starting interval with the accented strokes.

Timpani gestures have been studied for the purpose of modeling and animating a physical virtual character capable of interacting with a physical sound synthesis system [1], [2]. The timpani gesture sequences were cut into elementary beat units, and edited so that a new score could be used to animate the virtual character. Two evaluations were conducted. A qualitative evaluation of synthesized timpani performances, following instructed exercises [5]. For the quantitative evaluation of the synthesis system, a classification approach was developed to recognize the different attacks under different expressive variations [3]. As an intuitive hypothesis states that percussionists more specifically control the motion of mallets over time, mallet extremity trajectories were used to conduct the analysis, and more particularly discrete local extrema extracted from position trajectories during beat-to-beat phases [4].

A more recent study was conducted [7] on Taiko, a Japanese form of ensemble drumming. Various categories of expressive Taiko performances were captured, recording both drumstrokes (position of the wrist) and sounds. Using machine learning methods, the authors have classified key aspects of Taiko technique, and showed that gesture and sound classification share similar results.

In this paper, we propose to refine our previous work on timpani gestures, and to compare various features of skeletal motion to analyze timpani performances, without any a priori on the kinematic features. Hence we evaluate through a classification approach different sets of joints, of kinematic measures (3D position, velocity, acceleration, etc.), and also how the starting and ending of the beat influences the performance. Furthermore, we propose a methodology to automatically evaluate the different students, using professional gestures as training set, and students gestures as test sets, with a ranking as output.

TIMPANI GESTURES DATABASES AND MOTION REPRESENTATION

Within the three main classes of timpani gestures categorization, i.e. – Attack, Dynamics, Striking position –, we consider the following expressive variations.

- Attack: three playing modes in attack are considered in our study: $legato(l)$, $accent(a)$ and $vertical\ accent(v)$. In music performance and notation, legato indicates that musical notes are played smoothly and connected. An accent adds emphasis to a particular note, requiring that it should be played louder than unaccented notes at the same dynamic level. A vertical accent is considered as accented tenuto, which means to hold the note along its full length and play it slightly louder. In terms of gestures, a performer plays legato accent and vertical accent with different positions, velocities and accelerations of the mallet extremity.

- Dynamics: dynamic indications in music are graduated between p or $piano$, meaning soft, and f or $forte$, meaning loud. More subtle degrees of loudness or softness are indicated by: mp, standing for $mezzo\text{-}piano$, meaning moderately soft ; mf, standing for $mezzo\text{-}forte$, meaning moderately loud. The classes of dynamic levels covered in our study are p, mf, and f.

- Striking positions: the basic playing area on a timpani head is at least 4 to 5 cm from the edge of the bowl directly in front of the player, which is nearly $1/3$ to the rim. In our study, we request the participants to hit the timpani at the center location (c), at $1/3$ to the rim ($1/3$), or at rim (r) of the timpani.

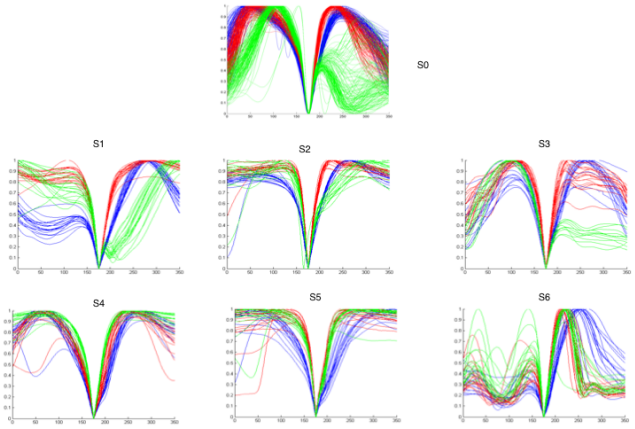

Figure 2. z-trajectories along time for the subjects S_0 (FM) to S_6.

Timpani Data Bases

In this work two databases of captured motion (MoCap) are considered. The first one, called DB10, was the recording in 2010 of a professional timpani player, Prof. F. Marandola at Schulich School of Music (IDMIL lab, McGill University). This subject, named S_0, will be considered as the referent subject in the rest of the paper. A high definition camera-based tracking system (Vicon 460) was used with an acquisition rate of 250 Hz. The Vicon Upper-body markers setup augmented with stick markers was chosen for capturing the movements of the timpani performers. Three classes of gestures were considered: Attack, Dynamics and Striking position. Under attack, there are 3 playing modes: $legato(l)$, $accent(a)$, $vertical\ accent(v)$; under dynamics, there are 3 intensity variations: p, mf, f; while under striking positions, we have 3 variations: $center(c)$, $1/3$, $rim(r)$. We captured 10 sequences for each class and variations, each sequence containing 10 strokes.

The second motion data base (DB13) was recorded in December 2013 with a Qualisys Oqus MoCap system (250 Hz). Six participants, named S_1 to S_6, were involved in this MoCap session. All of them are students from Schulich School of Music, McGill University. Four are percussion students in Prof. Marandola's studio; 2 of them have a bachelor degree in percussion and study Music Technology and Composition at the Master level at McGill University. They are 5 males and 1 female, right-handed, and have different training timpani experiences (from 6 to 13 years of practice). The same marker setup was used for both databases. DB13 also contains timpani beats with 3 classes and 3 variations within each class.

For each class and each variation there is one sequence of 30 strokes. For both databases the performances were executed at the same tempo of 80 bpm, and a pause occurred between the sequences.

Motion Representation

Each recorded sequence is composed of repeated beat units which are manually segmented, each beat being centered on the striking position. This means that one beat contains several frames before the stroke, corresponding to the preparatory movement, and several frames after the stroke, corresponding to the retractive movement. Each beat is a multi-dimensional time series, which is represented as a sequence of k upper-body postures including the stick configuration (with k around 350). Each posture, $X_i = \{x^1, x^2, ..., x^m\}$, corresponds to the 3D Cartesian positions or the angular values of $m = 18$ markers located at major joints in the human upper-body plus the position of 3 other markers located at the stick's extremity. Throughout the paper the following notation will be used: each beat B is represented by the triplet (Attack, Dynamics, Striking position), let $B = (A, D, S)$, with A $\in \{l, a, v\}$, D $\in \{p, mf, f\}$, and S $\in \{r, 1/3, c\}$.

Beat Descriptors

Once the beats are segmented, they can be represented by spatio-temporal descriptors. However, as a beat is represented by a matrix of k postures $X_i, i = 1..k$, it is necessary to find out a reduced-dimensional representation that still contains the expressive kinematics and its variation along the beat. In this paper, inspired by knowledge in musicology and percussion teaching, we will assume that the position and velocity of the end position of the mallet in the vertical plane contain sufficient information to characterize timpani performances. In order to validate our beat descriptors, we use a classification approach. This classification is carried out under various features sets: (i) the position $X = (x, y, z)$ of the mallet extremity; (ii) the position and velocity $(dx/dt, dy/dt, dz/dt)$ of the mallet extremity; (iii) the position, velocity, and acceleration $(d^2x/dt^2, d^2y/dt^2, d^2z/dt^2)$ of the mallet extremity.

We will also focus on the temporal aspects of the beat, i.e. we will analyze which part of the beat before and after the striking position mostly influences the classification results. Figure 2 illustrates the z-position trajectories of several superimposed beats achieved for different attack conditions for the referent S_0 (top) and the trainees (S_1 to S_6). In this figure we can observe the regularity of the different strokes executed for a given attack, and the changing position of the velocity peaks according to the attack. In the next section we will analyze the intra-subject classification of the main variations of attack (both for the referent S_0 and trainees S_1 to S_6). We will also analyze for inter-subject classification the influence of the center position of the beat, and the influence of the window's length taken from either side of the striking position.

Classification

We adopt a simple non-parametric classification method, i.e. k-Nearest Neighbors, also called k-NN, commonly used in pattern recognition. The data set is divided into a training set and a test set (remaining data). Each sample input from the test set is used to predict the class in which this sample falls, by finding the k closest training examples. The output of the classifier is the class that obtains the majority vote of its neighbors. The nearest neighbors are computed according to a distance between the test beat and the example beats from the training set.

As the timpani beats are produced at the same tempo, and are all segmented around the striking moment, they are temporarily aligned and have the same length. Therefore, to evaluate the similarity between the beats, a simple Euclidian distance can be applied on the time series representing them. As experimentally verified, more sophisticated elastic similarity measures such as Dynamic Time Warping do not perform better than the straightforward Euclidean distance. As results of the classification tasks, we will obtain confusion matrices, which show the success rate of the k-NN classifier for each class.

INTRA-SUBJECT EVALUATION

The aim of the first study is to find an effective subset of motion features that describes the kinematic of timpani gestures and is sufficient to classify the main variations of attack for the referent data S_0. We first classify the whole set of beats according to the 3 main variations of attack ($legato$, $accent$, $vertical\ accent$) executed in different conditions of intensity and striking position. The beat descriptors are the z-position of the extremity of the right hand mallet expressed in the shoulder coordinate frame. The k-NN classifier outputs a score rate of 100% for this task.

We then refine this classification task by analyzing the influence of sub-variations related to the intensity (p, mf, f) or the striking position of the mallet ($center$, $1/3$, rim). The results are showed in Table 1 and 2 with an average classification rate of 98.40% for intensity sub-variations, and 99.83% for striking position sub-variations.

The classification task carried out on subject S_0 applied on the Attack beats (with intensity mf and striking position $1/3$) is also validated for the other 6 subjects (S_1 to S_6), with the score rates presented in table 3, ranging from 86% to 100%. Hence, the results obtained are very good for intra-subject accuracies. This can probably be explained by the fact that percussionist players have a very long period of learning, resulting in extremely regular skilled gestures.

INTER-SUBJECT EVALUATION

As a result of intra-subject evaluation, we proved that for a large amount of data provided by a performer, the trace of the right hand mallet is able to represent the expressiveness of the gesture. In this section we propose to compare different percussion students in relation to a professional percussionist. The basic idea of inter-subjects evaluation is to use the data of the professional performer S_0 as a training set, and the students data (from S_1 to S_6) as a test set, and to classify

		Legato			Accent			Vertical		
		p	**mf**	**f**	**p**	**mf**	**f**	**p**	**mf**	**f**
Legato	**p**	100	0	0	0	0	0	0	0	0
	mf	0	100	0	0	0	0	0	0	0
	f	0	0	100	0	0	0	0	0	0
Accent	**p**	0	0	0	94.66	0	0	5.34	0	0
	mf	0	0	0	0	100	0	0	0	0
	f	0	0	1.47	0	0	95.59	0	0	2.94
Vertical	**p**	0	0	0	1.55	0	0	98.45	0	0
	mf	0	0	0	0	0	0	0	100	0
	f	0	0	0	0	0	3.125	0	0	96.87
Average classification rate: 98.40 %										

Table 1. Confusion matrix under 3 attacks and 3 sub-variations of intensity: p, mf, f.

		Legato			Accent			Vertical		
		center	**1/3**	**rim**	**center**	**1/3**	**rim**	**center**	**1/3**	**rim**
Legato	**center**	100	0	0	0	0	0	0	0	0
	1/3	0	100	0	0	0	0	0	0	0
	rim	0	0	99.24	0	0	0.76	0	0	0
Accent	**center**	0	0	0	100	0	0	0	0	0
	1/3	0	0	0	0	100	0	0	0	0
	rim	0	0	0.74	0	0	99.26	0	0	0
Vertical	**center**	0	0	0	0	0	0	100	0	0
	1/3	0	0	0	0	0	0	0	100	0
	rim	0	0	0	0	0	0	0	0	100
Average classification rate: 99.83 %										

Table 2. Confusion matrix under 3 attacks and 3 sub-variations of striking position: $center$, $1/3$, rim.

Subjects	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Rate	**98.11**	**100**	**95.35**	**93.75**	**86.95**	**97.87**

Table 3. Score rates for intra-subject classification for Attack beats (mf, $1/3$).

Subjects	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Rate	**41.51**	**67.35**	**88.37**	**35.42**	**39.13**	**42.55**

Table 4. Score rates for inter-subject classification for whole Attack beats.

these data according to attack variations. The evaluation result turns out to be a ranking of how well the students perform comparatively to a professional performer. However, the classification scores fall down to an average of 52.4% when using only the trace of the right hand mallet, as shown in table 4. Therefore, in order to improve the classification, it is necessary to refine the classification experiments.

Following studies on percussion playing [6], we consider as descriptors not only the trace of the trajectory of the mallet extremity over time, but also time series expressing the kinematics of the movement (velocity, acceleration, etc.). Hereinafter we will experiment two feature sets, one with the mallet extremity position (x,y,z), and one with both position and velocity of the mallet extremity, i.e. the 6 dimensional time series ($x(t)$,$y(t)$,$z(t)$,$vx(t)$,$vy(t)$,$vz(t)$).

Furthermore, as observed previously, the beat unit of each subject contains not only sound-producing gesture but also ancillary gesture, which stylizes each performer and tends to introduce a large variability in the performance. Therefore we apply a time window to every stroke, so as to restrict the temporal size of the beat, and thus to focus on the sound-producing gesture while minimizing the effect of the ancillary gesture.

Influence of the Window's Size

In this experiment we consider that the classification is applied on a window centered on the frame when the mallet hits the skin, i.e. frame 175 (middle of the 350-size beats), and we change the size of the window (from 50 frames to 350 frames around the Center 175). The experiment is first conducted with the position of the mallet extremity as feature set, and then with the position and velocity of the mallet extremity.

In the first case (position only), the classification rates are given in table 5 for each subject and window's size, as well as the average for all subjects. We observe that the best results are obtained for a window's size of 150 or 200 for most students (with an average classification rate between 65.8% and 66.6%).

In the second case (position and velocity), the classification scores are given in table 6. We observe that they are sensibly improved (from an average rate of 65.8% for position only to 71.9% for both position and velocity). This supports the hypothesis of Cook, wherein the velocity of the mallet plays an important role in the control of percussive gestures [6].

Influence of the Window's Center

To study the importance of the preparatory gesture versus the retractive one, we also tested dissymmetrical windows, i.e.

SZ	S1	S2	S3	S4	S5	S6	AVG
50	54,72	65,31	88,37	25,00	54,35	51,06	56,47
100	58,49	91,84	86,05	35,42	60,87	**59,57**	65,37
150	60,38	**95,92**	**95,35**	**37,50**	**65,22**	40,43	65,80
200	**67,92**	93,88	86,05	35,42	**65,22**	51,06	**66,59**
250	71,70	85,71	83,72	33,33	**65,22**	48,94	64,77
300	60,38	75,51	88,37	33,33	60,87	42,55	60,17
350	41,51	67,35	88,37	35,42	39,13	42,55	52,39

Table 5. Score rates for inter-subject classification using position features for various window sizes, with a window center = 175.

SZ	S1	S2	S3	S4	S5	S6	AVG
50	84,91	55,10	72,09	35,42	52,17	53,19	58,81
100	77,36	93,88	90,70	**37,50**	**65,22**	**70,21**	**72,48**
150	71,70	**95,92**	**95,35**	**37,50**	**65,22**	65,96	71,94
200	69,81	**95,92**	93,02	**37,50**	**65,22**	65,96	71,24
250	**83,02**	93,88	83,72	**37,50**	67,39	55,32	70,14
300	**83,02**	81,63	81,40	39,58	67,39	53,19	67,70
350	69,81	75,51	76,74	37,50	52,17	42,55	59,05

Table 6. Score rates for inter-subject classification using position and velocity features for various window sizes, with a window center = 175.

CTR	S1	S2	S3	S4	S5	S6	AVG
100	30,19	46,94	41,86	29,17	56,52	31,91	39,43
125	35,85	53,06	46,51	33,33	60,87	36,17	44,30
150	**75,47**	91,84	74,42	33,33	**67,39**	44,68	64,52
175	60,38	95,92	95,35	**37,50**	65,22	40,43	65,80
200	62,26	**97,96**	**97,67**	**37,50**	65,22	**65,96**	**71,10**
225	52,83	91,84	97,67	**37,50**	56,52	51,06	64,57
250	56,60	93,88	97,67	39,58	30,43	30,43	58,10

Table 7. Score rates for inter-subject classification using position features for various window centers, with a window size = 150.

CTR	S1	S2	S3	S4	S5	S6	AVG
100	30.19	44.90	46.51	**39.58**	47.83	34.04	40.51
125	39.62	46.94	44.19	33.33	58.70	31.91	42.45
150	**88.68**	89.80	69.77	35.42	**67.39**	59.57	68.44
175	71.70	95.92	95.35	37.50	65.22	**65.96**	**71.94**
200	62.26	**100.00**	**97.67**	37.50	63.04	63.83	70.72
225	60.38	**100.00**	**97.67**	37.50	50.00	53.19	66.46
250	60.38	**100.00**	**97.67**	37.50	41.30	44.68	63.59

Table 8. Score rates for inter-subject classification using position and velocity features for various window centers, with a window size = 150.

we analyzed the influence of the center of the window. We tested our classification rates for both feature sets (position and position + velocity), while varying the window's center (from 100 to 250).

The results are respectively given in tables 7 and 8. When selecting position as feature set, the best results are obtained for a window centered around frame 200 (i.e. during the mallet rebound), whereas they are obtained for a window centered around frame 175 for position and velocity (i.e. when the mallet hits the skin). They are also slightly better for the second feature set (72% against 71%).

Besides the improvement of the scores, we also observe more stability within the subjects when selecting the second feature set (position + velocity). Note that we also experimented another feature set comprising position, velocity and acceleration, but the results were about the same as the ones with position and velocity. As acceleration did not improve classification accuracy, we therefore kept this last feature set in the remaining experiments.

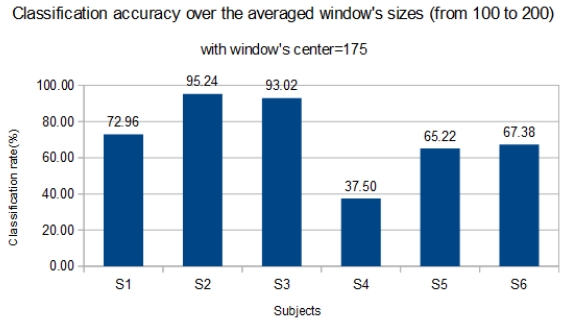

Figure 3. Inter-subject average accuracy according to the subjects, using position and velocity features

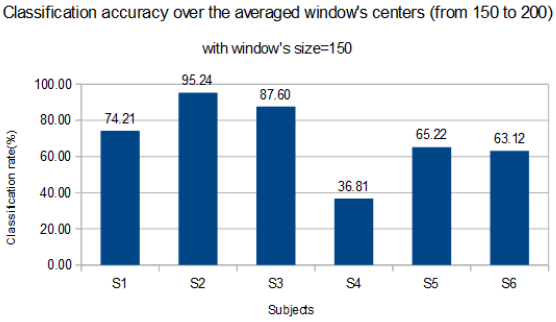

Figure 4. Inter-subject average accuracy according to the subjects, using position and velocity features

Quantitative Ranking of the Students

For position and velocity selected as descriptors, the scores of each subject are respectively illustrated in Figures 3 in the case of averaging window's sizes of 100, 150, and 200 (with window's center = 175), and in 4 in the case of averaging window's centers of 150, 175, and 200 (with window's size = 150). They can be interpreted as follows: for both sets of features, students are mostly sub-divided in three groups. S_2 and S_3 have the best score, (respectively 95.24% for S_2 and between 87.6% and 93% for S_3), S_1, S_5 and S_6 have a score around 63-74%, while S_4 has the lowest score (around 37%).

EVALUATION THROUGH COMPARISON WITH GROUND TRUTH RANKING

Perceptual evaluations have been conducted by Prof. Marandola to provide a ground truth on the quality of the timpani's performances. This evaluation was carried out on the basis of the videos of the database DB13 showing sequences of beats executed with intensity mf and striking position $1/3$. The evaluation focused on both quality of the produced sounds and gestures performed by each student, with a score rating the following sound/gesture attributes (score from 1 to 10). For the sound was rated 1.1) the fulness of sound; 1.2) the quality of the attack (does the sound correspond to the expected attack?). For the gesture was rated 2.1) the balance between Left and Right Hand; 2.2) the regularity of the dynamics of the strokes; 2.3) the muscular relaxation (absence

Subjects	S1	S2	S3	S4	S5	S6
Ground truth rate Gesture	5	1	3	6	4	2
Ground truth rate Gesture + Sound	5	1	2	6	4	3
Classification rate Position + Velocity	3	1	2	6	5	4

Table 9. Intra-subject ranking: ground truth and automatic classification.

subjects		1.1	1.2	2.1	2.2	2.3	Gesture Average	Sound Gesture Average	Gesture Ranking	Sound Gesture Ranking
S1	Legato	6	5	5	7	7	6.33	6		
	Accent	6	7	6	6	4	5.33	5.8		
	Vertical	7	7	7	7	7	7	7		
							6.22	6.27	5	5
S2	Legato	7	8	8	8	8	8	7.8		
	Accent	7	8	7	7	6	6.67	7		
	Vertical	7	8	8	8	8	8	7.8		
							7.55	7.53	1	1
S3	Legato	8	8	7	7	8	7.33	7.6		
	Acent	8	7	7	7	6	6.67	7		
	Vertical	7	6	7	7	7	7	6.8		
							7	7.13	3	2
S4	Legato	4	5	5	5	6	5.33	5		
	Accent	5	5	6	6	7	6.33	5.8		
	Vertical	5	6	6	6	5	5.67	5.6		
							5.78	5.47	6	6
S5	Legato	7	7	7	7	7	7	7		
	Accent	7	7	7	7	6	6.67	6.8		
	Vertical	6	6	7	6	6	6.33	6.2		
							6.67	6.67	4	4
S6	Legato	7	7	8	7	8	7.67	7.4		
	Accent	6	6	7	7	7	7	6.6		
	Vertical	6	7	7	7	6	6.67	6.6		
							7.11	6.87	2	3

Table 10. Perceptual evaluation: scores (from 1 to 10) using gestural and sound attributes for students S_1 to S_6.

of tensions in the neck, shoulders, arms, etc.). For each student separately, and for each category of attack, the different attributes were rated on a scale of 1-10. The results are given in Table 10. The average rates are also given per subject on the basis of the gestural attributes, and the sounds and gestures attributes. We also output a general ranking associated to the average scores.

A comparison between ground truth and automatic classification ranking is then possible. As seen in Table 9, we observe that the classification using both position and velocity is closer to the ground truth established on the conjunction of Gesture plus Sound (for a window size of 150, centered on 175). Moreover, the main tendencies are respected: the highest scores for the ground truth, respectively (7.53 and 7.13) are obtained by subjects S_2 and S_3 (we got 95.92 and 95.35 for the classification accuracies), and the lowest classification rate by subject S_4 (5.47) which corresponds to an accuracy of 37.5. The 3 other subjects S_1, S_5 and S_6 obtain very close scores for the ground truth, respectively (6.27, 6.67 and 6.87), which is about the same for the classification accuracies (71.70, 65.22 and 65.96). According to these results, it can be argued that the raw kinematic features (position + velocity) calculated from either side of the striking moment give a good estimation of the overall performance. It seems not necessary to use the whole ancillary gesture (lasting 1750 ms), since it seems sufficient to observe what happens during the sound-producing gesture around the stroke (the observation interval being 750 ms).

This experiment shows that focusing the analysis of the trajectory of the mallets on a small window centered on the stroke moment improves the correlation between the classification rates and the perceptual evaluation of the quality of the gestures produced by a subject. One can argue that since this analysis windows corresponds to the most effective part of the sound-producing gesture, it integrates some information that is highly correlated to the sound that is produced and which is used as a perceptual cue by the evaluator.

CONCLUSION

In this paper different sets of raw kinematic features were experimented for describing timpani gestures, and a study about how these features precisely characterize variations of the attack in various contexts was conducted. An automatic classification approach was used for both intra-subject and inter-

subjects in varied conditions: different sets of time series (positions, velocities), different window sizes, symmetrical or not around the striking moment.

Our results confirm musical performance hypotheses stressing out the importance of the combination of position and velocity of the mallet trajectories when executing percussion gestures. Note that the 3D trajectories give much better results than only the vertical ones, especially for musicians whose performance is not so regular. In addition, we showed the importance of the window's size for evaluating the quality of the beat attacks: the classification results are more stable and significant around the striking hit (150 frames) than for the whole beat. This result stresses the fact that the most important part of a percussive gesture, during sound production, does not necessarily include ancillary aspects. Ancillary gestures still play an important role, as they ensure that the striking target is reached with the appropriate kinematic, and the wide variation from one performer to another could then be explained by the necessity, for each player, to find the best way to reach the ideal balance between velocity and position, given his/her particular morphology and playing style.

The results are also consistent with the ground truth made from a perceptual evaluation which has led to a ranking of different students on the basis of gestural and sound attributes. This has led us to conclude that raw kinematic trajectories around the striking hit contain most of the information needed for a subjective evaluation. Our approach gives also some insight into a possible quantitative evaluation of students, by classifying their gestures using professional gestures as training set. Therefore, showing that the quality of the students gestures (according to their professor) matches some quantitative score is highly interesting. This could lead to innovative and self-guided ways of learning, that could be adapted to different playing styles pre-recorded from different reference interpreters.

A follow up will be to directly evaluate the individual gestural performance though distance or similarity measures with an expert performer. Concerning the methodology, we intend to use both measures of Precision / Recall to improve the interpretation of the results (threshold effect). We also consider to extend our data set by taking all the beats in various execution situations (varying dynamics and striking positions). Our study could however be extended, taking into account right and left hands, and various performing conditions (variations in intensity, striking position, tempo, etc.). We might also use physical measures about the timing impact with the skin, or compute biomechanical features estimating the relaxation of gesture. Finally, it would be interesting to link this study to a classification of the sound-related signals.

REFERENCES

1. Bouënard, A. *Synthesis of Music Performances: Virtual Character Animation as a Controller of Sound Synthesis*. PhD thesis, Université Bretagne Sud, France, 2009.
2. Bouënard, A., Gibet, S., and Wanderley, M. M. Hybrid inverse motion control for virtual characters interacting with sound synthesis - Application to percussion motion. *The Visual Computer 28*, 4 (2012), 357–370.
3. Bouënard, A., Wanderley, M., and Gibet, S. Virtual Gesture Control of Sound Synthesis: Analysis and Classification of Percussion Gestures. *Acta Acustica united with Acustica 96*, 4 (2010), 668–677.
4. Bouënard, A., Wanderley, M. M., and Gibet, S. Analysis of Percussion Grip for Physically Based Character Animation. In *Proc. of the International Conference on Enactive Interfaces* (2008), 22–27.
5. Bouënard, A., Wanderley, M. M., Gibet, S., and Marandola, F. Virtual Control and Synthesis of Music Performances: Qualitative Evaluation of Synthesized Timpani Exercises. *Computer Music Journal 35*, 3 (2011), 57–72.
6. Cook, G. *Teaching percussion*. Wadsworth Publishing Company, 1997.
7. Cuykendall, S., Michael, J., Amanzadeh, M., Tcheng, D., Wang, Y., Schiphorst, T., Garnett, G., and Pasquier, P. Hearing movement: How taiko can inform automatic recognition of expressive movement qualities. In *Proceedings of the 2Nd International Workshop on Movement and Computing*, MOCO '15 (2015), 140–147.
8. Dahl, S. The playing of an accent–preliminary observations from temporal and kinematic analysis of percussionists. *Journal of New Music Research 29*, 3 (2000), 225–233.
9. Dahl, S. Playing the Accent: Comparing Striking Velocity and Timing in Ostinato Rhythm Performed by Four Drummers. *Acta Acustica united with Acustica 90*, 4 (2004), 762–776.
10. Dahl, S. *On the beat: Human Movement and Timing in the Production and Perception of Music*. PhD thesis, KTH Royal Institute of Technology, Sweden, 2005.
11. Dahl, S. Striking movements: A Survey of Motion Analysis of Percussionists. *Acoustical Science and Technology 32*, 5 (2011), 168–173.
12. Kapur, A., Essl, G.and Davidson, P., and Cook, P. The Electronic Tabla Controller. *Journal of New Music Research 32*, 4 (2003), 351–360.
13. Larboulette, C., and Gibet, S. A Review of Computable Expressive Descriptors of Human Motion. In *Proceedings of the 2Nd International Workshop on Movement and Computing*, MOCO '15 (2015), 21–28.
14. Tindale, A. R., Kapur, A., Tzanetakis, G., Driessen, P., and Schloss, A. A Comparison of Sensor Strategies for Capturing Percussive Gestures. In *New Interfaces for Musical Expression* (2005), 200–203.