draft version - The Augmented String Quartet: Experiments and Gesture Following
F Bevilacqua, F Baschet, S Lemouton
Journal of New Music Research 41 (1), 103-119, 2012

final version available at the JNMR website
http://www.tandfonline.com/doi/abs/10.1080/09298215.2011.647823

The augmented string quartet:

experiments and gesture following

Frédéric Bevilacqua , Florence Baschet, Serge Lemouton

STMS Ircam-CNRS-UPMC

1 Place Igor Stravinsky,

75004 PARIS, France

{frederic.bevilacqua}@ircam.fr

February 7, 2012

Abstract

We present interdisciplinary research undertaken for the development of an "augmented" string quartet. Hardware and software components were especially designed to enable mixed acoustic/electronic music where bow gestures drive digital sound processes. Precisely, inertial motion sensors and a bow force sensor were added to each musician's bow, and dedicated modules allowed for the wireless data transmission to an on-line gesture analysis system. Prior to the performance, a research phase was performed to evaluate qualitatively the variability of the gesture data. Recording sessions of both gesture and audio data were carried on with a professional string quartet. The

music material included a set of prototypical musical phrases containing various bowing styles and playing techniques as well as a complete music composition. The analysis of the recorded sessions allowed us to compare the consistency within and between players. While a given player was found to be generally consistent, the comparison between players revealed significant gesture idiosyncrasies. These results helped us to adapt a real-time gesture analysis system called the *gesture follower*. This tool was successful to automatically synchronize the live performance with electronic sound transformation in two concerts. A quantitative assessment is reported on a specific section of the piece, illustrating the accuracy and the types of errors encountered.

1 Introduction

This paper reports on a interdisciplinary research project aiming at the development of an "augmented string quartet". This project was actually a follow-up of the "augmented violin" project (**?**). The "augmented violin" refers to a mixed acoustic/digital instrument where motion sensors are added to the violin bow, allowing for the direct control of sound processes by the bow movements. The "augmented quartet" was thus first designed as an extension of the "augmented violin" technology to an entire string quartet, equipping the bows of the two violins, the viola and the cello with sensors. Moreover, a major goal was also to develop further the analysis framework that was built for the augmented violin, which was essentially

able to only recognize standard bowing styles (e.g. *Détaché*, *Martelé*, *Spiccato*) based on the bow acceleration. This was considered as too limited for most contemporary music where a larger variety of playing techniques must be considered.

Therefore, the specific aims of the project were threefold. The first aim was to replace the bow measuring unit with extended motion sensors and to add a bow force sensor previously developed within a research context (**?**). The second aim was to assess qualitatively, in a given artistic context, the variability of gesture data measured from different string players. This was thought as a necessary step for our third aim, which was to adapt and assess a real-time gesture recognition system, called the *gesture follower*(**??**). This system allows for the on-line recognition of continuous time profiles. It was initially developed for cases such as dance or conducting but had never been used with string instruments before this project. This third part concerned the implementation of the *gesture follower*in the case of the augmented quartet, in order to synchronize bowing gestures with digital sound processes in a concert setting.

From a research point of view, the whole project was considered as a case-study: studying instrumental gesture in a specific artistic context. Nevertheless, the scope of this project was actually larger, expecting that the results and the tools developed could branch out to further investigations in the development gesture-based interactive systems.

This paper is structured as follows. First, we recall important related works. Second, we explain the general methodology and the artistic and musical context of this research. Third, we describe the gesture capture systems and the analysis system. Finally we discuss a series of results we obtained with the "augmented string quartet". Particularly, we describe a quantitative assessment of the *gesture follower*for a specific music section.

2 Related works

The interdisciplinary work presented here is related to several research fields. We give here a selection of references related to the fields of music interfaces, the study of instrumental gestures and more generally in human-machine interactions.

2.1 Augmented and alternative string instruments

Experiments with augmented instruments, i.e. adding sensors to existing instruments, have been reported since the beginning of electronic music (see Poepel for a review of several string related electronic instruments (**?**)). For example, at the end of the eighties Machover developed several instruments he called *hyperinstruments*, including the *hypercello* (**?**). Over the last ten years, several types of novel string instruments were proposed and used in music performances. Young created the *hyperbow* using various sensors to measure acceleration, force and position of a violin bow (**???**). Freed

created software and hardware enhancements to an electric 6-string cello (**?**). Overholt entirely built an unique electronic cello, with various sensors on the bow and on the violin body (**?**). Several other works concerned the development of augmented bows using inertial sensors that could work with any type of acoustic string instruments (**???**). Several pieces written specifically for such instruments were also documented (**???**). Note that a commercial product, the KBow has recently appeared on the market (**?**).

Other related works include interfaces directly inspired by bowing techniques. For example, Nichols developed a virtual violin bow haptic human-computer interface, which senses bow position to drive bowed-string physical model synthesis (**?**). Trueman and Cook developed the BoSSA, a Bowed-Sensor-Speaker-Array that includes violin's physical performance interface and its spatial filtering audio diffuser (**?**). We also note that several systems have been developed for music pedagogy (**??**).

2.2 Studies of string gestures

Other researchers have been studying violin gestures using motion capture systems. Unlike the previous cited works that are aimed to artistic performances, these studies are directed towards the understanding of acoustics and/or instrumental performances. These studies typically requires accurate capture systems that are generally not compatible with concert settings (**??????**). Instrumental gesture measurements are also carried on for the

improvement of sound synthesis based on physical modelling (**??**). More generally, such gesture studies are also directly related to broader gesture research in music (**??**).

2.3 Interactive systems and machine learning techniques

The *gesture follower*system presented here is close, from a technical point of view, to existing score following systems that allow for real-time automatic synchronization of audio with a symbolic score. Score following has a long history and has increasingly integrated machine learning techniques over the years (**????**). The *gesture follower*system is also related to the various techniques that were developed for the recognition or classification of bow strokes using gesture data (**????**).

Note that gesture recognition systems are increasingly used in media and performing arts. For example, the system described in this paper has been implemented in other artistic practices such as dance (**?**). Moreover, such artistic practices tend to overlap with the community of Human-Computer Interactions (**???**).

3 Research Context and Methodology

The research presented here was performed in the context of a collaboration with Florence Baschet (2^{nd} author), who was granted a research residency to experiment on gesture interfaces for electroacoustic music and commissioned

a composition for string quartet.

This piece was designed as a mixed acoustic-electronic piece with real-time sound transformation of the strings sound. The specific approach was to use bowing gestures to control the sound transformation. The research and tools development started in 2007 and the piece (called *StreicherKreis*) was premiered in Paris on November 2008. The composer collaborated closely with researchers and sound designers, and the whole interdisciplinary research process is documented in (**?**).

We worked closely with a professional string quartet of high international recognition. Their usual repertoire varies from classical, romantic and contemporary music. They were commissioned to participate in recording sessions for experimentation, and to perform two concerts (including the premiere). As explained below, the work plan was set in two phases.

3.1 Phase 1

The first phase included eight recording sessions of three hours (approximatively one every month) with various members of the string quartet (two sessions were with the full quartet). The goal of this phase was twofold. The first goal was to test the hardware system and to confirm its playability by the musicians (considering possible disturbances due to the sensors). The second goal was to provide us with a large set of data representative of the musical material sought by the composer. This part was primarily

dedicated to off-line analysis, in order to assess *qualitatively* the consistency within and between players with identical musical material and to test a first version of the *gesture follower*.

- Sessions 1-3. The first three sessions were dedicated to test musical materials used in our previous "augmented violin" project, namely the composition "Bogenlied" (by Florence Baschet). This composition, initially written for violin, was adapted for cello (session 1) and viola (session 2) . In session 3, both violin players also played music excerpts from "Bogenlied" as well as other music phrases prepared by the composer.

- Sessions 4-6. These sessions were dedicated to eight prototypic phrases that were specifically composed to include various playing techniques (*gettato, spiccato, marcato, détaché, flautando, tremolo, écrasé/high pressure*), various sets of dynamic marks and other specific bow motion indications. These phrases were recorded several times by the two violin players in session 4, and by the viola and cello players in session 5. At least two satisfactory versions of each phrase, according to the composer and musicians, were kept. The two violin and cello players recorded again these phrases in session 6, in order to evaluate the variations possibly occurring between different sessions (apart from 3 to 7 weeks). Complementary materials, conceived as variations of the eight prototypical phrases, were additionally prepared by the composer

and recorded.

- Sessions 7-8. Sessions 7 (with 1st and 2nd violin) and 8 (with viola and cello) were dedicated to record new composed material and to test further interpretation intentions. Precisely, the musicians were specifically asked to play the same phrases with different interpretations.

3.2 Phase 2

The second phase was dedicated to record various sections of the composition *StreicherKreis*, which creation benefited from the knowledge gained in the first experiment phase. The composer structured this twenty-five minutes musical piece in eleven sections (A to K), based on artistic criteria .

The recording sessions allowed for the testing of the *gesture follower* and for the adjustment of different parameters (described in section Analysis Methods and Results).

These recordings were also necessary for the composer to create the electronic part of the piece. Direct mapping between the sensors values and sound transformation were designed and saved as presets (the complete description of the sound mapping and the sound transformation would be out of the scope of this paper). Each mapping preset was activated by the *gesture follower*at a specific synchronization marker. All markers were written in the score (Figure 3), and imported in the *gesture follower*.

At least two satisfactory versions of each section were recorded at each

session, organised as follows:

- Sessions 1 and 2. First recording of sections B, C, D and I by the whole quartet, and first test of the *gesture follower*with the violin players one month later (September-October 2007).

- Sessions 3 and 4. Full recording of the piece "StreicherKreis" (only few changes were introduced in the written score after these recording sessions). Sections A to F were recorded in January 2008, G to K in February 2008.

- Session 5-7. Additional recordings of sections F to K (October 2008), sections A, D, E, I, J, K (November 2008), sections C and K (November 2009, one week before the premiere).

4 Sensing Technology

The technology for the augmented quartet was specified based on our previous experience with the augmented violin (**?**) . Similarly to the previous system, an inertial measurement unit was mounted on the bow. Nevertheless, instead of using two single axis accelerometers, we upgraded our measurement unit to a module combining a 3-axis accelerometer (Analog Device ADXL335) and a dual-axis gyroscope (InvenSense IDG500). All sensors were sampled with a precision of 10 bit at a sampling frequency of 200 Hz, using wireless modules developed by Emmanuel Flety and Nico-

las Leroy described in (**?**). These modules are based on the *XBee* (from MaxStream), which is a small form factor OEM module with simple communication means operating with the 802.15.4 IEEE standard (also called ZigBee). Each emitter module is linked to a receiver with a specific channel (one for each instrument), which enables the use of four wireless modules in parallel (this was not possible with the previous version of the augmented violin). All receivers were connected to a local Ethernet network, transmitting data to a central computer using the Open Sound Control protocol (OSC).

We will refer in the text to the accelerometer data as the values given by the accelerometer sensors (the X, Y and Z axis are indicated in Figure 1). It is important to note that the accelerometer raw values depend on both the acceleration and the orientation of the bow in regards to the gravity. Thus, the term "acceleration" used in this article must be understood as the *raw sensor values* and not as acceleration absolute values.

An additional "bow force sensor" was used, that was developed by Matthias Demoucron (see Figure 1a). The sensor is made of strain gauges that are pressed on the bow hair, close to the frog. This sensor can measure indirectly the force of the bow hair normal to the string (after calibration). A complete description is available in (**?**). In our case, the data were sampled and transmitted by the same device as the motion data.

It is important to note that this sensor can only report on the actual bow

force if the value is corrected by the bow position. Since our setup could not provide the bow position, we used the raw value of this "bow force sensor" without calibration. This value is thus not an absolute measure of the bow force but a value that increases with the actual bow force and decreases with the distance between the bow frog and the bridge. For the sake of simplicity, we refer to it as "bow force sensor".

Compared to the inertial unit, the bow force sensor was found to be more invasive and cumbersome to attach to the bow. Specifically, the sensor covered a short distance of the bow hair, and added few millimeters to the frog height. The different recording sessions allowed us to adjust its design and to reduce significantly the length of covered hair to approximately 1 cm (i.e. shorter than shown in Figure 1a, taken on the first session) . This reduced the sensitivity range of the sensor, and this configuration did not allow to sense bow force in soft bowing. However, the sensor remained sufficiently sensitive to observe bow force changes in many playing styles when played at sufficiently high dynamics or in playing mode naturally played with high "pressure" such as *marcato* or *écrasé*. Its utility was found very valuable from an artistic point of view, particularly when mapped to digital sound processes.

On the contrary to the early augmented violin design, we chose to place the sensors on the bow and to have the wireless emitter worn on the wrist, as shown in Figure 1. This limits the weight added to the bow, which

was 6.4 g (2.6 g for the inertial measurement unit and 3.8 g for the force sensor). Moreover, this separation simplifies the mounting of the motion sensors on the bow, since they can be attached by a simple elastic strap on the side of the bow frog. After several tests, this was found acceptable by the musicians. With this design, the sensing system can be installed on any musician's personal bow.

5 Off-line analysis and gesture following

Both off-line and on-line analysis were carried on during the project. In all cases, we considered all the sensor data combined with the audio energy of each instrument, i.e. time sequences of 7 measured values per instrument : 5 inertial measurements, 1 force sensor and 1 audio energy. We describe below the different methods applied during the different phases of the project.

5.1 Off-line Analysis

In the first phase of the project, all the recorded data were annotated to align the recorded data with the score. This implied to manually associate data features to specific events in the score such as notes and articulations. This task allowed us to qualitatively inspect the data, as discussed in the Results section.

Further quantitative evaluations were performed by data alignment using Dynamic time-warping (DTW), which is a well-known method for the off-

line alignment of two similar time-signals. We use DTW to automatically align the accelerometer data of some of the recorded phrases, employing the Matlab implementation of Ellis (**?**). Complementary to DTW, the *gesture follower*, described below, was also used to quantify differences between recorded data based on likelihood values.

5.2 On-line Data Synchronization: the *gesture follower*

For the concerts, the aim was to develop a generic tool that could analyze the musician gesture data in real-time, and synchronize them to sound processes. The software called *gesture follower*(**?**) appeared as an appropriate tool but had never been tested with strings. Therefore, this project represented an opportunity to improve this software and assess its validity for string performances.

Specifically, the *gesture follower*is a generic software tool that allows for on-line warping of a live performance data-stream to template data. In other words, the algorithm allows for the estimation of the time progression of the piece, given prerecorded data. The *gesture follower*can be seen as a derivative of a score-following system (**??**), where the sensors data themselves are used to represent the performance instead of a symbolic score.

The *gesture follower*can be considered as a non-standard implementation of Hidden Markov Models and has already been fully described in (**??**). The *gesture follower*is particularly suitable to continuous sensor data. We

present here the system that was implemented at the time of this project (which was generalized in more recent versions).

5.2.1 Markov Models and Decoding Algorithm

The *gesture follower*is easily described using the formalism of the Hidden Markov Models (HMM) as described by (**?**), but it could actually be considered as an hybrid approach between HHM and DTW methods. The first step, the *learning* procedure corresponds to set a Markov Model based on recorded data. Because of the limited training data available, our approach is similar to the "template" approach of DTW: we use a single example of the recorded data for the learning procedure. The state structure is therefore directly set from a "template" , which is a time sequence called $E_{1:2N}$:

$$E_{1:2N} = E_1 \; E_2 \; \ldots \; E_{2N} \tag{1}$$

where $2N$ is the length of the sequence.

Each element E_i is a data vector of size M measured at time $t = i/f$ where f is the sampling frequency, and M corresponding of the number of recorded "channels", i.e. sensor data and/or audio descriptors).

As illustrated in Figure 2, the Markov model is set after downsampling the data sequence by a factor two, leading to the sequence $E'_{1:N}$:

$$E'_{1:N} = E'_1 \; E'_2 \; \ldots \; E'_N \tag{2}$$

As this will appear clearer, the downsampling is necessary to model correctly

the initial sequence $E_{1:2N}$ with the state structure described below (note that this constrain was relaxed in a later version by taking into account more complex state structures).

The sequence $E'_{1:N}$ is then used to set a left-to-right Markov chain $S = S_1, S_2, \ldots S_N$. We choose only two possible transitions: *self* a_{ii} and *next* $a_{i(i+1)}$, a_{ij} being the state transition probability distribution from state i to j. Since the data is regularly sampled in time, the transition probabilities must be set such as $a_{i(i+1)} = 1 - a_{ii} = a$. A value of $a = 0.5$ corresponds to an average transition time equivalent to the original sequence.

The probability $b_j(O)$ of an observation O in state j are set to Gaussian distributions centered on the vector E'_j (O is a measured vector of length M). Obviously, using a single template is not enough for an complete estimation of these distributions. Heuristically, we choose a simplified form given by the following function:

$$b_j(O) \propto exp[-\sum_{m=1}^{M} w_m^2 \frac{(O_m - E'_{j_m})^2}{2\sigma^2}] \tag{3}$$

The σ value can be interpreted as an "average" standard deviation between the measured and template data. The w_m values are interpreted as *weights* for each data channel m. A value $w_m = 0$ suppresses the effect of the channel m. Both the σ and the w_m values are adjusted by the user, as it will be discussed in the results section.

Once the Markov model is set from a given template, we can run the decoding algorithm during a performance, i.e. on a growing sequence of

observation $O_{1:t}$

$$O_{1:t} = O_1\ O_2\ \ldots\ O_t \tag{4}$$

The decoding scheme corresponds to estimate the probability distribution $\alpha_t(i)$, which is the probability of the observation sequence $O_{1:t}$ and state S_i at time t (for the given model). $\alpha_t(i)$ is directly estimated using the well-known forward procedure (**?**).

From the $\alpha_t(i)$ distribution, we compute three different quantities:

1. The likelihood L_t of the observation sequence $O_{1:t}$

$$L_t = \sum_{i=1}^{N} \alpha_t(i) \tag{5}$$

   L_t which can be used as measure of similarity between the observation and template sequences.

2. The first moment μ_t of the normalized distribution $\alpha_t(i)/L_t$

$$\mu_t = \sum_{i=1}^{N} i\ \alpha_t(i)/L_t \tag{6}$$

   is used to estimate the *time progression index (t)* which is the essential output parameter of the system: it enables the real-time alignment of the observation to the template sequences:

$$time\ progression(t) = 2\mu_t/f \tag{7}$$

   The factor two in the last equation is necessary to correct the initial downsampling. Note that, due to the chosen Markov structure, the

maximum speed of the time progression index is twice the original speed of the template.

3. The variance of normalized distribution $\alpha_t(i)/L_t$ is also useful to calculate, as it will be discussed in the results section:

$$Var_t = \sum_{i=1}^{N} (i - \mu_t)^2 \, \alpha_t(i)/L_t \tag{8}$$

For efficiency, the forward procedure is calculated on a sliding window as described in (**?**):

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N \tag{9}$$

$$\alpha_{t+1}(j) = k[\sum_{i=i_{inf}}^{i_{sup}} \alpha_t(i) a_{ij}] b_j(O_{t+1}) \quad 1 \leq t \leq T-1, 1 \leq j \leq N \tag{10}$$

where a_{ij} is the state transition probability distribution,

π_i is the initial state distribution,

and

i_{inf} and i_{sup} are the inferior and superior index of the sliding window of length $2p$.

The k value is a renormalization factor due to the truncation of the sum. In practice, this factor can be ignored if p is large enough. Note that since a_{ij} has an extremely simple form, the computation of $\alpha_{t+1}(j)$ can be very efficient.

The i_{inf} and i_{sup} values are set as functions of the index μ_t as described

below:

$$\begin{aligned} i_{inf} &= 1, & i_{sup} &= 2p+1 & \text{if} \quad & 1 < \mu_t \leq p \\ i_{inf} &= \mu_t - p, & i_{sup} &= \mu_t + p & \text{if} \quad & p < \mu_t \leq N - p \\ i_{inf} &= N - 2p, & i_{sup} &= N & \text{if} \quad & N - p < \mu_t \leq N \end{aligned} \tag{11}$$

5.2.2 Implementation

For this project, we used a *gesture follower*version implemented as a set of the Max/MSP modules integrated in the toolbox MnM of the library FTM (**?**). It takes advantages of the data structure of FTM for Max/MSP such as matrices and dictionaries, and a set of tools for data visualization. A dedicated graphical interface (using the FTM editor) was designed to display the recorded data of the string quartet, as illustrated in Figure 3. The timeline can be annotated with various markers. A cursor indicates the time progression of the performance (see Figure 3).

A second optional graphical tool allows for the real-time display of the time-warping (Figure 7). All the parameters can be displayed separately. This feature is particularly useful during tests and rehearsals since it allows for the visualization of the differences between live and template data.

The *gesture follower*input was composed of all the sensor data and audio energy values of the four instruments, corresponding to an input vector of 28 elements (4 times 7 parameters). All sensors and audio energy were normalized to a range between 0 and 1. It was possible to vary in real-time the normalization or to choose a smaller set of input data by adjusting

the weights w_m using a graphical interface. Using the complete set of 28 parameters was equivalent to consider the whole quartet as a "dynamic system" with a single time reference. In this case, the underlying assumption is that every musician should remain synchronous to a master tempo, which was consistent with the compositional approach of the piece.

The data sampling frequency was 50 Hz and the windows parameter p was set to 250 samples (after downsampling), thus corresponding to a total temporal windows of 20s (always centered around the estimated progression index). Various tests demonstrated that this value was low enough to guarantee sufficiently low CPU consumption and large enough to have a negligible influence on the following accuracy.

Please note that a more recent implementation of the *gesture follower*has been rewritten as a C++ library (by Bruno Zamborlin), offering more flexibility in the Markov structure, and can be used in Max/MSP with the MuBu data container and visualization tools (**?**).

6 Results and Discussions

6.1 First phase: inspecting prototypical data

The first part of experiments was dedicated to the study of bowing gestures. Each string player was asked to perform short musical phrases, and the sensor data were synchronously recorded with the sound. We recall that the

main aim in this phase was not to obtain quantitative results, but to proceed to a *qualitative* assessment of data measured in a setting that matches constrains found in a concert setting.

Figure 4 presents a typical example of a recording, displaying the 3D accelerometer data, the bow-force sensor data and the audio waveform (the gyroscope data were removed here for clarity). In this example, as indicated in the score, most of the playing techniques are *spiccato* and *pizzicato*. Some of the notes are indicated *col legno battuto* (literally "hit with the wood") meaning that the bow stick must be used (i.e. using the wood part).

As shown in Figure 4, the different parts of the phrases can be clearly identified on the accelerometer and bow-force sensor data, and as expected, the variations of the data amplitude are consistent with the dynamics.

As found in previous studies (**?**), the x-axis (along the bow) is generally the most significant signal of the accelerometer sensor, with a clear peak at each up-bow and down-bow change, at sufficiently high dynamics (*mf*). For the *pizzicati*, the peaks are also very clear on the y-axis acceleration, which is expected since these movements are essentially perpendicular to the strings.

The bow-force sensor data show very distinct peaks. For example, the *spiccati* can be clearly distinguished on the bow-force sensor data, when played with dynamics larger than *mf*.

Preparation gestures, defined as the gestures performed before producing

the sound, are also noticeable. In particular, the preparation for the *pizzicato* is clearly identifiable from the accelerometer data, on the x-axis (see arrow in Figure 4).

The first question concerns the consistency of these gesture data for a given player. Figure 4 shows two successive recordings of the same phrase by the first violin during the same session. These data were recorded after few trials allowing him to practice the phrase (which contain technical difficulties). The second performance is approximatively 20% slower, but a strong consistency is found between the recordings. For example, a correlation of 0.939 is found on the accelerometer x-axis after time-warping the two recordings (see Table 1). Note that even small peaks in the accelerometer signals, which might be considered at first sight as "noise", are reproduced with high precision.

The general finding illustrated in Figure 4 were confirmed for each player (violin, viola and cello). Typically a high consistency for a given musician and a recording sessionwas observed for the eight musical prototypic phrases, involving various playing techniques (*gettato, spiccato, marcato, détaché, flautando, tremolo, écrasé (high pressure)*).

Generally, the largest discrepancies were found in the "bow force sensor" data. Several facts could explain the observed differences. First, we often found drift over time in the data, making difficult to achieve a constant calibration over time. This was related to the sensor length on the bow hair

that was minimized to reduce its invasiveness but reduced its sensitivity. Second, as explained previously, the raw value depends on both the bow force and the bow position (distance between the contact point on the bow and the sensor at the frog), and thus the variations could be also related to differences in the bow position.

As expected, much greater differences were found when comparing different interpretation contexts (e.g. playing solo vs in quartet) and different musicians. Let us examine now inter-players differences. Figure 5 shows the recording of the same phrase as in Figure 4, but played by the second violin (top) and the viola (bottom). A pitch transposition was necessary for the viola, which did not affect significantly the bow motion.

All the correspondences between the score, sound and gestures we commented for Figure 4 remain valid. Particularly, the different playing techniques appear with their specific motion features. This is confirmed quantitatively by the correlation values in Table 1: the correlation values associated to different players remain relatively high (between 0.8 and 0.9).

Precisely, the comparison between the first and second violin (Figure 4 vs 5, top) shows that, from the *pizzicati* to the end, the time profiles are similar. Most of the differences in the x-axis accelerometer are present in the first part of the phrase (see symbol A in Figure 4-top and A' in Figure 5-top). The second violin appear to have "articulated" the soft bowing gestures with more details than the first violin. Interestingly, such differences were often

found between these two players, which could be related to personal stylistic playing from a "gestural" perspective.

Comparing the two violins and viola players reveals even greater differences (Figure 4 vs Figure 5, bottom). These findings are again confirmed by the correlation values; the lowest values are found for the violin vs viola data (see Table 1). The differences in the physical/acoustic properties of these instruments could partially explain the observed results. Nevertheless, we should also point out the influence of the recording context. In contrast to the two violins players who were recorded during the same session, and thus could listen to each other, the viola was recorded on a separate session and he could not be influenced by the violins playing. This could explain important interpretations differences between the viola and the violins, reflected for example in the playing of the dynamics by viola (Figure 5-bottom).

These results motivated us to further investigate gesture differences induced by the musical interpretation. For example, Figure 6 illustrates recordings of two different interpretations that were specifically proposed to the first violin: "quiet" (top) or "energetic" (bottom). The results show that the "quiet" interpretation is significantly slower than the "energetic" one (approximatively 14 s and 8s respectively), and that the signal amplitudes are lower. Interestingly, the bow-force sensor data remains remarkably stable.

Globally, the x-axis accelerometer time profiles between the "quiet" in-

terpretation and the "energetic" are similar except for the speed and amplitude scaling (the correlation after time warping is 0.86). The largest differences appear for the 32^{th} notes in the second part of the phrase (see symbol B and B' in Figure 6). In this part, the player seems to adopt a different motion strategy which might be related to the different pace of these interpretations. Similar effects were found in a previous study on accelerando/decelerando in violin (**?**), and could be explained by biomechanical constrains forcing the player to use different motion strategies depending on the tempo.

Nevertheless, the gesture differences shown in this last example appear less important than expected. This can be compared to the study of (**?**) that showed that each piece might involve a particular motion strategy (consciously or not) that can be very stable for a given player.

In summary, we observed several types of gesture variations (time stretching, amplitude scaling, different motion strategies, etc) which was highly depending on context and seem difficult to predict with a unique model. The most important variations occur when we compare different musicians.

These large data set also allowed us to test off-line a first version of the *gesture follower*. In agreement with the qualitative findings reported here, preliminary tests showed that the *gesture follower*was working when the live and template data were taken from the same player. However, using the template data recorded from another musician was not always

reliable. Therefore, we decided at this point to use and evaluate the *gesture follower*only in cases where players are analyzed based on their own recorded template data, as described in the next section.

6.2 Assessment of the *gesture follower*

The second phase of the study concerned the use of the *gesture follower*. Figure 7 illustrate two examples of time-warping operated in real-time with the complete string quartet, as displayed on the user interface. The top figure corresponds to the same excerpt as shown in 3 (section C). The live performance (black) is superimposed on the template data (color/grey) As the input is composed of the whole quartet data, the time-warping is based on an "average" time progression of the four musicians. Therefore, differences of the musician synchronizations, occurring between the performance and the template, can be directly observed. For example, Figure 7-top clearly shows that, at marker 25, the cello plays the *flautando* later relatively to the template recording. Similarly, after the marker 27, the viola also plays later relatively to the template. Figure 7-bottom illustrates another time-warping example where the two performances appear to be very close, except to some different articulations in the first violin (see symbol C in Figure 7-bottom).

The various recordings of the piece (spanning over more than one year see section 3.2) allowed us to set different parameters of the *gesture follower*and to assess its accuracy. This assessment was performed using annotated sound

and data recordings. Precisely, markers were manually added at specific data features (e.g. acceleration peaks) in recordings being used either as a performance simulation (called test) or as a template. During a simulation, the *gesture follower*reports a timing for each marker which can be directly compared to the annotated timing. The time difference corresponds to an error value, which can be reported along a section as illustrated in Figure 8 for section C. This is one of the section we evaluated in details, since it contained a representative set of playing techniques used thorough the piece.

Figure 8 reports the errors related to two different recordings of section C, both recorded ten months apart from the used template. On average, 80% of the errors are inferior of 240 ms, corresponding to less than 6 sample intervals (40 ms). Nevertheless, Figure 8 shows that the errors are not constant over the section (of 3 minutes). This is due to a highly non-linear dependency of the warping errors to signal differences between the "test" and "template" data.

Larger errors, sometimes superior to 1s are found at specific locations. In such cases, the data similarity is very low between the test and template data. In extreme cases, the system can even stop and is then referred as "lost" (e.g. the system was lost around time 170 s in Figure 8-bottom due to a significant discrepancy between the test and template data). Interestingly, in Figure 8, the large errors found in the two recordings often occur at the same part of the phrase. Errors are actually more prone to appear at specific

spots, typically where there are insufficient features in the data to uniquely set the time-warping.

The large errors could be categorized as "misses" or "false positive" as formalized for score following (**?**). They occur when the system misaligns a feature with another one, making a feature to appear as "added" or "missing" when comparing the test and template data. As already suggested, such large errors occur when, from a statistical point of view, there are different possibilities for the time-warping. This is evidenced in the Figure 9 where the absolute errors are plotted along the variance of the $\alpha_t(i)$ distribution. A large variance indicates that the estimation of the *progression index* value might be ambiguous, leading to a large error. A small variance indicates that the *progression index* is statistically well defined, leading generally to small errors.

The information provided by the variance of the $\alpha_t(i)$ distribution is available in real-time, and could be used to invalidate (or weight) some of the output of the *gesture follower*. The variance value is complementary of the *likelihood* value (Equation 5) which provides information on the similarity between data.

The recordings set was also useful to adjust the σ and the weight parameters w_m (see Equation 3). Nevertheless, the dataset we collected was insufficient to run a general optimization procedure as generally defined in machine learning techniques (using much larger databases). The risk resided

in "over-fitting", i.e. optimizing these parameters for a limited number examples and losing generality.

The influence of the σ and the weight parameters w_m are plotted in Figure 10, estimated using the section C (same data as for Figure 8 top). For the sake of simplicity we report here the errors for only three w_m configurations, but which illustrate well our findings. The first configuration takes into account all sensors and audio energy values (28 input channels). The second configuration takes into account the sensors data only (24 input channels) and the third configuration takes into account the audio energy only (4 input channels). In each of these cases, we varied the σ values and reported the mean errors, and the maximum errors and the workable range (i.e. the values working for the whole section). The mean errors were calculated only for errors less than 1 s to avoid the disturbance of outliers (which effects are reported in the maximum errors). This allows for the separation of the two types of errors we discussed previously.

Figure 10 shows that the best results are obtained when combining both sensor and audio energy data. The use of the "sensor only" gives the smallest mean errors but the largest maximum errors. Moreover, the operation is restricted to a small range of possible σ values. This is due to the fact that the sensors data can be seen as a series of narrow peaks, favoring an high accuracy but also provoking large errors if these features are modified (for example if one peak is missing). On the contrary, the audio energy profiles

feature larger peaks and generally smoother data, leading to more robust time-warping (i.e. larger range of operation and smaller maximum errors) but with a lower precision (larger mean errors). The combination of sensor and audio energy data is thus complementary: Figure 10 shows that both the mean and maximum errors are minimized in this case. Moreover, there is optimal region for the σ values between 0.4 and 0.6 . We finally choose a value of $\sigma = 0.5$ which was confirmed to work efficiently in all the sections of the piece.

We also investigated the differences when operating the *gesture follower*with the whole quartet or with a single musician. For the music material we tested we found that globally the system was always more robust using the data of the four musicians combined. This was explained, considering the specific musical material we used, that the superposition of each musician's data contained more information for the *gesture follower*than taking each musician data individually.

Globally, the *gesture follower*results presented here were found satisfying for the specific application it was intended for. The accuracy was judged sufficient for the targeted artistic application (mean errors typically less than 200 ms), and the largest error we discussed were avoided by making appropriate choices on the markers locations, avoiding sparse difficult zones.

Finally, the *gesture follower*was successfully used in two concerts, using a set of heterogenous template data recorded from one week to ten months

earlier. This confirmed that the system was flexible enough as long as each player is associated to its own template data. Precisely, the system was able to report continuously the time progression during the whole piece (approx. 25 min long), and synchronize the digital sound processing. Only few errors occurred at the transition between sections, not affecting the electronic music being produced. The errors were due to the fact the section transition were played significantly faster in the performance than in the recording sessions.

7 Conclusions

We presented interdisciplinary research on technology developments for an "augmented" string quartet. This work was conducted in a specific artistic context and should be considered as a case study. Nevertheless, we believe that the results provide important insights for music instrumental studies and more generally for a large range of applications in gesture-controlled media.

First, we reported on sensing technology, applied to a string quartet, providing us with real-time continuous data related to bow motion and force. A large set of recordings with professional musicians showed that these data could be put in direct relationship with the score and the interpretation markings. The results were found to be very consistent considering each musician separately, revealing also some gesture idiosyncrasies.

Second, we presented a specific implementation of the *gesture follower*, a system to synchronize musician gestures to electronic processes. We reported an evaluation of the system and found that the best accuracy was obtained when combining gesture and audio features. The *gesture follower*was successively used in two concerts of the augmented quartet.

This project opened valuable perspectives for interactive media performances that we are currently pursuing. First, the technology has already been used in other music and dance performances. Moreover, the experience we gain with this project motivated us to further develop the *gesture follower*as a more generic tool. As we reported here, important challenges remain to characterize the "interpretation" of performers. This project represented a step toward this problematic that, we believe, is central for the advance of human-machine interaction.

8 Acknowledgements

We warmly thank the Danel quartet (Marc Danel, Gilles Millet, Vlad Bogdanas, Guy Danel) for their extraordinary patience and enthusiasm during this project. The authors are grateful the various people that were involved in this project and greatly helped us for the technology : Emmanuel Fléty, Nicolas Leroy, Matthias Demoucron, Nicolas Rasamimanana, Riccardo Borghesi, Norbert Schnell. We thank Nicolas Donin, Samuel Goldszmidt, Maylis Dupont for their documentation work and the useful discus-

sion around this work. We acknowledge support from the EU-IST project i-Maestro.

List of Figures

Table 1: Correlation values

Recording	correlation
Violin 1: ver. a vs ver. b	0.938
Violin 1 (ver. a) vs Violin 2	0.886
Violin 1 (ver. b) vs Violin 2	0.895
Violin 1 (ver. a) vs Viola	0.856
Violin 1 (ver. b) vs Viola	0.877
Violin 2 vs Viola	0.840

The data is relative to the x-axis acceleration data measured with the prototypic phrase *spiccato*. The correlation is calculated after time-warping.

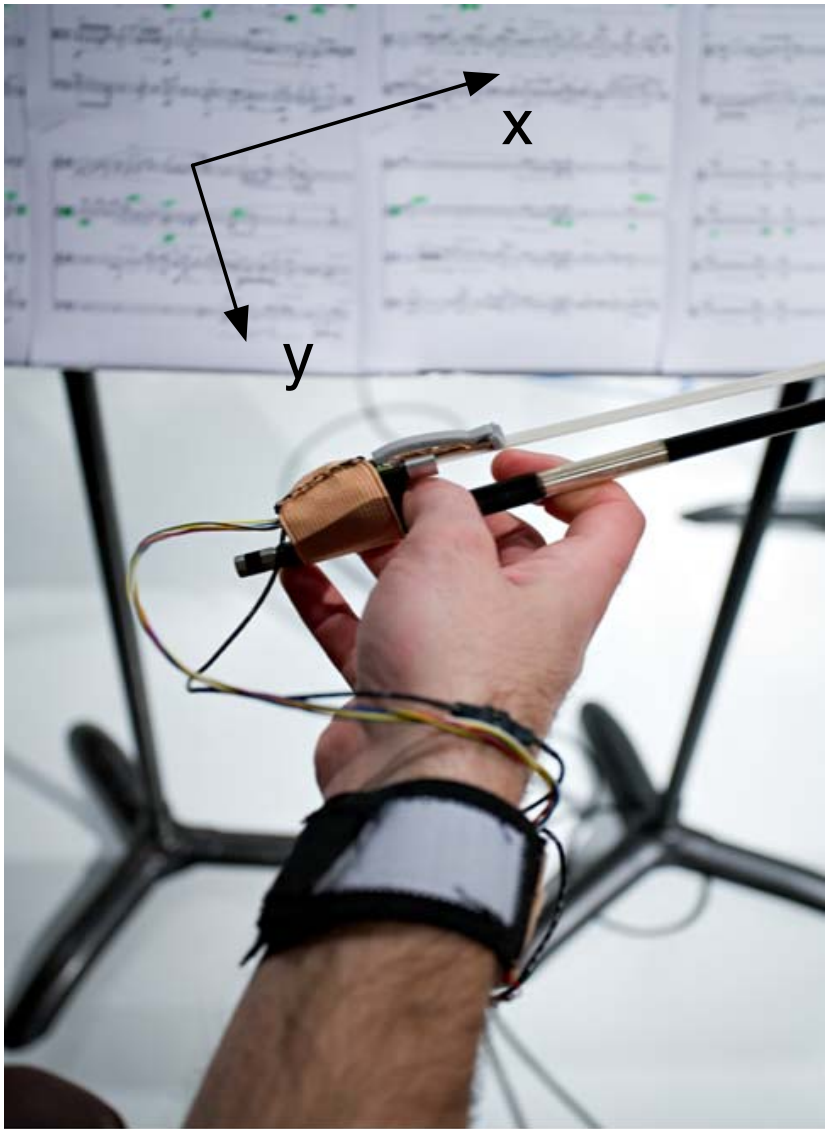

Figure 1: Pictures of the violin bow equipped with sensors. The part over the bow hair is the bow-force sensor (which has been shortened in the latest version). The module attached to the wrist is the radio emitter transmitting the data to the computer (photos by Klenefenn)

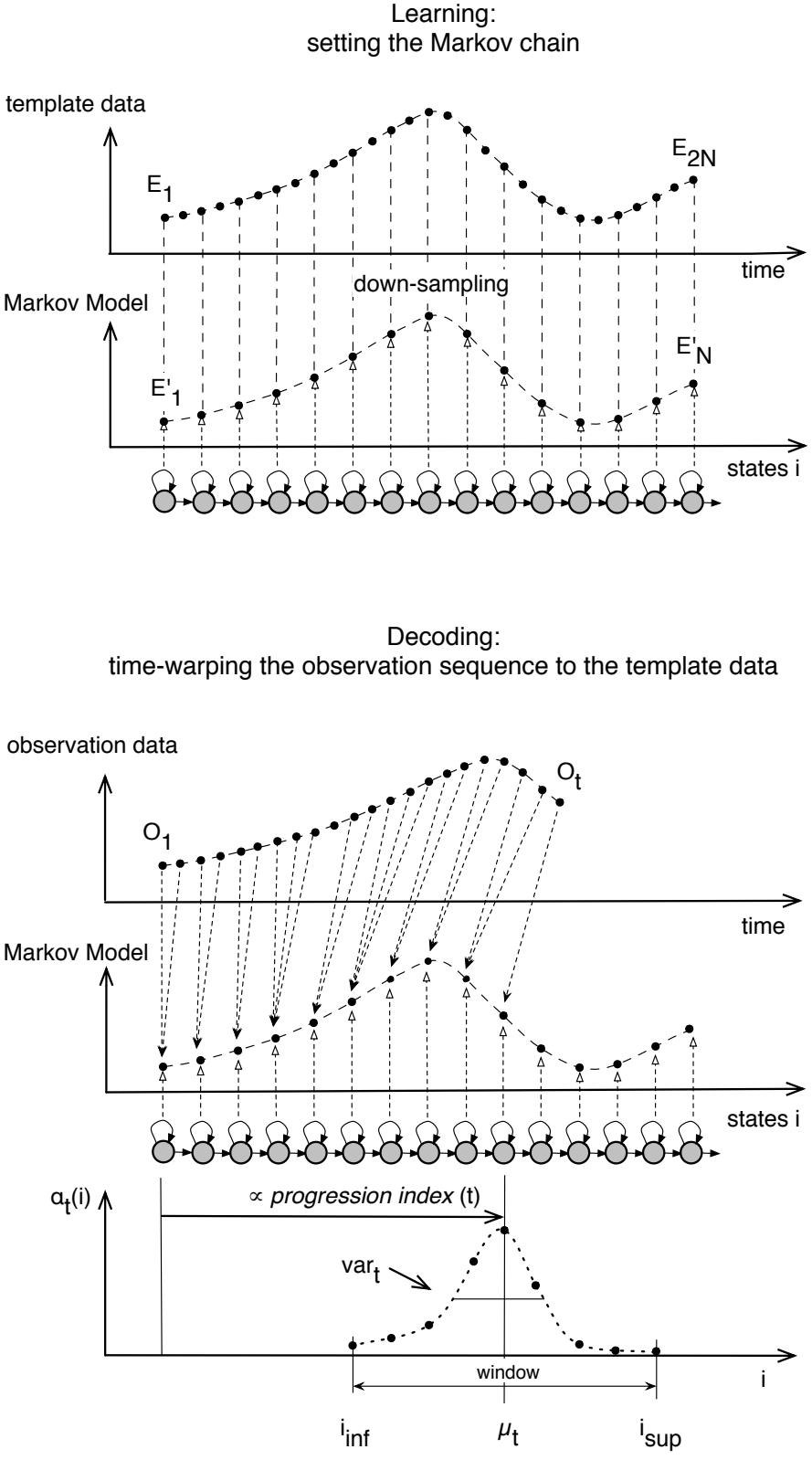

Figure 2: Top: graphical illustration of the Markov model built from a recorded example (learning procedure). Bottom: graphical illustration of the time-warping operated during the decoding procedure, and the associ-

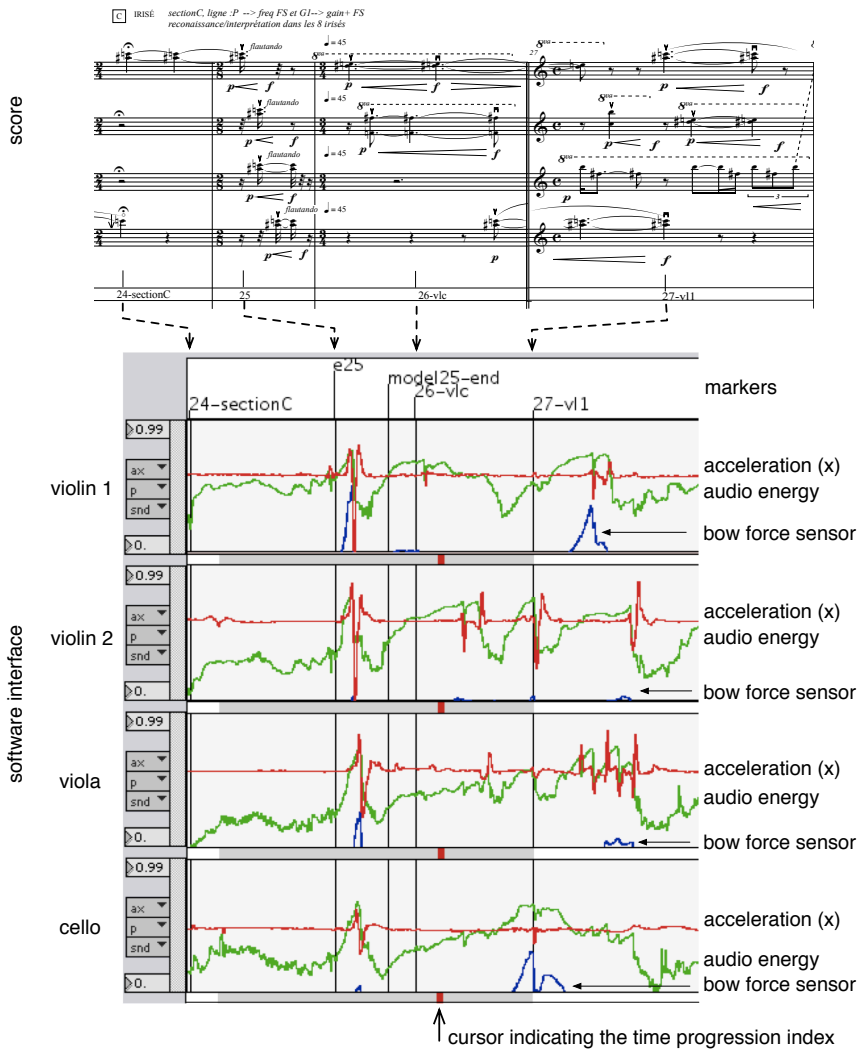

Figure 3: Example showing the relationship between the score and the data annotated with markers, as displayed on the computer interface

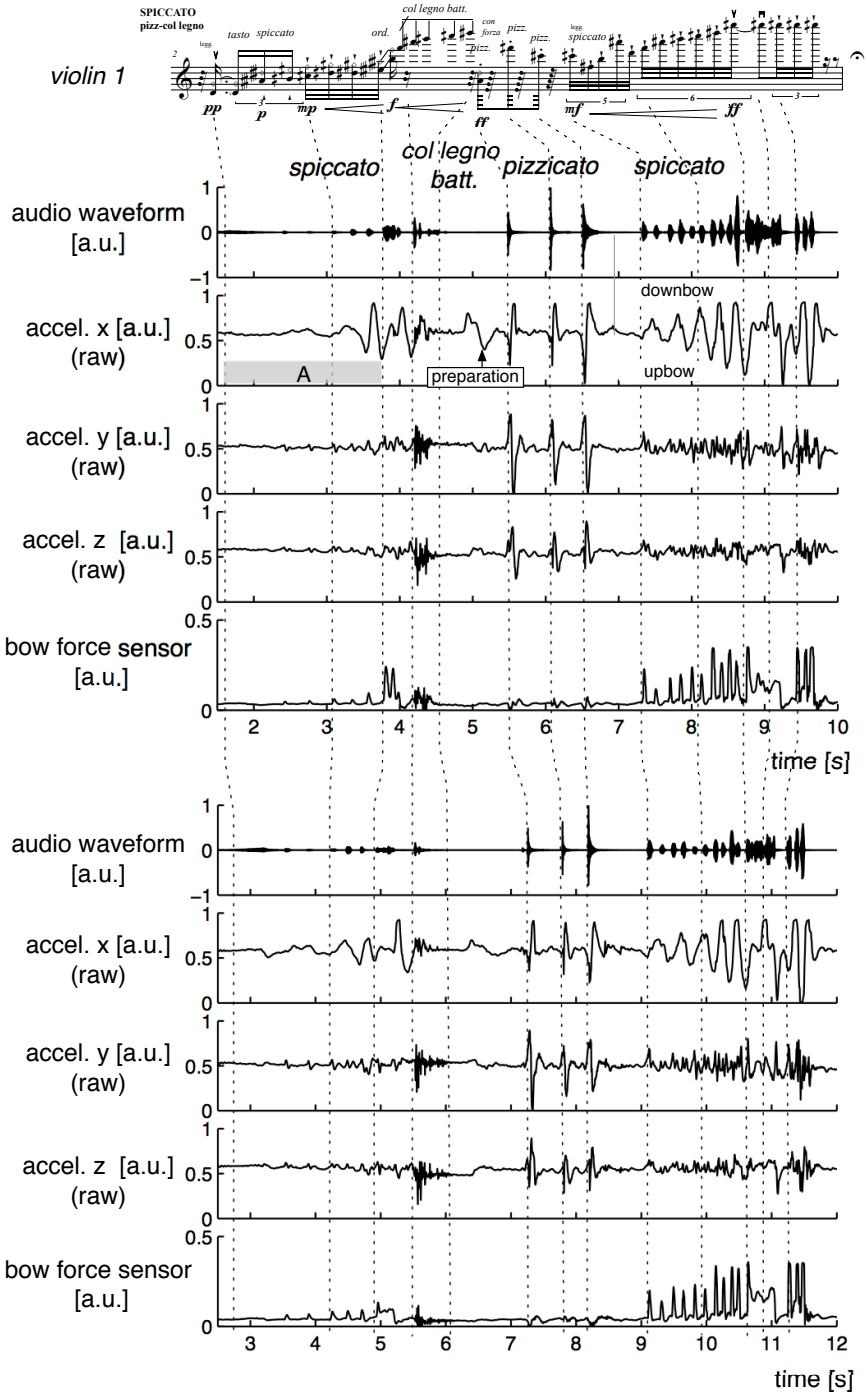

Figure 4: Score, audio waveform and sensor data for the prototypic phrase called *spiccato*. Two versions were successively recorded by the first violin. The gyroscope data were removed from the figure for clarity.

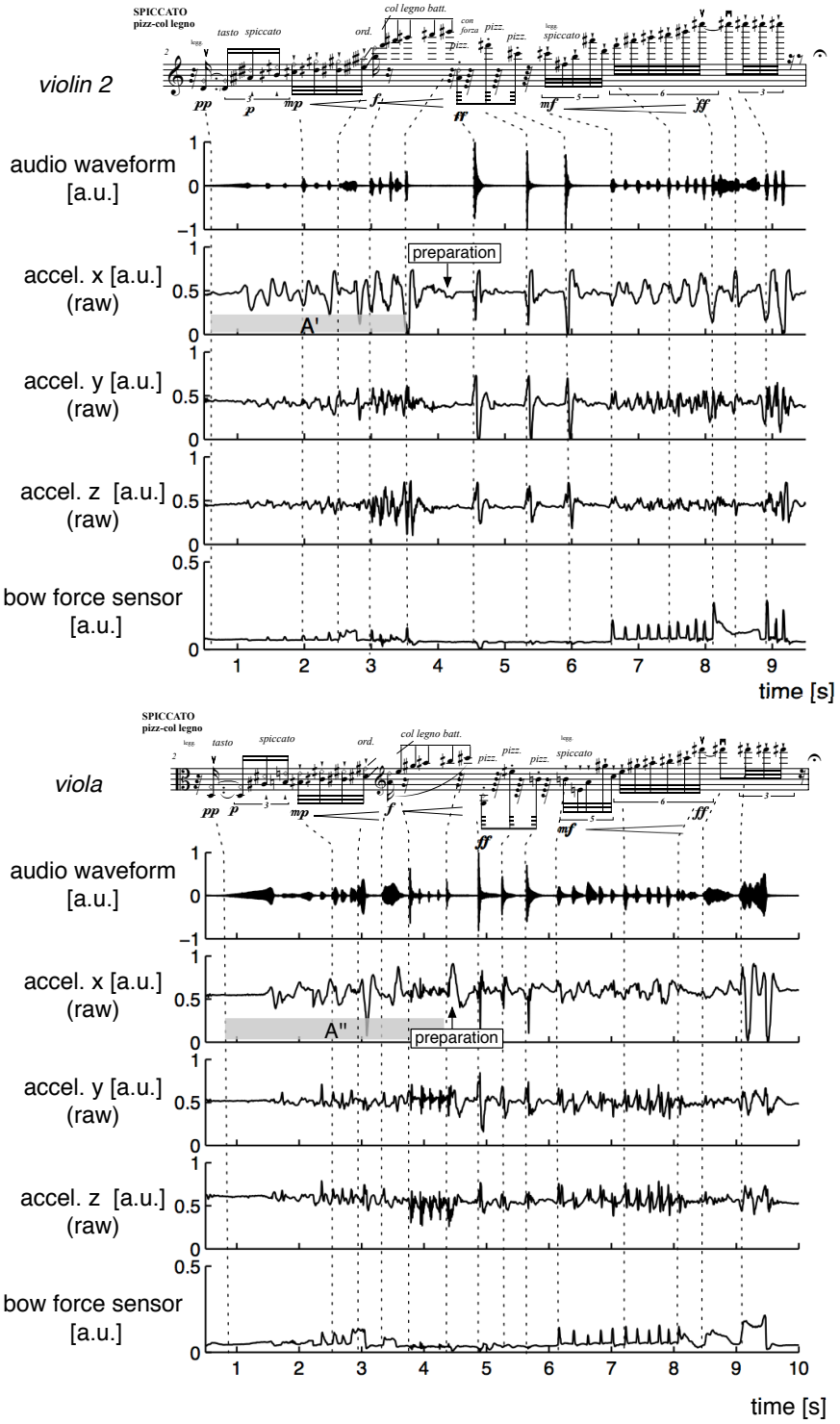

Figure 5: Score, audio waveform and sensor data for the prototypic phrase called *spiccato*. Top version was recorded by the second violin. Bottom version was recorded by the viola (transposed score). The gyroscope data were removed from the figure for clarity.

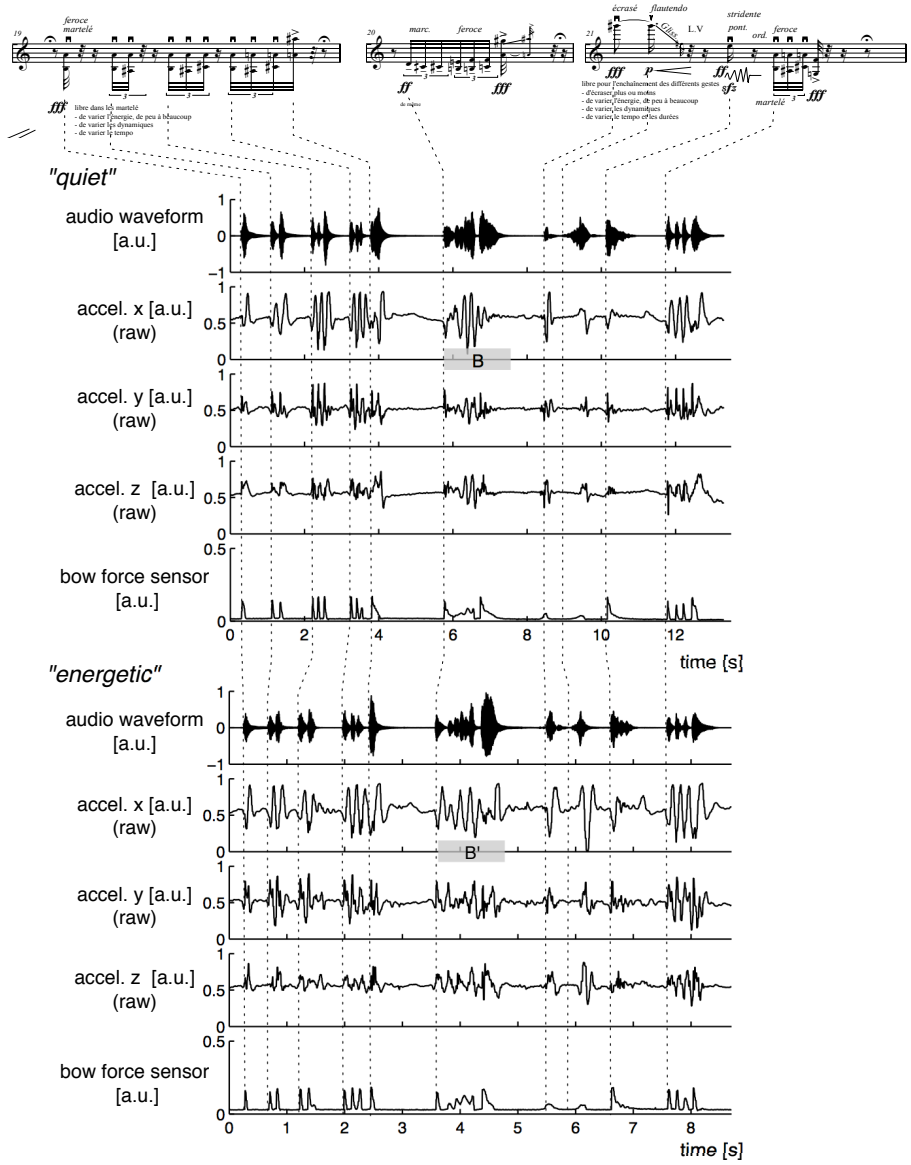

Figure 6: Score, audio waveform and sensor data for two different interpretation of short phrases, recorded by the first violin. The indication were "quiet" (top) and "energetic" (bottom). The gyroscope data were removed from the figure for clarity.

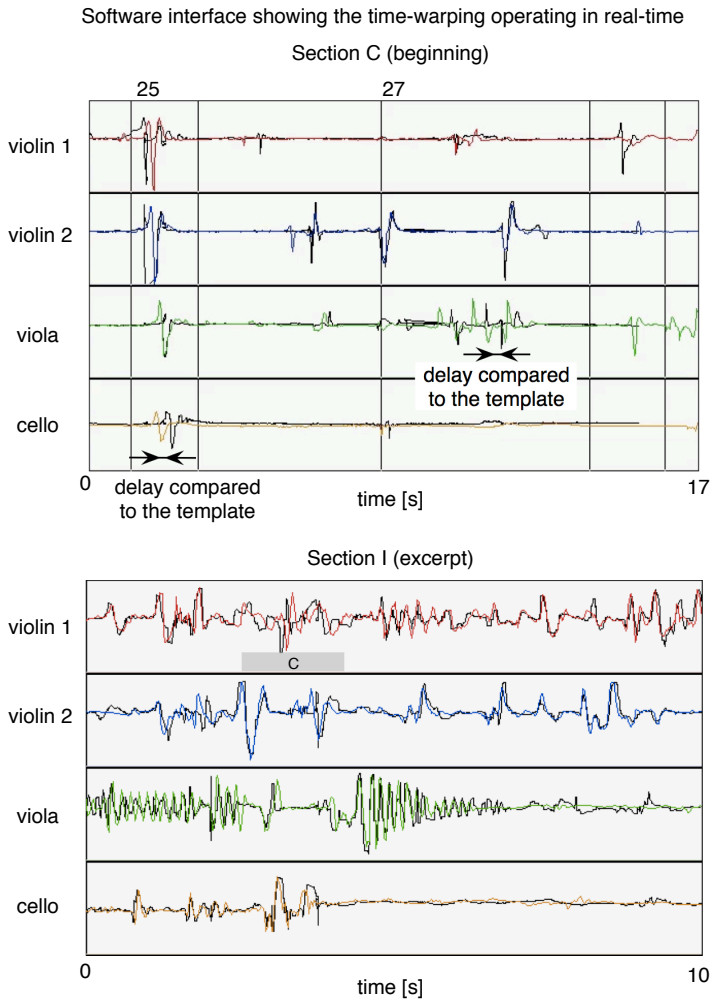

Figure 7: Examples of the time-warping operated in real-time as shown on the computer interface. The black lines correspond to the live data that is time-warped to match the recorded data (color/grey lines). Top figure is the section C (also shown in Figure 3). Bottom figure is an excerpt of section I

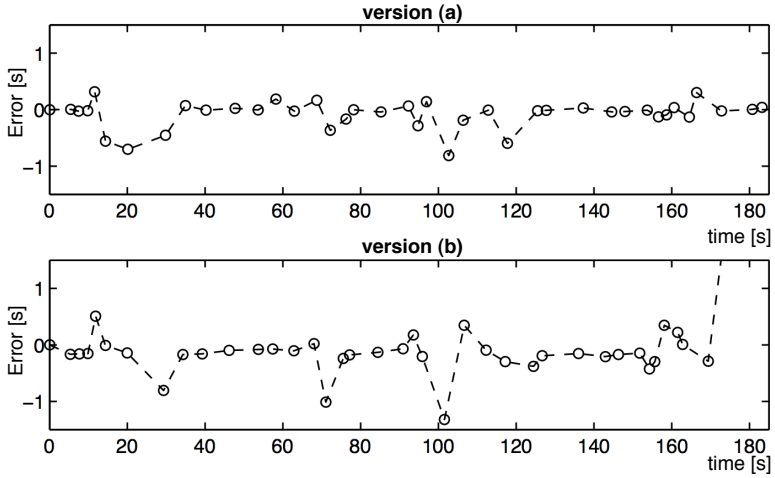

Figure 8: Error of the progression index with two different recordings of section C. (σ=0.5 and all 28 input parameters were used). The dot correspond to error values associated to markers

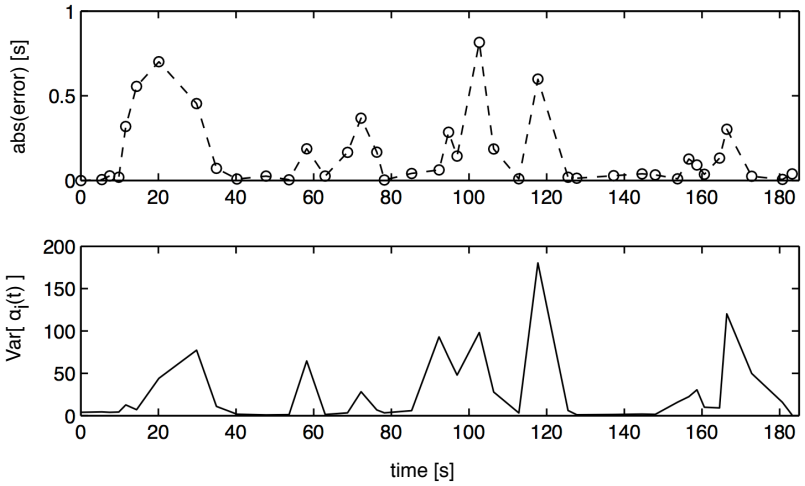

Figure 9: Absolute error of the *time progression index* plotted along with the variance of distribution $\alpha_t(i)$. The data are the same as reported in Figure 8 (top) (σ=0.5 and all 28 input parameters were used.)

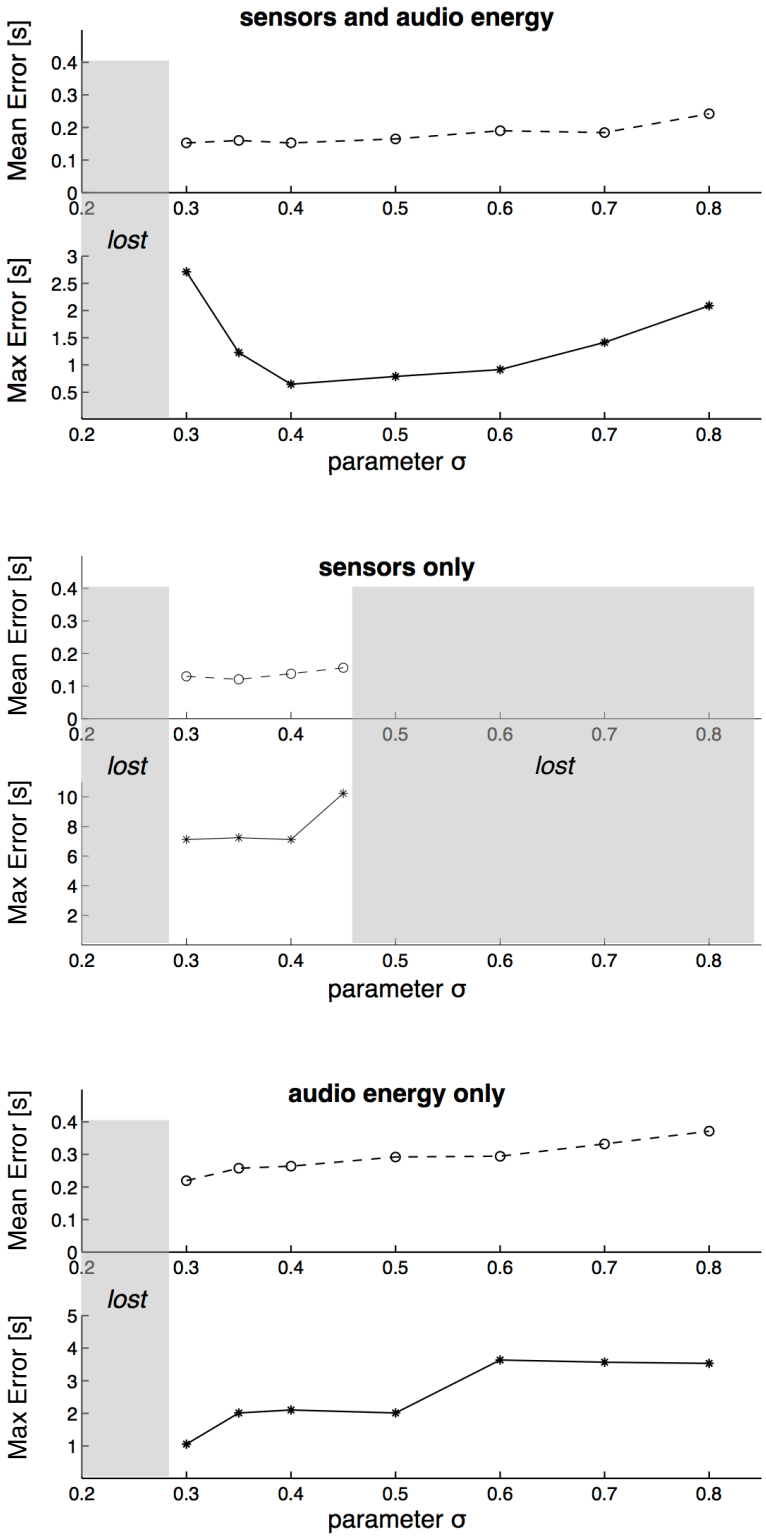

Figure 10: Error of the time progression index as a function of the σ value in the decoding algorithm (see Equation 3), for three different sets of incoming data: sensors and audio energy, sensors only and audio energy only