Learning and Extraction of Violin Instrumental Controls from Audio Signal

Alfonso Perez Carrillo
IDMIL and CIRMMT
Schulich School of Music, McGill University.
555 Sherbrooke Street West, Montreal, Canada
alfonso.perezcarrillo@mail.mcgill.ca

Marcelo M. Wanderley
IDMIL and CIRMMT
Schulich School of Music, McGill University.
555 Sherbrooke Street West, Montreal, Canada
marcelo.wanderley@mail.mcgill.ca

ABSTRACT

Acquisition of instrumental gestures in musical performances is an important task used in different fields ranging from acoustics and sound synthesis to motor learning or electroacoustic performances. The most common approach for acquiring gestures is by means of a sensing system. The direct measurement involves the use of usually expensive sensors with some degree of intrusivity and generally entails complex setups. Indirect acquisition is based on the processing of the audio signal and it is usually informed on acoustical or physical properties of the sound or sound production mechanism. In this paper we present an indirect acquisition method of violin controls from an audio signal based on learning of empirical data that is previously collected with a highly accurate sensing system. The learning consists of training of statistical models with a database of multimodal data from violin performances. The database includes audio spectral features and instrumental controls (bow tilt, bow force, bow velocity, bowing distance to the bridge and played string) and is designed to sample most part of the violin performance control space. We expect that once the indirect acquisition system is trained, no sensors should be required, so the indirect acquisition becomes a low-cost and non-intrusive acquisition method.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*parameter learning*; I.6.1 [**Simulation and modelling**]: Model Validation and Analisys; J.5 [**Arts and Humanities**]: Music

General Terms

Experimentation, Algorithms

Keywords

information retrieval, musical gesture, indirect acquisition, violin instrumental controls

MIRUM'12, November 2, 2012, Nara, Japan.

1. INTRODUCTION

During a musical performance a performer transforms a musical idea or score into a sequence of instrumental gestures that control the instrument, which in turn, produces the sound. In this manner, the musical idea is transformed into different representation domains: the musical score, the gesture and the sound domain (see Figure 1). In the field of music computing, the translation from score to gestures or sound is known as synthesis (gesture and sound synthesis), and the other way round is called (music) information retrieval (IR). In this work, we are predicting instrumental controls (i.e. gesture domain) from an audio recording and we denominate it 'performance information retrieval'.

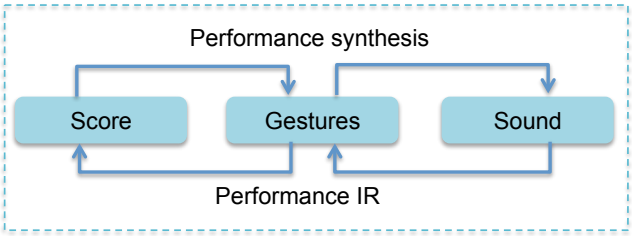

Figure 1: Typical representation domains for a musical performance: score, gestures and sound. Prediction of sound (or gestures) from a musical score is known as sound (or gesture) synthesis and extraction of information from an audio signal is commonly referred to as music information retrieval. In this work, we are predicting instrumental controls (i.e. gesture domain) from an audio recording and we denominate it 'performance information retrieval'.

The acquisition of musical gestures and particularly of instrumental gestures, from a musical performance is a field of increasing interest with applications in performance transcription [23], performance modelling [11], mapping strategies between gestures and sound [22] or sound synthesis [14] among others. The direct way for the acquisition of such gestures is by measurement of physical variables with sensors on the instrument or on the performer and we have seen an important development of technologies related to sensors and gestural interfaces in the recent years. The direct measurement involves the use of usually expensive sensors with some degree of intrusivity and generally entails complex setups.

An alternative way is by indirect acquisition from analy-

sis of an audio signal as it is our case. Indirect acquisition has a handful of advantages such as the simplicity and non-expensiveness of the acquisition, the possibility of extracting features from old audio recordings and it is a not intrusive technique. The main difficulty is to be able to build robust detection algorithms to be as accurate as the sensors. There are different approaches for indirect acquisition from audio signal, all of them having in common their need for measurements and observation previous to the acquisition. A typical approach is by informing the audio analysis with physic-acoustical properties of the instruments (physically based), as for example a technique for estimating the reed pulse from the pressure signal recorded at the bell of a clarinet [17] or the use of digital waveguides to exploit the asymmetry of the guitar body's admittance in order to provide an estimate of the plucking angle of release [16]. Other common approach is based on previous knowledge about the perceptual influence of instrumental gestures on the sound (perceptually based). For instance, overblown flute fingerings can be detected from the residual noise spectrum [20]. In the case of the violin, there are reports for the extraction of the notes, the string it was played on, whether the instrument was bowed or plucked and the location of the bowing or plucking point [9]. For the bass guitar, the plucked string [1] and the plucking style [2] can be automatically extracted. Also mixed approaches combining physical and perceptual knowledge there can be found, as in [18], where authors estimate the plucking position in guitar playing. Finally, there can be found methods based on data mining of empirical data without using any previous knowledge. For instance, blowing pressure in a recorder can be estimated from spectral descriptors (pitch and MFCCs)[21] by learning data with feed-forward neural networks and also flute fingerings corresponding to harmonic notes can be extracted [8].

In this paper we present a method to extract gestural information from an audio signal. More specifically, we are extracting violin instrumental controls, namely, string being played, finger position in the string being played, bowing force, bowing velocity, bowing distance to the bridge and bow tilt. The research is based on the mapping study from gestural to sound parameters by Perez [14]. Analogously, here we propose to do the mapping in the opposite direction, that is, from sound features to control parameters. In fact, it should be easier to predict a few control parameters from a large set of spectral parameters than the other way round.

The main contribution of this work is that we are predicting continuous instrumental controls, including bowing controls, which are directly related to the sound production. In the literature we find methods able to acquire discrete basic control parameters such as pitch, fingering, plucking position, string being played [9], bowing technique [3] or vibrato features [10].

In this work, we are using our own developed sensing system to collect control and sound signals [12]. A clear advantage of our approach is the use of sensors for the learning as in [19] as we do not rely on previous knowledge or measurements from others and large amounts of data can be collected. Once the acquisition is learned we do not need the sensors anymore.

The training is based on machine learning techniques, among them Tree-based algorithms and Multilayer Perceptrons. We used the WEKA [7] framework for the learning process.

2. LEARNING DATABASE

The learning process it based on the acquisition of data based on sensors. A set of musical scores was designed with the aim of sampling most part of the violin controls space (i.e. combinations of bowing force, velocity, and bow distance to the bridge, at different strings). Performances of the scores are recorded with a measuring system able to capture gestural and audio data. This data is aligned and segmented into notes and a set of spectral and gestural descriptors is extracted to train the models. The dataset consists of around $9x10^5$ analyzed temporal frames.

Audio is recorded by means of a commercial pickup consisting of a transducer built into the bridge (Yamaha VNP1[1]) instead of a microphone, that is, we are analyzing the signal captured by that specific pickup. The use of a pickup provides some advantages: 1) it is more convenient for the data collection as we do not have to be concerned with microphone or room effects (position, direction, orientation, reverberation, etc.); and 2) we obtain a *clean* signal that is close to the string vibration, minimizing the violin body resonances, which allows for more accurate prediction.

Motion data was collected by means of a commercial 3D tracking system [2] that consists of an electro-magnetic field source and a set of small sensors. It tracks sensor position and orientation inside the magnetic field, with the coordinate axes being given by the source. The sample rate is 240 Hz and accuracy around $8 * 10^{-4}$ m and 0.15 deg. Bowing force was measured with strain gages fixed on a metallic surface under the hair ribbon in the frog of the bow and capture the strain of the hairs. More details related to the setup and calibration are detailed in [5, 6, 12].

3. FEATURE EXTRACTION

3.1 Audio Features

In the first stage of the analysis of the data, basic sound parameters are extracted from the acoustic signal, through frequency-domain analysis. The sample rate of the recordings is 44100 Hz and the analysis window is a Blackman-Harris window of size 2048 samples. The data acquisition rate is determined by the Polhemus system (240 Hz). Each instance in the dataset is represented by a vector containing all the spectral and control features in a temporal frame.

Two sets of spectral features are being computed and compared. Both sets contain additionally the fundamental frequency in Hz (*pitch*), which is estimated from the signal. The first set is composed of low-level parameters consisting of harmonic and residual energy in frequency bands computed as follows: the audio signal is separated into harmonic and residual components in the frequency domain, then a spectral envelope is estimated for both component and finally, the energy of the envelopes is computed in 40 overlapping frequency bands with centers following a logarithmic scale. The selection of the bands is inspired by perceptual models such as the Mel or Bark scale. More details are found in [14].

The second set of descriptors is defined in order to take advantage of the existing knowledge on the effect of the variation of the instrumental gesture parameters on timbre [14, 15, 4]. This set of parameters include widely used

[1] http://www.yamaha.co.jp/english/product/strings/v_pickup/index.html
[2] http://www.polhemus.com/?page=Motion_Liberty

spectral features [13], among them, spectral centroid, high-frequency-content (*hfc*), zerocrossing rate, kurtosis, skewness and spread of bark bands, spectral energy, spectral energy in four bands (high, low, middle-high, middle-low), spectral flatness, spectral rolloff, spectral strongpeak, zero-crossing rate, inharmonicity, odd-to-even harmonic energy ratio and tristimulus.

It is known [14, 15, 4] that an increase of bowing force boosts high frequency, so it is affects the spectral decay, which is correlated to the features spectral flatness, spectral rolloff, spectral centroid, and the relative distribution of energy in bands. An increment of bowing velocity increases energy independently of frequency, so it is strongly correlated with the spectral energy. The main role of the bowing distance to the bridge is to determine the range of bowing force and velocity. Regarding the bow tilt, it is very related to dynamics and also is seems to be highly correlated to the bowing transversal position [3]. The features tristimulus, odd-to-even-harmonics, inharmonicity and zero-crossing rate seem to have little or no correlation with the gestures and were discarded in many of the trainings.

3.2 Instrumental controls

The instrumental controls that we expect to extract from the audio, include the parameters that more influence the sound produced as found in the classic literature about bowed strings [4, 15] (i.e. bow force, bow velocity and bowing distance to the bridge) as well as the string being played, the finger position (on the playing string) and the bow tilt. The bowing parameter's computation is based on an initial estimation of the string being played as explained in [12]:

- *velocity* (v_B). This parameter is the bowing speed in m/s in absolute values. It is computed as the absolute value of the smoothed derivative of the bow position (x_B), $v_B = |\frac{d}{dt}x_B|$. Bowing speed is positive when playing downbow and negative when playing upbow, however bowing direction should not affect the sound [4]. In this study we use the absolute values of the bowing speed to make the learning independent of the bowing direction. This feature is highly correlated with sound energy [14].

- *force* (f_B). This parameter is the force in Newtons exerted by the bow on the string. It is obtained with strain gages as explained in [5]. The main effect of the force on the sound is that is boosts high frequencies, so it is highly correlated with the energy, the spectral centroid and the spectral decay [14].

- *tilt* (t_B). The tilt is the angle in degrees between the plane defined by the bow hairs and the string being bowed. It is a meassure very correlated to the width of bow-hairs in contact with the string and in performance terms it is very correlated to dynamics. In our database it is also highly correlated on the bowing position [3] [14] . This width is very difficult to measure, but can be indirectly estimated by measuring the bow tilt.

[3] The training database was recorded by a single violinist, so the correlation between the tilt and the bow position could be a particularity of that performer.

- *fingerpos* (x_F). Finger position is the length of the playing string (in cm) between the bridge and the finger stopping it. It is obtained as

$$x_F = \frac{L_s f_s}{f_0}, \quad (1)$$

where L_s is the string length and f_s is the fundamental frequency in Hz of the open string being played.

- *beta* (β). This parameter represents the bow-bridge distance relative to string-length in vibration (stopping a string with the finger makes it shorter). It is calculated as

$$\beta = x_{BB}/(L_s - x_F), \quad (2)$$

where L_s is the length of the string, x_F is the position of the finger in the playing string and x_{BB} is the bow-bridge distance, that is, the length (cm) of the segment of the playing string between the bridge and the bow hairs.

4. DATA MINING

Two different training datasets were compared for the learning of the instrumental controls with different types of classifiers. The first dataset is based on the low-level set of descriptors (40 harmonic and 40 residual energy bands plus the pitch). The second approach makes use of the perceptual descriptors set plus the pitch. The training is based on widely known machine learning techniques, obtaining higher prediction results with pruned tree-based algorithms (J48, Random Forest) and Multilayer Perceptrons (*MP*) with one hidden layer containing half the neurons as the number of inputs. We used the WEKA [7] framework for the learning process. The scheme for the prediction of the instrumental controls is shown in Figure 2 and is as follows (numbering refers to steps in the figure): There are 17 models, the first one 'String prediction' (step 1) predicts the playing string given the spectral features as input. After the string prediction a hysteresis function is applied in order to avoid rapid fluctuations in the string prediction caused by an error (i.e. a change in the *string* that last only a few consecutive frames and then goes back to the previous *string* value indicates an error in the prediction). This smoothness is making the temporal evolution of the prediction stable and corrects most of the errors. This is shown in Figure 3. In blue is the prediction of the model, in black with a thick-dashed line is the corrected prediction after the hysteresis and in red we can observe the actual value of the *string*. The function is implemented as an averaging filter of size 25 samples (it computes the closest integer of the average). Once the string is predicted, finger position is directly computed from the pitch (step 2) as described in equation 1. The blocks in the figure labeled String1 to String4 (steps 3,4,5,6), represent each four models (so, in total the 16 remaining models), which predict the four remaining instrumental controls (*force*, *velocity*, *beta* and *tilt*) given a specific string.

5. RESULTS

Here we present the numerical results for the prediction of the models, comparing the two training datasets and different classifiers. A summary of the results for the best performing classifier (*MP*) is shown in table 1. Models trained

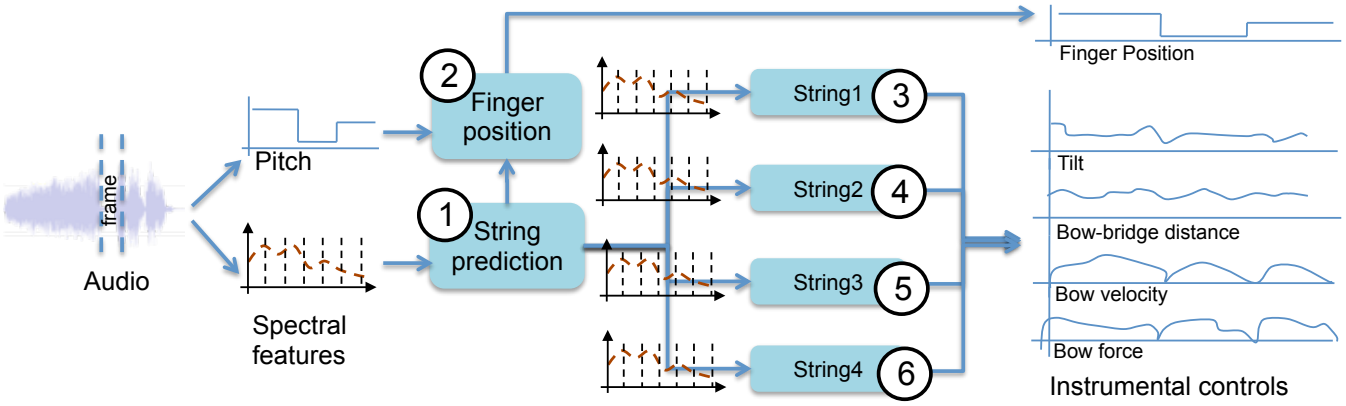

Figure 2: Schema for the prediction of the instrumental controls. Audio is analyzed in a frame-by-frame basis and spectral features computed. Then, the bowed string is predicted (1) and corrected by means of a hysteresis function from which the finger position is computed (2). Finally, *force*, *velocity*, *beta* and *tilt* are predicted (3, 4, 5 and 6) using as inputs the spectral features and the string predicted in step (1).

with the low-level descriptors show higher prediction rates than the perceptually informed. For the correct interpretation of the results some details have to be taken into account: 1) the results correspond to isolated models, so for an error estimation of the whole process of prediction (Figure 2), we have to take into account the error propagation (i.e. if there is an error in the string extraction, then the other 4 predicted controls in the following steps will have an error too); 2) the high prediction rate is due to the high percentage of frames from note sustains, where prediction works better; 3) the models are static in the sense that they do not take into account time evolution of the parameters, except for the hysteresis function applied to the string prediction; and 4) the controls are predicted independently, although it is known that there is a correlation among them [15].

Prediction accuracy for the string is given in percentage of correctly classified instances (temporal frames) and for the other controls we indicate the correlation coefficient. Models for *velocity*, *force*, *beta* and *tilt* correspond to a specific string, so the indicated correlation values correspond to the average among the four strings. The evaluation is computed as an average over 10-folds of the training data.

Applying an attribute selection before the training, e.g. *Principal Component Analysis*, is a convenient way to make models simpler (e.g. from 40 energy bands we obtain a set of 5 eigenvectors, covering 95% of the data variance). However, prediction result rates decrease.

An graphical example of the instrumental control prediction for a performance of the *preludium* (two first phrases) of the *Czardas* by Monti is shown in Figures 3 and 4. Finger position is not shown as its prediction is exactly as the string prediction if we assume that pitch computation is always correct.

- *string*. The highest string prediction results are obtained with the low-level dataset descriptors (including the pitch) by means of a *MP* with a 98% of correctly classified instances. If only harmonic bands are used (and pitch) we obtain a rate of 97% and if only residual bands (and pitch) we achieve the 97.74%. With a Random Forest algorithm we obtain 97.34% and with a J48-Tree 96.1%. Using the perceptual features training dataset (and pitch) with a *MP* we obtain a 90%. The most relevant perceptual features for the string prediction are the spectral centroid, the spectral flatness and the *hfc*, which can correctly classify a 89% of the instances.

- *fingerpos*. Finger position on the played string is obtained as described in equation 1, so prediction results are the same as for the string, assuming that pitch is always correct.

- *velocity*. Prediction of the *velocity* with the low-level training dataset (harmonic and residual bands plus pitch for a specific string) and a *MP* classifier achieves an average (among the 4 strings) correlation coefficient of 95%. Using only harmonic bands and pitch we can get a 92% and with only the residual bands a 90%. If we predict the signed velocity (sing indicating the direction of the bow) the prediction rate decreases until a 82%. By using perceptual features (spectral centroid, spectral flatness, *hfc* and spectral energy) as the training dataset with a *MP* we can reach the 87%.

- *force*. Prediction with the low-level training set and a *MP* classifier we can achieve an average correlation coefficient of 93.5 %. If we only train with the harmonic bands (and pitch) only 83% is reached and with only the residual bands a 84%. If the training is done with the perceptual attributes, we obtain only a 80%.

- *beta*. With the low-level training set and a *MP* classifier we obtain a correlation coefficient of 97.8%. With only harmonic bands 88% and only taking into account the residual 96%. By training with perceptual parameters we decrease to a 88%.

- *tilt*. Based on the low-level training set and a *MP* classifier we obtain a correlation coefficient of 89%. By using only the harmonic bands the correlation is the 73% and with only the residual bands a 65%. Tilt prediction can be reinforced by including in the training dataset the bow position, obtaining until a 95%. Bow position can be computed as the integral of the *velocity*

plus a constant representing the position at time zero (the start) but it is not yet implemented. By training with perceptual parameters we obtain around 50% and if we reinforce with the bow position, we can achieve a 85%

Table 1: Prediction rates for the models isolated with a *MP* classifier. String rate is in percentage of correctly classified instances and for the rest of controls the correlation coefficient is indicated. Finger position prediction is equal to the string assuming that pitch is always correct. Tilt prediction rate is for the model reinforced with bowing position in the training.

Control	Low-level	Perceptual
string, fingerpos	98%	89%
velocity	95%	87%
force	93.5 %	80%
beta	97.8%	88%
tilt	95%	85%

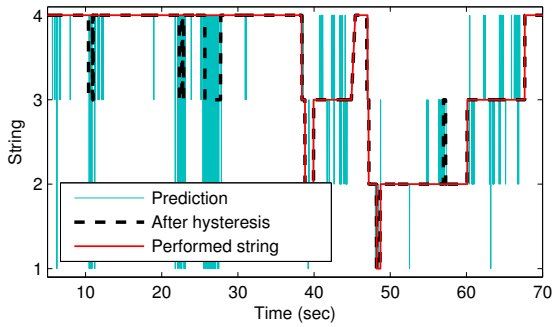

Figure 3: *string* extraction (blue) of a recording of the *preludium* of the *Czardas* by Monti, correction after the hysteresis function (thick-black-dashed line) and actual string (red). The hysteresis function helps having a more stable and accurate string prediction.

6. CONCLUSIONS

We presented a method to extract violin instrumental controls from an audio signal. The extraction is based on learning by means of direct measurement with sensors. The extracted descriptors are bowing velocity, bowing force, bow tilt, bowing distance to the bridge, string being played and finger position in the played string. The training is carried out by commonly used machine learning techniques. Two types of spectral training descriptors are compared, low-level spectral descriptors and a set of higher level (perceptual) descriptors on which the instrumental controls have a known direct effect. The combination of low-level descriptors with pitch and multilayer perceptron as the classifier seems to give the better prediction results.

Sound analysis is based on a signal captured with a violin pickup, which is a signal close to string vibration and is almost not affected by the resonances of the violin body. The main application of this work is for the indirect acquisition of violin controls of an already trained violin without the need for the sensors. Applications of this work are many, for instance, as a tool for score transcription, during artistic performances or for the acquisition of gestures in special environments where the possibility of having the sensors is not possible (e.g. cost, sensor interferences with the environment, availability). Automatic acquisition from any violin recording would be more difficult as the specific violin and recording would have particular spectral properties so there should be necessary a calibration.

Some steps for the future are 1) study a similar procedure with acoustic recordings, 2) explore time-aware algorithms such as feedback networks, Markov models or dynamic programming that would bring out more reliable and stable models and 3) take into account the interdependence of the predicted variables.

7. ACKNOWLEDGMENTS

The authors would like to thank Esteban Maestre, Merlijn Blaauw, Enric Guaus and Jordi Bonada for the data aquisition stage. This research has been funded with a fellowship *Beatriu de Pinos* granted by the catalan research agency (AGAUR).

8. REFERENCES

[1] J. Abeßer. Automatic string detection for bass guitar and electric guitar. In *Proc. 9th International Symposium on Computer Music Modelling and Retrieval*, London, June 2012.

[2] J. Abeßer, H. Lukashevich, and G. Schuller. Feature-based extraction of plucking and expression styles of the electric bass guitar. In *Proc. ICASSP*, 2010.

[3] I. Barbancho, C. de la Bandera, A. Barbancho, and L. Tardon. Transcription and expressiveness detection system for violin music. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 189 –192, april 2009.

[4] L. Cremer. *Physics of the Violin*. The MIT Press, November 1984.

[5] E. Guaus, J. Bonada, E. Maestre, A. Perez, and M. Blaauw. Calibration method to measure accurate bow force for real violin performances. In *Int. Computer Music Conf.*, pages 251–254, Montreal, Canada, 16/08/2009 2009.

[6] E. Guaus, J. Bonada, A. Perez, E. Maestre, and M. Blaauw. Measuring the bow pressing force in a real violin performance. In *Int. Symposium on Musical Acoust.*, Barcelona, Spain, 2007.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[8] C. Kereliuk, B. Scherrer, V. Verfaille, P. Depalle, and M. M. Wanderley. Indirect acquisition of fingerings of harmonics notes on the flute. In *33rdInt. Computer Music Conf.*, volume 1, pages 263–6, Copenhagen, Denmark, August 2007.

[9] A. Krishnaswamy and J. O. Smith. Inferring control inputs to an acoustic violin from audio spectra. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2*, ICME '03, pages 733–736, Washington, DC, USA, 2003. IEEE Computer Society.

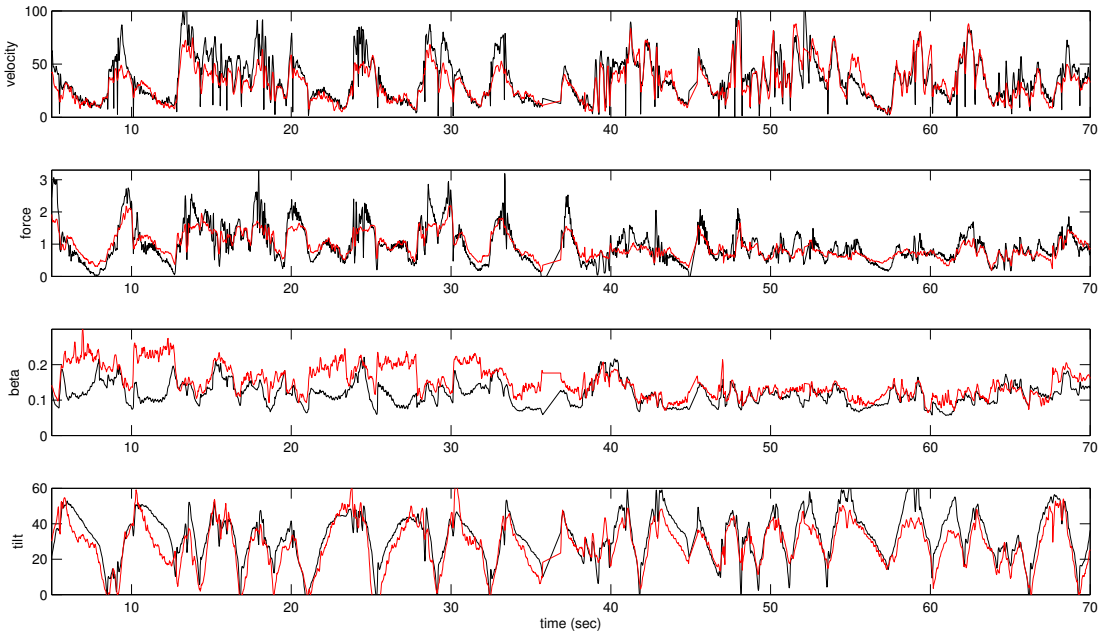

Figure 4: ***velocity*, *force*, *beta* and *tilt* prediction (black) of a recording of the *preludium* of the *Czardas* by Monti and actual parameters(red). Note that tilt prediction is reinforced by using the strored bow position as input.**

[10] A. Loscos. Low level descriptors for automatic violin transcription. In *In ISMIR*, pages 164–167, 2006.

[11] E. Maestre, M. Blaauw, J. Bonada, E. Guaus, and A. Pérez. Statistical modeling of bowing control applied to sound synthesis. *IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Virtual Analog Audio Effects and Musical Instruments.*, 2010.

[12] E. Maestre, J. Bonada, M. Blaauw, A. Pérez, and E. Guaus. Acquisition of violin instrumental gestures using a commercial EMF device. In *Proc.Int. Computer Music Conf.*, Copenhagen, Denmark, 2007.

[13] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, Paris, France, 2004.

[14] A. Perez Carrillo, J. Bonada, E. Maestre, E. Guaus, and M. Blaauw. Performance control driven violin timbre model based on neural networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):1007 –1021, march 2012.

[15] J. Schelleng. The bowed string and the player. *The Journal of the Acoustical Society of America*, 53(1):26–41, 1973.

[16] B. Scherrer and P. Depalle. Extracting the angle of release from guitar tones: preliminary results. In *2012*, Nantes, France, 2012.

[17] T. Smyth and J. S. Abel. Toward an estimation of the clarinet reed pulse from instrument performance. *The Journal of the Acoustical Society of America*, 131(6):4799–4810, 2012.

[18] C. Traube, P. Depalle, and M. Wanderley. Indirect acquisition of instrumental gesture based on signal, physical and perceptual information. In *Proceedings of the 2003 conference on New interfaces for musical expression*, NIME '03, pages 42–47, Singapore, Singapore, 2003. National University of Singapore.

[19] G. Tzanetakis, A. Kapur, and A. Tindale. Learning indirect acquisition of instrumental gestures using direct sensors. In *8th IEEE Workshop on Multimedia Signal Processing*, pages 37 –40, oct. 2006.

[20] V. Verfaille, P. Depalle, and M. M. Wanderley. Detecting overblown flute fingerings from the residual noise spectrum. *The Journal of the Acoustical Society of America*, 127(1):534–541, 2010.

[21] L. Vinceslas, F. García, A. Pérez, and E. Maestre. Mapping blowing pressure and sound features in recorder playing. In *Proceedings of the International Conference on Digital Audio Effects*, 2011.

[22] M. M. Wanderley and P. Depalle. Gestural control of sound synthesis. In *Proceedings IEEE*, pages 632–644, 2004.

[23] B. Zhang and Y. Wang. Automatic music transcription using audio-visual fusion for violin practice in home environment. Technical Report TRA7/09, Shool of Computing, National University of Singapore., 2009.