GESTURE SAMPLING FOR INSTRUMENTAL SOUND SYNTHESIS: VIOLIN BOWING AS A CASE STUDY

Esteban Maestre, Alfonso Pérez, Rafael Ramírez

Music Technology Group
Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain
esteban.maestre@upf.edu

ABSTRACT

This paper presents a framework in which samples of bowing gesture parameters are retrieved and concatenated from a database of violin performances by attending to an annotated input score. Resulting bowing parameter signals are then used to synthesize sound by means of both a digital waveguide violin physical model, and an spectral-domain additive synthesizer.

1. INTRODUCTION

In the context of instrumental sound synthesis, generating realistic, natural-sounding performances from an annotated input score represents a difficult task, especially for excitation-continuous musical instruments [4]. On one hand, although synthesis techniques based on sound sampling provide in general higher fidelity, eventual discontinuities in generated sound limit the naturalness of the synthetic performance. On the other hand, synthesis methods based on physical or spectral models present a more continuous behavior, but the necessity of appropriate input controls keeps them from raising more success.

These two general problems reflect a clear limitation that performance modeling traditionally presented: the focus has been put on the sound and not on how the performer controls the instrument. In fact, the unavailability of reliable measurement techniques for acquiring those physical actions directly involved in the sound production process has derived into certain lack of attention to the important role of instrumental gestures in music performance.

When the acquisition becomes feasible, the possibility of sampling instrumental gesture parameters from real performance recordings arises as an attractive opportunity for driving physical or spectral sound models. In spite of the complexity of violin practice, the measurement and analysis of bowing parameters has received attention during the last years [9, 3, 4, 7, 8], but no attempt has been done to bring the classical ideas of sample-based synthesis to the domain of bowing gestures. Recently, an interesting work dealing with percussive gestures brought up the possibility of retrieving gesture signals for reconstructing a natural performance by using physical models [2].

As a natural continuation of previous works by the authors [6, 7, 5], this paper presents a case study on gesture sampling for instrumental sound synthesis. Samples of bowing parameter signals are retrieved from a multi-modal database of real violin performances by attending to an annotated input score. Synthetic bowing parameter signals are obtained by stretching and concatenating retrieved samples, and then used as input controls both to a violin physical model, and to a violin spectral model based on additive synthesis. An overview of the framework is depicted in Figure 1.

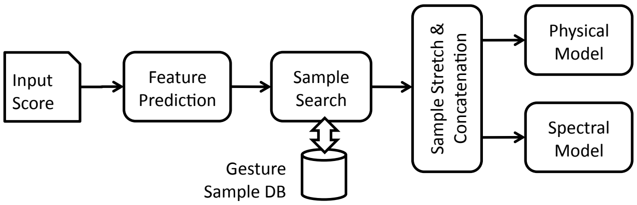

Figure 1. Schematic view of the proposed system.

The rest of the paper is structured as follows. Section 2 describes the construction of the gesture sample database. Then, the process of rendering (bowing) gestures by sample retrieval and concatenation is introduced in Section 3. An overview of sound synthesis is given in Section 4. The paper closes by pointing out some conclusions and possible improvements.

2. DATABASE CONSTRUCTION

The performance database used for this study is composed by annotated samples of gesture parameter signals. Each sample, corresponding to a note, contains the segmented streams of bowing parameters plus a number of annotations to be used during sample retrieval.

Recording scripts (including both exercises and short musical pieces) were designed to cover four different articulation types (détaché, legato, staccato, and saltato), three different dynamics, and varied note pitch and duration values in different performance contexts (attending to bow direction changes and rest segments).

2.1. Bowing data acquisition

Acquired bowing parameters are bow velocity, bow force, and β ratio. The β ratio is defined the proportion between the bow-bridge distance and the effective string length (distance from the finger to the bridge). Bowing data acquisition is performed by means of a commercial electromagnetic field sensing device, and using strain gauges for extracting bow force. See [4] for a detailed description.

Recorded performances are segmented at note level by means of a dynamic programming procedure (based on the Viterbi algorithm) as introduced in [4]. In a post-processing step, minor manual corrections are applied for ensuring the appropriate segmentation of around 10K notes.

2.2. Database annotation

Each note is first classified into one of several note groups by attending to score annotations. Secondly, four note features are attached to the note: the note duration, the string played, and the effective string length (obtained from the string played and the fundamental frequency), and the dynamics. Finally, the contour of the bowing gesture parameters of each note is characterized, and a feature vector is extracted. Next we provide an scheme containing the different annotations that accompany each i-th sample in the database:

– **Note group** G^i.

– **Note features**: duration D^i, string played s^i, effective string length l_e^i, and dynamics d^i.

– **Contour feature vector** $\mathbf{f}^i$.

2.2.1. *Note classification*

Attending to the score annotations, segmented notes are classified into different groups attending to both two intrinsic characteristics, and two contextual characteristics.

Regarding intrinsic characteristics, we considered the bowing technique (*détaché*, *legato*, *saltato*, or *saltato*), and the bow direction (*down* or *up*).

As contextual characteristics, we considered its position within a slurred note sequence, and its position within a sequence of consecutive notes (no rest segments in between). In both cases, a note can be labelled as *init*, *mid*, or *end*, if it respectively appears at the beginning, middle, or end of the sequence; or as *iso* for a single-note sequence. Each feasible combination of the four characteristics above leads to a note group. Detailed information on note classification can be found in [6, 4].

2.2.2. *Gesture feature extraction*

The contours of the gesture parameter signals segmented for each note in the database are automatically analyzed by attending to a predefined contour model based on sequences of cubic Bézier segments. For each note group, a specific contour model is defined by attending to the temporal signature of bow velocity, bow force and β ratio. Roughly, a contour model consists on the number of Bézier segments, and a series of slope constrains to be respected when applying the model. Out of this analysis, introduced in [6, 4], contours of gesture parameter signals are represented by concatenation of Bézier curve segments whose parameters are used for constructing a vector $\mathbf{p}^i$ for each i-th note in the database. Vectors corresponding to notes of each group present a different dimensionality, typically ranging from 25 to 40.

In a second step, principal component analysis is applied to the contour parameter vectors $\mathbf{p}$ of samples of each note group, obtaining a lower dimensionality feature vector $\mathbf{f}^i$ for each i-th sample in the database. Depending on the note group, a different dimensionality reduction is applied, with ratios typically ranging from 4 to 5 [5]. For each note group G_k, a normal distribution is estimated from the feature vectors $\mathbf{f}$ of its corresponding samples, having the resulting covariance matrix Σ_{G_k} annotated. Both the feature vectors and the covariance matrices are used during sample search, as it is described in Section 3.2.

3. CONCATENATIVE GESTURE RENDERING

In this section we give details on gesture sample selection and concatenation. Out of this process, bowing control parameter signals (bow velocity, bow force, and β ratio) are rendered from an annotated input score. Roughly, sample search is based on computing a distance (or cost) between each note in the input score and a number of sample candidates. Each m-th note in the input sequence is first assigned a note group G^m as described in Section 2.2.1. Then, for each m-th note a sample candidate list is generated by collecting those database samples matching the assigned note group.

3.1. Gesture feature prediction

In a preprocessing step, a predictive model based on inductive logic programming is constructed for each note group, using the database sample examples that were previously analyzed. Each model is trained with the samples belonging to the corresponding note group, having the performance context features used as inputs, and the contour feature vectors used as outputs. The bowing gesture prediction models (they have been built by means of the Tilde algorithm [1]) achieve mean absolute errors of around $\overline{MAE} = 0.2$, as reported in a previous work by the author which uses the same database and gesture features for a different application (see [5] for a detailed description).

For each m-th note in the score, the note's group model T_{G^m} is applied, and a vector f_p^m of gesture features is predicted out of the score annotations (duration D^m, effective string length l_e^m, string played s_m, and dynamics d^m).

3.2. Sample Search

For an input sequence of M notes, the task of sample selection is set as finding the vector $\mathbf{q}^* = \{q^{1,*} \ldots q^{M,*}\}$ of sample candidate indexes which minimizes a total cost $C(\mathbf{q})$, expressed as

$$\mathbf{q}^* = \underset{\mathbf{q}}{\operatorname{argmin}}\, C(\mathbf{q}) = \underset{\mathbf{q}}{\operatorname{argmin}} \sum_{m=1}^{M} C^{m,q^m}, \tag{1}$$

where the value for q^m will iterate over the m-th note's possible candidates, and C^{m,q^m} represents the cost associated the m-th note when choosing the candidate pointed by q^m. The solution is found using dynamic programming (in particular the Viterbi algorithm).

The cost C^{m,q^m} is decomposed as a weighted sum of two different terms, as expressed in equation (3.2). The first one, referred to as *basic cost* and labelled as C_b^{m,q^m}, is computed by attending to database annotations not regarding specific gesture features. The second one, referred to as *gesture cost* and labeled as C_g^{m,q^m}, is computed from a vector of gesture-specific features (see Section 2.2.2) predicted from the score.

$$C^{m,q^m} = w_b C_b^{m,q^m} + w_g C_g^{m,q^m} \tag{2}$$

Basic cost. The cost C_b^{m,q^m} is computed as a weighted sum of five different sub-costs:

$$C_b^{m,q^m} = w_D C_D^{m,q^m} + w_{l_e} C_{l_e}^{m,q^m} + w_d C_d^{m,q^m} + w_s C_s^{m,q^m} + w_c C_c^{m,q^m} \tag{3}$$

The duration cost $C_D^{m,q^m} =\mid \log_2 \frac{D^m}{D^{q^m}} \mid$ is computed from the duration ratio, and is equal to one for duration ratios of two or point five. The cost $C_{l_e}^{m,q^m} =\mid \log_2 \frac{l_e^m}{l_e^{q^m}} \mid$, associated to the effective string length, is computed in an analogous manner: its value equals one for length ratios of two or point five. Given the three different dynamics levels present in the database, the dynamics cost C_d^{m,q^m} is given a value of zero when dynamics match, a value of point five when there is one-level difference, and a value of one for two-level difference. Similarly, the string cost C_s is given a value of $1/3$ if the candidate corresponds to a note played at a string next to the string that was scripted in the score, a value of $2/3$ for two-strings differences, and a value of one for samples played at the string E and score notes annotated with G string (or viceversa). Finally, the continuity cost C_c^{m,q^m} (computed for each note and its predecessor) is set as a penalty for strong sample-to-sample discontinuities in bowing parameter signals.

Gesture cost. Based on the predicted gesture features, the gesture cost C_g^{m,q^m} is computed as the Mahalanobis distance between the sample candidate's feature vector $f_s^{q^m}$ and the predicted vector f_p^m, using the covariance matrix Σ_{G^m} corresponding to the m-th note's group distribution (see Section 2.2.2). This is expressed as:

$$C_g^{m,q^m} = \sqrt{(f_p^m - f_s^{q^m})^T \Sigma_{G^m}^{-1} (f_p^m - f_s^{q^m})}. \tag{4}$$

3.3. Time-stretch and concatenation

Retrieved samples are first transformed in duration by applying time-stretch to the bowing parameter signals, so that the duration matches that scripted in the score. Time-stretch factors are equally applied to the three signals (bow velocity, bow force and β ratio) in a non-linear fashion, so that sample edges remain unaffected. Then, sample concatenation is carried out for each bowing parameter, by means of smoothly connecting retrieved samples' signals around note junctions.

4. SOUND SYNTHESIS

The synthetic bowing parameters obtained through the process described above are used to drive both physical and spectral-domain synthesizers of violin sound. Due to the focus on right-hand gestures, no fingering controls are considered. Thus, the input signals to the sound synthesis algorithms are bow velocity, the bow force, and the β ratio. The scripted pitch leads to a step-wise constructed pitch signal as input control, together with the string being played. Given the importance of left-hand control in violin playing (e.g., vibrato), the naturalness of synthetic sound falls still upon the inclusion of an appropriate left-hand control model, especially for *legato* bowing technique. To our subjective perception, the obtained preliminary sounds nevertheless showed a degree of realism that demonstrates the potential of considering gesture samples as a basic source for instrumental sound synthesis.

Physical modeling synthesis. Physical modeling synthesis is carried out by means of the *Synthesis Toolkit in C++* (STK[1]) implementation of Smith's waveguide bowed-string

[1] http://ccrma.stanford.edu/software/stk/

model. A single string model with a low-pass one-pole implementation of the loss filter has been chosen as a proof-of-concept physical model for this work. For obtaining the violin sound from the string velocity signal, we used impulse responses computed through different methods [4].

Spectral-domain additive synthesis. Violin additive synthesis is achieved trough the violin timbre model reported in [7]. The model is based on neural networks that are trained with audio and bowing descriptors from real performances. The model is able to predict the spectral envelopes of the harmonic and residual components of the string vibration signal given the instantaneous values of bowing parameters. A simplified model was built for the three bowing parameters being considered here. Predicted harmonic envelopes are filled with the harmonics corresponding to the pitch in the score, while and residual envelopes are filled with white noise, following the SMS model[2]. After overlap-add of resulting frames, the output sound is convolved with a violin body impulse response previously estimated by deconvolution of a microphone signal by a pickup signal, as described in [7].

Tuning of a sample-based selection process is a difficult task. In our experiments, we used different weight tunings of the two main costs involved in sample selection: the *basic cost* and the *gesture cost* 3.2. For both sound synthesis methods, the elimination of the *gesture cost* led to obtain sounds presenting a less consistent timbre continuity, mainly due to certain disparity in note's dynamics annotated the performance database. Also, biasing weights towards the *basic cost* usually leads to selecting samples that present a worse segmentation, especially at bow direction changes.

5. CONCLUSION

In this paper we have introduced gesture sampling and concatenative gesture rendering in the context of violin sound synthesis. In our preliminary application case study, samples of bowing gesture parameters (bow velocity, bow force, and β ratio) are retrieved and concatenated from a database of performance recordings, by attending to an input score. Synthetic bowing controls are then used for driving sound generation through physical and spectral-domain synthesis.

Although the fidelity of obtained performances is limited by the quality of the sound synthesis methods, synthetic sounds show in general a significant degree of realism. The promising results lead us to believe that important improvements can be achieved in less preliminary implementations that will include a better balanced database, left-hand controls, and cost weight tuning based on formal subjective tests. So far, no gesture-specific sample transformation was studied, but the inclusion of higher level gesture transformations (for example, changing the dynamics of a sample), appear as a promising post-processing step that should account for keeping resulting signals within the limits of the playable space [7]. Moreover, the incorporation of a physical model considering the arm movements could be an interesting tool for giving coherency to gesture concatenation and transformation [2].

In the future, larger-scale subjective tests could be carried out in order to determine whether gesture sampling might represent an important improvement over generative models trained to emulate the temporal evolution gesture parameter signals out of a score [4].

6. REFERENCES

[1] H. Blockeelm, L. De Raedtm, and J. Ramon, "Top-down induction of clustering trees," in *Proc. of the Intl. Conf. on Machine Learning*, Wisconsin, 1998.

[2] A. Bouënard, M. M. Wanderley, and S. Gibet, "Advantages and limitations of simulating percussion gestures for sound synthesis," in *Proc. of the ICMC*, Montréal, 2009.

[3] M. Demoucron, "On the control of virtual violins: Physical modelling and control of bowed string instruments," Ph.D. dissertation, Univ. Pierre et Marie Curie (Paris 6) and KTH (Stockholm), 2008.

[4] E. Maestre, "Modeling instrumental gestures: an analysis/synthesis framework for violin bowing," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, 2009.

[5] E. Maestre and R. Ramírez, "An approach to predicting bowing control parameter contours in violin performance," *Intelligent Data Analysis (In Press)*, 2010.

[6] E. Maestre, "Data-driven statistical modeling of violin bowing gesture parameter contours," in *Proc. of the ICMC*, Montréal, 2009.

[7] A. Pérez, "Enhancing spectral synthesis techniques with performance gestures using the violin as a case study," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, 2009.

[8] N. H. Rasamimanana and F. Bevilacqua, "Effort-based analysis of bowing movements: evidence of anticipation effects," *Journal of New Music Research*, vol. 37, no. 4, pp. 339–351, 2008.

[9] E. Schoonderwalt, "Mechanics and acoustics of violin bowing," Ph.D. dissertation, KTH (Stockholm), 2008.

[2] http://mtg.upf.edu/technologies/sms/