PROBABILISTIC MODELING OF BOWING GESTURES FOR GESTURE-BASED VIOLIN SOUND SYNTHESIS

Akshaya Thippur[1] **Anders Askenfelt**[2] **Hedvig Kjellström**[1]
[1]Computer Vision and Active Perception Lab, KTH, Stockholm, Sweden akshaya,hedvig@kth.se
[2]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden andersa@speech.kth.se

ABSTRACT

We present a probabilistic approach to modeling violin bowing gestures, for the purpose of synthesizing violin sound from a musical score. The gesture models are based on Gaussian processes, a principled probabilistic framework. Models for bow velocity, bow-bridge distance and bow force during a stroke are learned from training data of recorded bowing motion. From the models of bow motion during a stroke, slightly novel bow motion can be synthesized, varying in a random manner along the main modes of variation learned from the data. Such synthesized bow strokes can be stitched together to form a continuous bowing motion, which can drive a physical violin model, producing naturalistic violin sound. Listening tests show that the sound produced from the synthetic bowing motion is perceived as very similar to sound produced from real bowing motion, recorded with motion capture. Even more importantly, the Gaussian process framework allows modeling short and long range temporal dependencies, as well as learning latent style parameters from the training data in an unsupervised manner.

1. INTRODUCTION

The aim of the current study is to develop natural sounding violin sound synthesis, which includes the characteristics of human performance. Our thesis is that this is accomplished by modeling the music-production process as accurately as possible: The player reads the musical score and interprets the piece as a sequence of events linked by the musical structure. The interpretation involves planning a sequence of control gestures, each producing a single note or a short sequence of notes.

Two aspects on sequences of sound-producing gestures can be noted.

I) The exact shape of the control gestures depend on the individual interpretation of the musician, based on the knowledge of the style of the composition. It follows that it is desirable to be able to control performance style in a synthesized performance (e.g., from baroque to romantic violin playing style).

II) Each control gesture, corresponding to a single note or a short sequence of notes, depends on a range of other gestures preceding and following after the current gesture.

Both these aspects are captured by a probabilistic framework which can represent a set of generic bow motion gestures which together define a performance of a piece of music as well as important modes of variation. This is further discussed in Sec. 2.

As a basis for this framework, we propose to use Gaussian processes (GP) [1], see Sec. 3. In this paper we present a pre-study where GPs are trained with recorded bow motions in the same manner as Bezier curves in related work [2–4]. The results indicate the GPs have the same descriptive power as the Bezier curves. A listening test presented in Sec. 4 shows that the violin sound produced from synthetic bow motions is perceived as very similar to the sound produced from real bow motion, recorded with motion capture.

Furthermore, we suggest in Sec. 5 that GP provides a solid mathematical framework for addressing the issues of individual performances and the style of playing in a principled manner. Our thesis is that a GP framework will make the best use of recorded bow motion gestures. Such dependencies reflect variations in player interpretation and modes of performance based on composition style.

2. BACKGROUND

Bow motion studies. Recently, Demoucron and Schoonderwaldt have studied bow motion in violin performance using motion capture methods and a robust method for sensing bow force [5–7]. Their work resulted in a large database of calibrated motion capture measurements of main and secondary bowing parameters of the performances of professional violinists (see Figure 1). Major results of their analyses were to establish the bow control strategies that players use when changing the dynamic level and timbre, and playing on different strings [8, 9].

Demoucron also developed a parametric description of bouncing bowing patterns (spiccato) based on the recorded data. Bow control data generated by his analytical models were used to drive a physical model of the bowed string [5]. The results showed that the realism of the synthesis increases significantly when the variation in control parameters reflect real violin performance. Again the coordination of the variations in the control parameters is a key to realistic violin sound. The conclusion drawn was that modeling

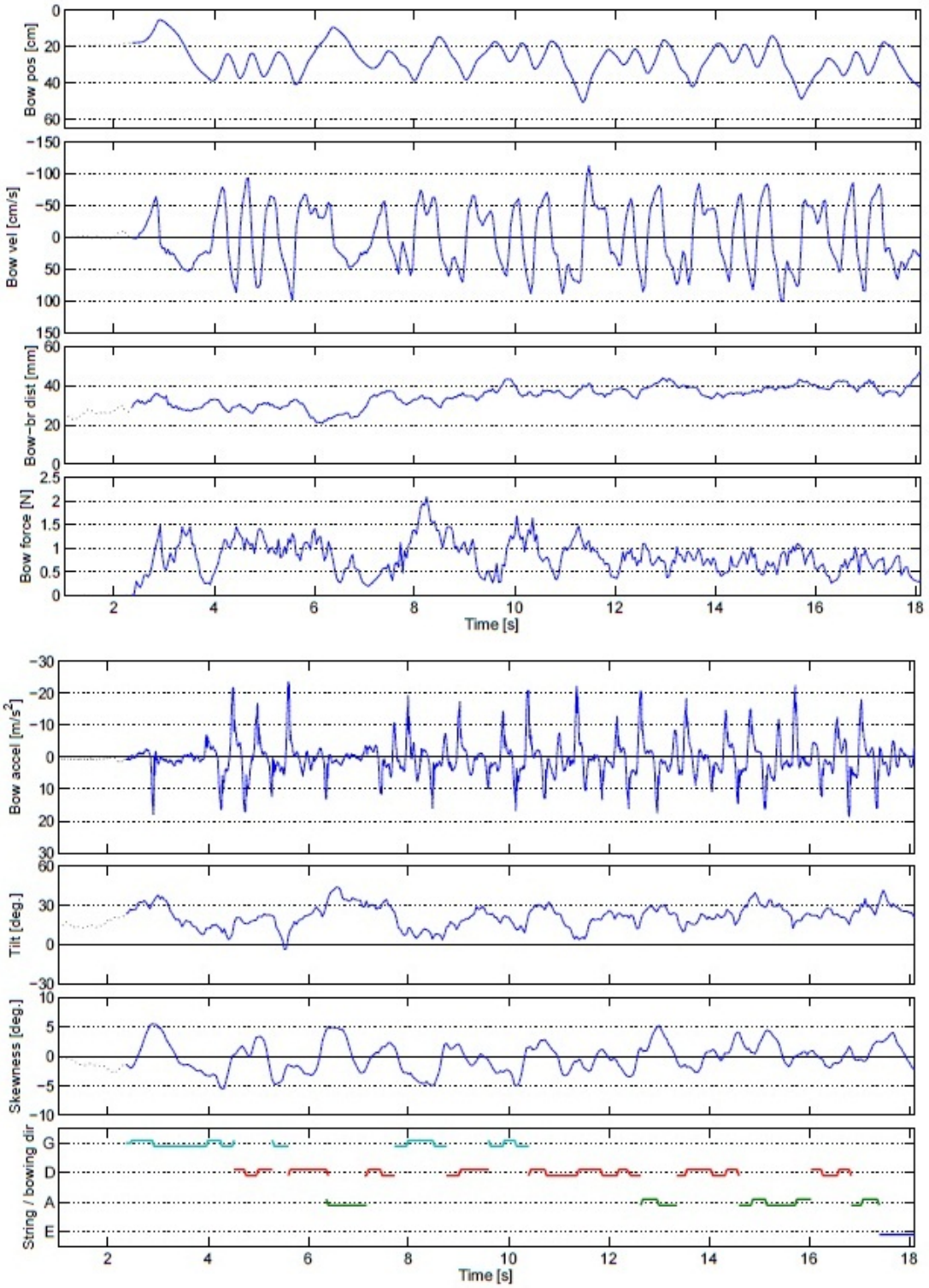


Figure 1. Example of a full set of bow control gestures for the first 18 notes of a Bach partita. From top: Transversal bow position, bow velocity, bow-bridge distance, bow force, bow acceleration, tilt, inclination, skewness, string played/bowing direction. From [6].

instrument control parameters is as important as modeling the instrument itself.

Violin synthesis from score information. A next step is using the acquired knowledge to learn computer models that produce violin sound (or rather, violin bowing gestures) given a musical score. Two approaches can be found in the literature: Reinforcement learning, where the computer model learns to perform bowing gestures (i.e., produce time sequences of bow motion parameters) under supervision of a human teacher, and supervised learning, where the computer model is trained with a set of recorded gestures, correlated with a musical score.

A reinforcement learning approach has been reported recently [10], where a generative model of bow motion is trained much in the same way as children learn to play according to the Suzuki school: The bow velocity and bow-bridge distance are preset using plain score information, while the range of bow forces producing a successful sound is learned using discriminative classifiers with human feedback judging the tone quality. Although a very interesting approach – the model can after four hours of training play like a Suzuki student with one year of experience – this is not a feasible approach to professional level violin sound synthesis.

A more time-efficient alternative is thus to directly show the model successful gesture examples, using a supervised learning approach. A recent, very ambitious collection of works from the Music Technology Group at Universitat Pompeu Fabra, Barcelona, deals with the task of automatically generating naturally-sounding violin performances from an annotated score [2–4]. Their approach is to retrieve samples of control gesture parameters from a database and concatenate them according to the score including instrument-specific annotations. The database is obtained from motion capture recordings of real violin performances, which have been segmented and classified into different groups for single notes with specific features (bowing style, dynamic level, context). All bow control gesture samples are parametrized using Bezier curve segments. For synthesis, the task of selecting the proper gesture sample in the database for each note in the score is based on an elaborated cost function which takes into account the deviations in duration and dynamic level between the stored samples and the specifications for the note in the score. The selected samples are stretched and concatenated and used to drive a simple physical model of the violin.

The obtained degree of realism in sound demonstrates the potential of gesture control of violin synthesis – it is indeed possible to simulate this extremely complex dynamical process. In the work presented here, we continue this path, using a supervised learning approach.

However, two aspects remain unaddressed in the Barcelona work, corresponding to aspects *I* and *II* discussed above.

I) No means exist to steer the performance style during synthesis. The grouping of the gesture examples according to bowing style (legato, spiccato etc.), dynamic level, and context give some possibility of style control, but more ubiquitous style variations (e.g., baroque vs. romantic playing style) are not captured – the model simply generates the "mean" performance in the database, given a certain score. This is however possible to accommodate in a probabilistic framework, such as the one proposed in this paper. The GPs capture the entirety of the numerous training curves comprehensively. It not only captures the average curve shape but also captures the different modes of variation in the training set. From a trained GP, slightly novel instances of synthetic bow motion can be generated, preserving the general shape and variances in the training set.

II) No means exist to model long-term gesture correlations. The curves are stitched together so that a coherent and physically plausible motion is generated. However, there is no way of selecting curves depending on gestures more than one time-step away. This is however possible to accommodate by adding extensions to our proposed framework. This is further discussed in Sec. 5.

3. MODELING BOWING GESTURES

Figure 2 gives an overview of the sound generating process. The score is viewed as a sequence of notes, each belonging to a note class defined by note value (duration), pitch, dynamic level, articulation, and bowing style. The strategy is to transform the notes in the score to a sequence of generic bow control gestures, each representing one note. The control gestures are then concatenated and used to drive a physical model of the bowed violin.

Gesture modeling using Gaussian processes. We model mapping 2 in Figure 2 using GPs in a manner similar to how Maestre et al. [3] use Bezier curves. The added value of the GPs is that not only the mean curves are captured by the model, but also the typical variations among training examples. This enables learning of style parameters, as discussed above. Furthermore, the GP is non-parametric, meaning that no analytical form is imposed on the data – in other words, we do not risk introducing erroneous assumptions in the model [1].

The models are trained with motion capture recordings from the database of Schoonderwaldt and Demoucron [7]. We use the bow velocity (V), bow-bridge distance (D), and bow force (F) data from two different professional violinists playing sequences of detaché notes in forte (*f*), mezzo-forte (*mf*), and piano (*p*), on each of the four strings. The recorded sequences are segmented into strokes (by detecting bow velocity zero crossings), and all segments are resampled to a length of $n = 125$ points, equal to 2.08 s with a sampling frequency of 60 Hz. Figure 3, left graph in each subfigure, show segmented and length-normalized *mf* curves corresponding to down-bow and up-bow, respectively.

In total, 6 models are learned: *f* down-bow, *f* up-bow, *mf* down-bow, *mf* up-bow, *p* down-bow, and *p* up-bow. There are $m = 16$ training examples for each model. Each model

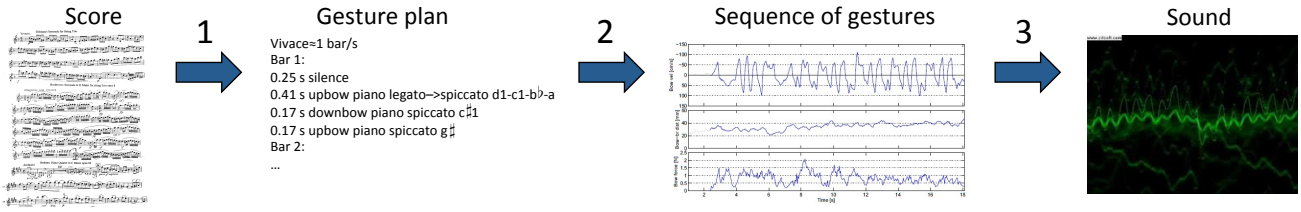


Figure 2. The violin sound generating process. Mapping 1 is well defined, the musical score is simply a coding scheme for the sequence of gestures in the plan. For mapping 3, we use the system developed by Demoucron [5]. The focus of the work here is mapping 2, the generative process of continuous bow motion from the sequence of gestures extracted from the score.

has three dimensions (V, D, F) which are modeled independenty from each other – in practice, there are three separate GPs for each model; for V, D, and F, respectively (for *mf* examples, see Figure 3(a,c,e) or (b,d,f)).

A GP is defined as follows (for a more detailed explanation, see [1]): View the training curves for a certain GP (for examples, see the left graph in each subfigure of Figure 3) as an array of tuples $[(x_i, t_i)]$ where $i \in [1, mn]$ (i.e., one tuple for each point on each training curve). Assume the data to be spatio-temporally Gaussian distributed:

$$[x_1, \ldots, x_{mn}] \in \mathcal{N}(\mu([t_1, \ldots, t_{mn}]), k([t_1, \ldots, t_{mn}], [t_1, \ldots, t_{mn}])) \quad (1)$$

where $\mu(t) = \frac{1}{m} \sum_{j:t_j=t} x_j$, the mean value of x at time step t, and $k(t, t')$, the covariance between values x at timesteps t and t'. We use a stationary squared exponential covariance function:

$$k(t, t') = \exp(-\frac{\|t - t'\|}{2\sigma^2}) \quad (2)$$

where σ is the characteristic time dependency length in the GP $x(t)$, and is learned from the data. A natural extension would be to use a non-stationary function, with a time-varying characteristic length $\sigma(t)$. For example, the velocity is by definition 0 at the beginning and end of each bow stroke; this would be learned with a time-varying $\sigma(t)$. We use the GP implementation by Lawrence [11]. Further extensions are discussed in Sec. 5.

Figure 3, right graph in each subfigure, show the GPs learned from the training data in the same subfigure. From these GPs, novel curves with the same characteristics, but with some stochasticity, can be sampled from the learned mean $\mu(t)$ and covariance $k(t, t')$.

The output of the mapping 2 model is a sequence of synthetic bow control curves. The choice of dynamics (*f* – *mf* – *p*) and the bowing, are selected according to the notation in the sheet. One V, D, F curve is then sampled for each note (or sequence of notes played with one bow) in the musical sheet, and stretched to the right duration as indicated in the sheet. The curves are then stitched together, forming a coherent synthetic bow motion.

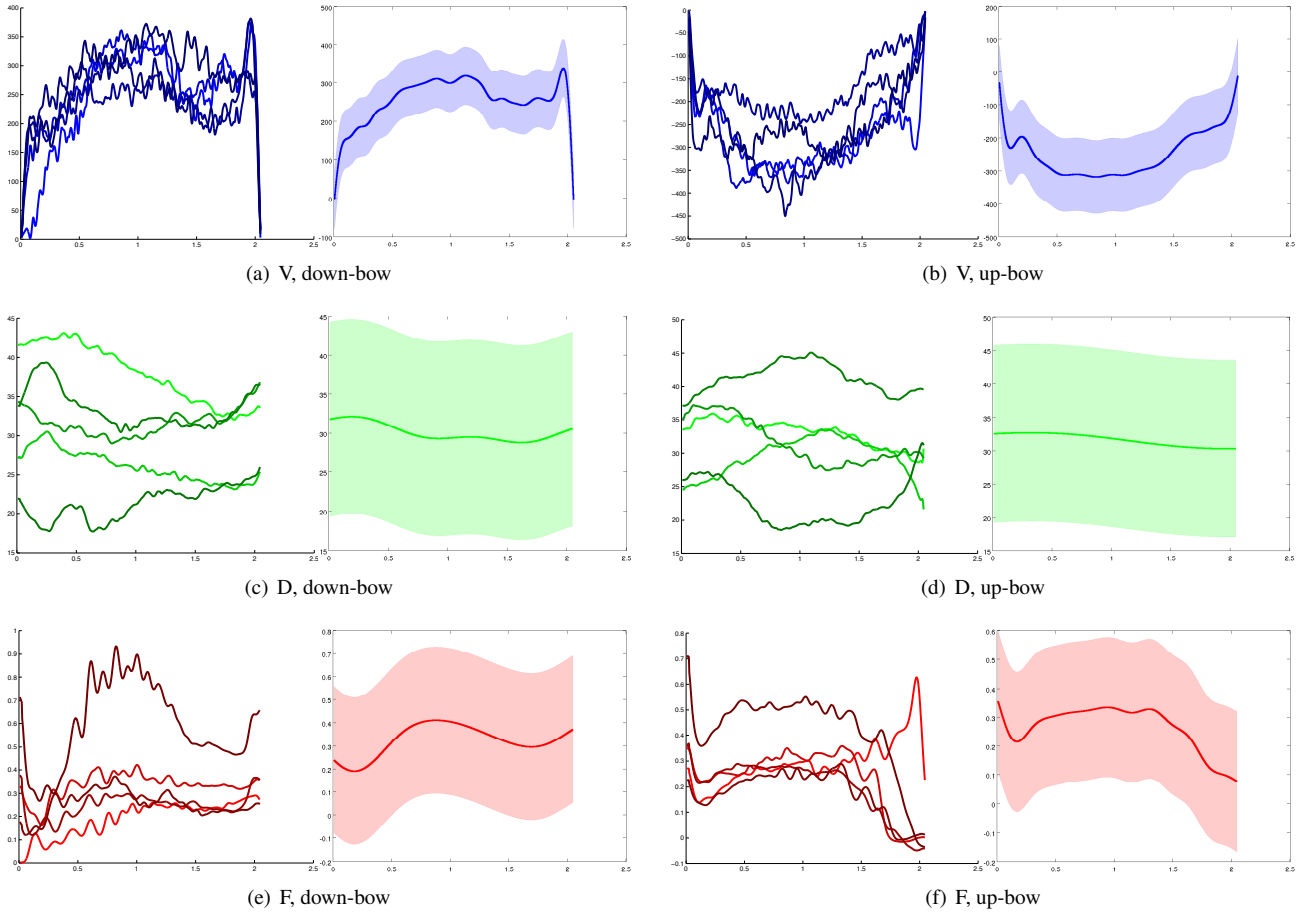
4. LISTENING TESTS

The naturalness of the curves generated from the Gaussian processes was evaluated using a listening test. Violin notes were synthesized using real bow control gestures from the database [7], and artificial gestures from the Gaussian processes, respectively, and compared to check if they were perceived as significantly different. The focus of the evaluation was not on the realism of the generated sounds as such, rather on the naturalness of the underlying bow motion. This aspect required listeners with extensive own experience of string playing. In order to make a fair comparison, all violin sound stimuli were synthesized in an identical manner (see Figure 2, mapping 3), using the bowed-string model developed by Demoucron [5]. The model, which is implemented using modal synthesis, gives a realistic bowed string sound when controlled by calibrated bow control gestures.

Stimuli. Bow control gestures from the Gaussian processes were compiled for pair of half notes played detaché down-bow-up-bow (or up-bow-down-bow), and the artificial V, F and D curves were fed into the bowed-string model. The length of the stimuli were $2 \times 2.08 = 4.16$ s. These stimuli were compared with sounds generated by feeding real motion capture recordings of V, F, D sequences of half notes, also of length 4.16 s, from the database into the same model. Two pitches were used, C4 and G5, played on the G and D string, respectively, combined with two dynamic levels (*mf* and *f*). No vibrato was included. Two independent samples of bowing gestures for each of the four cases were synthesized. A corresponding set of stimuli was generated played up-bow-down-bow. In all, 16 stimuli were generated from the GPs, and 32 from the database by including recordings of two players.

A selection of four down-bow-up-bow cases (and corresponding four up-bow-down-bow cases) from the Gaussian process stimuli was made after selective listening. The purpose of the selection was to limit the size of the listening test, and to include stimuli with different qualities of the attack which normally occur in playing; perfect, choked (prolonged periods) and multi-slip attacks. The 2×4 stimuli from the Gaussian processes were combined with the corresponding cases from two players. The listeners judged each of the 3×8 stimuli three times, in all 72 responses. The stimuli were presented in random order, different for each listener.

Procedure. Eight string players participated in the test; one professional and seven advanced amateurs. Their musical training as violin players ranged between 7 and 19 years, and they had between 12 and 32 years of experience of playing string instruments in general.



(a) V, down-bow (b) V, up-bow

(c) D, down-bow (d) D, up-bow

(e) F, down-bow (f) F, up-bow

Figure 3. GPs trained with motion capture recordings of half notes played detaché from the database in [6]. A bow stroke consists of bow velocity (V), bow-bridge distance (D), and bow force (F) curves. The V, D, F curves are modeled independently from each other in three GPs. There are separate models for up- and down-bow, and for forte (*f*), mezzoforte (*mf*), and piano (*p*). The figures show models for *mf*: (a,c,e) The three GPs for down-bow. (b,d,f) The three GPs for up-bow. Left in each subfigure: Examples of training data. Right in each subfigure: The GP learned from this training data. The shaded region indicates the standard deviation $\sqrt{k(t,t)}$. Note that the covariance is stationary (time independent). A natural extension would be to use a non-stationary function, with a time-varying characteristic length $\sigma(t)$.

The task of the listeners was to rate the naturalness of the bow motion. They were explicitly instructed to not pay attention to the general quality of the violin sound, but to focus on the underlying bow motion by responding to the question "How natural is the bow motion that produced the notes you heard?" The response was given on a scale from 0 ("artificial") to 1.0 ("like a human player") using a slider on a computer screen. The stimuli could be repeated as many times as desired, but that feature was rarely used. A short familiarization with the task including four stimuli initiated each test session. The listeners were informed about that the sounds could contain attacks of different quality and other noises which normally occur in playing. They were neither informed about the origin of the stimuli, nor about the purpose of the test.

Results. The results are summarized in Figure 4, showing average ratings across all 72 stimuli for each of the eight listeners. It is clear that the listeners had different opinions about the general level of naturalness of the bow gestures. Most listeners, however, gave an average response midway between "artifical" (0) and "like a human" (1.0), with a notable exception for Listener 7. The important result, however, is that the bow gestures generated by the Gaussian processes were judged to be more natural than the real gestures from the database by all but two listeners (5 and 7). For Listeners 1 and 2, the preference for the Gaussian processes was quite marked. The consistency and repeatability in judgements appeared satisfactory as indicated by the error bars.

A conservative interpretation of the results is that six out of eight experienced string players did not hear any difference between synthesized violin notes generated by bow gestures from Gaussian processes and real performances, respectively. Two listeners perceived the Gaussian processes bow gestures as more natural than the corresponding real ones.

5. CONCLUSIONS

We presented a probabilistic approach to modeling violin bowing gestures, for the purpose of synthesizing violin sound from a musical score. The gesture models were based on GP, a principled probabilistic framework. Models

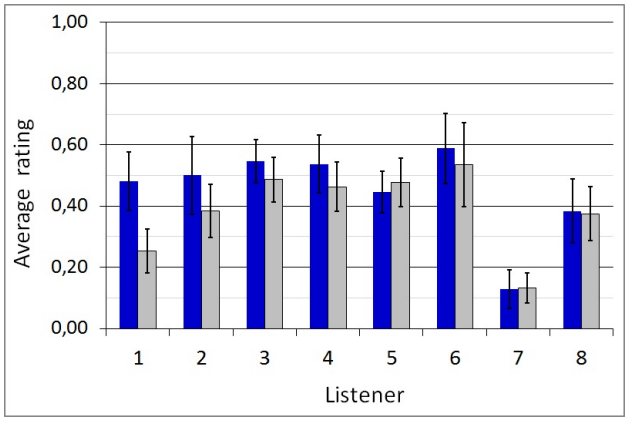


Figure 4. Result of the listening test. Average scores for eight listeners across all stimuli generated by bow gestures from the Gaussian processes (dark blue) and from real bow gestures in the data base (light grey). Error bars correspond to ± 0.5 standard deviation.

for bow velocity, bow-bridge distance and bow force during a stroke were learned from training data of recorded bowing motion. From the models of bow motion during a stroke, slightly novel bow motion could be synthesized, varying in a random manner along the main modes of variation learned from the data. Such synthesized bow strokes could be stitched together to form a continuous bowing motion, which was used to drive a physical violin model, producing naturalistic violin sound. Listening tests showed that the sound produced from the synthetic bowing motion was perceived as very similar to sound produced from real bowing motion, recorded with motion capture.

5.1 Future Work

The proposed framework built on GP allows for principled extensions to address aspects *I* and *II* in Sec. 2. Capturing aspect *I* requires models where style modes can be learned from data. We propose to use Gaussian process latent variable models (GPLVM) [11] which are an extension of GP, and have been used extensively for modeling human behavior. Long-term dependencies (aspect *II*) can be modeled using grammar-like models. We propose to use dynamic Bayesian networks (DBN) [12] where each node is a GPLVM representing a gesture.

Fewer, parameterized, gesture classes. The number of conceivable note classes is very large due to the combinations of score characteristics: duration, pitch, dynamic level, articulation, and bowing style. An example of a note class would be [A4, quarter note, forte-level, sforzando (accented), staccato (short) , preceded by a long note at piano-level and followed by a rest (silence)]. One could also add instrument-specific information, e.g., that the note should be played in a down-bow on the D string. A combinatorial explosion of cases will emerge; the task of assigning a set of bow control gestures to each note class will not be scalable, when going from a basic division into few note classes based on a couple of broad characteristics (e.g., high pitch, long note, loud) to a more detailed description as in the example above.

In [3], 102 note classes were used even without including pitch and duration among the note class properties. These were handled later in the selection of a suitable sample of bowing gestures from the generic gestures. A particular concern in music performance is the strict timing imposed by the note values and tempo given in score. The attack and termination of a note cannot be stretched or compressed much without changing the perceived quality [13].

We propose to use the experience of expert players to investigate to what extent the number of note classes can be restricted. Bowing styles like detaché, legato, spiccato are examples of note characteristics which definitively define different note classes. Pitch, duration, dynamic level are examples of characteristics which are possible to encode as latent parameters in the GPLVM models. The context dependence – which notes come before and after – may also be possible to handle to a certain extent by controlling the end constraints when sampling from the processes.

Learning ubiquitous style modes. The linear and well-defined modes of variation described above are possible to train in a supervised manner, since the training examples could be labeled with objective measures of volume (dB), duration (s), pitch (Hz). However, style variations such as baroque vs. romantic violin playing style are not apparently observable in recorded bowing parameters. As discussed in aspect *I* above, a highly desirable property of a violin synthesizer is the possibility to control high-level performance style parameters.

It is however possible to learn unobservable latent parameters from data using GPLVM [11]. Any level of supervision can also be included if such is available; for example, a mode of variation corresponding to music style could be learned from the data given that the examples were labeled with baroque – Viennese classic – romantic – jazz etc. It will be necessary to collect a wider range of data examples.

Learning phrases, bow planning, and other long time-range dependencies. Addressing aspect *II* above, we will then proceed to modeling dependencies between gestures that are separated in time. This is necessary in order to be able to represent phrase-based music interpretation (see Sec. 2). Moreover, on a slightly shorter time scale, the finite length of the bow needs to be taken into account. This will require a preplanning which takes many notes ahead into account so that bow changes can take place at musically motivated instances, and that notes are played using a "natural" bowing direction (down-bow/up-bow). Related to this question is the modeling of sound feedback in the gesture production [5, 14]. Sound feedback is very important for small modulations in bowing motion, e.g., during spiccato.

To represent hierarchical dependencies and dependencies between a whole sequence of gestures – a gestural "grammar" – we will employ Dynamic Bayesian Networks (DBN) [12] which is the mathematically principled way of representing probabilistic dependencies between data segments over time.

6. REFERENCES

[1] C. Rasmussen, *Gaussian Processes in Machine Learning*, Springer, 2004.

[2] A. Perez, *Enhancing Spectral Synthesis Techniques with Performance Gestures using the Violin as a Case Study*, PhD Thesis, Universitat Pompeu Fabra, Spain, 2009.

[3] E. Maestre, *Modeling Instrumental Gestures: An Analysis/Synthesis Framework for Violin Bowing*, PhD Thesis, Universitat Pompeu Fabra, Spain, 2009.

[4] E. Maestre, M. Blaauw, J. Bonada, E. Guaus, and A. Perez, "Statistical modeling of bowing control applied to violin sound synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 4, pp. 855–871, 2010.

[5] M. Demoucron, *On the Control of Virtual Violins: Physical Modelling and Control of Bowed String Instruments*, PhD Thesis, KTH, Sweden, 2008.

[6] E. Schoonderwaldt, *Mechanics and Acoustics of Violin Bowing: Freedom, Constraints and Control in Performance*, PhD Thesis, KTH, Sweden, 2009.

[7] E. Schoonderwaldt and M. Demoucron, "Extraction of bowing parameters from violin performance combining motion capture and sensors," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2695–2708, 2009.

[8] E. Schoonderwaldt, "The violinist's sound palette: Spectral centroid, pitch flattening and anomalous frequencies," *Acta Acustica united with Acustica*, vol. 95, no. 5, pp. 901–914, 2009.

[9] E. Schoonderwaldt, "The player and the bowed string: Coordination and control of violin bowing in violin and viola performance," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2709–2720, 2009.

[10] G. Percival, N. Bailey, and G. Tzanetakis, "Physical modeling meets machine learning: Teaching bow control to a virtual violinist," in *Sound and Music Computing Conference*, 2011.

[11] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2004.

[12] K. P. Murphy, *Dynamic Bayesian networks: Representation, Inference and Learning*, PhD Thesis, University of California at Berkeley, USA, 2002.

[13] K. Guettler and A. Askenfelt, "Acceptance limits for the duration of pre-Helmholtz transients in bowed string attacks," *Journal of the Acoustical Society of America*, vol. 101, pp. 2903–2913, 1997.

[14] K. Guettler and A. Askenfelt, "On the kinematics of spiccato and ricochet bowing," *Catgut Acoustical Society Journal*, vol. 3, no. 6, pp. 9–15, 1998.