Estimation of violin bowing features from Audio recordings with Convolutional Networks

Alfonso Perez-Carrillo
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
alfonso.perez@upf.edu

Hendrik Purwins
Audio Analysis Lab
Aalborg University Copenhagen
Copenhagen, DK
hpu@create.aau.dk

Abstract

The acquisition of musical gestures and particularly of instrument controls from a musical performance is a field of increasing interest with applications in many research areas. In the last years, the development of novel sensing technologies has allowed the fine measurement of such controls. However, the acquisition process usually involves the use of expensive sensing systems and complex setups that are generally intrusive in practice. An alternative to direct acquisition is through the analysis of the audio signal. So called indirect acquisition has many advantages including the simplicity and low-cost of the acquisition and its non-intrusive nature. The main challenge is designing robust detection algorithms to be as accurate as the direct approaches. In this paper, we present an indirect acquisition method to estimate violin bowing controls from audio signal analysis based on training Convolutional Neural Networks with a database of multimodal data (bowing controls and sound features) of violin performances.

1 Introduction

Acoustical studies of sound production in bowed string instruments show that there is a close relationship between bowing controls and the sound produced. During note sustains the three major bowing parameters that determine the characteristics of the sound are the bowing force (bforce), the bowing distance to the bridge (bbd) and the bowing velocity (bvel). We are interested in modeling the correspondence between bowing controls and sound but in the opposite direction, i.e., mapping sound features extracted from an audio recording to the original bowing controls used. This inverse process is usually called indirect acquisition of gestures [1] and it is of great interest in different fields ranging from acoustics and sound synthesis to motor learning or augmented performances.

This work presents an *indirect* method able to extract continuous violin bowing controls from audio signal analysis by means of Convolutional Neural Networks and we compare the obtained results with past work [2, 3] .The predicted controls are *which string is played*, *bowing velocity*, *bowing force* and *bowing distance to the bridge* and they are estimated as continuous signals taking into consideration their evolution in time. The proposed methods are trained with empirical data previously collected with a highly accurate sensing system [4]. The training database contains synchronized streams of audio signals captured with a vibration transducer and instrumental controls measured with sensors [5].

2 The Method

The algorithm is based on Convolutional Neural Networks that are trained with empirical data previously collected with a highly accurate sensing system. The training database contains synchronized

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

streams of audio signals captured with a vibration transducer and bowing controls (string, velocity, force and *bbd*) measured with electro-magnetic field sensors [4].

2.1 Audio Features: Auditory EnergyGram

The audio signal is divided into harmonic and residual components through SMS [9] analysis. All pronounced spectral peaks other than the harmonics corresponding to the actual pitch are removed (e.g., sympathetic harmonics and other tones). The harmonic part is composed of the spectral peaks around harmonic frequencies and few bins around those peaks. The residual component is made up of the other bins between the harmonics. Then, energy of each component (i.e. harmonics and residual) is computed in spectral domain at 40 overlapping frequency bands with 50% of overlap factor. The band centers are distributed on a logarithmic scale based on Auditory models such as the *Bark* or *MEL* Scales. The energy of each residual band is estimated as the average of the corresponding frequency bins, weighted by a triangular function and for the harmonic part, the amplitude at each bin is determined by a harmonic envelope obtained at each frame by interpolating harmonic peaks using a third-order spline. More details regarding the specific algorithm for the computation of the bands can be found in[10]. Figure 1 shows an example of harmonic and residual Auditory EnergyGrams corresponding to a fragment of audio signal.

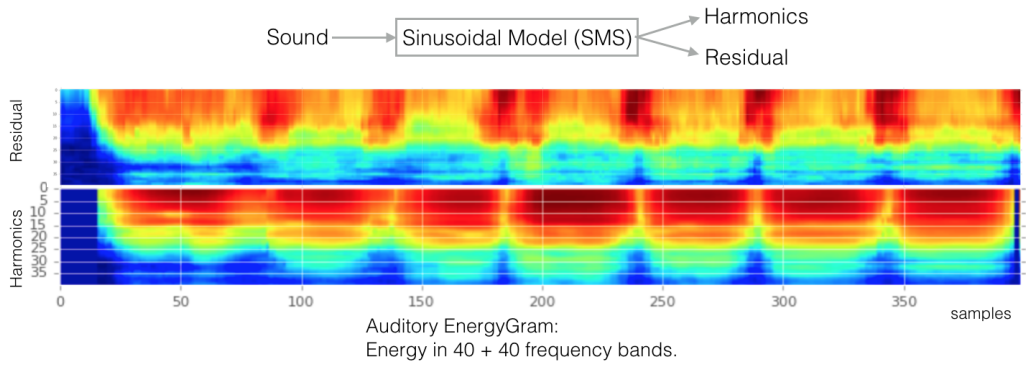

Figure 1: Example of harmonic and residual Auditory EnergyGrams corresponding to a fragment of audio signal.

2.2 Convolutional Networks

A convolutional neural network (CNN, or ConvNet) generally processes 2D data as it happens in simulated vision. In the case of audio, an adaptation needs to be done in order to transform 1D continuous data of the audio signal into 2D feature data. In this study, we use as inputs 2D fragments of the Auditory EnergyGrams (Figure 2).

2.3 Network Architecture

The proposed network architecture is shown in Figure 2. The input to the network is a 2D fragment of an Energygram of size 40x18. The first dimension (40) corresponds to the 40 energy bands previously computed (Section 2.1) and the second dimension (18) corresponds to audio frames, that is around 75ms at a sampling frequency of 240Hz for the control features.

The original Energygram is filtered out with 9 different kernels of size 2x2, applying a horizontal and vertical stride of 2 points. The result (*h1*) is a matrix containing 9 filtered versions of the original EnergyGram. The filtered versions are smaller in size (20x9), due to the convolution with the kernels. Convolutional layers with equal parameters are applied two more times. The result (h3) are nine versions of the original fragment, filtered out several times and compressed into 2D data of size 3x2 points. After the convolutions, an LSTM layer may be applied, which helps keeping the internal state of the network through time. The output of the LSTM layer is flattened and connected to two fully connected layers, whose output is the estimation of the bowing controls, one value per input frame.

The main advantage of using ConvNets is that we do not need to define other spectral features than the Energygrams. Instead, the features are self-computed in the form of convolution kernels during the back-propagation in the training process.

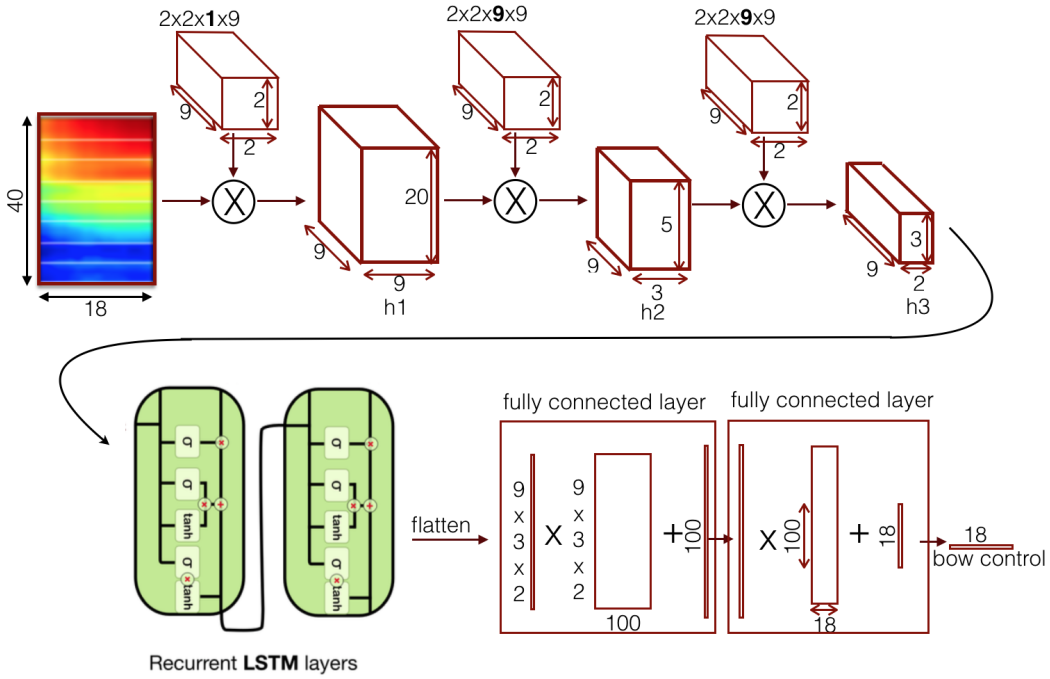

Figure 2: Architecture of the proposed network. Three convolutional layers, one LSTM layer, and two fully connected layers.

3 Evaluation and Results

The evaluation measure is the Correlation Coefficient (CC), which estimates the correlation of the predicted controls against the recorded ground truth by ten-fold cross-correlation of the training dataset. String detection achieves almost 100% correct estimation, which corresponds to a CC of 0.99, bowing velocity achieves a CC of 0.91, bowing force scores of 0.87, and bow-to-bridge distance scores of 0.83. Figure3 shows numerical results of the evaluated methods (average across all predicted features).

4 Conclusion

These results demonstrate that it is in fact possible to estimate the bowing controls from audio analysis based on statistical methods with a relatively small error. However, it should be noted that the current system is trained from recordings of a single violin and with a very specific audio signal, acquired with a vibration transducer built into the violin bridge. The signal captured with such a transducer is cleaner and easier to process than that of a microphone as it minimizes the resonances of the violin body, it avoids room acoustic reflexions, and it is not affected by the motion of the performer or the sound radiation patterns of the violin. The possibility of successfully performing indirect acquisition independently of the microphone and violin still remains uncertain. In the future, training of the models with more general databases including different violins and sound capturing devices will be performed in order to test the generality of the proposed models.

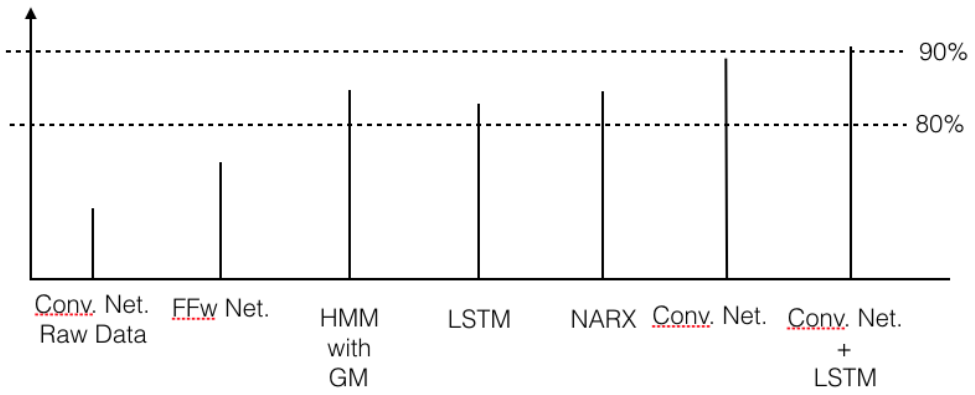

Figure 3: Average correlation coefficient of the estimation of the bowing parameters for the presented models, as well as a comparison with past works [2, 3]. From left to right: ConvNet with raw 1D audio Data, Feed-fordward Neural Networks [2], Hidden Markov Models with Gaussian Mixtures [3], LSTM, NARX [2], ConvNets and ConvNets with LSTM layers.

Acknowledgments

This work has been partly sponsored by the European Union Horizon 2020 research and innovation programme under grant agreement No. 688269 (TELMI project), and the Spanish TIN project TIMUL (TIN2013-48152-C2-2-R).

References

[1] M. M. Wanderley and P. Depalle, "Gestural control of sound synthesis," in *Proc. of the IEEE*, pp. 632–644, 2004.

[2] A. Pérez-Carrillo and M. Wanderley, "Learning and extraction of violin instrumental controls from audio signal.," in *In proc. of the MIRUM Workshop, ACM Multimedia Conference*, (Nara, Japan), November 2012.

[3] A. Pérez and M. M. Wanderley, "Indirect acquisition of violin instrumental controls from audio signal with hidden markov models," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, pp. 932–940, 2015.

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

[5] A. Perez-Carrillo, *Enhancing Spectral Synthesis Techniques with Performance Gestures using the Violin as a Case Study*. PhD thesis, Universitat Pompeu Fabra, 2009.

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

[7] J. Schelleng, "The bowed string and the player," *The Journal of the Acoustical Society of America*, vol. 53, no. 1, pp. 26–41, 1973.

[8] K. Guettler and A. Askenfelt, "On the creation of the helmholtz motion in bowed strings," *Acta Acustica united with Acustica*, vol. 88, no. 6, pp. 970–985, 2002.

[9] X. Serra, *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[10] A. Perez-Carrillo, J. Bonada, E. Maestre, E. Guaus, and M. Blaauw, "Performance control driven violin timbre model based on neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1007 –1021, March 2012.

[11] E. Schoonderwaldt, K. Guettler, and A. Askenfelt, "An empirical investigation of bow-force limits in the schelleng diagram," *AAA*, vol. 94, no. 4, pp. 604–622, 2008.

[12] D. . L. Lab., "Convolutional neural networks (lenet) – deeplearning 0.1 documentation."

[13] A. Pérez-Carrillo, J. Bonada, J. Pätynen, and V. Välimäki, "Method for measuring violin sound radiation based on bowed glissandi and its application to sound synthesis," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 1020–1029, 2011.