Gestural input effects on the spectral envelope of violin sounds

Quim Llimona

MUMT605 - Final Project

Contents

Introduction

The goal of this project is to explore the changes on the sound of a violin when the gestural input from the player changes. It is well-known that changes in the bowing parameters (bow velocity, bow force, etc.) affect mainly the spectral envelope of the sound (including the total energy, which is the spectral bias), and that is why we will focus on that.

This project is a preliminary step towards building gesture-controlled spectral processing tools for bowed strings, similar to what is presented in (Perez et al. 2007). We will make use of the databased generated by the author and detailed in (Llimona 2014), consisting on over 20 thousand notes played by different players on different instruments and with different bowing parameters, following a score designed to cover as much as possible of the playable range of the violin. In the dataset, each note contains audio data as well as performance gestures measured with a Motion Capture device. Bow force is estimated using a method derived from (Marchini et al. 2011), incorporating some of the suggested improvements.

In the first part of the report, we will compare different techniques that extract a one-dimensional signal that represents the spectral envelope of a violin sound.

In the second part, we will focus on how to extract the spectral envelope of a violin note, and on how to whiten it so that it is only affected by the bowing parameters and not by the violin being played. We will already report there clear relationships between these whitened spectral envelopes and some of the bowing parameters.

Finally, in the last part we will seek a low-dimensional representation of the changes in spectral envelope, which will allow an easier visualization of the effect of bowing parameters. This representation will be based in decomposing the spectral envelope into elementary filters in the log-spectral domain using non-negative matrix factorization.

Spectral envelope estimation

The first step towards modeling how bowing parameters affect the sound of a violin is to obtain a representation of its spectrum that accurately depicts the characteristics of the sound using a pitch-independent representation.

Auto-regressive methods: LPC

On of the first algorithms we considered for spectral envelope estimation was LPC (Atal and Hanauer 1971), for its simplicity. The MATLAB environment already provides an implementation of the Levinson-Durbin algorithm, so computing the autoregressive filter from the signal was straightforward.

Its main drawback is that it computes a spectral envelope with the same total energy as the original signal, and therefore the envelope has less maximum power than the harmonics. It also includes the background noisy component in the estimation, which we don't want.

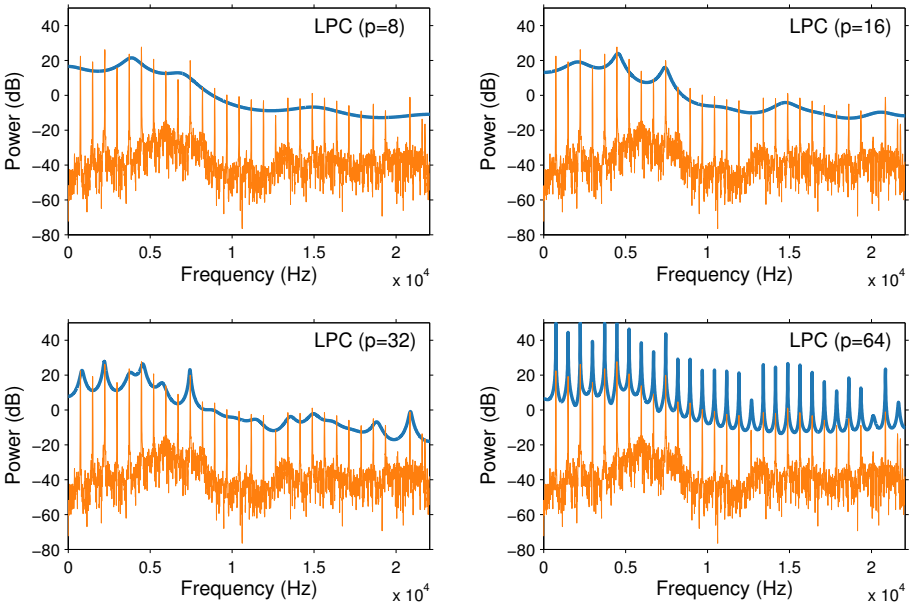

Figure 1: Envelope estimation using LPC of varying order.

Max-pooling

In order to get something that approximates the envelopes of the *spectral peaks* rather than the spectrum as a whole without having to tune a peak picker, we

devised the following algorithm:

1. We compute the power spectrum in dB.
2. Using a sliding window slightly wider than the distance between peaks, we substitute each bin by the maximum found inside the window.
3. The resulting envelope is smoothed with a gaussian kernel to get rid of the discontinuities between partials, using a window width related to the distance between peaks.

Since in this recordings an approximate pitch is known beforehand, it is trivial to get the expected inter-peak distance.

Notice that if the window is too narrow the algorithm picks the individual peaks, as happened with LPC when the order was too large, and if the window is too wide there are flat regions that correspond to the largest peak in the neighborhood.

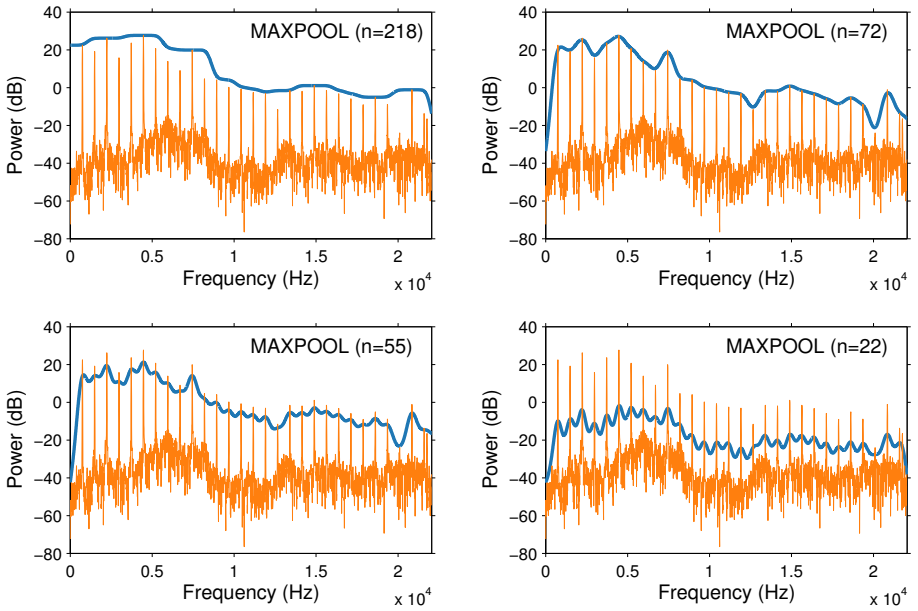

Figure 2: Envelope estimation using MAXPOOL of varying window length.

Peak picking with polynomial interpolation

Finally, we tried an algorithm that finds the individual spectral peaks explicitly and then interpolates a generic function between them. The peak picker was based on a minimum inter-peak distance, set similarly to the max-pooling window length, and on a minimum energy threshold.

Since the spectrum is exponentially decaying in high frequencies, we applied a blind pre-whitening of the signal. An exponential decay in magnitude corresponds to a linear decay in dB, so we removed the linear trend component from the overall spectrum. Then, we shifted it so that the median (which is close to the average noise floor energy, since peaks represent a small number of the overall bins) was at 0 dB. From this, meaningful threshold made sense: we only keep peaks 10 dB above the average noise energy.

There are multiple algorithms that fit more adequate models to the peaks, such as (Röbel and Rodet 2005); it would be interesting to see how well it performs in future research. In the figure, we used 3rd order polynomial interpolation.

In the project, however, we used the MAXPOOL estimator for its computational simplicity.

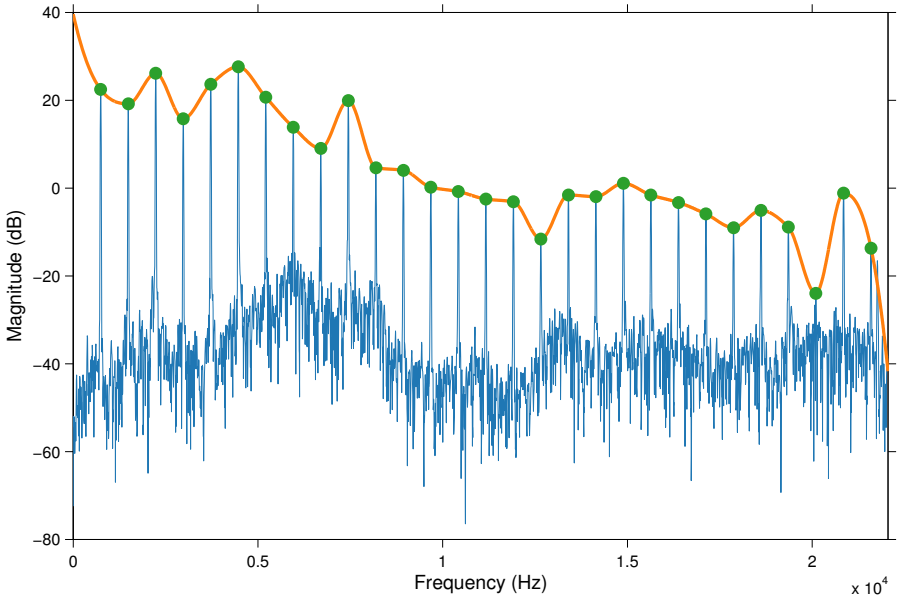

Figure 3: Envelope estimation using explicit peak picking

Spectral whitening

From a source-filter perspective, the sound of a violin can be characterized as:

1. An excitation signal coming from the self-sustained oscillation at the bow-string contact, which depends on the bowing parameters.
2. An LTI filter given by the frequency-dependent losses of the string.
3. An LTI filter given by the fact that the bridge is not completely rigid.

4. An LTI filter given by the transducer, be it a pickup attached to the filter, the body radiation plus a microphone, etc.

If we use statistical tools to study the relationship between bowing parameters and spectral envelope and we include sounds from different violins or on different strings, we will get noisy results because of that. If the system is unsupervised, we may even catch the differences in envelope due to the instrument or the string rather than to the bowing parameters themselves, because the change can be quite notable.

In the first part of this report, we used a linear regression and the median of the spectrum to help balance the effect of these filters in order to extract peaks. While that had the advantage of not requiring anything beside the spectrum envelope itself (it was a completely blind whitening), it is possible to get better results by introducing some prior knowledge.

Deconvolution-based whitening

A very sound approach from a theoretical perspective is to identify these filters independently of the recordings, invert them, and perform a deconvolution. One of the points where we had easy access to was what happens at the bridge.

At the bridge, an incoming force wave from the string results in part of the energy moving the body and, along with it, the body of the instrument. This energy component is what makes the body plates vibrate and radiate energy into the air, so it is crucial in the violin emanating sound. The remaining energy is reflected back to the string, and is what allows the movement to have a stable oscillation period. This energy scattering is frequency-dependent, and depends on the modes of vibration of the violin body.

It is possible to characterize the transfer function of this filter by inputting a force into the bridge with a hammer, measuring it with an attached accelerometer, and measure the bridge velocity with a laser Doppler vibrometer. This method is described in one of the sections of (E. Maestre, Scavone, and Smith 2013).

In the log-spectral domain, convolutions convert into sums, and therefore inverse filtering can be implemented as a simple subtraction (Oppenheim, Schafer, and Stockham Jr 1968).

We tried this, but the results were not very promising. The spectrum indeed changes, but there are still big effects due to other components such as string losses or the pickup transfer function. Actually, the deconvolved spectrum was flatter than the original – but it looked even less like an ideal sawtooth-like wave, which is the ideal bowed string motion.

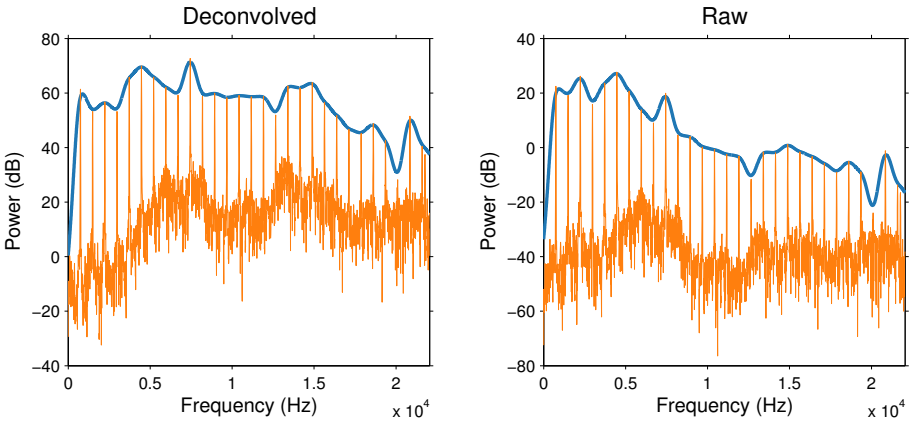

Figure 4: Raw and whitened spectral envelopes by inverse filtering the measured driving-point admittance at the bridge.

Data-driven whitening

Another approach is to take all the samples where the underlying filter is assumed constant, such as notes played on the same violin and on the same string, compute the average spectral envelope in the log-spectral domain for that set, assume that is the filter to remove, and deconvolve it by subtraction, as first suggested in (Otis and Smith 1977) for geophysical applications and then applied to audio restoration in (Stockham Jr, Cannon, and Ingebretsen 1975). Stockham suggested averaging lots of recordings to get the playback medium filter, and since he did not have many, he averaged different parts of the recordings.

In our case, the notes were all played on the same string and by the same player, but on 3 different violins, so we had 3 different whitening groups.

Analysis of spectral envelopes

Once we had the spectral envelope for the notes in the database, both raw and whitened, we ordered the notes according to the magnitude of different bowing descriptors and plotted the spectral envelopes in that order to check for differences.

Since the bow parameter sampling was not uniform but rather score-driven, the number of spectra in a given range of bowing parameter values may be grater than in another range. To account for that, we re-sampled the resulting 2D matrix of ordered spectra into a linearly increasing bow parameter magnitude scale.

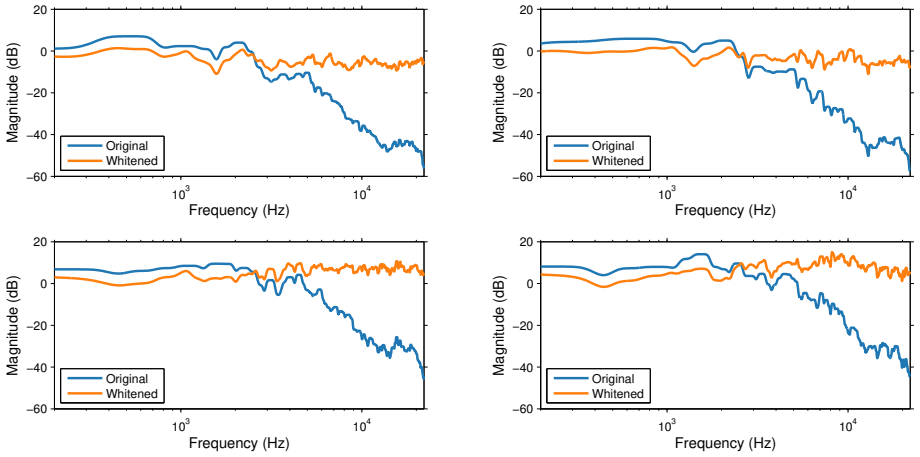

Figure 5: Raw and whitened spectral envelopes by group-wise mean subtraction in the log-spectral domain.

Figure 6 shows these spectral matrices for both raw and whitened spectral envelopes. The matrices have been smoothed using a gaussian kernel to remove artifacts and prevent spatial aliasing when rendering them in a small screen, since they originally had thousands or rows and columns. The histograms at the top show the original parameter distribution; only values above the 1th percentile and below the 99th percentile are interpolated and displayed. The colors have been chosen so that each image makes use of the full scale available after removing outliers, defined as values below the 1th percentile of above the 99th percentile of the overall image.

In the raw spectra version, the overall spectral shape (with less energy at high frequencies) dominates and it is hard to tell differences apart. Moreover, in this picture there are spectra from different violins, and it could be that the difference due to the violin itself is larger than the difference due to the gestures in some cases. This is not apparent because of the gaussian smoothing, but contributes to hiding useful features.

In the whitened plots, it becomes obvious that all bowing parameters have an effect on the overall energy (velocity and force with positive slope and bow-bridge distance with negative), as well as some frequency-dependent effects.

Non-negative matrix factorization of log-spectral domain filters

While this whitened spectrum visualization provides visually discriminable features when comparing different bowing gestures, it would be convenient to have

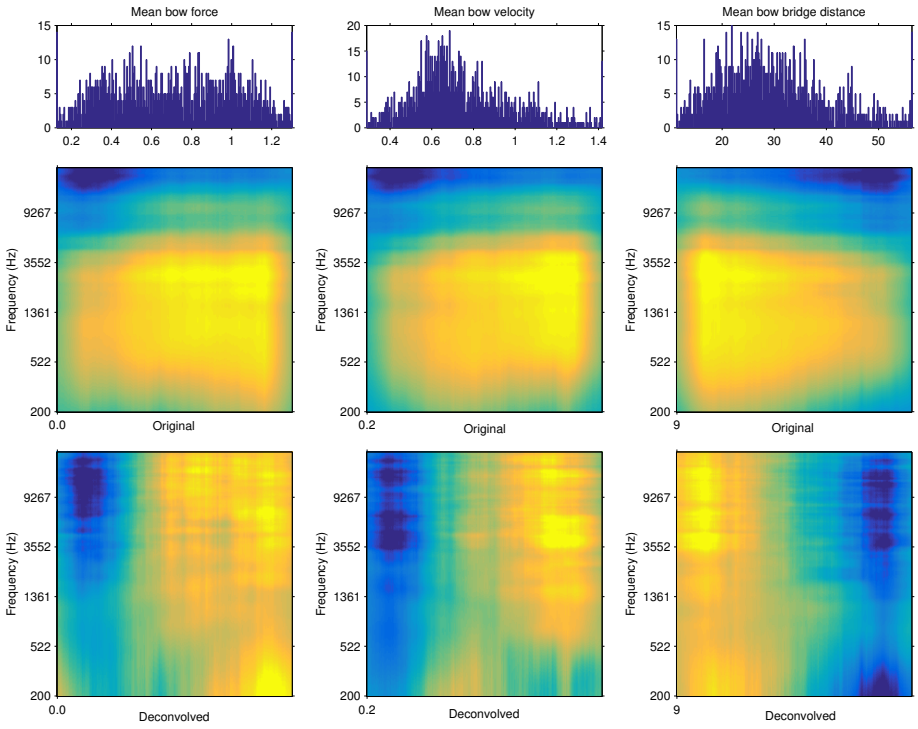

Figure 6: Raw and whitened spectral envelopes for different bowing gestures.

something more explicit that captures more subtleties in the form of a few, more orthogonal descriptors.

As mentioned before when discussing whitening techniques, a sound passed through a filter chain can be represented as an excitation plus the sum of the transfer functions of the filters in a log-spectral domain such as the spectral energy in a dB scale.

If we assume that the spectral modifications due to changes in bowing parameters can be decomposed as a chain of elementary filters that have more or less weight depending on the specific gesture, we can find an optimal decomposition of the measured spectral envelops as a linear combination of these elementary filters in the log-spectral domain.

If we had as many filters as spectral bins, then we would achieve perfect reconstruction; we seek to find a much smaller number of filters that preservers as much of the original envelope as possible. This problem has traditionally been approached using Principal Component Analysis, but in our case there is an additional constraint: the coefficients that determine how much each filter weights into creating a given envelope cannot be negative. While addition in our domain corresponds to convolution, subtraction would correspond to deconvolution, which we do not want.

Therefore, we chose to use Non-negative Matrix Factorization (NMF), which decomposes a matrix with as a product of two "narrow" matrices with positive entries. If the original matrix contains an observation at each column from a number of variables, the two outputs can be thought of as a dictionary of "templates" and their activations that produce the input by a linear combination of them.

We took this approach for simplicity, although similar frameworks exist in the literature and have further refinements (Liang, Hoffman, and Mysore 2013).

In order to have everything positive, we shifted the whitened spectrum so that the minimum is near 0 dB.

Each one of the spectral envelopes from the notes we analyzed can be approximated as a linear combination of the templates. Therefore, the envelopes can be parametrized by the two weights, and we can get a much more compact representation.

If we compare the activation weights in different bow-bridge distances, we see first of all that the distribution of the weights themselves is not very symmetric; the orange template (high-pass) spans a much larger coefficient range than the blue one. Therefore, the orange template accounts for most of the variance in spectral envelope, and the blue template is something more like a constant offset.

In this case, the blue template seems to have a slight negative correlation with bow-bridge distance, which would indicate that at high bow-bridge distance the low frequency components have less energy.

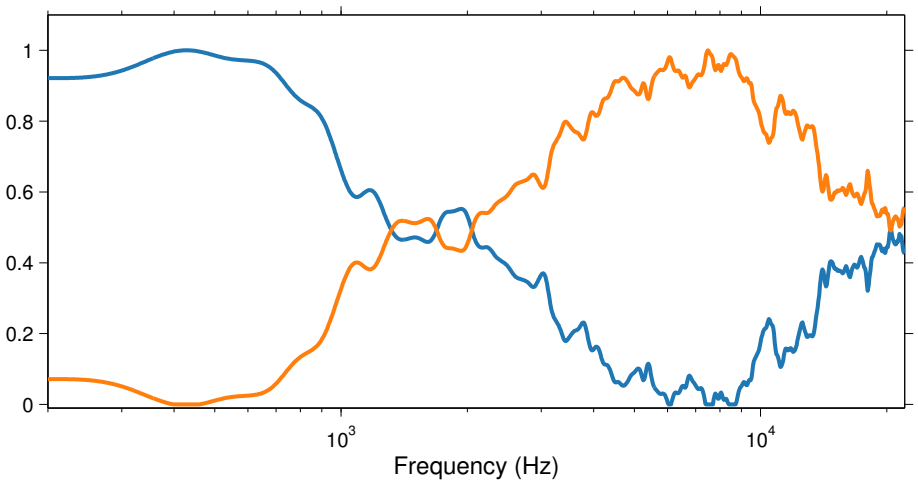

Figure 7: NMF templates learnt from the data.

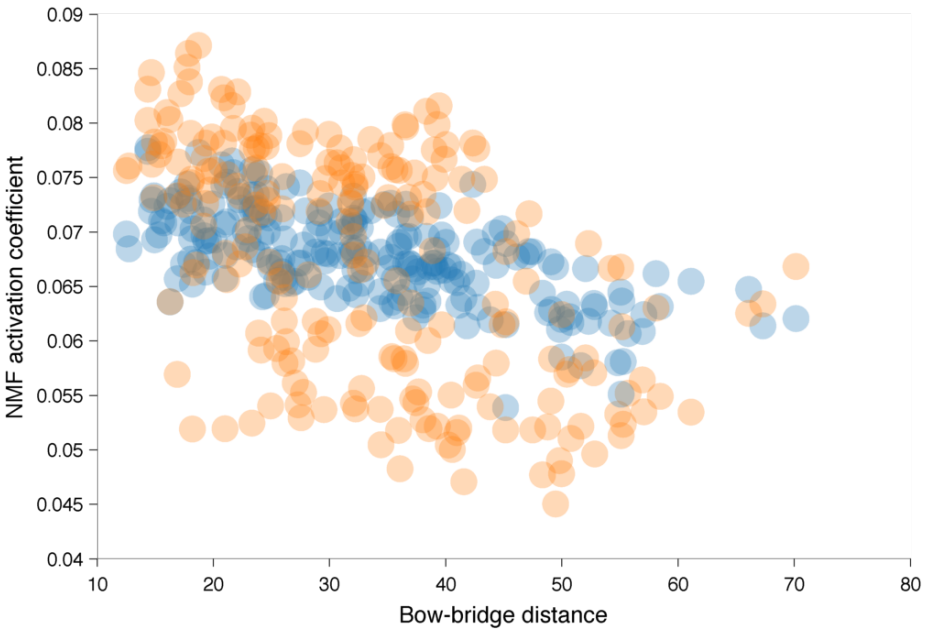

Figure 8: NMF template activations for different bow-bridge distances.

By plotting against bow velocity we see that there is again some correlation, although not as clear and this time positive.

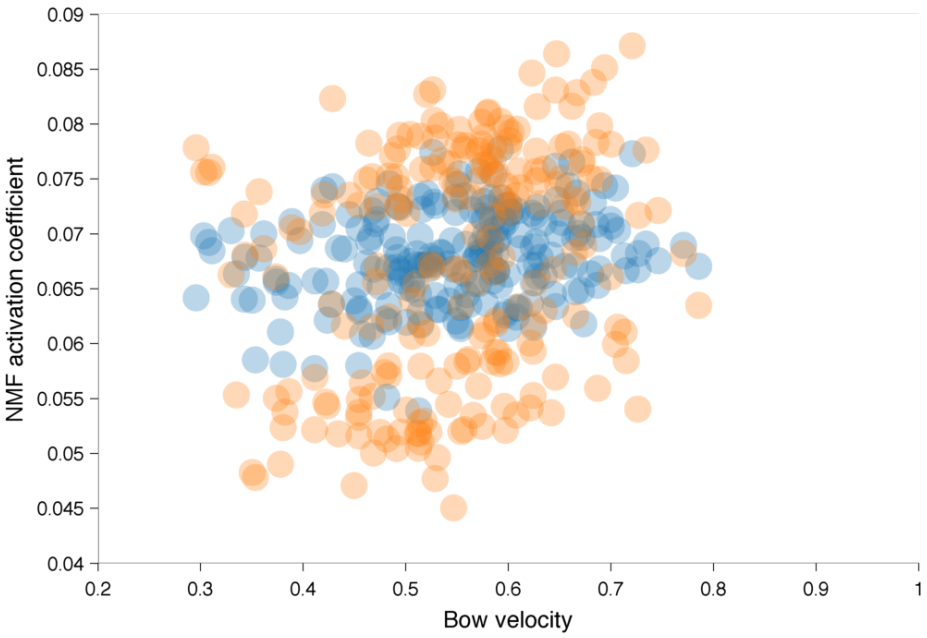

Figure 9: NMF template activations for different bow velocities.

The same kind of analysis but focused on bow force reveals a much stronger correlation than bow velocity. The orange template displays some bi-modality, suggesting that the sound changes a lot on either sides of the force threshold 0.8; this could be due to errors in the force estimation method.

It is possible to go even further by analyzing two bowing parameters at the same time. In this case, we plot on a 2D projection of the gesture space points with the color tone indicating which template is more prominent, and with the point size indicating the total energy represented by the sum of the two activations.

This projection of features on a log-log space containing bow force and bow-bridge distance is a very well-known representation of bowed string behavior, the Schelleng diagram (Schelleng 1973). In the diagram, good-sounding oscillatory regimes only appear in a triangular region, similar to the one that shows up in the figure.

In his paper, Schelleng explains that the left part of the triangle corresponds to brighter sounds, and the right part to softer. Similarly, the part closer to the top boundary corresponds to louder sounds. Some correlations with that can be seen on our data, where in the left and upper parts there is more energy overall and the high-pass template dominates.

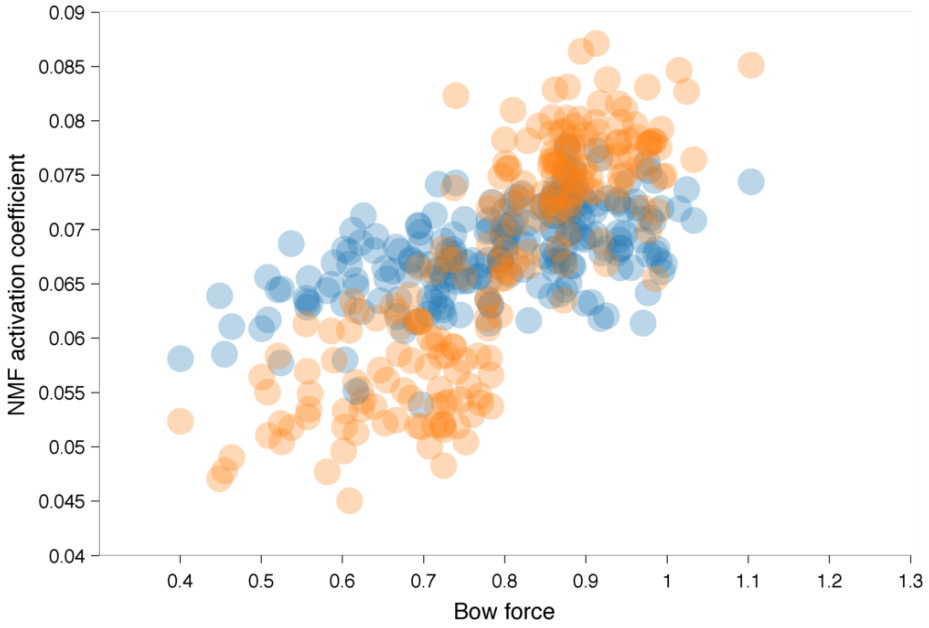

Figure 10: NMF template activations for different bow force values.

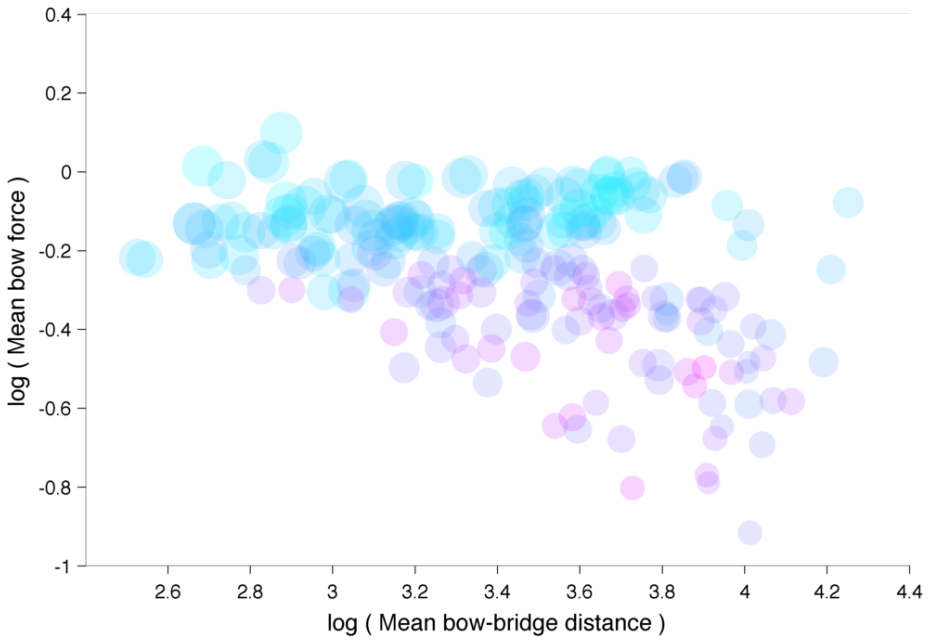

Figure 11: Ratio (color) and sum (size) of NMF template activations on the Schelleng space.

Outlook

In this project, we have devised a method for the extraction of the spectral envelope of violin notes, by comparing LPC, max-pooling and peak picking with polynomial interpolation. Max-pooling seems to be a good trade-off of quality and simplicity for this study. Then, we have deconvolved common filters such as the effect of the instrument body, both by inverse filtering the measured admittance and by blind log-spectral mean subtraction. Mean subtraction works better, because we lack additional information such as the effects of the pickup transducer or the string losses. After visual inspection of spectral envelope differences for different bow gestures, we have extracted two filters that form a basis in the log-spectral domain. By linear combination of these filters, the different envelopes can be re-generated; this combination is equivalent to chaining the two filters by convolution. The filters are respectively low-pass and high-pass, approximately, and some clear correlations show up when comparing their activations to the bow gestures.

Several aspects could be improved, such as:

- **Adaptive chunk selection**: Right now, a fixed portion of the notes is selected for analysis; we could take into account the transient duration of each note.
- **Short-time feature extraction**: Since bow parameters can vary a lot within a note, we could extract multiple short-time frames from them.
- **Better spectral envelope extraction**: As mentioned in the report, there are better methods for spectral envelope approximation of filtered harmonic signals.
- **Improved guided whitening**: If we had more data, such as the bridge force to pickup transfer function, the deconvolution method based on inverse filtering external measurements could work better.
- **More robust spectral descriptors**: The ultimate goal of this project is to be able to predict the spectral envelope from bow gestures; for that, better spectral envelope descriptors are needed.

One of the possibly more difficult aspects of making the system robust will be parameters that affect the spectrum beyond the specific bow parameters we are considering and the deconvolved parts, such as left-hand finger position (shorter string segments have less losses) or left-hand finger pressure (with less pressure there are more losses at that junction).

Bibliography

Atal, Bishnu S, and Suzanne L Hanauer. 1971. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave." *The Journal of the Acoustical Society of America* 50 (2B): 637–655.

Liang, D., M. D. Hoffman, and G. J. Mysore. 2013. "A Generative Product-of-Filters Model of Audio." *ArXiv E-Prints*.

Llimona, Quim. 2014. "Bowing the Violin: a Case Study for Auditory-Motor Pattern Modelling in the Context of Music Performance." Universitat Pompeu Fabra.

Maestre, E., G. Scavone, and J. O. Smith. 2013. "Digital Modeling of Bridge Driving-Point Admittances from Measurements on Violin-Family Instruments." In *Proceedings of the Stockholm Music Acoustics Conference*.

Marchini, Marco, Panos Papiotis, Alfonso Pérez, and Esteban Maestre. 2011. "A Hair Ribbon Deflection Model for Low-Intrusiveness Measurement of Bow Force in Violin Performance." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2011), Oslo, Norway*. Vol. 30. Citeseer.

Oppenheim, Alan V, Ronald W Schafer, and Thomas G Stockham Jr. 1968. "Nonlinear Filtering of Multiplied and Convolved Signals." *Audio and Electroacoustics, IEEE Transactions on* 16 (3): 437–466.

Otis, Robert M, and Robert B Smith. 1977. "Homomorphic Deconvolution by Log Spectral Averaging." *Geophysics* 42 (6): 1146–1157.

Perez, Alfonso, Jordi Bonada, Esteban Maestre, E Guaus, and Merlijn Blaauw. 2007. "Combining Performance Actions with Spectral Models for Violin Sound Transformation." In *Proceedings of the International Congress on Acoustics*.

Röbel, Axel, and Xavier Rodet. 2005. "Efficient Spectral Envelope Estimation and Its Application to Pitch Shifting and Envelope Preservation." In *Proc. DAFx*.

Schelleng, J. C. 1973. "The Bowed String and the Player." *Journal of the Acoustical Society of America* 53: 26–41.

Stockham Jr, Thomas G, Thomas M Cannon, and Robert B Ingebretsen. 1975. "Blind Deconvolution Through Digital Signal Processing." *Proceedings of the IEEE* 63 (4): 678–692.