MAPPING BLOWING PRESSURE AND SOUND FEATURES IN RECORDER PLAYING

Leny Vinceslas, Francisco García, Alfonso Pérez
Music Technology Group
Universitat Pompeu Fabra, Barcelona (SPAIN)
leny.vinceslas@gmail.com
pul.editions@gmail.com
alfonso.perez@upf.edu

Esteban Maestre
Music Technology Group
Universitat Pompeu Fabra, Barcelona (SPAIN)
esteban.maestre@upf.edu
Center for Computer Research in Music and Acoustics
Stanford University, Stanford (USA)
esteban@ccrma.stanford.edu

ABSTRACT

This paper presents a data-driven approach to the construction of mapping models relating sound features and blowing pressure in recorder playing. Blowing pressure and sound feature data are synchronously obtained from real performance: blowing pressure is measured by means of a piezoelectric transducer inserted into the mouth piece of a modified recorder, while produced sound is acquired using a close-field microphone. Acquired sound is analyzed frame-by-frame, and features are extracted so that original sound can be reconstructed with enough fidelity. A multi-modal database of aligned blowing pressure and sound feature signals is constructed from real performance recordings designed to cover basic performance contexts. Out of the gathered data, two types of mapping models are constructed using artificial neural networks: (i) a model able to generate sound feature signals from blowing pressure signals, and therefore used to produce synthetic sound from recorded blowing pressure profiles via additive synthesis; and (ii) a model able to estimate the blowing pressure from extracted sound features.

1. INTRODUCTION

Studying and modeling instrumental music performance can be considered among the most challenging topics in the field of computer-aided analysis and generation of musical sound. In instrumental music practice, the performer transforms a musical score into control parameters used for driving sound production. Indeed, modeling how instrumental control signals are related to sound production appears as a pursuit with different application possibilities (e.g. performance, sound synthesis).

Regarding sound synthesis, the naturalness of synthetic sound greatly relies on how input controls are represented and mapped into sound features, both for the case of physical models or in spectral-domain sample-based frameworks. In general, physical models suffer from a lack of appropriate input control parameters while systems combining spectral modeling with pre-recorded samples do not explicitly consider instrumental gestures [1].

When aiming at low-intrusiveness measurement of instrumental gesture parameters, the idea of inferring instrumental control parameter signals from recorded sound appears as an attractive alternative. Indirect acquisition of instrumental gesture parameters in music playing represents a broader access to computationally studying music performance.

In this paper we present our recent work on sound-instrumental gesture mapping in recorder playing. Mapping models relating sound features and blowing pressure are constructed from real data acquired in practice scenario, by means of a low-intrusiveness pressure acquisition system and a close-field microphone. On one hand a first mapping model able to generate sound features from blowing pressure is used to generate synthetic recorder sound from blowing pressure profiles using additive synthesis. On the other hand, a second mapping model is able to estimate the blowing pressure from features extracted from acquired recorder sound.

The rest of the paper remains as follows. Section 2 provides background and related work. In Section 3 we outline data acquisition, while Section 4 presents a spectral-domain representation used for analysis and re-synthesis of recorder sound. The construction and evaluation of mapping models is outlined in Section 5. Finally, Section 6 concludes by summarizing results and pointing out future directions.

2. BACKGROUND

Within the woodwind instrument family, the recorder can be considered as one of the simplest instruments, mainly due to the fact that its geometry fixes a number of input control parameters that for other instruments of the family are controllable by the performer (like for instance the distance between the mouth and the labium, which are not fixed for the transverse flute). Although the influence of the configuration of the mouth and the vocal tract has traditionally been a very controversial issue to which research efforts have been devoted [2], we assume here that only two control parameters enable the performer to achieve sound modulations during performance: the speed of the air jet (derived from the blowing pressure), and the fingering. Each fingering introduces distinct set of minima in the acoustic impedance that correspond to resonances. Blowing pressure, as opposed to fingering, presents a continuous nature while allowing the control of dynamics, timbre and fundamental frequency. By adjusting his blowing pressure, the performer is also able to select which impedance minimum controls the oscillation, and so cans reach different notes while keeping the same fingering. In this work, we will focus on blowing pressure in the first oscillation mode assuming a fixed fingering.

During performance, blowing pressure is exponentially related to fundamental frequency [3], and linearly related to dynamics, as it has been shown also for the transverse flute [4, 5]. With regards to timbre in particular, Fletcher [4] pursued a study of the amplitude and of the harmonics as a function the intensity of pressure exerted by musicians, and showed two main phenomena (both related and expressed qualitatively): (i) the amplitudes of the second

and third harmonics, as related to the amplitude of the fundamental harmonic, increase with blowing pressure; and (ii) the high frequency content grows with blowing pressure. In this work we are aiming at providing model able to represent such relations in a flexible manner.

In general, there exist two main approaches for the acquisition of instrumental gesture parameters from real performance: direct acquisition and indirect acquisition. While direct acquisition methods are devised to measure the actual physical parameter from sensors conveniently placed in the instrument (and thus providing relatively straightforward and reliable measuring capability), indirect acquisition is based on the analysis of acquired sound signals and the extraction of parameters from which the actual instrumental control parameters used for produced such sound [6]. Given that intrusive measurement setups often keep performers from playing naturally for a prolonged period of time, the idea of being able to estimate instrumental gesture parameters (in this case blowing pressure) represents a motivation for our modeling approach to data-driven estimation of blowing pressure from sound features.

Conversely, it is also possible to find a mapping function that goes from control gestures to sound features. A first study [7] shows how statistical methods for non-linear prediction can be used for sound synthesis. A probabilistic inference technique (cluster-weighted modeling) is used to map performance controls of a bowed monochord to the amplitude of the first 25 harmonics of the sound.

A similar approach is used in [8] where a timbre model based on neural networks is presented. The neural networks are able to predict harmonic and residual spectral envelopes corresponding to violin bowing controls. The model is coupled to an additive synthesizer that fills the envelopes with harmonic and noisy content in order to produce synthetic violin sounds. This method has been successfully implemented for real time synthesis of violin sounds [9] where the spectral model is driven by bowing parameters captured using an ad-hoc control interface.

This approach for generating synthetic instrument sounds offers the possibility of substituting the habitual physical models by trained generative models, resulting particularly attractive for excitation-continuous musical instruments, such as bowed strings or wind instruments.

3. DATA ACQUISITION

Basically, data acquisition consisted in the synchronous acquisition of blowing pressure and sound signals from a real recorder practice scenario. A set of scripts was designed in order to cover a number of performance contexts when constructing a multi-modal database including aligned sound feature signals and blowing pressure signals. The scripts consisted on (i) a series of scripts covering different fingerings, articulations, note durations, and dynamics; and (ii) recording repetitions of a long crescendo (increasing blowing pressure) for each fingering. While sound was acquired by means of a close-field microphone attached to the body of the instrument, a special measurement system was designed for acquiring blowing pressure while maintaining a low impact on the performer's comfortability.

The mouthpiece block of an alto recorder was modified by enabling the connection of a pressure sensor without altering the timbre of the original instrument, as described in [10], leading to a significant reduction of the intrusiveness as compared to using a plastic catheter in the mouth of the musician [5, 11].

4. SOUND ANALYSIS

Acquired recorder sound is analyzed in the spectral domain with the aim of extracting a lower-dimensionality sound feature representation that enables (i) training a model able to learn mapping functions between extracted spectral parameters and blowing pressure, and (ii) to use such lower-dimensionality representation for reconstructing the original sound with enough fidelity so that those parameters can be used for synthesis purposes.

4.1. Some observations on the spectrum of the recorder sound

It was a good exercise for us to observe the time-varying spectra of a number of the crescendo recordings in the database, along with blowing pressure and fundamental frequency. Apart from the expected correlation between fundamental frequency and blowing pressure (as well observed in all the other recordings) [3], the overall amplitude of both harmonic and stochastic components increases with blowing pressure, having mid-range harmonic components to raise more prominently. Up to the expected change of oscillation mode starting to take place at higher pressure values (and in agreement with [12]), a differentiated behavior is observed for odd and even harmonic components, as it happens in other instruments of the same family [4]. This behavior leads to foresee a representation of odd and even harmonics independently. Therefore, we decided to represent each sound frame by the fundamental frequency, the noise floor, and the odd and even harmonic components. In this work we did not focus on higher modes of oscillations or chaotic, multiphonic transition states between oscillation modes.

4.2. Spectral-domain representation

In a first step, the audio signal is divided into overlapping frames and, for each of the frames, it is decomposed into deterministic (harmonic) and stochastic (residual) components via the spectral modeling synthesis (SMS) technique [13]. The algorithm implementation is based on a spectral peak-picking and tracking algorithm making use of the two-way mismatch (TWM) fundamental frequency estimation algorithm [14]. Given that notes have been segmented and fingerings (pitch values) have been annotated, octave errors are easily avoided by restricting the possible values of fundamental frequency candidates.

In the past, it has been shown that directly using harmonic amplitudes and frequencies in a timbre modeling pursuit leads to a highly complex and often inaccurate model [7], so with the idea of representing each frame by a reduced number of parameters, spectral envelope approximation is applied in a second step to each of the two extracted components. Given (i) our observations on the harmonic envelopes of recorded sound, (ii) the different modeling errors obtained during spectral envelope prediction (see next Section), and (iii) perceptual quality of obtained sound, it resulted best for our modeling and re-synthesis purposes to split the harmonic component into two different subcomponents: the odd harmonic envelope and the even harmonic envelope.

The next step is to approximate the spectral envelopes into consideration. Aiming at a compact representation of the harmonic spectral envelopes, harmonic peaks of each of the two sets (odd and even) are first interpolated by using a piecewise Hermite function. As for the stochastic spectral envelope (noise floor), a similar procedure is carried out starting from spectral peaks. Two example of the three envelopes are represented in Figure 1.

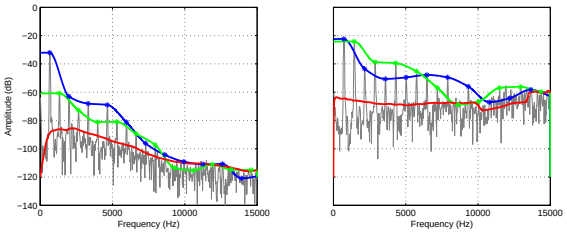

Figure 1: Spectral evolution of an E5 crescendo with its odd harmonic envelope (BLUE), even harmonic envelope (GREEN) and residual envelope (RED). Left: resulting spectrum for a blowing pressure of 197 Pa. Right: resulting spectrum for a blowing pressure of 1513 Pa.

Finally, each of the interpolated envelopes is encoded as a set of 15 mel-frequency cepstral coefficients (MFCCs), which for our experiments resulted to be a good compromise between size and fidelity. Out of the encoded three envelopes plus the value of fundamental frequency, it is possible to satisfactorily reconstruct the original sound via additive synthesis.

5. MAPPING MODELS

From encoded spectral envelopes and acquired blowing pressure signals, two frame-by-frame mapping functions are built for each fingering by means of artificial neural networks, using Matlab's neural network toolbox functions `newcf` and `newff`. The first model translates blowing pressure into sound features (spectral shape plus fundamental frequency), while the second one realizes the inverse mapping. Models are evaluated by means of 10-fold cross-validation. In both cases, separating harmonic envelopes into odd and even components led to obtain lower prediction errors; therefore our decision (also based on early observations by Fletcher [12]) to split harmonics and work from separated envelopes.

5.1. Mapping blowing pressure to sound features

A generative spectral model predicts, from the blowing pressure signal given in a frame-by-frame fashion, the three spectral envelopes (odd harmonics, even harmonics and residual) and fundamental frequency required to reconstruct sound. The prediction is carried out using two neural networks working in parallel. A first two-layer cascade-forward backpropagation network takes as input the blowing pressure and its derivative, and predicts the fundamental frequency plus the 15 MFCCs corresponding to each of the three spectral envelopes (2 inputs to 46 outputs). This network is trained with a subset of the training dataset which is composed of frames with harmonic structure. Since this model provides a coherent prediction for harmonic frames, an additional binary-output network (three-layer feed-forward backpropagation, 2 inputs to 1 output) capable of predicting harmonic state (whether a frame is voiced or not) was used as an activation to the first network. This last network was trained using the full dataset.

The results of spectral envelope prediction showed a highly robust and reliable performance, reaching a averaged correlation coefficient of 0.95 in the MFCC space. In Figure 2 it can be observed how predicted envelopes (dashed lines) almost match actual envelopes (continuous lines) for one frame of the signal. With respect to the prediction of fundamental frequency, in some cases we

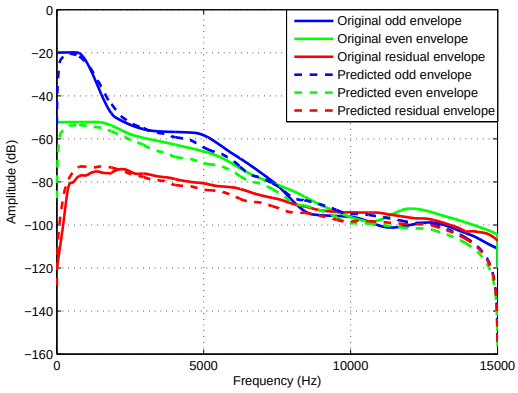

Figure 2: Example of spectral envelope prediction for a frame of sound in E5.

observed a maximum absolute error of 5Hz for some fingerings in the lower register. An averaged correlation coefficient of 0.83 was obtained when evaluating the performance of the second network. A preliminary additive synthesizer was implemented. It takes as input the note durations and fingerings from a segmented performance of a known score, together with the blowing pressure profiles from a previous recording of the same score. The synthesizer predicts envelopes and fundamental frequency and generates sound by means of additive synthesis: the harmonic sequence is generated from the fundamental frequency, partial amplitudes are set to the values in predicted harmonic spectral envelopes, and finally the residual component is generated as white noise filtered in the spectral domain by attending to the predicted residual spectral envelope. Although produced sound is perceived as very realistic, when subjectively comparing it with original sound, transients show in general a worse behaviour, mainly because only harmonic frames were used for the training. Example sounds can be listened on-line[1].

5.2. Estimating blowing pressure from sound features

The second model is able to estimate blowing pressure value from the MFCCs and fundamental frequency extracted for a given frame. A two-layer cascade-forward backpropagation network is trained with data from analyzed frames (46 inputs to 1 output) of each fingering in the database, leading to one different model per fingering. An averaged correlation coefficient of 0.91 was obtained for all fingerings. Figure 3 shows a prediction example: it is compared a recorded blowing pressure profile (red line) with the predicted signal (dashed blue line) of blowing pressure of one of the recordings, legato articulation. It can be observed that predicted signal suffers from fluctuations, however general shape is preserved and types of articulation can be identified.

6. CONCLUSION

We have presented the construction of mapping models relating sound features and blowing pressure in recorder playing. The

[1] `http://ccrma.stanford.edu/~esteban/recsyntspec.html`

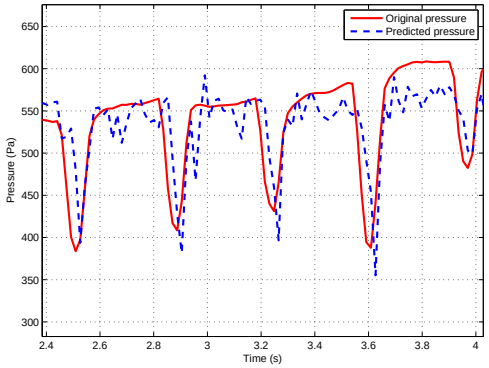

Figure 3: Comparison between the real blowing pressure (RED) and the predicted blowing pressure (BLUE) for E5.

models are built using artificial neural networks, which have been trained from blowing pressure and sound features extracted from real performance data. Blowing pressure is acquired by means of a low-intrusive-ness pressure acquisition system based on a mechanical modification of the block of an alto recorder, to which a pressure sensor was attached in a low-invasive setup that allowed to play naturally. A first mapping model is trained to generate, given a blowing pressure signal, spectral envelopes (odd and even harmonic, plus noise floor) and fundamental frequency, so that sound can be faithfully synthesized via additive synthesis. A second model is able to estimate the blowing pressure from audio attributes extracted from acquired recorder sound. Both models showed a fairly good performance. For the case of sound feature predictoin, an additive sound synthesizer was implemented as a proof-of-concept system able to generate realistic sound from blowing pressure and fingering information. Sound quality was evaluated only qualitatively by the authors and collaborators, but more formal subjective studies are to be carried out along with a number of improvements.

Results on spectral-domain sound generation from instrumental gesture controls shed light on the idea of using gesture-driven spectral models for enhancing sample-based sound synthesis, or even in real-time applications as in [9] where a Max/Msp allowed a real-time synthesis from a virtual instrument interface. The spectral envelope representation based on separating odd and even harmonics as two different envelopes showed to be a good approach for this instrument. Regarding Some improvements need to be carried out, especially for the case of transient modeling, for which the performed worse: time-domain signal analysis and synthesis from blowing pressure could result to be crucial, or even a hybrid model combining simple physical models (for transient parts) and spectral models (for steady-state parts). Another possibility would be to use this synthesis framework with automatically generated gesture parameter signals, as it was carried out in [1] for the case of violin bowing.

Preliminary achievements obtained for blowing pressure estimation showed promising possibilities for indirect acquisition from microphone in real performance contexts, leading to the easier construction of multi-modal databases more suited to study higher-level playing styles. Automatic detection of fingering information [15] is a very interesting addition to future developments of the system.

7. ACKNOWLEDGMENTS

We are extremely grateful to Luthier Josep Tubau for his great help in designing the recorder, and to recorder performance Professor Joan Izquierdo who kindly collaborated during the recordings.

8. REFERENCES

[1] E. Maestre, *Modeling instrumental gestures: An analysis/synthesis framework for violin bowing*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2009.

[2] J. Chen, D. Laurin, J. Smith, and J. Wolfe, "Vocal tract interactions in recorder performance. 19 th," in *International Congress on accoustis Madrid*, 2007, pp. 1–6.

[3] J. Martin, *The acoustics of the recorder*, Moeck, 1994.

[4] NH Fletcher, "Acoustical correlates of flute performance technique," *Acoustical Society of America*, vol. 57, no. 1, 1975.

[5] N. Montgermont, B. Fabre, and P. de la Cuadra, "Flute control parameters: fundamental techniques overview," in *International Symposium on Music Acoustics*, 2007.

[6] 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*. Citeseer, 2003, pp. 42–47.

[7] B. Schoner, C. Cooper, C. Douglas, and N. Gershenfeld, "Data-driven modeling of acoustical instruments," *Journal of New Music Research*, vol. 28, no. 2, pp. 81–89, 1999.

[8] A. Perez, *Enhancing spectral synthesis techniques with performance gestures using the violin as a case study*, Ph.D. thesis, 2009.

[9] A. Perez and J. Bonada, "The bowed tube: a virtual violin," in *Proceedings of the 2010 conference on New interfaces for musical expression*, 2010.

[10] F. García, L. Vinceslas, J. Tubau, and E. Maestre, "Acquisition and study of blowing pressure profiles in recorder playing," in *Proceedings of the 2011 conference on New interfaces for musical expression*, 2011.

[11] P. de la Cuadra, B. Fabre, N. Montgermont, and C. Chafe, "Analysis of flute control parameters: A comparison between a novice and an experienced flautist," *Acta Acustica united with Acustica*, vol. 94, no. 5, pp. 740–749, 2008.

[12] NH Fletcher and L.M. Douglas, "Harmonic generation in organ pipes, recorders and flutes," *J. Acoust. Soc. Am*, vol. 68, no. 3, pp. 767–771, 1980.

[13] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[14] R.C. Maher and J.W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[15] 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*, vol. 127, pp. 534, 2010.