TOWARDS DIFFERENTIABLE PIANO SYNTHESIS BASED ON PHYSICAL MODELING

Hans-Ulrich Berendes[1] **Simon Schwär**[1] **Maximilian Schäfer**[2] **Meinard Müller**[1]
[1]International Audio Laboratories Erlangen, Germany
[2]Institute for Digital Communications, FAU Erlangen-Nürnberg, Germany
simon.schwaer@audiolabs-erlangen.de

ABSTRACT

We explore the concept of combining physical modeling of the piano with deep learning using methods from differentiable digital signal processing. The core of our proposed approach is a modal synthesis model for the piano string, which is combined with a linear filter to approximate the acoustic properties of a grand piano. In a preliminary experiment, we train a neural network to estimate an excitation signal for a string in an autoencoder setting and show that the system can match the spectral content of a given target note. Our differentiable piano model could be utilized in a multitude of music processing tasks, including sound matching, signal enhancement, or source separation.

1. INTRODUCTION

Differentiable Digital Signal Processing (DDSP) [1] provides a flexible toolkit to include domain knowledge into the design of machine learning models for audio synthesis. While general-purpose methods like spectral modeling synthesis (SMS) [2] make it possible to reproduce a wide range of sounds, they can also generate unrealistic outputs and artifacts when being used for specialized tasks like piano synthesis. Previous work has focused on controlling the parameters of SMS to avoid such issues [3], enabling to match the sound characteristics of piano recordings given aligned MIDI annotations [4]. More physically inspired synthesis models allow for a better control of the possible outputs and thereby can introduce stronger inductive bias to the system. This has recently proven useful, e.g., for unsupervised source separation of multiple voices in a cappella ensemble music [5], or for singing voice reconstruction from mel-spectrograms [6].

In this work, we propose to replace the *sinusoidals-plus-noise* synthesis of SMS with a differentiable and specialized model for piano synthesis, which is inspired by physical modeling [7]. In particular, we use modal synthesis to model the vibrating strings of a grand piano with a bank of second-order IIR filters and emulate the sound radiation characteristics of the instrument with a subsequent FIR filter. In an initial experiment, we use pre-defined parameters for these filters and train a neural network to generate the *excitation signal* corresponding to the force that the hammer exerts on the strings when a key is pressed, showing that we can match the spectral content of a known reference excitation signal with this method. Audio examples can be found on a supplemental website [1].

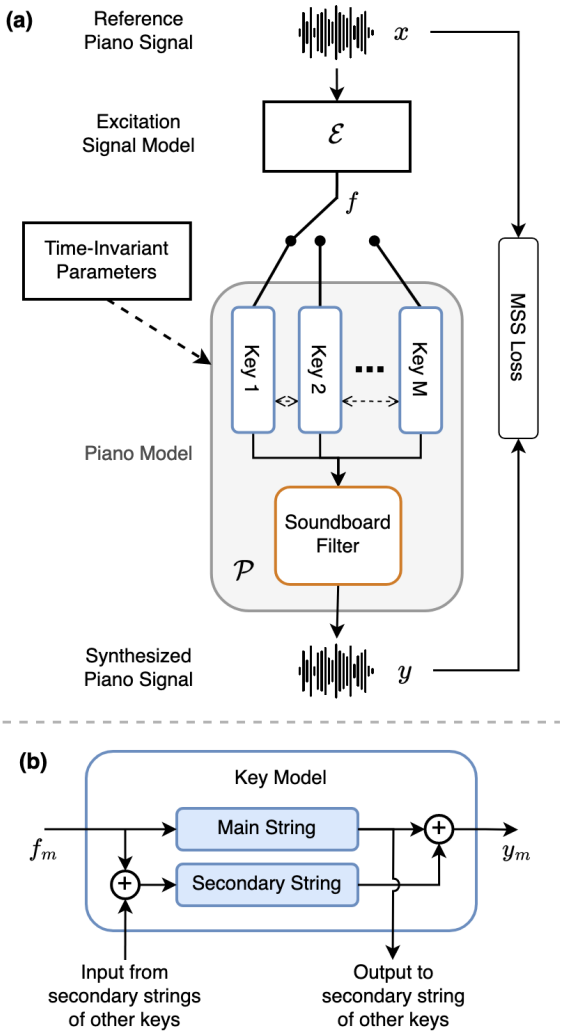

Figure 1. Overview of the differentiable piano model.

[1] https://audiolabs-erlangen.de/resources/MIR/2023-piano-synth-lbd

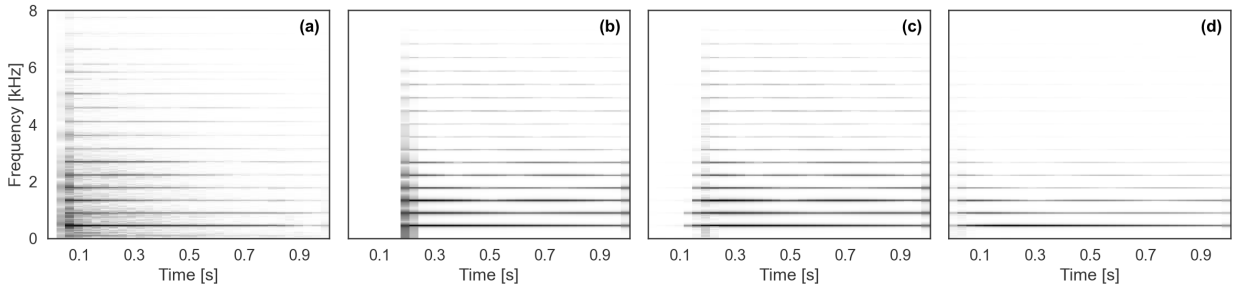

Figure 2. Results of the excitation learning experiment for an A4 (440 Hz) visualized as spectrograms. **(a)** Recording of a real piano tone. **(b)** Model output using the reference excitation signal. **(c)** Model output using the learned excitation signal using the signal from (b) as input. **(d)** Model output using a different reference excitation signal.

2. DIFFERENTIABLE PIANO MODEL

For the design of the piano model, we aim for a balance between perceptual accuracy and computational simplicity. The system consists of an *excitation signal model* $\mathcal{E}$ and a *piano model* $\mathcal{P}$, which itself contains M *key models* and a *soundboard filter*, as shown in Fig. 1a.

The task of $\mathcal{E}$ is to emulate the function of the hammer that excites the corresponding strings when a certain key of the piano is pressed. Since an accurate physical description of the hammer–string interaction is a difficult task [8], we rely on a perceptual modeling approach (see Section 3). The output signal of $\mathcal{E}$ is routed to the respective key model depending on the played note.

The model for a single key (shown in Fig. 1b) comprises two separate banks of parallel second-order IIR filters, the *main* and *secondary string* [7], which are excited by the output signal of $\mathcal{E}$. Each individual filter represents a transversal mode of vibration of a physical piano string, with one second-order filter for each mode below the Nyquist frequency. While the individual filter coefficients may also be learned, we calculate the filter coefficients for our experiment from the known or estimated physical properties of the strings and keep them fixed. With slightly different parameters for the secondary string, the key model can efficiently replicate many of the perceptually relevant effects of using multiple strings for a single key, including beating and the characteristic two-stage decay [9]. Furthermore, the M keys are interconnected to emulate the effect of *sympathetic resonance* [10], i.e., string vibrations induced by striking neighboring keys.

The output signals from each key are summed and fed to the soundboard filter. This FIR filter models the (approximately linear) radiation characteristics of the piano, which can further be combined with the transfer function between instrument and a microphone, possibly including room acoustics, similar to the reverb in [1]. While the filter characteristics may also be learned from recordings, we fix the FIR coefficients to a measured soundboard impulse response (IR) of a Yamaha C3 grand piano. Measurement details and IRs are available on the supplemental website.

The proposed model has several limitations. First, it does not account for the effect of the pedals. In particular the sustain pedal influences the characteristics of the whole piano, which should be considered in a full model. Second, the current system does not account for key release damping of the strings [11], as well as the noises of hammer and damping mechanisms [12]. Third, the soundboard characteristics change over the range of the keyboard [13], which is not accounted for by a single time-invariant filter. We plan to address these issues in future work.

3. EXCITATION LEARNING EXPERIMENT

The combination of $\mathcal{E}$ and $\mathcal{P}$ can be interpreted as a DDSP autoencoder, where $\mathcal{E}$ is a neural network trained to estimate the latent excitation signal f from an input signal x, and $\mathcal{P}$ is a fixed decoder that generates an output signal y. The training objective for $\mathcal{E}$ is to make y as similar as possible to x, measured by a multi-scale spectral loss [1]. We choose the *ResNet18* [14] architecture with 12M parameters using a magnitude spectrogram of x as input. To ensure efficient back-propagation of gradients, we use frequency sampling of the IIR filters [15] during training. In our experiment, f has the same duration and sampling rate as x and y (one second at 16 kHz) and we train on a dataset of 44000 synthetic piano tones (plus 11000 test samples). We generate this synthetic dataset using $\mathcal{P}$ with the same time-invariant filter parameters as during training and synthetically generated reference excitation signals which are varying in attack time and pulse shape [16]. This method allows for an analysis whether the learned f resembles the reference excitation signal in terms of timing and spectral content.

A comparison of Fig. 2b and c shows that the system in our preliminary experiment can indeed match the reference excitation signal in terms of approximate onset timing and the intensity of individual partials. However, we observe some temporal "blurring" of f compared to the reference, creating an unnatural attack phase in y.

4. FUTURE WORK

In addition to addressing the current model limitations, a next step is to train the system on real piano recordings, increasing the complexity of the learning task. Since $\mathcal{E}$ is not limited to the physical interpretability of a hammer force signal and $\mathcal{P}$ is linear, it can theoretically also compensate for some signal characteristics that are not accurately modeled by $\mathcal{P}$. Furthermore, we plan to explore various applications for this model, including the enhancement of corrupted piano recordings and unsupervised transcription.

5. ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (MU 2686/10-2 and MU 2686/13-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

6. REFERENCES

[1] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[2] 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.

[3] L. Renault, R. Mignot, and A. Roebel, "Differentiable Piano Model for Midi-to-Audio Performance Synthesis," in *Proceedings of the 25th International Conference on Digital Audio Effects (DAFx)*, Vienna, Austria, 2022.

[4] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019.

[5] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Unsupervised audio source separation using differentiable parametric source models," *Computing Research Repository (CoRR)*, vol. abs/2201.09592, 2022. [Online]. Available: https://arxiv.org/abs/2201.09592

[6] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, "DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022, pp. 76–83.

[7] B. Bank, S. Zambon, and F. Fontana, "A modal-based real-time piano synthesizer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 4, pp. 809–821, 2010.

[8] X. Boutillon, "Model for piano hammers: Experimental determination and digital simulation," *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 746–754, 1988.

[9] B. Bank, "Accurate and efficient modeling of beating and two-stage decay for string instrument synthesis," in *Proceedings of the MOSART Workshop on Curr. Res. Dir. in Computer Music*, Barcelona, Spain, 2001.

[10] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. New York, NY: Springer, 1998.

[11] A. Stulov, V. Välimäki, and H.-M. Lehtonen, "Modeling of the part-pedaling effect in the piano," in *Proceedings of the Acoustics Conference*, Nantes, France, 2012.

[12] A. Askenfelt, "Observations on the transient components of the piano tone," *STL-QPSR*, vol. 34, no. 4, pp. 15–22, 1993.

[13] K. Ege and X. Boutillon, "Vibrational and acoustical characteristics of the piano soundboard," *Computing Research Repository (CoRR)*, vol. abs/1212.3068, 2010. [Online]. Available: https://arxiv.org/abs/1212.3068

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[15] S. Lee, H.-S. Choi, and K. Lee, "Differentiable artificial reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2541–2556, 2022.

[16] S. A. Van Duyne and J. O. Smith, III, "A linear filter approximation to the hammer/string interaction for use in a commuted synthesis piano model," *The Journal of the Acoustical Society of America*, vol. 97, no. 5 Supplement, p. 3390, 1995.