Industrial Applied Data Science
18-19 October 2018
Lugano @SUPSI – Palazzo dei Congressi
www.ftal-conference.net

Deep Neural Yodeling

Topic: Life Sciences and Healthcare

Daniel Pfäffli, Andrea Kammermann, Marc Pouly, Tim vor der Brück
School of Information Technology / School of Music
Lucerne University of Applied Sciences and Arts
Email: daniel.pfaeffli@hslu.ch

In recent years, generative models were applied for music syntheses like for pop music [5], orchestra [4], Irish folk [1] or classical piano [7]. We continue this strand of research by investigating how generative models can learn and reproduce Swiss yodel music [6]. Yodel shares strong and measurable characteristics such as a choir accompaniment and the transition from chest voice to falsetto with an audible glottal stop that can be recognized even by laymen. In contrast to the state-of-the-art research with only technical analyses conducted, e.g., spectrum analysis of frequencies, these characteristics allow for a qualitative, semantically meaningful analysis and can hopefully guide us towards a better understanding of the structures that a generative model is capable of to reproduce.

Background

Existing approaches to music synthesis can be grouped into lead-sheet generation models ([1]) and audio waveform generation models ([5],[7],[4]). This work focuses on audio waveform because natural yodel is usually transmitted orally. [5] proposed a long short-term memory-based approach with one hidden layer network consisting of 2048 neurons trained on pop-music. They observed that generated sequences more than three times larger than the initial seed tend to get stuck in generation loops. In [7] the authors proposed WaveNet consisting of 50 dilated convolutional layers with dilation factor 1,2,4,…,512. A μ-law transformation converts the input signal into 8-bit. For music synthesis of piano music, the authors concluded that the samples were often harmonic and aesthetically pleasing although the genre, instrumentation, volume and sound quality varied from second to second. Finally, SampleRNN proposed by [4] utilizes a 3-tier model consisting of two one layer Gated Recurrent Units models with 1024 neurons, followed by a multi-layer perceptron with 1024 neurons. For Beethoven's piano sonatas, their model outperformed the WaveNet approach reaching 1.076 negative log-likelihood (NLL) (WaveNet: 1.464). In a human evaluation, listeners preferred SampleRNN over WaveNet.

Experiments

We used the free available WaveNet implementation on GitHub[1]. At first, the Adam optimiser [4] was used with a learning rate of 10^{-3} with no regularization. The sample size had to be set to 5s with a sample rate of 16000 Hz and a batch size of 1. The model performed 1.5 Mio. training steps on yodel with and without instrumental accompaniment, in total 14h. The data corpus was then switched to pure natural yodel music of Unterwalden, and the

[1] https://github.com/ibab/tensorflow-wavenet, accessed 15.01.2018

optimizer RMSprop [2] was applied, to increase the learning speed using momentum of 0.9. The model was trained for additional 2'143k training steps and reached a final mean NLL of 2.04 (see Table 1 and for samples at http://bit.ly/2siW16P).

Steps (in Mio.)	NLL Mean	NLL Std.	NLL Min	NLL Max	Audio Sample Name
0.45 – 0.55	2.38	0.42	1.12	3.22	a_yodel_mixed.wav
1.45 – 1.50	2.29	0.52	0.97	3.15	b_yodel_mixed.wav
3.65 – 3.68	2.04	0.37	1.04	2.89	c_yodel_uw.wav

Table 1: Mean, Variance, Minimum and Maximum are derived from 200 random samples drawn from the interval of the training steps given in the first column.

Discussion

For qualitative analysis of the final model, four samples were generated and assessed with the involvement of experts for yodel music. In all evaluated samples of the final model, the tunes feature an audible, polyphonic chorus which accompanies a solo yodeling voice. The model further succeeded in reproducing the correct vocalization of high notes with an "u" and low notes with an "o". Both are main characteristics of yodel music. Unfortunately, however, the glottal stop characteristic is not detectable. Overall, the music is much too fast, restless, and it is interrupted here and there by a very dominant slapping (sounds like an abrupt thunk, clapping). Other characteristics such as enduring notes and the coziness of yodel music are utterly missed.

Unlike [5] this WaveNet-based model achieved to create samples with voice-like singing which was not found in published GRUV-samples. Analysing the results of WaveNet proposed by [7] or SampleRNN by [4], indications for melody structure are not explicitly identifiable. Experts pointed out that the published samples are almost too short to recognize a melody line. Hence, it must be concluded that autoregressive approaches for audio waveform fail so far to compose music. Future work will be focused on an interleaving Variational Autoencoder ([3]) approach to include long-term melody structures.

In summary, the WaveNet model succeeded in producing the timbre of yodel music but definitely failed at composing yodel music.

[1] Florian Colombo, Samuel P. Muscinelli, Alexander Seeholzer, and Gerstner Wulfram, Eds. 2016. *Algorithmic Composition of Melodies with Deep Recurrent Neural Networks*. 1st Conference on Computer Simulation of Musical Creativity. University of Huddersfield, Huddersfield, UK. DOI: https://doi.org/10.13140/RG.2.1.2436.5683.

[2] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2014. *Overview of mini-batch gradient descent* (February 2014). Retrieved February 20, 2018 from http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

[3] Diederik P. Kingma and Max Welling, Eds. 2014. *Auto-Encoding Variational Bayes*. 2nd International Conference on Learning Representations. ICLR, Calgary, CAN.

[4] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, Eds. 2016. *SampleRNN. An Unconditional End-to-End Neural Audio Generation Model*. 5th International Conference on Learning Representations. ICLR, Toulon, France.

[5] Aran Nayebi and Matt Vitelli. *GRUV. Algorithmic Music Generation using Recurrent Neural Networks*. Retrieved February 20, 2018 from https://arxiv.org/abs/1601.03642.

[6] Daniel Pfäffli. 2018. *Master Thesis - Deep Neural Yodelling*. Zenodo. School of Information Technology. Lucerne University of Applied Sciences and Arts, Lucerne.

[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. *WaveNet. A Generative Model for Raw Audio* (2016). Retrieved February 20, 2018 from http://arxiv.org/pdf/1609.03499.

☒ If my contribution is selected for the post-conference publication, I will accept to send a 6-8 pages full version that will be peer-reviewed to be included as an article.