TŌKYŌ KION-ON: QUERY-BASED GENERATIVE SONIFICATION OF ATMOSPHERIC DATA

Stefano Kalonaris

RIKEN AIP
Music Information Intelligence Team
stefano.kalonaris@riken.jp

ABSTRACT

Amid growing environmental concerns, interactive displays of data constitute an important tool for exploring and understanding the impact of climate change on the planet's ecosystemic integrity. This paper presents *Tōkyō kion-on*, a query-based sonification model of Tokyo's air temperature from 1876 to 2021. The system uses a recurrent neural network architecture known as LSTM with attention trained on a small dataset of Japanese melodies and conditioned upon said atmospheric data. After describing the model's implementation, a brief comparative illustration of the musical results is presented, along with a discussion on how the exposed hyper-parameters can promote active and non-linear exploration of the data.

1. INTRODUCTION

The planet's rising temperature is an indisputable trend, with catastrophic effects for the ecosystemic balance and all life on Earth as we know it. With growing consensus and awareness that a tipping point of this fragile dynamical, complex, and adaptive system has perhaps already been reached (or that is inevitably close), what has been dubbed as the "climate crisis" is polarizing opinions, actions, and stakeholders' interest across the globe. It has been argued [1] that alternative ways to convey environmental data can include auditory display methods, which can help overcome the issues related to engagement with graphical representations. However, in cases where the data is not sufficiently multidimensional and/or complex to require alternative means of display and insight, mapping methods which explicitly couple the range of data variables to that of musical parameters (e.g., pitch, dynamics, etc.) risk offering an equally predictable and unengaging representation of the phenomena under consideration. While general consensus argues that "mappings will follow simple intuitive basics, e.g., temperature to pitch, location to spatial rendering, etc., and the sound design shall evoke climate associations that are straightforward, e.g., as known from weather conditions" [2, p.11], this can be problematic in the case of temperature data, which follows a linear growth trend; the aftermath of a few degrees in the planet's temperature has irrevocable consequences, but to appropriately convey this via audiovisual display almost necessarily involves re-scaling the temperature range so that such a modest difference value can be appropriately perceived and communicate the sense of urgency that it warrants. This, in turn, might result in unwanted sound caricatures or contour exaggerations, to the detriment of musicality.

This paper, on the other hand, foregrounds aesthetic concerns and presents a query-based model of auditory display where a more subtle correspondence is sought. To this end, mapping of the data is applied to a neural network's internal representation of the musical corpus' parametric encoding. Given the *black box* nature of neural networks, the sonification model confers a degree of opacity and surrogacy [3] which might elicit meaning-making processes rather than meaning-finding or meaning-carrying.

Furthermore, the environmental data of interest constitutes but the initial inspiration behind the adoption of the author's chosen sonification design and procedure. Along the continuum between concrete to abstract sonification as outlined in [4], *Tōkyō kion-on* is mostly focusing on the latter, and should be intended primarily as a generative task rather than an empirically-based sonification. In accordance to more current trends in the broader field of aesthetics, where the sensibility has continued to move towards a pragmatist stance, the author establishes affinity with a notion of interpretation that welcomes and embraces the role of indeterminacy [5], fully aware that "by casting the definition in terms of science and objectivity alone, the more complex interrelationship between the arts and using data as a direct determinate is lost" [6, p.208]. While sequential/chronological auditory display of the data is certainly possible, much like the majority of environmental data sonification systems that focus on the temporal evolution of the phenomenon of interest, the author's model affords a non-linear approach, allowing the user to single out years of interest, change the length of the generated melody, experimenting with different priming seeds, and, ultimately, procedurally combine results for compositional or purely explorative goals.

2. RELATED WORK

Different methods are normally used to auditorily display data. These include *Earcons* or *Auditory Icons* (sound aliases for events or actions, typically used as auditory aids), *Audification* [7] (direct sonification of data as a series of sound pressure values, using re-sampling and filtering), *Parameter-Mapping Sonification* (PMSon) [8] (mapping and transfer functions reveal latent structures in the data), *Model-Based Sonification* (MBSon) [9] (a virtual instrument model built from the data of interest is played via user interaction), and *Wave Space Sonification* (WSSon) [10] (a data-driven trajectory scans a scalar field).

Sonification of environmental and atmospheric data is certainly not a novel endeavor, and the above methods have been explored in [11, 12, 13, 14]. An accomplished example of artistic-led

https://doi.org/10.21785/icad2022.039

sonification of weather data which manages to integrate disparate concerns and viewpoints is the *Locus Wrath* [15], a multi-modal interactive system for dance performance, where aesthetic considerations and high ecological validity are coupled with the concurrent usage of different sonification methods.

In the realm of auditory display, the use of artificial neural networks (ANNs) is still relatively a novelty, unlike in other subfields of music and sound computing where these are increasingly becoming the norm. For the most part, precedents in this direction tend to deal with data of interest other than climate-related. Notwithstanding, the recent work of Herrmann [16] is an important reference when reviewing deep learning-based sonifications. This approach is followed in [17] and there, too, the sonification process auditorily displays inner layers of the network's activation functions activity.

One example of ANNs application to atmospheric data is documented in [18], where a variational auto-encoder (VAE) model released by Google's Magenta team is used to sonify smart city data including weather data. The VAE is employed to blend a source and a target musical motif with a morphing coefficient proportional to the change in rainfall. In the same work, temperature data and wind speed are also mapped to an LFO's fundamental frequency and amplitude. This work is perhaps the most similar to the system presented in this paper, so far as the use of ANNs is concerned.

Tōkyō kion-on uses a recurrent neural network (RNN) known as long short-term memory (LSTM). This architecture has been the most widely used in tasks involving temporal dependencies (e.g., time series) before the advent of newer models using self-attention [19], with the music domain not being an exception in this regard [20, 21, 22]. However, while there are countless examples of generative music systems built using LSTMs [23, 24, 25, 26], this approach remains under-explored in sonification tasks.

3. MODEL

Tōkyō kion-on can be considered both a MBSon and a PMSon model in the currently accepted taxonomy of auditory display methods. On one hand it is a model where the data of interest is explored via means of user interaction. However, the data model is not the atmospheric data itself but the representation of a musical corpus learned by an ANN. On the other hand, an arbitrary mapping between the data and sound parameters is established so that the sonic output is, virtually, a surrogate of the phenomena under consideration. In the form described in this paper, and as of the time of writing, *Tōkyō kion-on* is available online at `https://gitlab.com/skalo/tokyokionon`. Before delving into the details, a brief description of both the training musical corpus and the atmospheric data is given.

3.1. Data

The neural network was trained on a small dataset of public domain Japanese melodies in .musicxml format hosted online,[1] excluding those arranged for piano, and all melodies were "normalized" to the same key (it is possible, if desired, to re-train the model with the dataset augmented to all 12 keys).

Atmospheric data was obtained from the Japan Meteorological Agency's website,[2] using the monthly mean daily maximum temperature table. The specific architecture of ANN (see Section 3.2) used allows to control the randomness of predictions with a hyper-parameter which is, incidentally, known as *temperature* (one must be careful to disambiguate this from the atmospheric data homonymous measure). The atmospheric data was pre-processed to obtain two year-indexed normalized vectors that serve as the ANN's temperature hyper-parameter for a given year's sonification query. One of these vectors is used for the randomness control of the pitch predictions, the other for the randomness control of the notes' duration (see Section 3.3 for a detailed explanation).

The pitch temperature vector is obtained by calculating the cosine similarity between a given year's forward difference of the monthly air temperature average and the corresponding forward difference for the reference year's (i.e., 1876) and subtracting the resulting value from 1. This is because cosine similarity returns a value between $[0, 1]$ where 1 is the identity but, for the ANN's temperature hyper-parameter, 1 should correspond to the highest probability of sampling from unlikely candidates in the prediction (see Section 3.3). As for the duration temperature, the normalized vector of the annual mean daily maximum temperature is used.

3.2. Architecture & Training

Notwithstanding ample room for experimentation with different models, the LSTM architecture was deemed appropriate because it is apt for tasks involving time series modeling (thus it works well both for atmospheric data and melodic sequences, alike). One known shortcoming of LSTMs is the ability to model time dependencies at meso and macro level, which in music would translate to relational structures beyond musical periods (e.g., sections, repetitions, and long-term structure, in general). To address these levels, models using self-attention such as the Transformer [19] are more useful. However, there are two main reasons why this was not the traveled path: 1) the corpus used in training comprises exclusively short traditional melodies, whose local dependencies are well catered for by LSTMs and 2) Transformers require vast amounts of training data which, for the musical genre of choice, is not currently available.

Just like RNNs, LSTMs are chains of connected modules called units. Unlike RNNs, an LSTM unit comprises 4 layers (three sigmoid and one hyperbolic tangent), as shown in Figure 1. *Tōkyō kion-on* uses 256 such units and employs the *attention* mechanism, which overcomes RNNs limitations in the encoder-decoder architecture by allowing the network to learn where to attend/focus most in the output sequence. *Tōkyō kion-on* takes two input sequences, one for the pitch and one for the duration values, embeds these into vectors and concatenates these into a single input vector to the recurrent layers. After training, it produces two outputs, a prediction for the notes and another for their duration.

Given the modest number of samples (46) in the training dataset, the model was trained without using GPU on a 8-core AMD processor and converged after 205 epochs with *Early Stopping* and *patience* of 10.

3.3. Generation

Users can query a year in the range of the atmospheric data at the time of writing (i.e., 1876 to 2021). Without any further argument,

[1] `http://www.daisyfield.com/music/htm/-genres/japan.htm/`

[2] `https://www.data.jma.go.jp`

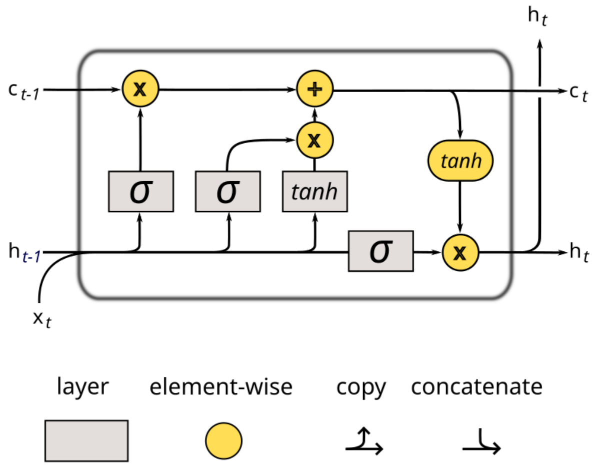


Figure 1: One unit of an LSTM network. Gray blocks are neural network activation layers, yellow ellipses are point-wise operators, and directed lines are vector transfers, copies or concatenations. C is the cell state, x is the input, and h the output, at a given time t.

the generative part of the system "predicts" a melody based on the weights learned during training, on the corresponding pitch and duration temperatures in the year-indexed vectors obtained during data pre-processing (see Section 3.1), and on default parameters that dictate the output's sequence length. A brief explanation of how temperature works in an LSTM is now due.

Neural networks use a logit vector z where $z = (z_1, \ldots, z_n)$ and apply an activation function to produce a probability vector $q = (q_1, \ldots, q_n)$ over predicted output classes by comparing z_i with the other logits. *Softmax* is normally the activation function of choice for the last layer of a neural network. In LSTMs, the temperature hyper-parameter T controls the randomness of the predictions, effectively being a scaling factor for the logits, when computing the softmax output. Probability qi, in this case, would then be calculated as shown in Equation 1.

$$q_i = \frac{\exp(\frac{z_i}{T})}{\sum_j^n \exp(\frac{z_j}{T})} \tag{1}$$

In the specific implementation of *Tōkyō kion-on*'s LSTM, T is a value in the range $[0, 1]$. When equal to 1, the softmax is computed directly on the logits, whereas for temperatures < 1 the softmax will yield larger values, making the LSTM more confident in the prediction (since less input can activate the output layer) albeit more conservative, less prone to sample from unlikely candidates.

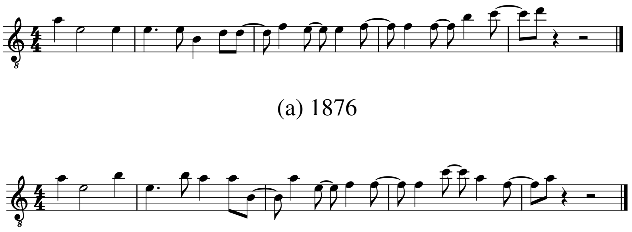
4. OUTPUT EXAMPLES

In terms of user interaction, and as for the design stage discussed in this paper at the time of writing, a convenience script has been included in the repository that hosts the code. This script can be executed directly by providing arguments of choice for the exposed hyper-parameters. An example of shell command is shown below, with abbreviated arguments (i.e., `-y` instead of `--year`, etc.):

Listing 1: Example of invoking the sonification script via command-line.

```
>>> python tokyokionon.py -y 1984
-s [[‘A4’],[0.5]] -mxx 8 -mxl 16 -sql 16
```

Using default values for the generation allows for direct comparison between different query years, since the priming seed is kept invariant (pitch: $A4$, duration in quarter length: 1.0) and the output sequence length is set to 16 tokens. Figure 2 shows the generated melodies for the reference year (1876) and for one of the years that are most dissimilar to it according to the monthly forward difference cosine similarity metric (see Section 3.1), which affects the note randomness. As one can verify, the notes vary, whereas the duration pattern is the same. In fact, 1980's duration temperature value stands at 0.37, a value perhaps not sufficiently large to trigger variations in the notes' duration values.



(a) 1876

(b) 1980

Figure 2: Comparing melodies generated for 1876 and 1980 using default parameter values.

In addition to the year query, the user has access to additional parameters. For example, querying 2021 with an empty seed yields the melody shown in Figure 3. This time, the duration values are considerably different from the reference year, since the normalized value of the annual mean temperature for 2021 is the maximum attainable value, 1.

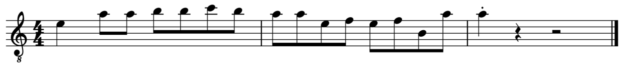


Figure 3: Melody generated for data points corresponding to 2021, without a priming seed.

Figure 4, instead, shows a query for 2004, the second warmest with a duration temperature of 0.826, as well as a relatively high note temperature of 0.761. In this instance, the query parameter for maximum extra notes was increased to 32 and the priming seed was reverted to the default, for comparison with the melodies shown in Figure 2.

As expected, both the notes and their duration differ from the 1876 melody. It is also possible to inspect both the attention matrix and the note candidates distribution of a generated melody, for added insight. The former shows how the neural network attends every hidden state from each encoder node at every time step, determining which of these are considered more informative for making predictions. The latter portrays the likelihood of notes being

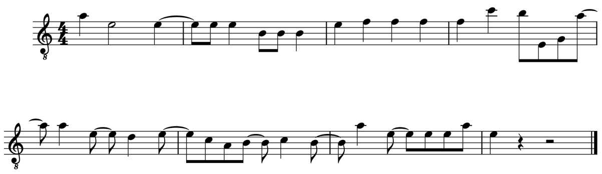


Figure 4: Melody generated for data points corresponding to 2004, with default priming seed and maximum extra notes length of 32.

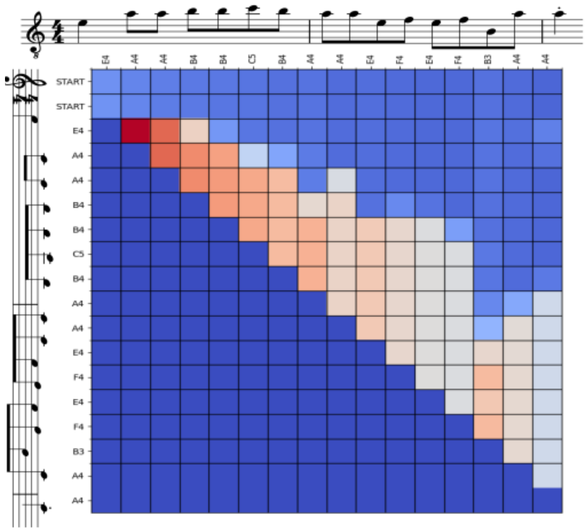


Figure 5: Attention matrix for the 2021 example. The amount of "redness" of a cell is proportional to the attention given to the hidden state of the model corresponding to the y position, when predicting the note on the x position.

selected over time. Examples of these representations, which are visually akin to heat maps, are shown in Figures 5 and 6.

If, instead of single-year queries, one wishes to obtain a sonification of given range of years in the atmospheric dataset, this could be easily done by, for example, concatenating melodies for all desired entries. One would perhaps use a shorter sequence length, as demonstrated in Figure 7 which shows the melodies for the first and the last ten years in the dataset (1876-1886 and 2011-2021) generated using an empty seed and a sequence length of 4.

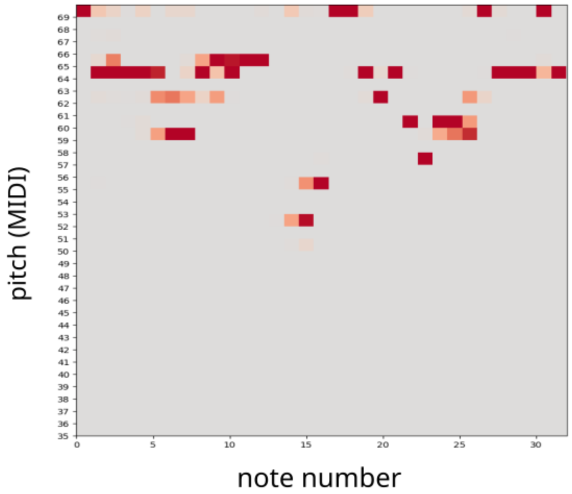
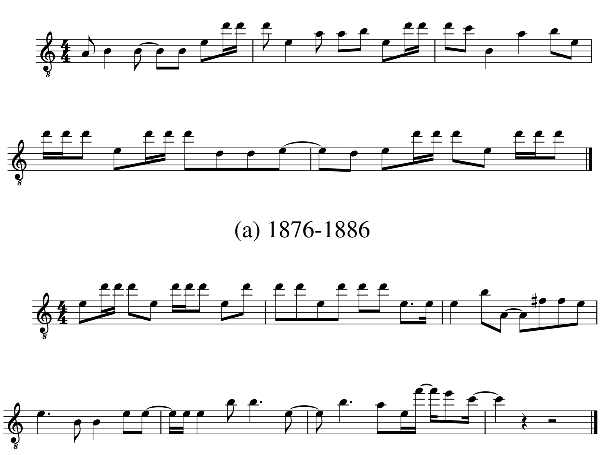


Figure 6: The distribution of the note candidates for the 2004 example. The darker (red) is a note at any given time step, the more likely it is that the model will select it.



(a) 1876-1886

(b) 2011-2021

Figure 7: First and last ten-year range in the air temperature dataset, sonified using a sequence length of 4 and an empty priming seed.

5. DISCUSSION

This paper presented *Tōkyō kion-on*, an interactive sonification model of Tokyo's air temperature from 1876 to 2021 based on a popular recurrent neural network architecture, and trained on a dataset of Japanese melodies. *Tōkyō kion-on* as presented here is still at a prototypical stage, and a formal evaluation has yet to be carried out. However, in the author's informal experimentation with the model's hyper-parameters, this system provided an interesting tool for querying environmental data and melodic generation alike. By exploiting the intrinsic affordance of the network to control the randomness of the output candidates, *Tōkyō kion-on* offers a simple yet effective interface to explore both single year entries as well as entire ranges of years, and to easily compare different years auditorily. Unlike the leading trend in sonification, whereby the phenomenon under consideration is rendered sequentially so as to display its development and temporal unfolding in linear time, this work allows non-linear exploration of the data of interest, by querying specific years of choice. It is of course possible if trivial to simply concatenate all years' generated melodies in succession to obtain a more mainstream rendition, if so desired. Furthermore, *Tōkyō kion-on* can be used creatively and beyond the sonification task, to generate sequences (melodies) of arbitrary length primed with different seeds (providing these exist in the learned vocabulary).

The system could be improved greatly if trained on a larger corpus, but this was hindered by the endemic lack of Japanese music datasets. More experimentation should be carried out with different ANNs models and it would be valuable to implement an interactive web-based version of the system. Moreover, it would be interesting to expand the scope of the system to include different pre-trained models (on different corpora) or allowing the users to upload different environmental data to condition the generative output.

6. REFERENCES

[1] S. S. George, D. Crawford, T. Reubold, and E. Giorgi, "Making Climate Data Sing: Using Music-like Sonifications to Convey a Key Climate Record," *Bulletin of the American Meteorological Society*, vol. 98, no. 1, pp. 23 – 27, 2017.

[2] V. Goudarzi, "Contextual Inquiry for a Climate Audio Interface," in *Human-Computer Interfaces and Interactivity: Emergent Research and Applications*, P. Isaas and K. Blashki, Eds. IGI Global, 2014, pp. 1–13.

[3] S. Kalonaris and I. Zannos, "High-Order Surrogacy for the Audiovisual Display of Dance," in *Proceedings of the International Conference on Auditory Display*, 2021.

[4] P. Vickers and B. Hogg, "Sonification Abstraite/Sonification Concrète: An 'Aesthetic Perspective Space' for Classifying Auditory Displays in the Ars Musica Domain," in *Proceedings of the International Conference on Auditory Display*, 2006, pp. 63–68.

[5] R. Shusterman, "The Pragmatist Aesthetics of Richard Shusterman: A Conversation - Interviewed by Günter Leypoldt," *Zeitschrift für Anglistik und Amerikanistik: A Quarterly of Language, Literature, and Culture*, vol. 48, no. 1, pp. 57–71, 2000, online; accessed 27 February 2020.

[6] S. Gresham-Lancaster, "Relationships of sonification to music and sound art," *AI & Society*, vol. 27, pp. 207–212, 2012.

[7] F. Dombois, "Using Audification in Planetary Seismology," in *Proceedings of the International Conference on Auditory Display*, 2001, pp. 227–230.

[8] D. Worrall, "Parameter-Mapping Sonification of Tick-Data," in *Sonification Design: From Data to Intelligible Soundfields*. Cham: Springer International Publishing, 2019, pp. 237–252.

[9] T. Hermann and H. Ritter, "Sound and meaning in auditory data display," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 730–741, 2004.

[10] T. Hermann, "Wave Space Sonification," in *Proceedings of the International Conference on Auditory Display*, 2018.

[11] M. Quinn and L. D. Meeker, "Research Set to Music: The Climate Symphony and Other Sonifications of Ice Core, Radar, DNA, Seismic and Solar Wind Data," in *Proceedings of the International Conference on Auditory Display*, 2001, pp. 57–61.

[12] J. H. Flowers, L. E. Whitwer, D. C. Grafel, and C. A. Kotan, "Sonification of Daily Weather Records: Issues of Perception, Attention and Memory in Design Choices," in *Proceedings of the International Conference on Auditory Display*, 2001, pp. 222–226.

[13] T. Hermann, J. M. Drees, and H. Ritter, "Broadcasting Auditory Weather Reports - A Pilot Project," in *Proceedings of the International Conference on Auditory Display*, 2003, pp. 208–211.

[14] A. Polli, "Atmospherics/Weather works: A multi-channel Storm Sonification Project," in *Proceedings of the International Conference on Auditory Display*, 2004.

[15] P. Lindborg, "Interactive Sonification of Weather Data for The Locust Wrath, a Multimedia Dance Performance," *Leonardo*, vol. 51, no. 5, pp. 466–474, 10 2018.

[16] V. Herrmann, "Visualizing and sonifying how an artificial ear hears music," in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, ser. Proceedings of Machine Learning Research, H. J. Escalante and R. Hadsell, Eds., vol. 123. PMLR, 2020, pp. 192–202.

[17] F. C. Halac and M. Delgadino, "DreamSound: Deep Activation Layer Sonification," in *Proceedings of the International Conference on Auditory Display*, 2021.

[18] S. Roddy and B. Bridges, "The Design of a Smart City Sonification System Using a Conceptual Blending and Musical Framework, Web Audio and Deep Learning Techniques," in *Proceedings of the International Conference on Auditory Display*, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS17. Red Hook, NY, USA: Curran Associates Inc., 2017.

[20] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, and D. Eck, "Music Transformer: Generating Music with Long-Term Structure," *arXiv preprint arXiv:1809.04281*, 2018.

[21] ——, "An Improved Relative Self-Attention Mechanism for Transformer with Application to Music Generation," *CoRR*, vol. abs/1809.04281, 2018.

[22] C. Hawthorne, A. Huang, D. Ippolito, and D. Eck, "Transformer-NADE for Piano Performances," in *Proceedings of the NIPS 2018 Workshop on Machine Learning for Creativity and Design*, 2018.

[23] I. Simon and S. Oore, " Performance RNN: Generating Music with Expressive Timing and Dynamics ," https://magenta.tensorflow.org/performance-rnn, 2017.

[24] F. T. Liang, M. Gotham, M. Johnson, and J. Shotton, "Automatic Stylistic Composition of Bach Chorales with Deep LSTM," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 2017.

[25] B. Sturm, "What do these 5,599,881 parameters mean? : An analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer," in *Proceedings of the 6th International Workshop on Musical Metacreation*, 2018.

[26] J. Thickstun, Z. Harchaoui, D. P. Foster, and S. M. Kakade, "Coupled Recurrent Models for Polyphonic Music Composition," *CoRR*, vol. abs/1811.08045, 2018.