Generating Music with Emotion Using Transformer

1st M. Aqmal Pangestu
School of Computing
Telkom University
Bandung, Indonesia
aqmalpangestu@student.telkomuniversity.ac.id

2nd Suyanto Suyanto
School of Computing
Telkom University
Bandung, Indonesia
suyanto@telkomuniversity.ac.id

2021 International Conference on Computer Science and Engineering (IC2SE) | 978-1-6654-0046-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IC2SE52832.2021.9791928

***Abstract*—The Transformer model has gained popularity because of its capabilities for solving multiple tasks. One of them is automatic music generation. Many studies have proven that this model can generate music with a consistent structure, but the pieces that are generated still lack emotion in it. In this paper, we extend Transformer-based model capabilities to generate music with controllable emotion. The emotion is divided into three categories: negative, neutral, and positive. We train the model using 120 MIDI files from our new piano datasets. The dataset has been labeled based on their emotion. The labeling process is done manually by hearing. The total MIDI files available in the dataset is 210 but we filter it so that only 120 remains. We also add a new token to represent emotion on REvamped MIDI-derived event (REMI). The experimental results show that human subject agreed that Transformer-XL model using REMI and emotion token is able to generate emotion-based music. We also compare our generated pieces with other datasets. The result show that the majority of respondents prefer pieces that are generated using our datasets.**

***Keywords*—music generation, controllable music generation, transformer**

I. INTRODUCTION

Music consists of a repeated sequence of melody, harmony, tempo, timbre that can create series of emotions [1]–[3]. One of the reliable deep learning models to generate such sequence is the Transformer [4]. In this model there is a self-attention mechanism that allows the model to access any part of the previously generated output at every step of generation [5], [6]. With this mechanism Transformer model is the intuitively suitable deep learning model for generating musical sequences. The transformer is a deep learning model that consist of an encoder and a decoder. Encoder serves to generate a value in the form of a vector (attention score) [5]. This vector is used to find information from a long sequence. The encoder consists of a multi-head self-attention layer module and a point-wise fully connected feed-forward network [6]. Point-wise means applying a linear transformation to every element in the sequence. Each module has a residual connection and a normalization layer. Each module produces output with the same dimensions. The decoder of the transformer serves to receive the encoded representation, and provide further predictions. The decoder architecture is similar to the encoder, the difference is that the decoder has two multi-head self attention modules. The first multi-head attention module applies masking to some of its elements to prevent an unknown element's position from appearing prematurely, and the second is the same as the encoder [6].

Deep learning models, including Transformer, need to learn musical features from many musical data to generate new music. The data used for training such model need to be converted in the form of number that can be interpreted as musical elements. This process of converting is called tokenization. Some tokenization methods for music generation are based on the MIDI protocol. With MIDI, the musical event can be represented as a text-like event. With the text-like representation, we can create the vocabulary for music. Hence music generation can become a lot like a text-generation task [4].

However, not every musical element can be represented in MIDI as humans represent music [7]. REvamped MIDI-derived event (REMI) and recently proposed Compound Word (CP) claimed to be able to represent musical elements more efficiently than MIDI [7], [8]. REMI used Transformer-XL as a backbone model while CP uses linear Transformer. Both are using a pop piano music dataset for training the model. This paper proves that a human-friendly representation of music improves Transformer capability to generate a coherent piece.

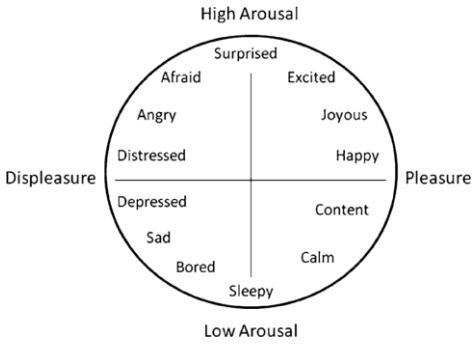


Fig. 1. Emotion dimensional model

In this paper, we try to extend the Transformer-XL capabilities to generate controllable music with three emotions, neutral, positive, negative, using various genres of piano music as training data. Positive emotion means that the music can makes the listener feels happy, excited, and joyful. Negative emotion means that the music can makes the listener feels sad and anxious. The neutral emotion is the middle line

between negative and positive. It supposes to makes the listener feel calm, chill, and melancholy. To do that, we added a new musical event called Emotion-Start and Emotion-End to represent the emotional element of the music. The definition of emotion in this paper is based on the valence-arousal emotion model [9] that can be seen in Figure 1.

We choose Transformer-XL as our backbone model because it has a recurrence mechanism to improve the baseline Transformer architecture [10]. We also introduce a new PianoEmotion MIDI (PEMIDI) dataset, consisting of 210 MIDI format piano songs from various artists and genres. These datasets are obtained from the internet, and the detail of the dataset will be explained in Section III.

II. RELATED WORK

A. Music Generation

Computer-generated music has been researched many times before. One of the research is able to successfully generate music using different instruments and combine the style of the music from a different artist [11]. To do that, they trained the GPT-2 model using thousands of different MIDI files from various artists [11]. There is also research to generate music using raw audio to generate a singing and instrument simultaneously [12]. There is a drawback to different musical data representation for music generation. With sound wave representation, the model can learn from multiple instruments, even singing [12]. However, it can take a lot of time to train the model because of the large file size [13]. On the other hand, a text-like representation such as MIDI can reduce training process time because of its smaller file size. However, the music-generated instrument is restricted and needs further processing before the model can use it for training [7], [8].

MIDI is a communication protocol for digital music [14]. It has a symbolic representation, transmitted serially that indicates Note On and Note Off events and allows for a high temporal sampling rate. The loudness of each note is encoded in a discrete quantity referred to as velocity (the name is derived from how fast a piano key is pressed) [14]. With MIDI representation for music generation, the duration of generated music can be a minute-long, even a hour-long. [8], [15]. It can be done because with MIDI, we can extract many musical element and create our own musical representation so that the music produced can be more expressive.

B. Emotion-Based Music Generation

One of the challenges in automatic music generation is to generate music as similar as possible to human-composed music. Human composed music tends to have emotion in it, and we can feel different emotions from listening to different styles of music. There are already papers that address music generation with a given emotion. One of these studies applies the LSTM model to evoke music with positive and negative emotions [16]. This paper uses a genetic algorithm to optimize the L1 layer's neuron weights. The weight of these neurons is the parameter that most influences emotional changes in the evoked music. This paper also classifies positive and negative emotions in the VGMIDI dataset using the LSTM model with 89.83% accuracy.

We tend to feel emotion from human facial expression. A research by Madhok try to combine facial expression with musical emotion [17]. They try to generate music based on the facial expression. The model used is LSTM for generating music, and CNN for facial expression classification. The result shown that the CNN model achieved 75% accuracy and the human evaluation result shown that most of the human subject correctly classify the music emotion based on it facial expression [17].

C. Transformer for Music Generation

Different deep learning models can produce a distinct musical structure. Reference [4] proved that the Transformer model could maintain a consistent musical structure compared to the LSTM model. It happens because of the characteristics of LSTM, which has limited memory. The use of relative attention in the Transformer model, on the other hand, makes the music generated to have a consistent musical structure [6].

From the best of our knowledge, the first paper that use Transformer model for music generation is Music Transformer by Huang et al. [4]. This paper use MAESTRO dataset to train the Transformer model. In this paper, a new relative self-attention algorithm used in addition to make the model able to generate music with long structure [4]. Since then, many research has come in music generation that using this model. One of the research is to generate music with different style [18]. A research by Mao prove that Transformer-XL model is able to generate music with three distinct style, baroque, classic, and romantic [18]. Pop Music Transformer introduced by Huang also use Transformer-XL model to generate a minute-long pop piano music [7]. The transformer model is making computer generated music has a better structure and a longer duration.

III. METHODOLOGY

A. Transformer-XL

Transformer is a deep learning model that is based on the use of an attention mechanism. Transformers were first proposed in 2017 [6]. This model can do well with sequence to sequence modeling without using recurrence. To replace recurrence, attention and positional encoding are used. It gives high performances in computational linguistic and text processing [19] and speech processing [20]. Moreover, it can beat the LSTM model in the case of translation [6]. The effectiveness of this model is not just for a sequence to sequence modeling but also for image recognition problems [21].

One of the variants of Transformers capable of processing huge data is Transformer-XL (Extra Long), which was first proposed in 2019 [10]. This model suggests using a recurrence mechanism in attention. Instead of calculating the new hidden state in each segment, Transformer-XL uses the hidden state from the previous segment. The hidden state acts as memory in each segment, thus creating recurrence between segments.

In this way, the model can model very long sequences because information can be propagated through recurrent connections. Transformers have a limited attention span and are not dynamic. This model can only handle elements in the same segment while computing, and no information can pass between separate attention spans. This creates context segmentation problems such as the model cannot remember very long segments, it is difficult to predict the first output without an input segment, and heavy computing. Every shift, all calculations have to start all over again.

Transformer-XL is able to solve the context segmentation problem of the original Transformer by reusing the calculated hidden state [10]. In order to reuse the hidden state in the new segment, the positional encoding needs to be changed because the positional encoding calculation used in the original Transformer is not relative (but absolute), so the hidden state cannot be reused in the new segment.

B. PEMIDI Dataset

We introduced a new dataset called PianoEmotion MIDI (PEMIDI). The PEMIDI dataset is a collection of 210 MIDI piano music varying from various artists and genres (classic, pop, jazz) labeled based on positive, neutral, and negative emotions. For the labeling, we assume that negative emotions are sad, fearful, and depressed. Neutral emotions are relaxed and calm. Positive emotions are happy and uplifting. This labeling is done manually by hearing and looking at the chord progression, mainly at the beginning part of the song. We assumed that positive emotion is a music with many major chords, neutral emotion is a song with jazz chords, and negative emotion is a song with a minor chord. This assumption are based on these research [2], [3], [22].

The time signature of the music varies. Some of the pieces obtained have 3/4 time signatures, and others have 4/4. The time signature is deliberately varied so that the model can generate music with various time signatures rather than generate in 4/4 beat. However, in this paper, we try to minimize the usage of the time signature other than 4/4.

We obtained the dataset from the internet, downloaded it, and then converted it into mp3 format. The mp3 data is then converted using the Convolutional Neural Network (CNN) method [23] into MIDI. Lastly, the MIDI data will be processed into a series of musical events using the REMI tokenization method [7]. Figure 2 explains the process of obtaining and converting the data into ready-to-use training data.

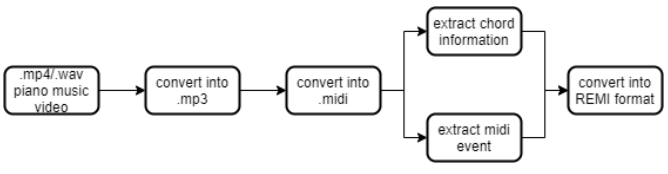


Fig. 2. How data processing works

C. REMI Tokenization

Tokenization is done by extracting MIDI events and then translate them into REMI events [7]. These REMI events are then grouped to create a music vocabulary. The music vocabulary will contain a musical event represented as a word token, for example, "Note on_13", the Note on is the musical event, the value 13 is the word token.

In this paper, we add a new event called Emotion-Start and Emotion-End to represent musical emotion. In the training dataset, these events will be added at the beginning of the sequence (for Emotion-Start) and in the middle (for Emotion-End). We choose to use the Emotion-End token in the middle of a sequence because one piece of music can have multiple emotions. To avoid the ambiguity of the multiple emotion music, we use the Emotion-End token in the middle of a sequence. Table I explains the original REMI token [7] and our new emotion token.

TABLE I
REMI TOKENIZATION

Event	Value
Bar	Indicate the beginning of a bar.
Chord	All of 60 possible chord available in music.
Note Duration	Ranging from 1-64.
Note On	All of 127 possible note in music.
Note Velocity	Represent how loud the note is being played. There are 32 different note velocity events.
Position	The possible position a note is placed in a bar ranging from 1-16.
Tempo Classes	Class of the tempo
Tempo Value	Ranging from 30-209 BPM.
Emotion-Start	Emotion of the song (neutral, positive, negative).
Emotion-End	Represent the closing emotion of the song (neutral, positive, negative).

D. Training Process

After the data has been processed into REMI events, the Transformer-XL model training is carried out. The parameters that need to be considered in the training process are as follows.

- Recurrence length = 512
- Training input events = 512
- Self attention layer = 12
- Attention heads = 8
- Optimizer = Adam opt$i_4$mizer
- learning rate = 2 x 10 , batch size = 2.

The training process will be evaluated by the loss obtained in every epoch until loss < 0.1. These parameters are based on other research [7], [24]. Figure 3 illustrates the training process flowchart. The sampling method that will be used in music generation is temperature-controlled stochastic [7]. This

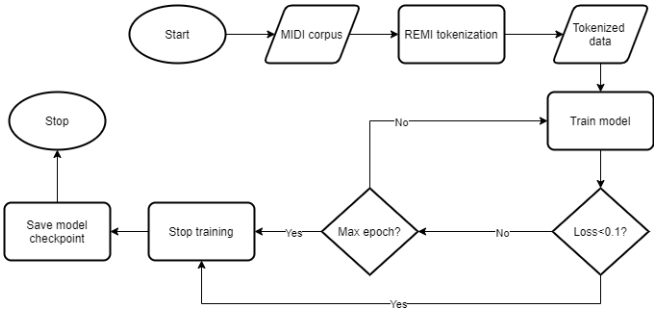


Fig. 3. Training process flowchart

paper only trains our model using 120 MIDI files from the PEMIDI dataset because we only use music with a 4/4 time signature. We also train our model with 120 VGMIDI datasets to compare our dataset's quality for music generation. The comparison process is going to be using the MOS metric.

E. Generating Music

We randomly choose the chord token as the first sequence for the generating process, and then the model will predict the next token based on that random chord. For example, if we want to generate positive emotion music, the model will choose a random major chord as its first sequence and then predict the next token based on the training data. For the negative emotion, the model will choose a random minor chord, and for the neutral emotion, the model will choose a random chord from any available chord. There is 60 chord available [7]. With the emotion token, we can generate multiple emotion music. To do that, we add a new emotion token at the middle of a sequence. The overall process of generating music is shown in Figure 4.

The explanation of Figure 4 is as follows. First,we initialize the target bar (desired number of music bars), random chord event and emotion start event. The model then predict the next event based on the randomly initialize chord event and emotion event. The generating process has a condition, if we want to generate music with more than one emotion in it, the system will append an Emotion-End event to stop the current emotion, and then replaced the emotion with a new given emotion. The position of Emotion-End event is at the middle of the sequence (equal to half the targeted bar). If we only want to generate music with single emotion, then the system will continue to predict next event. If generated bar is equal to target bar, then the generating process will be stopped and the MIDI will be written.

F. Evaluation Process

The generated music will be evaluated using the Mean Opinion Score (MOS) by doing an online survey. Respondents will judge whether the music has a positive, negative, or neutral emotion and then compare the quality of music generated from the Transformer-XL model using the PEMIDI dataset with another publicly available dataset called VGMIDI. Since the expected musical emotions are three, the rating scale is 0 to 10 with an interval of 0-4 means that the emotion is negative, 5 which means that the emotion is neutral, and 6-10 which means that the emotion is positive [17]. The survey process will be divided into two segments. In each segment, respondents will be given an assessment procedure and information about that segment. The division of survey segments can be described as follows.

- Respondents assess whether the musical emotions produced by this paper are as expected or not. Each emotion consists of three pieces, so that the total number of pieces in this segment is 6.
- Respondents compare the music produced by the PEMIDI dataset with the VGMIDI dataset. In order to compare it we train the Transformer-XL model using REMI tokenization method with the VGMIDI dataset [16]. Respondents will judge whether the newly trained model contains the desired emotion and it is more pleasant to hear or not. The generated music will be divided based on positive and negative emotions. The total pieces in this segment are four.

IV. RESULT AND DISCUSSION

The model is trained on a single NVIDIA GTX 1060 GPU with 6GB GPU memory, two batch sizes of input, segment length of 512. It takes around 48 hours to complete the training with 110 epoch and gained 0.1 average loss. The optimizer used is Adam optimizer with 2×10^{-5} learning rate. We use Mean Opinion Score (MOS) to evaluate whether our generated music has the desired emotion or not. Thirty people have listened to our ten generated samples and given a score based on the rating scale. The average loss for the model is shown in Figure 5.

A. Mean Opinion Score Result

From 10 samples, 8 of them use the PEMIDI dataset, and 2 of them use the VGMIDI dataset. We compare sample 7 with sample 10 and sample 8 with sample 9 to see which one is more enjoyable. The result is 67.7% are choose sample 10 and 54.8% choose sample 9. The comparison result may indicate that our PEMIDI dataset may be a better suit for music generation with given emotion. The overall average result can be seen in Table II.

TABLE II MOS RESULT

Sample	Expected Emotion (Score)	Dataset	Average User Rating Score
1	Negative (0-4)	PEMIDI	5
2	Neutral (5)	PEMIDI	9
3	Positive (6-10)	PEMIDI	5
4	Negative (0-4)	PEMIDI	1
5	Neutral (5)	PEMIDI	5
6	Positive (6-10)	PEMIDI	8
7	Positive (10)	VGMIDI	10
8	Negative (0)	VGMIDI	0
9	Negative (0)	PEMIDI	0
10	Positive (10)	PEMIDI	10

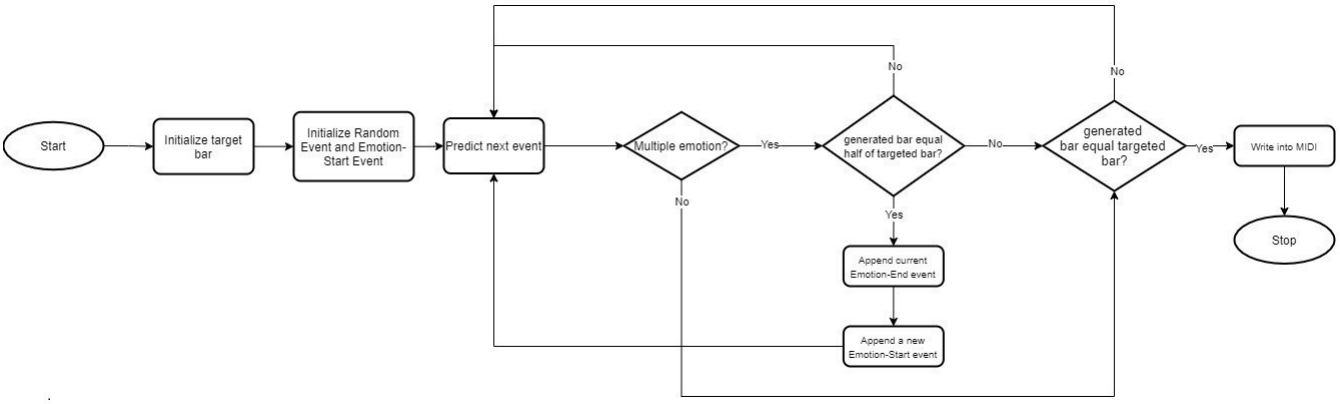


Fig. 4. Generating process flowchart

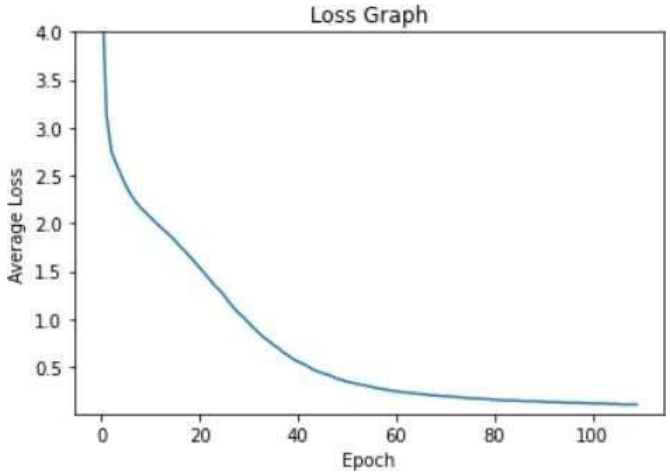


Fig. 5. Average loss graph plot

Based on the result from Table II, 70% of samples are correctly identified according to their given emotion. All of the samples that are using VGMIDI are correctly identified for their emotion. It can be concluded that Transformer-XL using REMI and emotion token can generate music with a given emotion. The rating scale is shown in Figure 6

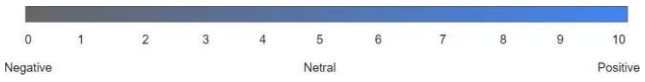


Fig. 6. MOS Rating Score

B. Multiple Emotion Generation

We also experimented by generating the music using multiple emotion tokens. For every half of the targeted bar (half of the piece), we add the Emotion-End token (to end current emotion) and Emotion-Start token (to add a new emotion). The model then will predict the next token based on the new Emotion-Start token. The results show that although this way of generating music with multiple emotions works [1], there are still many shortcomings in the quality of the music. The audio and MIDI example of our generated music can be seen in our public cloud storage [2].

C. Musical Structure Evaluation

The appearance of the repeated structure is why we tend to enjoy music and feel the emotion from it [25]. Musical structure is how melody, chord, tempo, and other musical elements are being placed and played. Based on our generated samples, it clears that our samples lack musical structure. We then analyze the *structureness* of multiple emotion samples using the fitness scape plot algorithm [26], [27]. After that, we visualize it using SM Toolbox [28] and compare it to human-composed music. The fitness scape plot is a matrix of $S_{N\times N}$, where $s_{ij} \in [0, 1]$ is the fitness, namely, the degree of repeat in the piece derived from the self-similarity matrix (SSM) [29], of the segment specified by (i, j). N is the number of frames sampled from the audio of a piece, the 1st axis represents the segment duration (in frames), and the 2nd axis represents the center of the segment (in frames) [24]. We chose the multiple emotion sample because it has the longest duration to be compared to full-length human-composed music. This method of analyzing *structureness* of music is based on the Jazz Transformer paper [24].

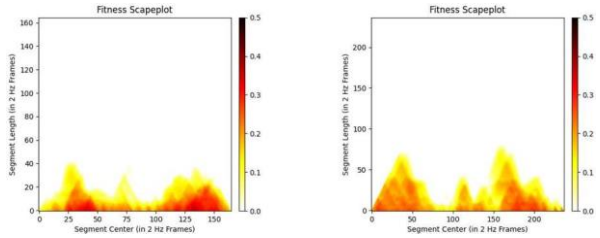
The result in Figure 7 shows that in our generated samples, some traces of repetitive structures disappear at the timescale of around 40, whereas in the human-composed music, not only do the fitness values stay high in longer timespans, but a clear sense of section is also present, as manifested by the large triangle in its scape plot. Based on that, it can be concluded that our generated sample still lacks musical structure.

V. CONCLUSION

Computer-generated music is still far behind human-composed music. Based on our survey, the PEMIDI dataset may have a better quality than VGMIDI, but it still needs further research because we cannot fully use all of the PEMIDI and VGMIDI datasets. The emergence of emotion is present, and the majority of the emotions are as expected. It can be

[1]The model can continue to generate the next token based on the newadded emotion token.

[2]https://bit.ly/300SjCk



(a) Generated sample (b) Human composed sample

Fig. 7. Fitness scape plots

concluded that the Transformer-XL model using REMI and emotion token is suitable for emotion-based music generation. The lack of *structureness* from the music generated is the cause of why the music is not as enjoyable as human-composed music. It can be due to our limited usage of PEMIDI dataset for training the model. For future works, more datasets can be used for training the model. MAESTRO dataset [30] can be included to add more data for the model. A new token can be added, like a time signature, to indicate the time signature used so the model may have a neater musical structure.

ACKNOWLEDGMENT

The author would like to express gratitude to family, friends, and colleagues from Telkom University that made this research possible.

REFERENCES

[1] Juslin, P. N. (2019). Musical Emotions Explained. Oxford University Press.

[2] Juslin, P. N., Harmat, L., Eerola, T. (2014). What makes music emotionally significant? Exploring the underlying mechanisms. Psychology of Music, 42(4), 599–623. https://doi.org/10.1177/0305735613484548

[3] Juslin, P. N., Sloboda, J., Handbook of Music and Emotion: Theory, Research, Applications, ser. Series in affective science. OUP Oxford, 2011.

[4] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., Eck, D. (2018). Music Transformer. 1–14. http://arxiv.org/abs/1809.04281

[5] Bahdanau, D., Cho, K. H., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December(Nips), 5999–6009.

[7] Huang, Y.-S., Yang, Y.-H. (2020). Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. http://arxiv.org/abs/2002.00212

[8] Hsiao, W.-Y., Liu, J.-Y., Yeh, Y.-C., Yang, Y.-H. (2021). Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. http://arxiv.org/abs/2101.02402

[9] Russell, James. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology. 39. 1161-1178. 10.1037/h0077714.

[10] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., Salakhutdinov, R. (2020). Transformer-XL: Attentive language models beyond a fixed- length context. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2978–2988. https://doi.org/10.18653/v1/p19-1285

[11] Payne, C. M. (2019). MuseNet. OpenAI Blog. https://openai.com/blog/musenet/

[12] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," arXiv [eess.AS], 2020

[13] S. Dieleman, A. van den Oord, and K. Simonyan, "The challenge of realistic music generation: modelling raw audio at scale," arXiv [cs.SD], 2018.

[14] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck and Karen Simonyan. This Time with Feeling: Learning Expressive Musical Performance, 2018; arXiv:1808.03715.

[15] Xianchao Wu, Chengyuan Wang and Qinying Lei. Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes, 2020; arXiv:2007.07244.

[16] Ferreira, L. N., Whitehead, J. (2021). Learning to Generate Music With Sentiment. http://arxiv.org/abs/2103.06125

[17] Madhok, R., Goel, S., Garg, S. (2018). SentiMozart: Music generation based on emotions. ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence, 2(Icaart), 501–506. https://doi.org/10.5220/0006597705010506

[18] Huanru Henry Mao, Taylor Shin and Garrison W. Cottrell. DeepJ: Style-Specific Music Generation, 2018; arXiv:1801.00887. DOI: 10.1109/ICSC.2018.00077.

[19] Suyanto, S., Romadhony, A., Sthevanie, F., Ismail, R.N., (2021). Augmented words to improve a deep learning-based Indonesian syllabifica- tion. Heliyon, Vol. 7, Issue 10.

[20] Suyanto, S., Arifianto, A., Sirwan, A., Rizaendra, A. P., (2020). End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language. In Proc. 8th International Conference on Information and Communication Technology (ICoICT).

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 1–21. http://arxiv.org/abs/2010.11929

[22] Cho, Y. H., Lim, H., Kim, D. W., Lee, I. K. (2017). Music emotion recognition using chord progressions. 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings, 2588–2593. https://doi.org/10.1109/SMC.2016.7844628

[23] Kong, Q., Li, B., Song, X., Wan, Y., Wang, Y. (2020). High-resolution piano transcription with pedals by regressing onsets and offsets times. ArXiv, 1–10.

[24] Wu, S.-L., Yang, Y.-H. (2020). The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures. Section 5. http://arxiv.org/abs/2008.01307

[25] Daniel J. Levitin. This is Your Brain on Music: The Science of a Human Obsession. Dutton, 2006.

[26] Meinard M üller, Peter Grosche, and Nanzhu Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In Proc. International Conference on Music Information Retrieval(ISMIR), pages 615–620, 2011.

[27] Meinard M üller and Nanzhu Jiang. A scape plot representation for visu- alizing repetitive structures of music recordings. In Proc. International Conference on Music Information Retrieval (ISMIR), pages 97–102, Porto, Portugal, 2012.

[28] Meinard M üller, Nanzhu Jiang, and Harald G. Grohganz. SM Toolbox: MATLAB implementations for computing and enhancing similarity matrices. In Proc. Audio Engineering Society (AES), 2014.

[29] Jonathan Foote. Visualizing music and audio using self-similarity. In Proc. ACM International Conference on Multimedia, pages 77–80, 1999.

[30] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. Z. A., Dieleman, S., Elsen, E., Engel, J., Eck, D. (2018). Enabling factorized piano music modeling and generation with the maestro dataset. ArXiv, 1–12.