2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP) | 978-1-6654-3288-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/MMSP53017.2021.9733502

Evaluating Deep Music Generation Methods Using Data Augmentation

Toby Godwin
GLAM
Imperial College London
London, UK
tobygodwin1@gmail.com

Georgios Rizos
GLAM
Imperial College London
London, UK
georgios.rizos12@imperial.ac.uk

Alice Baird
EIHW
University of Augsburg
Augsburg, Germany
alice.baird@uni-a.de

Najla D. Al Futaisi
GLAM
Imperial College London
London, UK
n.al-futaisi18@imperial.ac.uk

Vincent Brisse
GLAM
Imperial College London
London, UK
v.brisse@gmail.com

Björn W. Schuller
GLAM
Imperial College London
London, UK
bjoern.schuller@imperial.ac.uk

***Abstract*—Despite advances in deep algorithmic music generation, evaluation of generated samples often relies on human evaluation, which is subjective and costly. We focus on designing a homogeneous, objective framework for evaluating samples of algorithmically generated music. Any engineered measures to evaluate generated music typically attempt to define the samples' musicality, but do not capture qualities of music such as theme or mood. We do not seek to assess the musical merit of generated music, but instead explore whether generated samples contain meaningful information pertaining to emotion or mood/theme. We achieve this by measuring the change in predictive performance of a music mood/theme classifier after augmenting its training data with generated samples. We analyse music samples generated by three models – SampleRNN, Jukebox, and DDSP – and employ a homogeneous framework across all methods to allow for objective comparison. This is the first attempt at augmenting a music genre classification dataset with conditionally generated music. We investigate the classification performance improvement using deep music generation and the ability of the generators to make emotional music by using an additional, emotion annotation of the dataset. Finally, we use a classifier trained on real data to evaluate the label validity of class-conditionally generated samples.**

***Index Terms*—music generation evaluation, data augmentation, music genre classification, music emotion classification**

I. Introduction

Recent advances in generative algorithms have been able to produce [1]–[3] realistic sounding music in the waveform domain, with large scale models now able to generate full length songs, including comprehensible lyrics [1]. While these models have produced impressive results, their evaluation has tended to rely on human evaluation [2]–[6], which is inherently subjective and extremely costly to scale. Alternatively, evaluation has been approached by monitoring quantities related to reconstruction quality [1], or engineered measures for evaluating musical merit, which are currently inadequate due to the vastness and diversity of the space of all music [2].

The enjoyment of music is deeply personal. So instead of performing a potentially subjective, and difficult to scale human perception survey, or attempting to engineer a mathematical measure for comprehensively quantifying whether generated music is enjoyable or interesting to listen to, we want to perform a *functional evaluation of class-conditionally generated music samples*: i. e., we want to assess whether generated samples contain the targeted class information, focusing on mood/theme and emotion classes. We present a framework, based on data augmentation, for the evaluation of samples of generated music, that offers a homogeneous and fair treatment across generation methods. Specifically, we quantitatively assess and compare the extent to which generated samples contain meaningful information pertaining to music of a given mood/theme. We report the change in predictive performance of a machine classifier as an indicator of whether there is meaningful information in the generated samples related to the particular classification task. We examine the generated samples in an additional manner; by using a machine classifier trained on the real training set to classify the generated samples, in order to validate whether they carry features that are similar to the real training data. Finally, we perform the same experiments on an almost comprehensive subset of our dataset with a relabelling, which we introduce, pertaining to coarser arousal/valence emotion classes. We use this tofurther investigate the properties of the generated samples and gain insights on the augmentation experimental performance.

We analyse the generated samples from three deep music generation methods of different philosophies: the autoregressive SampleRNN [3], Vector Quantised VAE (VQ-VAE) [7] based Jukebox [1], and Differentiable Digital Signal Processing (DDSP) [4]. While the former two have been shown

This work is funded by the UK Economic Social Research Council (UK-ESRC) through the research Grant No. HJ-253479 (ACLEW), the Engineering and Physical Sciences Research Council (EPSRC) Grant No. 2021037, and the the DFG's Reinhart Koselleckproject No. 442218748 (AUDI0NOMOUS).
978-1-6654-3288-7/21/$31.00 ©2021 IEEE

to generate commercial music [1], [3], [8], the latter was proposed in the context of monophonic audio. Here we explore its ability to reconstruct *polyphonic music*. While each method was presented with some subjective evaluation (often by the authors), there has yet been little effort to quantitatively and objectively evaluate generated samples. We propose to do this via data augmentation. Finally, this is also the first time such music generation methods have been considered for data augmentation of musical classification.

In the following section we briefly discuss common methods for evaluating generated music as well as the generative models used in this study. In Section 3 we discuss the data we used. Then in Section 4 we outline the details of the evaluation framework and discuss results. In Section 5 make a further analysis of the generated samples and conclude in Section 6.

II. Background

The evaluation of music is an inherently challenging task given the subjective nature of music itself [9]–[11]. It is common to employ multiple human annotators to evaluate each sample [11] in order to elicit confident insights, something that hinders scalability. Recent work [12] proposes simple, musically informed metrics to evaluate the musicality of generated music. However these metrics do not capture abstract qualities of music such as its emotion or mood. In other research, evaluation of deep music generation tends to focus on subjective surveys [2], [4], or the reporting of reconstruction loss related quantities [1].

In the emotional speech synthesis domain (another subjective audio domain), the generative adversarial network based studies performed in [13]–[15] quantitatively evaluated synthesised speech using a classifier to demonstrate that the generated speech contained meaningful emotional information. We draw from these techniques to assess and compare the ability of three generative models from different deep learning paradigms to generate music that contains meaningful information relating to mood/theme or a listener's emotional response.

A. Generative Models for Music

There have been significant advancements in the field of music generation, particularly in the waveform domain [1], [2], [4]–[6], [16], [17]. We functionally evaluate the performance of three generative models, which subscribe to different generation paradigms.

SampleRNN. The autoregressive SampleRNN [3] uses tiers of RNN modules that work on different timescales of the signal. Higher level tiers process the signal at lower temporal resolutions to lower level tiers, and each tier is conditioned on a vector from the tier above, which contains higher level contextual information. The tiered architecture of the SampleRNN allows for different computational focus to be applied to different levels of abstraction of the audio, which allows long term dependencies to be modelled efficiently.

Jukebox. Jukebox [1] trains three separate convolutional VQ-VAEs [7] at different levels of abstraction. Each level learns a discrete latent codebook of the input, so Jukebox learns three separate representations of the input data. Prior distributions of the latent codes are approximated using autoregresive methods [18], and music is generated by sampling from the latent codebooks, upsampling the lower resolution codes and decoding a single high-resolution latent representation to produce music.

	Train	**Val**	**Test**	$\sum$
Mood/theme	160	22	22	204
Emotional	157	20	20	197

TABLE I
Train, Val(idation), and Test partition durations (hours) for mood/theme and emotional labels.

DDSP. DDSP [4] learns the parameters of deterministic digital signal processing techniques; used then to synthesise audio. The model itself is an autoencoder acting on Mel-spectrograms, where the latent representation is decomposed into: the time-varying fundamental frequency of the audio F_0, loudness l, and a latent vector that encodes the input z. All latent representations are time-varying and sequential, and since F_0 and l have interpretable, physical meanings. Here, we only consider the harmonic plus noise synthesiser model [19], [20].

III. MTG-Jamendo Dataset

For our experiments we use the MTG-Jamendo dataset [21], which is a large collection of labelled, high-quality commercial music. We use a subset of MTG-Jamendo, labelled according to mood/theme, which was used in the MediaEval 2019 music classification competition [22]. There are a total of 56 well-balanced classes ranging from 'epic' to 'dance'. In order to gain insights respective to which types of music the generation models can generate, we utilise a second label set: this is an almost fully overlapping subset of the first, on which we performed a more coarse relabelling based on a psychology-based interpretation of the original labels. A summary of the duration of data partitions is given in Table I for both sets of labels. The reader is referred to [21], [22] for statistics of data with mood/theme labels.

A. Music Emotion Labels

We derive the emotional labels from a theory of emotion that takes into account a person's categorical emotion responses in terms of arousal and valence [23]. Valence is a psychological term that describes the intrinsic pleasure that a person derives from something and arousal describes the amount of attention that a person pays to something. Therefore, the labels ('activated pleasant', 'activated unpleasant', 'deactivated pleasant', 'deactivated unpleasant') describe the expected arousal and valence responses of an individual when listening to the music. We label the music by grouping data such that the mood/theme labels are mapped to quadrants of the circumplex model of emotions [23] (such coarse binning of arousal and valence into classes is not uncommon [24]): e. g., since 'happiness' is mapped to activated pleasant according to the model, tracks

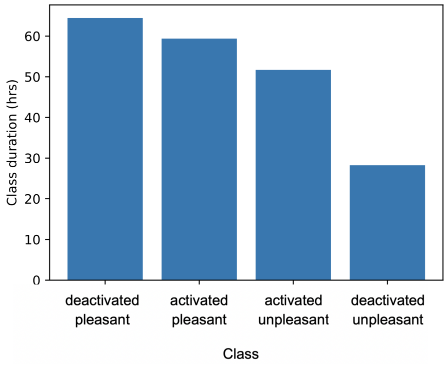


Fig. 1. Distribution of total duration of music for each emotional class for the dataset used in this study. Whereas there is a degree of class imbalance we note that it is not overwhelming.

with the ‘happy’ mood/theme label are labelled as ‘activated pleasant’. This process is extended for all mood/theme labels that are clearly mapped to evoked arousal and valence and we include the mapping thereof in the project webpage[1].

IV. Evaluation Framework

We assess whether the generated samples contain meaningful information pertaining to music of a given mood/theme or emotion by measuring whether their usage for data augmentation increases the predictive performance of a classifier on two tasks (mood/theme and emotional response). We compare classification performance after augmentation to the performance of a baseline classifier trained only on the training partition. The relatively strong baseline classification performance implies that the abstract features of thematic and emotional music are indeed modelled by the classifier. Therefore, we can say that an improvement over the baseline after augmentation indicates that, indeed, the samples contain meaningful information relating to music of a given mood/theme or emotion.

We measure classifier performance according to the three metrics that were used in the MediaEval 2019 competition: F1-score, area under the precision-recall curve (PR-AUC), and the area under the Receiver Operator Characteristics curve (ROC-AUC). We adopt two means of averaging each metric: a) micro-averaged, which is the harmonic mean of the overall metric scores and b) macro-averaged, which performs an unweighted average of class-specific metric scores. We evaluate performance on the test partition of each dataset with respect to micro and macro averaged metrics, calculated in the same manner as the MediaEval 2019 competition submission.

We use a classifier architecture [25] that came fourth in the MediaEval 2019 competition with respect to macro-averaged PR-AUC. The model was developed specifically for the data used in this work, and uses a pretrained MobileNetV2 [26] block in combination with a self attention block to classify music from its Mel-spectrogram representation. Since this classifier was developed specifically for the challenge, we used the same hyperparameters reported in its submission paper [25].

	Sample type	Sample length (s)	Prime length (s)	Sample rate (kHz)
SampleRNN	Primed	10	4	16
DDSP	Recons	4	n/a	16
Jukebox	Primed	24	8	44

TABLE II
Model properties pertaining to generating samples.

A. Augmentation Policy

For each generative method, we generate an equal duration of music per class, with the total length of generated music equal to 5 % of the duration of the train split of each dataset (8 hours for mood/theme, 7.85 hours for emotional). We then augment each class with the fixed duration of generated music. This gives a balanced treatment per class, increases the likelihood of successfully observing the augmentation effect, and allows for a fair comparison across generative methods. Since each method generates samples of different lengths (Table II), we adjust the number of samples accordingly to generate a fixed duration of music per class. We trained SampleRNN and DDSP on the training partition of each dataset and used performance on the validation partition to tune hyperparameters. We did not retrain Jukebox, and opted instead to use the pre-trained model from [1], trained on 1.2 million songs collected from the web.

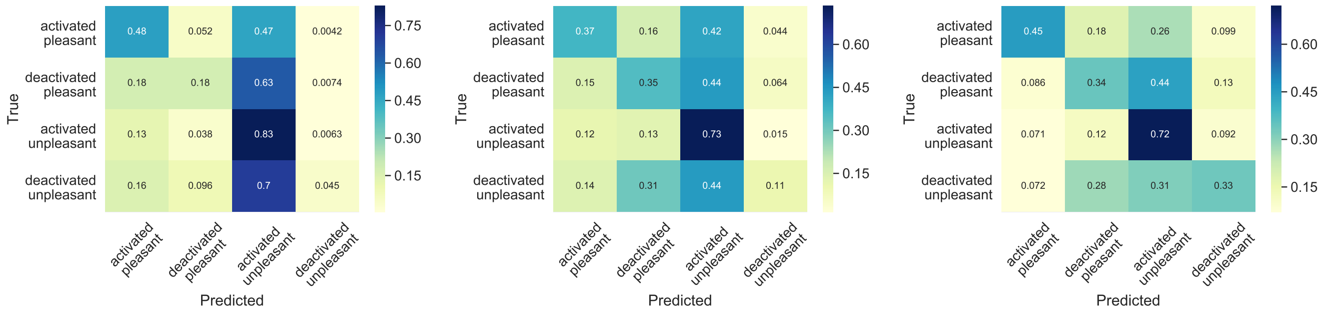
We generate samples with SampleRNN and Jukebox by ‘priming’ the model with an input length of real music to seed the sampling process. We select a prime length for each method based on its desired input and maximum capable sample length. Generated music comprises the majority of each sample, however, the priming sample length is not insignificant, which we believe justifies our assumption that the completed sequence may inherit the label from the input music for data augmentation. ‘Priming’ samples are used once, randomly sampled per class from the training partition without replacement.

As for DDSP, we reconstruct the input music sample and assume that the reconstructed sample has the same label as the original. The input music is again randomly sampled from the training partition without replacement. Although DDSP was designed for monophonic music, it can also reconstruct polyphonic music since it is trained to reconstruct a spectrogram.

V. Data Augmentation Experiments

Table III summarises our data augmentation experimental results on both label types. All values reported are averaged across two trials. We observe from Table III(a) that DDSP’s samples yielded a consistent increase in performance with respect to micro and macro averaged metrics; a quantitative indication that these samples contain meaningful information. Comparing our performance to the blind evaluation of the MediaEval 2019 competition would yield an absolute improvement to the baseline submission of 0.8 % with respect

[1] https://github.com/glam-imperial/Functional-Music-Generation-Evaluation



(a) SampleRNN

(b) Jukebox

(c) DDSP

Fig. 2. Normalised confusion matrices for classification of generated samples with emotional labels. The three music generation methods tend to produce samples that are perceived by the classifier as ‘activated-unpleasant’. DDSP has a stronger diagonal in the confusion matrix, and is the only one that achieves good performance in predicting the ‘deactivated-unpleasant’ class.

Model	F1		PR-AUC		ROC-AUC	
	Macro	Micro	Macro	Micro	Macro	Micro
SampleRNN	0.063	0.137	0.134	0.158	0.752	0.799
DDSP	0.066	0.146	0.134	0.159	0.761	0.807
Jukebox	0.063	0.133	0.125	0.151	0.744	0.795

(a) Mood/theme labels

Model	F1		PR-AUC		ROC-AUC	
	Macro	Micro	Macro	Micro	Macro	Micro
SampleRNN	0.483	0.523	0.515	0.565	0.764	0.774
DDSP	0.489	0.526	0.511	0.565	0.759	0.771
Jukebox	0.486	0.532	0.524	0.587	0.771	0.783

(b) Emotional (arousal/valence) labels

Labels	F1		PR-AUC		ROC-AUC	
	Macro	Micro	Macro	Micro	Macro	Micro
Mood/theme (ours)	0.064	0.134	0.125	0.148	0.745	0.794
Emotional	0.485	0.523	0.524	0.574	0.764	0.771

(c) Baseline classification performance.

TABLE III

Predictive performance of baseline classifier on test partition, and test partition classification results after data augmentation. We highlight the cells that have at least 1 % increase in performance relative to the performance of our replication of the baseline classifier.

to macro-averaged PR-AUC and ROC-AUC, promoting the method from 4th to 3rd place with respect to macro averaged ROC-AUC and PR-AUC. This is an important improvement, especially given that the performance achieved by our replication of the classifier proposed in [25] was lower than the authors’ reported *submission* performance, particularly with respect to macro-averaged ROC-AUC. Samples generated by SampleRNN and Jukebox failed to result in consistent performance increases to the same degree, although they were at least as good as the baseline in terms of PR-AUC and ROC-AUC, and competitive in terms of F1. One hypothesis for this behaviour is that these are both priming based generation methods: the final generated sample may deviate significantly from the priming sample class, impacting the augmentation behaviour in this numerous class setting.

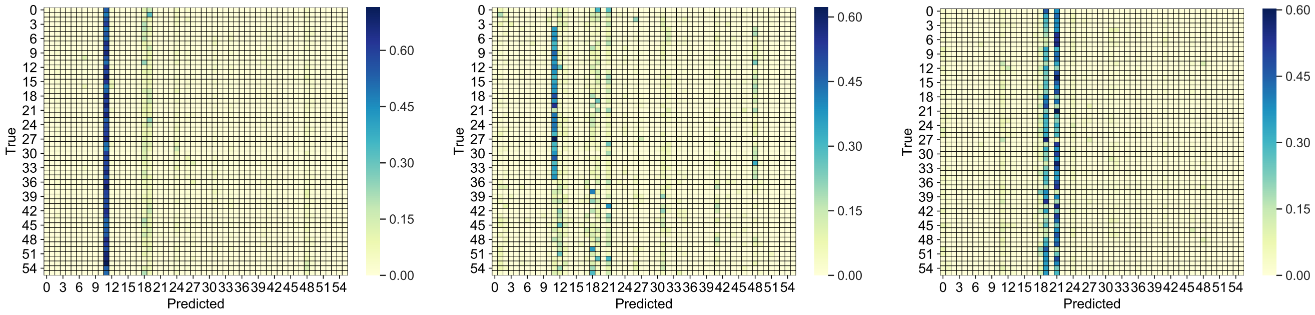
For emotional labels, we observe from Table III(b) that samples generated by SampleRNN and DDSP yield negligible performance increase. Only Jukebox samples yield consistent performance increase with respect to micro-averaged metrics, implying that performance was unequal across classes. We investigated and found that across all trials, only the ‘activated pleasant’ class consistently improved the classifier after augmentation. This implies that all methods could only generate music with meaningful content from ‘activated pleasant’ music.

VI. Classification of Generated Samples

We use the pre-trained baseline classifier to identify whether there is consistency between dominant features in the generated samples and characteristic features in the real training data. We measure the testing predictive performance of the classifier on the generated samples. We observe from Figure 2 that samples tend to be classified as ‘activated unpleasant’. All samples contain a degree of noise, and we hypothesise that this noise correlates with characteristic features of ‘activated unpleasant’ music. Similarly for mood/theme labels, Figure 3 shows that all samples from all methods tend to be classified as having a ‘dark’ mood or theme (‘11’ in Figure 3). The classifier predicts a greater variety of classes for both DDSP and Jukebox, implying that samples generated by these methods contain features that correlate to characteristic features of a range of classes. Figure 3 shows that DDSP’s mood/theme samples tend to be classified as ‘epic’ or ‘film’. Since DDSP reconstructs music from deterministic estimates of loudness, elements such as percussion, and changes in dynamics are often well captured in reconstructed samples, which we believe are characteristic features of ‘epic’ and ‘film’ music.

VII. Conclusion & Future Work

Figure 3 reveal that no model could generate music with meaningful features for *all* classes of each dataset. We further show that each model is particularly effective at generating meaningful music for certain classes, indicating that each model has a ‘character’ in terms of the music it generates. Figure 2 indicates that DDSP generalised best to all classes, implying that polyphonic music, reconstructed by DDSP, maintains much of its meaningful information. We demonstrate, by comparison to the MediaEval 2019 competition, that music genre classifiers are improved by using reconstructed samples



(a) SampleRNN

(b) Jukebox

(c) DDSP

Fig. 3. Normalised confusion matrices for classification of generated samples with mood/theme labels. The SampleRNN samples tend to be predicted as belonging to the ‘dark’ class. Greater range of possible predicted classes is observed for the Jukebox samples, albeit ‘dark’ is still somewhat dominant. The DDSP samples are predicted mostly as ‘epic’ or ‘film’. We see that existing music generation methods implicitly bias the generated samples with specific thematic indices.

from DDSP (less so from SampleRNN) to augment training data; however, this effect was not observed in the emotion classification experiment, where only JukeBox achieved some improvement in Micro averages of the performance measures used in MediaEval 2019.

We firmly believe that objective evaluation of music generation methods is key, in order to fairly monitor progress in this domain. Our experiments show that all models struggled to generate music with dominant features that were similar to characteristic features of a given theme/mood class, possibly due to artefacts present in the generated music.

This analysis depends on the behaviour of the classifier, so future work should explore the effect of different classifier architectures. Since the generated samples will inevitably be from a different distribution to the training data, using domain adversarial training [27], [28] would help to uncover domain invariant features in the generated samples, and is therefore an interesting avenue for further research. We have also assumed that generated samples strictly inherit the label from an original or priming sample: as we have seen, this is not necessarily the case, with only the DDSP method being consistent in its class-conditional generation, for the emotion prediction task. To that end, we believe that a more elaborate consideration of the confidence we have regarding the augmentation labels is required, and intend to explore this matter with regularisation methods like virtual adversarial training [29].

References

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” *arXiv preprint arXiv:2005.00341*, 2020.

[2] S. Dieleman, A. van den Oord, and K. Simonyan, “The challenge of realistic music generation: modelling raw audio at scale,” in *Advances in Neural Information Processing Systems*, 2018, pp. 7989–7999.

[3] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “Samplernn: An unconditional end-to-end neural audio generation model,” *arXiv preprint arXiv:1612.07837*, 2016.

[4] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, “Ddsp: Differentiable digital signal processing,” in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[6] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[7] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” in *Advances in Neural Information Processing Systems*, 2019, pp. 14 837–14 847.

[8] C. Carr and Z. Zukowski, “Generating albums with samplernn to imitate metal, rock, and punk bands,” *arXiv preprint arXiv:1811.06633*, 2018.

[9] H. Katayose, M. Hashida, G. De Poli, and K. Hirata, “On evaluating systems for generating expressive music performance: the rencon experience,” *Journal of New Music Research*, vol. 41, no. 4, pp. 299–310, 2012.

[10] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt, “Metrics for deep generative models,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1540–1550.

[11] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. W. Schuller, “The perception of emotion in the singing voice: The understanding of music mood for music organisation,” in *Proceedings of the 4th International Workshop on Digital Libraries for Musicology*, 2017, pp. 29–36.

[12] L.-C. Yang and A. Lerch, “On the evaluation of generative models in music,” *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[13] F. Bao, M. Neumann, and N. T. Vu, “Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition.” in *INTERSPEECH*, 2019, pp. 2828–2832.

[14] G. Rizos, A. Baird, M. Elliott, and B. Schuller, “Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[15] A. Baird, S. Amiriparian, and B. Schuller, “Can Deep Generative Audio be Emotional? Towards an Approach for Personalised Emotional Audio Generation,” in *Proc. International Workshop on Multimedia Signal Processing (MMSP)*. Kuala Lumpur, Malaysia: IEEE, 2019, 5 pages.

[16] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *arXiv preprint arXiv:1601.06759*, 2016.

[17] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, “Music transformer,” in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJe4ShAcF7

[18] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers,” *arXiv preprint arXiv:1904.10509*, 2019.

[19] X. Serra and J. Smith, “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.

[20] J. W. Beauchamp, *Analysis, synthesis, and perception of musical sounds*. Springer, 2007.

[21] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The mtg-jamendo dataset for automatic music tagging,” in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. [Online]. Available: http://hdl.handle.net/10230/42015

[22] D. Bogdanov, A. Porter, P. Tovstogan, and M. Won, “Mediaeval 2019: Emotion and theme recognition in music using jamendo,” in *MediaEval 2019 Workshop*, 2019.

[23] J. Posner, J. A. Russell, and B. S. Peterson, “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology,” *Development and psychopathology*, vol. 17, no. 3, p. 715, 2005.

[24] Z. Zhang, B. Wu, and B. Schuller, “Attention-augmented end-to-end multi-task learning for emotion prediction from speech,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6705–6709.

[25] M. Sukhavasi and S. Adapa, “Music theme recognition using cnn and self-attention,” *arXiv preprint arXiv:1911.07041*, 2019.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[27] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[28] M. Abdelwahab and C. Busso, “Domain adversarial for acoustic emotion recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[29] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, “Virtual adversarial training: a regularization method for supervised and semi-supervised learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.