Computer Evaluation Of Musical Timbre Transfer On Drum Tracks

Keon J. Lee[1], George Tzanetakis[1], and Philippe Pasquier[2]

[1] MISTIC Lab, Computer Science, University of Victoria,
[2] Metacreation Lab, Interactive Arts and Technology, Simon Fraser University
keon18@uvic.ca

Abstract. Musical timbre transfer is the task of re-rendering the musical content of a given source using the rendering style of a target sound, this task is sometimes also called interpretation. The source keeps its musical content, e.g., pitch, microtiming, orchestration, and syncopation. We specifically focus on the task of transferring the style of percussive patterns extracted from polyphonic audio using a MelGAN-VC model (Pasini, 2019) by training acoustic properties for each genre. Evaluating audio style transfer is challenging and typically requires user studies. We propose an analytical methodology based on supervised and unsupervised learning including visualization for evaluating musical timbre transfer that can be used instead of, or in addition to, user studies. The proposed methodology is used to evaluate the MelGAN-VC model for musical timbre transfer of drum tracks.

Keywords: Audio Style Transfer, GANs, Evaluation of Audio Style Transfer, Methodology

1 Introduction

Image style transfer (Gatys, Ecker, & Bethge, 2015) requires two images as input: a content (source) image and a target image. The goal of the neural style transfer model is to generate a new image representation which has the content of the first image rendered using the painting style of the target image. For example, the Wassily Kandinsky example[3] is a typical case of the Neural Style Transfer (NST). After the emergence of neural image style transfer, style transfer in audio domain was explored (Ulyanov & Lebedev, 2016; Mital, 2017; Grinstein, Duong, Ozerov, & Pérez, 2018; Verma & Smith, 2018). There is an increasing work for audio style transfer using Generative Adversarial Networks (GANs). Several GAN architectures have been proposed including CycleGAN (Zhu, Park, Isola, & Efros, 2017), WaveGAN (Donahue, McAuley, & Puckette, 2018) and MelGAN (Kumar et al., 2019). Music style transfer in particular is not well-established topic (Dai, Zhang, & Xia, 2018). One possibility for music audio style transfer is affecting the perceived genre of a source piece to be similar to a target musical

[3] https://www.tensorflow.org/tutorials/generative/style_transfer

genre. For example, one can transform a piece of hip-hop music to sound like it is played using heavy metal instrumentation. Another challenge in neural music genre transfer is that music genre does not have a well-defined definition and it is mostly based on subjective perception.

When it comes to music representations used for audio style transfer, there are two main possibilities: symbolic or audio. For symbolic music transfer, there were several successful cases mostly using GANs and VAEs (Variational Autoencoders) (Brunner, Wang, Wattenhofer, & Zhao, 2018; Brunner, Konrad, Wang, & Wattenhofer, 2018; Brunner, Moayeri, Richter, Wattenhofer, & Zhang, 2019; Cífka, Şimşekli, & Richard, 2020). With regards to the audio-based music genre transfer, there are only a few recent instances (Lu, Xue, Chang, Lee, & Su, 2019; Cífka, Ozerov, Şimşekli, & Richard, 2021) since generating audio is typically considered as a more difficult problem, in comparison to generating symbolic patterns especially when using GANs. In other words, audio-based GAN models tend to require more time to train and are more challenging as they frequently contain random noise and it is difficult to achieve high-quality audio.

2 Motivation

The motivation for this research is to explore how GANs can be used to perform music style transfer in the audio domain. More specifically, we focus on two problems: 1) experimenting with the MelGAN-VC for transforming drum parts of arbitrary polyphonic audio from one music genre to another musical genre, 2) proposing a computer-based methodology based on supervised and unsupervised learning that can be used to evaluate musical timbre transfer that does not require user studies. We decided to work on drum track style transfer using MelGAN-VC (Pasini, 2019) based on a wav-to-wav approach in the audio domain. There is no well-established quantitative evaluation methodologies for neural audio (music) style transfer, although there is a good example for evaluating style imitation corpora (Ens & Pasquier, 2018) and quantifying musical style in symbolic domain (Ens & Pasquier, 2020), respectively. Therefore, we propose new quantitative evaluation methodologies for this particular purpose. In most music generative systems, it is ideal to have human judgements who could tell the difference between music genres easily and evaluate the results; however, it is also good to have computational evaluation methods since it is more efficient and easier to integrate with a creative AI system. In addition, the subjectivity of music genre makes human listeners evaluation more challenging.

We study the musical timbre transfer problem in which transfer acoustic properties from hip-hop drum tracks are transferred to metal drum tracks, and vice versa. Temporal properties and syncopation are not considered in this case and all drum tracks are extracted by Spleeter on original polyphonic music tracks. The MelGAN-VC model is used to achieve the musical timbre transfer on drum tracks. For the computational evaluation, audio features are extracted from drum parts and analyzed based on the musical genre. Thus, each genre

has different acoustic properties and can be classified into genres using machine learning algorithms, such as supervised and unsupervised learning including dimensionality reduction. This task could be called percussive genre classification. And the output of the musical timbre transfer is holistically evaluated applying the percussive genre classification technique. Research contributions to the musical timbre transfer are:

1. Experiments using MelGAN-VC for musical timbre transfer of drum tracks.
2. Pipelines for quantitative evaluation of musical timbre transfer (Fig. 3.).
3. Exploration of audio feature importance for evaluating musical timbre transfer on drum tracks.
4. Evaluation for musical timbre transfer of drum tracks using supervised and unsupervised learning with visualization.

3 Neural Musical Timbre Transfer Training

The purpose for this neural style transfer is to achieve the transformation of timbre information from hip-hop (original genre) to metal (target genre) on drum tracks. Outputs are to retain timbral characteristics of the original source genre while also exhibiting timbral characteristics of the target genre. To evaluate the transfer effectiveness, we can use the probabilistic output of a trained binary classifier that discriminates between the source and target genre. If the transfer is effective, we expect that the probability of classification of the source genre will be reduced. And the transformed recording will exhibit timbral characteristics of the target genre. This change in probability of classification will depend on the specific drum track that is being transformed and the training epoch of the MelGAN-VC.

The procedure for MelGAN-VC training for musical timbre transfer is shown in Fig. 1. Musical timbre transfer on drum tracks was trained using the MelGAN-VC architecture which is based on the Transformation Vector Learning GAN (TraVeLGAN) (Amodio & Krishnaswamy, 2019). The TraVeLGAN architecture, which adds a siamese network to the generator and the discriminator, and trains to preserve vector arithmetic between points in the latent space of the siamese network. Thus, the architecture can preserve semantic information in spectrograms. For experiments, GTZAN dataset[4] (Tzanetakis, Essl, & Cook, 2001), which contains 100 tracks (30-second long each) for each 10 genres, was used. Source separated drum GTZAN stems by Spleeter (Hennequin, Khlif, Voituret, & Moussallam, 2020) were trained for 4000 epochs to transfer the style from hip-hop to metal. 4000-epoch was chosen when human listeners think the output sounds reasonable; however, proposed supervised method can provide the reasonable training epoch, when the output reach near 50% metal/50% hip-hop in this case. These two genres were selected as we expect the drum timbre for them to be distinct. We focus on the drum tracks as it is a more constrained

[4] https://www.tensorflow.org/datasets/catalog/gtzan

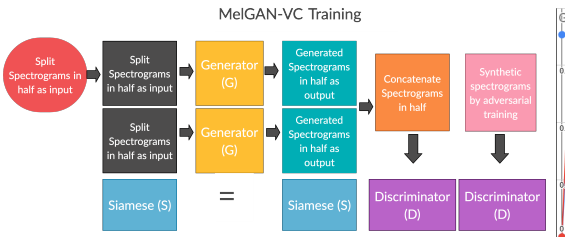

Fig. 1. MelGAN-VC training with generators, discriminators, and siamese networks. Siamese networks retain semantic information from spectrograms and assist training generators.

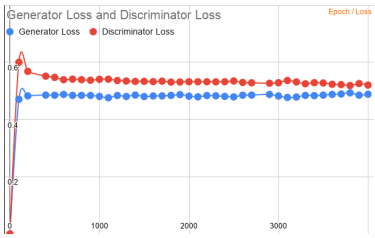

Fig. 2. MelGAN-VC generator and discriminator loss (x-axis: the number of training epochs, y-axis: loss values between 0.0 and 1.0.

task for timbre transfer than full music style which depends on a variety of factors such as vocals, instrumentation, etc. This musical timbre transfer can be viewed as analogous to style interpolation between metal and hip-hop genre. The training procedure (Fig. 1.) can be summarized as following:

1. Spectrograms (time-frequency 2-dimensional representation) were extracted from the drum tracks.
2. The spectrograms were split in half and used as input to the generator (G).
3. The style-translated halves from G were concatenated to the original shape and transferred to the discriminator (D).
4. The Siamese network (S) was added to keep semantic information (Amodio & Krishnaswamy, 2019) which captures music style.
5. Adversarial training (4000 epochs): D distinguished metal from hip-hop to improve style-transferred instances generated by G and S assisted G training by allowing translations between metal and hip-hop.
6. The missing phase information (spectrograms only include magnitude) was reconstructed by the Griffin-Lim (Griffin & Lim, 1984) algorithm.

The rationale for selecting the MelGAN-VC is that it can be trained with only 100 tracks for each genre (100: hip-hop and 100: metal) to avoid checkerboard artifacts (Odena, Dumoulin, & Olah, 2016), and it worked better than the CycleGAN especially for transferring audio style according to the original paper (Pasini, 2019).

The losses of the generator G and the discriminator D for the MelGAN-VC model after training for 4000 epochs are shown in Fig. 2. First, at least both G and D loss are not decreasing to 0.0 which is considered as a failure mode. Second, the losses are not changing drastically after 1000 epochs and the values look stabilized in a certain degree with relatively small fluctuations. Lastly, at least human listeners can tell there are differences between training epoch 100 and 2500, and the epoch 2500 sounds improved compared to the previous early epoch 100. This observation could be strengthened by performing a thorough user study but the goal is to investigate whether we can support this observation using supervised and unsupervised learning methods.

4 Evaluation Methods For Musical Timbre Transfer

Quantitative evaluation methods (Fig. 3.) including the audio feature importance (Fig. 4.) for evaluating style-transferred drum instances from the GAN model are discussed. The methodologies can be divided into two approaches: supervised and unsupervised method. Essentia (Bogdanov et al., 2013) is used to extract 110 audio features for both methods.

4.1 Audio Feature Selection

Selecting important audio descriptors for drum genre classification is an important task for this evaluation process. The evaluation models trained are dependant on the features and tend to overfit given the size of the data-set when using all 110 audio features. In addition, using 110 features is not efficient for computing evaluation pipelines. It is worthwhile to explore and get results for the importance of audio features by analyzing the binary music genre classification. Mutual information (MI) and Analysis of Variance F-value (ANOVA-F) were used for estimating the importance of each audio feature. 110 features (using Essentia extractors, excluding for temporal features and metadata, to evaluate holistically) were extracted from three genres (hip-hop, metal and rock from source separated drum tracks from GTZAN dataset) and each top 10 features (Fig. 4.) were selected based on MI and ANOVA-F, respectively. Therefore, for the supervised learning experiments, 17 features were selected. However, for the unsupervised learning experiments, we found that the ideal number was 6 and they were selected using ANOVA F-values. A noticeable point is that only one rhythm-related feature (as shown in Fig. 4.) was discovered according to the feature importance, although other audio low-level descriptors could be directly or indirectly related to rhythm.

4.2 Supervised Learning Method

Training: The supervised learning method for analyzing style transfer was trained with 17 audio features. There has been a trend that ensemble methods, such as random forest (RF) and gradient boosting algorithm, tend to have higher classification accuracy in comparison to linear models, such as Logistic Regression and Perceptrons. Therefore, the RF algorithm was chosen as the estimator for this case with the configuration of 118 trees in the forest and trained with two genres: hip-hop (label: 0) and metal (label: 1). As a result, the accuracy score turned out to be 91.0% for drum genre classification on hip-hop and metal for the GTZAN data-set. Note that the drum stems from Spleeter are used and therefore this is purely based on the separated drum track.

Testing: We tested the evaluation method with five randomly chosen instances from the style-transferred GTZAN drum tracks and one random hip-hop style drum-only track (without source separation) from outside of the GTZAN dataset.

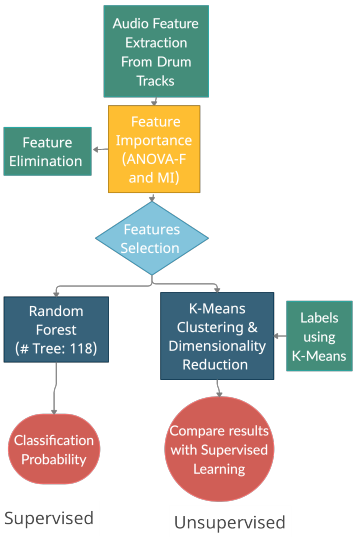

Fig. 3. Pipelines for quantitative evaluation methods of musical timbre transfer. Left: Supervised method, and right: unsupervised method.

Top 10 Audio features	ANOVA-F values
lowlevel.erbbands_skewness.stdev	50.89586639
lowlevel.loudness_ebu128.integrated	47.47364044
lowlevel.barkbands_spread.stdev	44.71256256
lowlevel.melbands_spread.stdev	37.2513504
lowlevel.spectral_rms.stdev	35.28242111
lowlevel.melbands_crest.stdev	34.80794144
lowlevel.barkbands_flatness_db.mean	32.10904312
lowlevel.melbands_kurtosis.mean	31.87611771
lowlevel.loudness_ebu128.loudness_range	30.59831238
lowlevel.barkbands_flatness_db.stdev	30.08287048

Top 10 Audio features	MI values
lowlevel.loudness_ebu128.loudness_range	0.2793418142
lowlevel.barkbands_spread.stdev	0.2785261127
lowlevel.pitch_salience.mean	0.2643023024
lowlevel.loudness_ebu128.momentary.stdev	0.2589722888
lowlevel.erbbands_skewness.mean	0.2476374769
rhythm.bpm_histogram_second_peak_weight	0.2160613488
lowlevel.loudness_ebu128.integrated	0.2113145765
lowlevel.melbands_crest.mean	0.1974811933
lowlevel.spectral_kurtosis.mean	0.1923467722
lowlevel.barkbands_skewness.mean	0.1883028758

Fig. 4. Top 10 audio feature importance for percussive genre classification. Top: ANOVA-F and bottom: Mutual information.

The results were analyzed at intervals of 500 epochs within the range of 500 to 4000 epochs from hip-hop to metal. The main purpose for this method is to reflect the trend for the progress of style transfer. Fig. 5. demonstrates evaluation results with the supervised learning pipeline for classifying six style-transferred tracks based on the RF classification probability. The ideal case for this result is where hip-hop and metal is near 50%, respectively, which proves the achievement of style interpolation between hip-hop and metal. Interpolated line graphs in Fig. 5 are demonstrating that the MelGAN-VC training is working because the testing instances are nearly converging to 50/50% as the number of training epochs increase. All line graphs are oscillating for both the discriminator and the generator loss. Top left scatter plot is showing the results for two style-transferred instances. The first test instance (hip-hop1/metal1) is the hip-hop style drum-only track and it was calculated hip-hop (50%) and metal (50%). The other five examples (hip-hop2-6/metal2-6) are more fluctuating compare to the first one. As we expected, the musical timbre transfer works better with the drum-only piece (track 1) because the source separated GTZAN drum stems still contain some portion of other stems, such as vocal and bass. When interpreting the interpolated line graphs, it is reasonable to ignore the intersection point (50/50%) before epoch 1500 because it is the early stage of the training MelGAN and tends to include more noise with music. The testing interval was 500 epochs (0, 500, 1000, etc) and the results between them, e.g., 300 and 700,

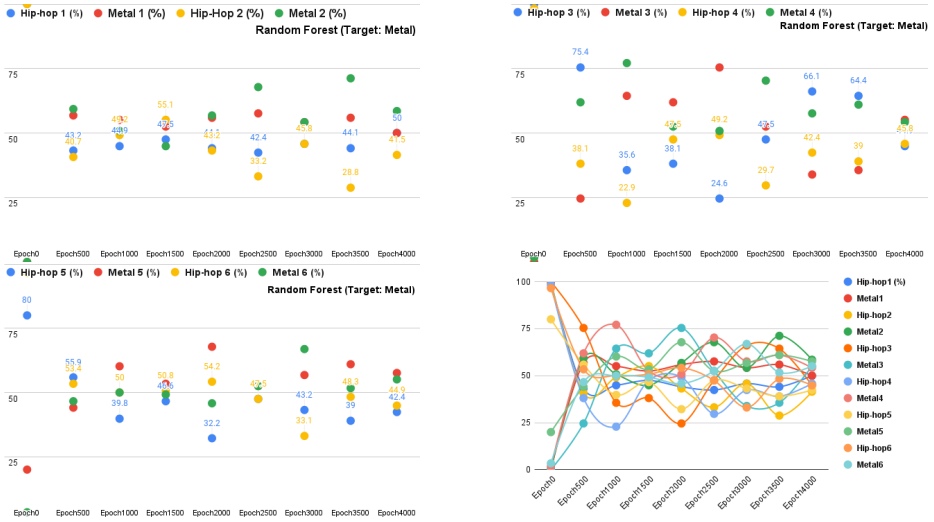

Fig. 5. Supervised method results: Track 1 & 2 (upper left), track 3 & 4 (upper right), track 5 & 6 (bottom left), and all tracks trend using interpolation (bottom right).

were interpolated; however, this could be solved by simply reducing the testing interval with the same method. According to the supervised evaluation pipeline, the best example instances[5] are listed in sequence: hip-hop/metal 1) epoch4000, 2) epoch3000, 3) epoch2500, 4) epoch2000, 5) epoch2500, and 6) epoch3500.

4.3 Unsupervised Learning Method

To investigate further how the style transformed instances relate to the original source and target genre, we employ clustering and visualization techniques. The K-means algorithm for discovering 2 clusters was used with the 200 GTZAN drum stems (100 metal and 100 hip-hop each) based on audio features. Top 15 audio features for consideration were selected by extracting feature importance using ANOVA-F and MI, respectively. The clustering was evaluated by examining how well it captures the original two genres. The K-means algorithm with ANOVA-F created clusters, where top 5 features were selected. 112 instances were classified as metal and 88 examples were labelled hip-hop. When K-means produced clusters with MI, it classified better than ANOVA-F. Top 6 features (in Fig. 4. MI table) were selected and the two genres were labelled almost evenly: 102 for hip-hop and 98 for metal. Therefore, we decided to select MI with K-means based on these six audio features.

In order to better understand the style transfer, dimensionality reduction is utilized. Principal component analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) were used to visualize 200 GTZAN drum stems

[5] https://drive.google.com/drive/folders/1dpn8NL0fhtXJ4JhzeAeIlDyxUzQWHttC?usp=sharing

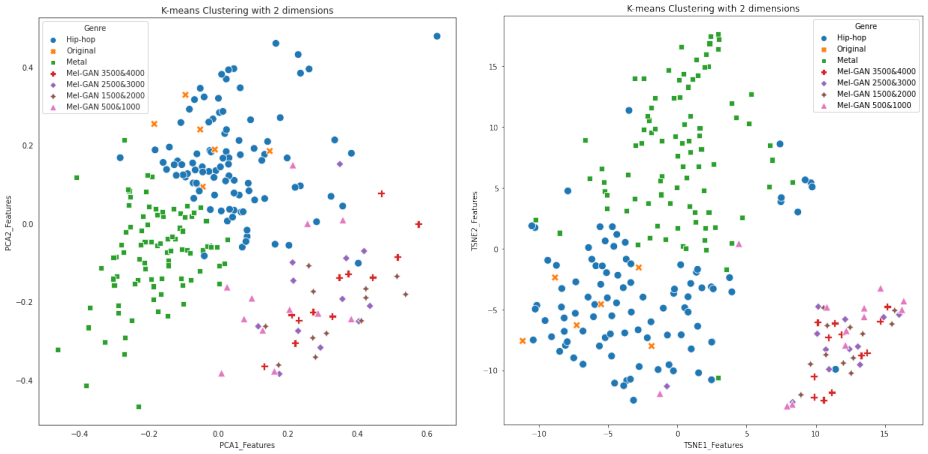

Fig. 6. Unsupervised method clustered by K-means and visualized by PCA (left) and t-SNE (right). Style-transferred outputs by MelGAN-VC is demonstrated depending on the training epoch (original legend denotes epoch 0 outputs). t-SNE works better and the outputs are clustered independent of original hip-hop and metal clusters.

along with 48 style-transferred instances: 6 transferred instances by MelGAN-VC at 8 training epoch points, 500, 1000, 1500, etc. Fig. 6. is demonstrating that three big clusters can be observed based on hip-hop, metal and Mel-GAN instances. t-SNE components worked slightly better because it generated more robust clusters based on the scatter plots in Fig. 6. And it is clearly representing that the Mel-GAN instances are moving away from the original dots. Mel-GAN instances where the epoch range between 2500 and 4000 are located outside of hip-hop and metal clusters. Moreover, the t-SNE plot provides that the instances of that particular epoch range (stronger instances) are clustered more densely compared to weaker instances (Mel-GAN 500 & 1000). Furthermore, the stronger instances (when greater than 2000 epochs) are style-transferred better.

5 Conclusion & Future Work

MelGAN-VC can be used for musical timbre transfer on drum tracks between two genres, particularly achieving style interpolation, which the output sounds similar to an in-between genre of hip-hop and metal. Evaluation methods for the musical timbre transfer are introduced by using supervised and unsupervised learning with visualization. For future works, the current model needs to be expanded to consider temporal information, microtiming, and syncopation. Comparisons of multiple audio GAN architectures using the same methodology and expansion of current examples to additional pairs of genre style transfers will be explored with bigger data sets. Use of a beat synchronous representation will be helpful to consider temporal and rhythmic information.

References

Amodio, M., & Krishnaswamy, S. (2019). Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8983–8992).

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., ... others (2013). Essentia: An audio analysis library for music information retrieval. In *Britto a, gouyon f, dixon s, editors. 14th conference of the international society for music information retrieval (ismir); 2013 nov 4-8; curitiba, brazil.[place unknown]: Ismir; 2013. p. 493-8.*

Brunner, G., Konrad, A., Wang, Y., & Wattenhofer, R. (2018). Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600*.

Brunner, G., Moayeri, M., Richter, O., Wattenhofer, R., & Zhang, C. (2019). Neural symbolic music genre transfer insights. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 437–445).

Brunner, G., Wang, Y., Wattenhofer, R., & Zhao, S. (2018). Symbolic music genre transfer with cyclegan. In *2018 ieee 30th international conference on tools with artificial intelligence (ictai)* (pp. 786–793).

Cífka, O., Ozerov, A., Şimşekli, U., & Richard, G. (2021). Self-supervised vq-vae for one-shot music style transfer. *arXiv preprint arXiv:2102.05749*.

Cífka, O., Şimşekli, U., & Richard, G. (2020). Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2638–2650.

Dai, S., Zhang, Z., & Xia, G. G. (2018). Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*.

Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*.

Ens, J., & Pasquier, P. (2018). Caemsi: A cross-domain analytic evaluation methodology for style imitation. In *Iccc* (pp. 64–71).

Ens, J., & Pasquier, P. (2020). Quantifying musical style: Ranking symbolic music based on similarity to a style. *arXiv preprint arXiv:2003.06226*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, *32*(2), 236–243.

Grinstein, E., Duong, N. Q., Ozerov, A., & Pérez, P. (2018). Audio style transfer. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 586–590).

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*(50), 2154.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., ... Courville, A. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*.

Lu, C.-Y., Xue, M.-X., Chang, C.-C., Lee, C.-R., & Su, L. (2019). Play as you like: Timbre-enhanced multi-modal music style transfer. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 1061–1068).

Mital, P. K. (2017). Time domain neural audio style transfer. *arXiv preprint arXiv:1711.11160*.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*. Retrieved from `http://distill.pub/2016/deconv-checkerboard` doi: 10.23915/distill.00003

Pasini, M. (2019). Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713*.

Tzanetakis, G., Essl, G., & Cook, P. (2001). *Automatic musical genre classification of audio signals.* The International Society for Music Information Retrieval. Retrieved from `http://ismir2001.ismir.net/pdf/tzanetakis.pdf`

Ulyanov, D., & Lebedev, V. (2016). Audio texture synthesis and style transfer. 2016. *URL https://dmitryulyanov. github. io/audio-texture-synthesis-and-style-transfer*.

Verma, P., & Smith, J. O. (2018). Neural style transfer for audio spectograms. *arXiv preprint arXiv:1801.01589*.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).