A CycleGAN for style transfer between drum & bass subgenres

Len Vande Veire [1] **Tijl De Bie** [2] **Joni Dambre** [1]

Abstract

In this work, we apply the CycleGAN image-to-image translation framework to Mel-scaled log-amplitude spectrograms, successfully realizing audio texture transfer between excerpts from two musically related genres. Such automatic musical transfer could provide music producers and DJs with new creative tools, e.g. to quickly prototype a remix of an existing song in another genre, or to use as an advanced effect during a live performance. We show that meaningful style transfer can be realized using only a limited amount of data and computational resources. A high-quality audio reconstruction is obtained from the generated amplitude spectrogram by simply using the phase of the original audio as an approximation for the phase of the generated spectrogram. This results in a significant quality improvement over traditional phase reconstruction methods.

1. Introduction

The advent of advanced generative models such as CycleGAN (Zhu et al., 2017a; Isola et al., 2017) has led to interesting musical applications, e.g. the transfer of timbre between individual instruments (Huang et al., 2019). In this work, we apply the CycleGAN framework to translate segments between two sub-genres of drum & bass music, namely *liquid* and *dancefloor* drum & bass. This effectively creates a rudimentary *remix* of the original segment in the other sub-genre, and hence offers a tool for DJs and music producers to quickly prototype remix ideas, or use this as an effect in live performances.

Our contribution is two-fold. Firstly, we show that one can obtain a sensible transfer of musical texture between music excerpts of two related musical sub-genres with relatively little training data and computational resources. Secondly, when reconstructing the spectral phase for the generated amplitude spectrogram, we avoid audio degradation as observed when using existing algorithms by simply using the phase of the original audio recording. This shows that this is a simple yet effective heuristic for this purpose.

[1]imec, IDLab, Department of Electronics and Information Systems, Ghent University [2]IDLab, Department of Electronics and Information Systems, Ghent University. Correspondence to: Len Vande Veire <len.vandeveire@ugent.be>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Methods

2.1. Dataset description

The spectrograms for training the CycleGAN model are extracted from 40 *liquid* and 40 *dancefloor* drum & bass songs. From each song, 3 segments of 4 downbeats (approximately 5.5 seconds) are selected, resulting in a total of 240 segments for training. The segments are downbeat-aligned and selected from the 'main' part of the song (similar to the chorus for vocal-based music). The tempo of each song is stretched to 175 BPM using WSOLA time stretching (Verhelst & Roelands, 1993) to ensure a consistent length for training. Note that after training, the model can be applied to extracts in any tempo.

2.2. CycleGAN model training

The CycleGAN model (with 9 residual blocks) (Zhu et al., 2017b) operates on Mel-scaled spectrogram images. This representation is obtained by first transforming each audio segment into a time-frequency spectrogram S using a STFT with 2048 frequency bins and a hop size of 512. Then, the squared amplitude of that spectrogram is calculated and converted to a decibel scale ($10 \log_{10}(|S|^2)$). This is then transformed to a Mel-frequency scale (Stevens et al., 1937), to offer a higher resolution in the lower frequency ranges. Finally, the spectrogram values are normalized between 0 and 1, and saved as grayscale images to be processed by the CycleGAN model. The network finishes training within a few hours on a single GeForce GTX 1080 GPU.

2.3. Audio reconstruction from amplitude spectrogram

Reconstructing the audio from a generated spectrogram first involves reverting the preprocessing steps: the initial pixel values are scaled back to the empirically observed average log-amplitude range of -15 dB to 65 dB, and the inverse Mel-frequency and log-amplitude transformations are applied to

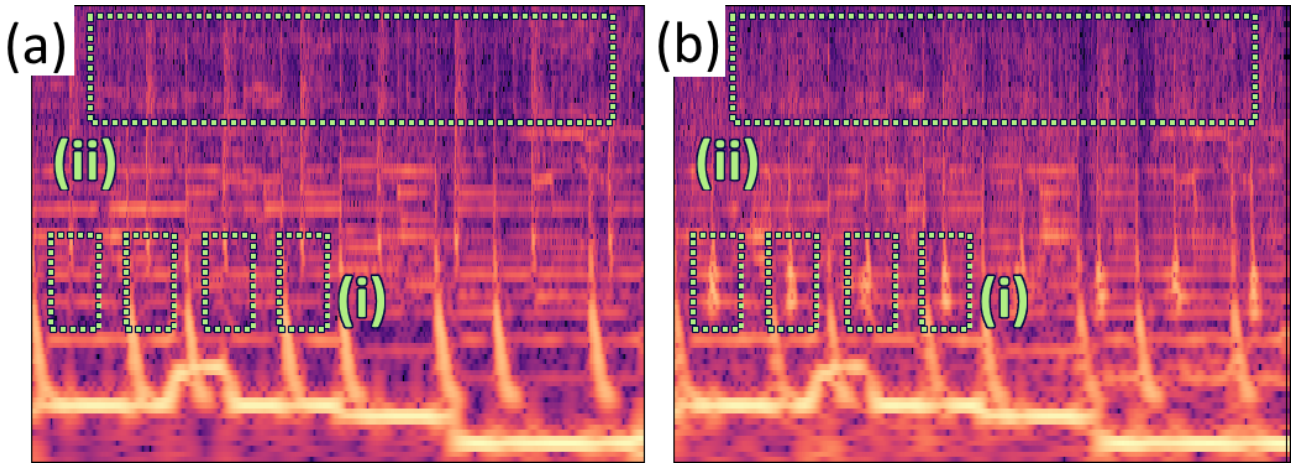

Figure 1. Side-by-side comparison of an input extract and the corresponding output. This figure shows CQT spectrograms instead of Mel-scaled spectrograms as the higher resolution of the CQT transform for lower frequencies illustrates the observed CycleGAN transformations more clearly. (a) input audio segment (genre: *'liquid'*). (b) output audio segment (genre: *'dancefloor'*). Annotated style transfer elements: (i) examples of the introduction of a snare drum; (ii) insertion of noise after hi-hats in high frequency regions.

obtain the STFT amplitude spectrogram. The phase of the spectrogram is approximated by the phase of the original audio segment. Applying the inverse STFT yields the audio reconstruction.

3. Results and discussion

Figure 1 shows the side-by-side comparison of two spectrograms: an original spectrogram, and the result after applying the trained CycleGAN. The corresponding audio fragments and additional examples are available online[1].

When translating a *liquid* segment to *dancefloor* drum & bass, several observations stand out. Firstly, the network inserts snare drums on every second beat of each measure, or changes the sound of already existing snare drums on those beats, giving them a harder and more 'punchy' sound ((i) in Figure 1). Secondly, noise is added in the high frequency regions, changing the timbre of the hi-hats ((ii) in Figure 1). We also observe a slight amplitude reduction of melody components in the mid/high frequency ranges.

In the reverse transformation (*dancefloor* to *liquid*), the reverse operations are observed – snare drums are less pronounced and they sound 'brighter', hi-hats are accentuated by reducing noise around them, and melody components are emphasized differently. We refer to the online examples for spectrograms and audio excerpts.

A major challenge in reconstructing high-quality audio is predicting the phase for the amplitude spectrograms. Our implementation simply uses the phase of the input spectrogram as an approximation. Existing phase reconstruction techniques (Griffin & Lim, 1984; Zhu et al., 2006) were found to introduce a significant low-frequency 'buzzing' sound in the audio. This appears to be an undesired consequence of using the Mel-scale transformation, which smears out the spectral energy between adjacent frequency bands. After applying the phase reconstruction algorithms, this leads to interfering frequency components and hence a degraded audio signal. Directly modeling the spectrogram phase (or rather, the instantaneous frequency) in the CycleGAN as in Engel et al. (2019) did not yet lead to the intended results, and is currently left as a path for future exploration.

4. Conclusion

We presented a method to translate segments between two sub-genres of drum & bass, by applying the CycleGAN framework to Mel-scaled log-amplitude spectrograms. To reconstruct the audio, we approximate the phase of the generated amplitude spectrogram by the phase of the input spectrogram, avoiding artifacts that are introduced by existing phase reconstruction methods. This work could be a first step towards automated remix generation, and exploring more genre combinations and alternative network architectures better tuned to the properties of music spectrograms offer interesting opportunities for further research.

Acknowledgements

Len Vande Veire is supported by a PhD fellowship of the Research Foundation Flanders (FWO). We would like to

[1] https://users.ugent.be/~levdveir/2019ML4MD/

thank Thomas Demeester for the insightful discussions on this topic, and Ira Korshunova for proofreading the paper.

References

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. GANSynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1xQVn09FX.

Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., and Grosse, R. B. Timbretron: A wavenet(cycleGAN(CQT(audio))) pipeline for musical timbre transfer. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1lvm305YQ.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Stevens, S. S., Volkmann, J., and Newman, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3): 185–190, 1937. doi: 10.1121/1.1915893. URL https://doi.org/10.1121/1.1915893.

Verhelst, W. and Roelands, M. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 554–557. IEEE, 1993.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Zhu, J.-Y., Park, T., and Wang, T. CycleGAN and pix2pix in PyTorch, 2017b. URL https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

Zhu, X., Beauregard, G. T., and Wyse, L. Real-time iterative spectrum inversion with look-ahead. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 229–232, July 2006. doi: 10.1109/ICME.2006.262424.