Time-Frequency analysis for neural synthesis of audio

Andrés Marafioti[1], Nicki Holighaus[1], Piotr Majdak[1], and Nathanaël Perraudin[2]
[1] *Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12–14, 1040 Vienna, Austria.*
[2] *Swiss Data Science Center, ETH Zürich, Universitätstrasse 25, 8006 Zürich*

Introduction

Despite the recent advance in machine learning and generative modeling, synthesis of natural sounds by neural networks remains a challenge. Recent solutions rely on, among others, classic recurrent neural networks (e.g., SampleRNN, 1), dilated convolutions (e.g., WaveNet, 2), and generative adversarial networks (e.g., WaveGAN, TiFGAN, MelGAN, 3; 4; 5). Especially, the latter offers a promising approach in terms of flexibility and quality. Generative adversarial networks (GANs, 6) rely on two competing neural networks trained simultaneously in a two-player min-max game: The generator produces new data from samples of a random variable; The discriminator attempts to distinguish between these generated and real data. During the training, the generator's objective is to fool the discriminator, while the discriminator attempts to learn to better classify real and generated (fake) data. Since their introduction, GANs have been improved in various ways (e.g., 7; 8). For images, GANs have been used to great success (9; 10). For audio, GANs enable the generation of signals at once even for duration in the range of seconds (3; 4).

Time-frequency (TF) domain representations of sound are successfully used in many applications and rely on well-understood theoretical foundations. They have been widely applied to neural networks, e.g., for solving discriminative tasks (11), in which they outperform networks directly trained on the waveform (12; 13). Further, TF representations are used to parameterize neural synthesizers, e.g., Tacotron 2 (14) or Timbretron (15). Despite the success of TF representations for sound *analysis*, why, one could ask, has neural *sound generation* via invertible TF representations only seen limited success?

In fact, there *are* neural networks generating invertible TF representations for sound synthesis. They were designed to perform a specific task such as source separation (16; 17), speech enhancement (18), or audio inpainting (19; 20) and use a specific and well-chosen setup for TF processing. While the general rules for the parameter choice are not the main focus of those contributions, these rules are highly relevant when it comes to synthesizing sound from a set of TF coefficients generated, e.g., by a neural network.

When both the TF representation and its parameters are appropriately chosen, we generate a highly structured, invertible representation of sound, from which time-domain audio can be obtained using efficient, content-independent reconstruction algorithms. In that case, we do not need to train a problem-specific neural synthesizer. Hence, in this article, we discuss important aspects of neural generation of TF representations particularly for sound synthesis. We focus on the short-time Fourier transform (STFT, e.g., 21; 22), the best understood and most widely used TF representation in the field of audio processing. We demonstrate the applicability of neural generation of STFT by presenting TiFGAN, a network which generates audio using a TF representation. We provide perceptual and numerical evaluations of TiFGAN demonstrating improved audio quality compared to a similar time-domain GAN for audio synthesis. Our software, complemented by instructive examples, is available at `http://tifgan.github.io`.

Time-Frequency Generative Adversarial Network (TiFGAN)

Here we present TiFGAN, which unconditionally generates audio using a TF representation. For the purpose of this contribution, we restrict to generating 1 second of audio, or more precisely $L = 16384$ samples sampled at 16 kHz. For the short-time Fourier transform, we select the window size $M = 512$ and the hop size $a = 128$, giving the minimal redundancy that we consider reliable, i.e., $M/a = 4$. For the analysis window g we chose a (sampled) Gaussian with time-frequency ratio $\lambda = 4 = aM/L$. Since the Nyquist frequency is not expected to hold significant information for the considered signals, we drop it to arrive at a representation size of 256×128, which is well suited to processing using strided convolutions.

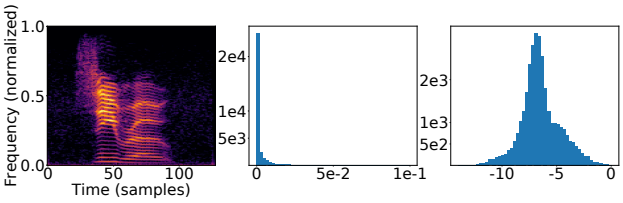

Abbildung 1: From left to right: log-magnitude spectrogram, distribution of the magnitude, distribution of the log-magnitude.

We generate log-magnitude STFT coefficients since its distribution is closer to human sound perception and, as show in Fig. 1, it doesn't have the large tail of the regular magnitude STFT coefficients. To do so, we first normalize the STFT magnitude to have maximum value 1, such that the log-magnitude is confined in $(-\infty, 0]$. Then, the dynamic range of the log-magnitude is limited by clipping at $-r$ (in our experiments $r = 10$), before scaling and shifting to the range of the generator output $[-1, 1]$, i.e. dividing by $r/2$ before adding constant 1.

To recover the time audio signal, first the phase derivatives are estimated from the generated log-magnitude.

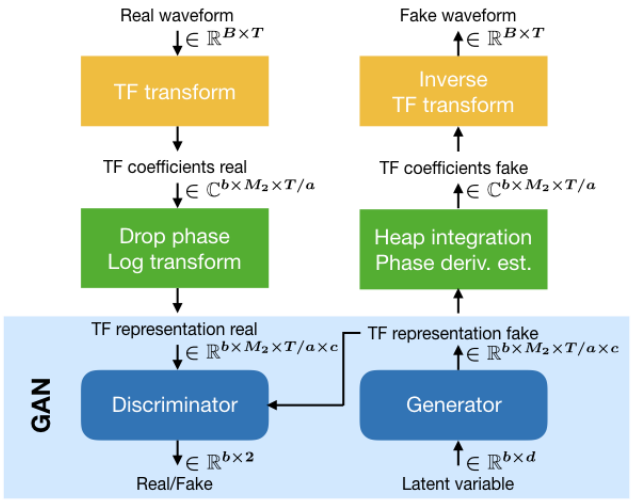

Abbildung 2: The general architecture with parameters $T = 16384$, $a = 128$, $M_2 = 256$ $c = 1, 3$, $d = 100$. Here $b = 64$ is the batch size. The orange and green steps describe the pre- and post-processing stages.

The phase is then reconstructed from the phase derivative estimates using phase-gradient heap integration (PGHI, 23), which requires no iteration, such that reconstruction time is comparable to simply integrating the phase derivatives. For synthesis from the STFT, we use the *canonical dual window* (24; 25), precomputed using the Large Time-Frequency Analysis Toolbox (LTFAT, 26), available at `ltfat.github.io`.

GAN architecture: The TiFGAN architecture, depicted in Fig. 2, is an adaptation of DCGAN (27) and similarly to SpecGAN and WaveGAN (3), we add one convolutional layer each to generator and discriminator to enable the generation of larger matrices. Moreover, we generate data of size $(256, 128)$, a rectangular array of twice the width and four times the height of DCGANs output, and twice the height of SpecGAN, such that we also adapted the filter shapes to better reflect and capture the rectangular shape of the training data. Precisely in comparison to SpecGAN, we use filters of shape $(12, 3)$ instead of the 31% smaller $(5, 5)$. To compensate, we further reduce the number of filter channels of the fully-connected layer and the first convolutional layer of the generator by a factor of 2. Since these two layers comprise the majority of parameters, our architecture only has 10% more parameters than SpecGAN's in total.

Training: During training of TiFGAN, we monitored the relative consistency γ (4) of the generated log-spectrograms in addition to the adversarial loss, negative critic and gradient penalty. In the optimization phase, networks that failed to train well could often be detected to diverge in consistency and discarded after less than $50k$ steps of training (1 day), while promising candidates quickly started to converge towards the consistency of the training data, i.e., $\gamma \to 0$, see Fig. 3. Networks with smaller γ synthesized better audio, but when trained for many steps, they were sometimes less reliable in terms of semantic audio content, e.g., for speech they were more likely to produce gibberish words than with shorter training.

Our networks were trained for 200k steps as this seemed to provide reasonably good results in both semantic and audio quality. We optimized the Wasserstein loss (8) with the gradient penalty hyperparameter set to 10 using the ADAM optimizer (28) with $\alpha = 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.9$ and performed 5 updates of the discriminator for every update of the generator. For the reference condition, we used the pre-trained WaveGAN network provided by (3)[1].

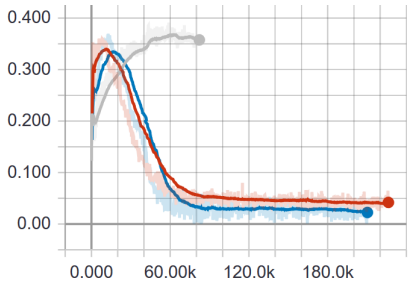

Abbildung 3: Relative consistency for three networks. Gray: failed network. Red and blue: TiFGAN.

Evaluation

To evaluate the performance of TiFGAN, we trained it using the procedure outlined above on two datasets from (3): (a) Speech, a subset of spoken digits 'zero' through 'nine' (sc09) from the Speech Commands Dataset (29). This dataset is not curated, some samples are noisy or poorly labeled, the considered subset consists of approximately 23,000 samples. (b) Music, a dataset of 25 minutes of piano recordings of Bach compositions, segmented into approximately 19,000 overlapping samples of 1 s duration.

Evaluation metrics: For speech and music, we provide audio examples online[2]. For speech, we performed listening tests and evaluated the inception score (IS) (30) and Fréchet inception distance (FID) (31), using the pre-trained classifier provided with (3). For the real data and TiFGAN, we additionally computed the consistency ϱ and the relative spectral projection error (RSPE) in dB, after phase reconstruction from the log-magnitude, i.e.,

$$10\log_{10}\left(\frac{\||\widetilde{\mathrm{S}}| - |\,\mathrm{S}_g(\mathrm{iS}_{\tilde{g}}(\widetilde{\mathrm{S}}))|\|}{\||\widetilde{\mathrm{S}}|\|}\right), \qquad (1)$$

where $|\widetilde{\mathrm{S}}| = |\,\mathrm{S}_g(s)|$ in the case of real data and $|\widetilde{\mathrm{S}}| = \exp(\widetilde{\mathrm{M}})$, with the generated log-magnitude $\widetilde{\mathrm{M}}$, for the generated data. Phase-gradient heap integration was applied to obtain $\widetilde{\mathrm{S}}$ from $|\widetilde{\mathrm{S}}|$.

Listening tests were performed in a sound booth and sounds were presented via headphones. The task involved pairwise comparison of preference between three conditions: real data extracted from the dataset, TiFGAN generated examples, and WaveGAN generated examples. In each trial, listeners were provided with two sounds from two different conditions. The question to the listener was "which sound do you prefer?". Signals were selected at

[1] `https://github.com/chrisdonahue/wavegan`
[2] `http://tifgan.github.io`

	vs TiFGAN	vs WaveGAN	Cons	RSPE (dB)	IS	FID
Real	86%	94%	0.70	-22.0*	7.98	0.5
TiFGAN	–	75%	0.67	-13.8	5.97	26.7
WaveGAN	25%	–	–	–	4.64	41.6

Tabelle 1: Evaluation results. First three left columns: Preference (in %) of the condition shown in a row over the conditions show in a column, obtained from listening tests. Cons: averaged consistency measure ρ. RSPE: as in Eq. (1). IS: inception score. FID: Fréchet inception distance. *These values were obtained by discarding the phase and reconstructing from the magnitude only. For the listening tests, the signals contained the full representation.

random from 600 pre-generated examples per condition. Each of the six possible combinations was repeated 80 times in random order, yielding 480 trials per listener. The test lasted approximately 45 minutes including breaks which subjects were allowed to take at any time. Seven subjects were tested and none of them were the authors. A post-screening showed that one subject was not able to distinguish between any of the conditions and thus was removed from the test, yielding in 2880 valid preferences in total from six subjects.

Results: The results are summarized in Table 1. On average, the subjects preferred the real samples over WaveGAN's in 94% of the examples given. The preference over TiFGAN decreased 86%. The large gap between generated and real data can be explained by the experimental setup that enables a very critical evaluation. Nonetheless, it is apparent that TiFGAN performed best in the direct comparison to real data by a significant margin. Additionally, subjects preferred TiFGAN over WaveGAN in 75% of the examples given.

The analysis of IS and FID leads to similar conclusions: TiFGAN showed a large improvement on both measures over WaveGAN, with still a large gap to the real-data performance. In summary, TiFGAN provided a substantial improvement over the similarly sized time-domain WaveGAN in unsupervised adversarial audio generation.

Conclusions

In this contribution, we considered adversarial generation of a well understood time-frequency representation, namely the STFT. We proposed machine learning motivated signal processing steps to overcome the difficulties that arise when generating audio in the short-time Fourier domain, taking advantage from the recent progress in phaseless reconstruction (23).

We presented TiFGAN, a GAN directly generating invertible STFT representations. TiFGAN outperformed a similar time-domain GAN both in terms of psychoacoustic and numeric evaluation, demonstrating the potential of TF representations in generative modeling.

In the future, further extensions of the proposed approach are planned towards TF representations on logarithmic and perceptual frequency scales (32; 33; 34; 35; 36).

Literatur

[1] 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," in *Proc. of ICLR*, 2017.

[2] A. van den 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," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[3] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[4] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "Adversarial generation of time-frequency features with application in audio synthesis," in *Proc. of the 36th ICML*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 4352–4362.

[5] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 14 910–14 921.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[9] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[10] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[11] R. H. Abbasi, P. Balázs, A. Noll, D. Nicolakis, M. Adelaide, Marconi, S. M. Zala, and D. J. Penn, "Applying convolutional neural networks to the analysis of mouse ultrasonic vocalizations," in *Proc. of the 23rd international congress on Acoustics*, 2019.

[12] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.

[13] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. of ISMIR*, 2018.

[14] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. of ICASSP*. IEEE, 2018.

[15] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, and R. B. Grosse, "Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer," *arXiv preprint arXiv:1811.09620*, 2018.

[16] Z.-C. Fan, Y.-L. Lai, and J.-S. R. Jang, "Svsgan: Singing voice separation via generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 726–730.

[17] J. Muth, S. Uhlich, N. Perraudin, T. Kemp, F. Cardinaux, and Y. Mitsufuji, "Improving dnn-based music source separation using phase features," *arXiv preprint arXiv:1807.02710*, 2018.

[18] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[19] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[20] A. Marafioti, N. Holighaus, P. Majdak, and N. Perraudin, "Audio inpainting of music by means of neural networks," in *Audio Engineering Society Convention 146*, Mar 2019.

[21] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[22] J. Wexler and S. Raz, "Discrete gabor expansions," *Signal Processing*, vol. 21, no. 3, pp. 207 – 220, 1990.

[23] Z. Průša, P. Balazs, and P. Søndergaard, "A noniterative method for reconstruction of phase from stft magnitude," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 5, pp. 1154–1164, 2017.

[24] T. Strohmer, "Numerical algorithms for discrete Gabor expansions," in *Gabor Analysis and Algorithms: Theory and Applications*, ser. Appl. Numer. Harmon. Anal., H. G. Feichtinger and T. Strohmer, Eds. Birkhäuser Boston, 1998, pp. 267–294.

[25] O. Christensen, *An Introduction to Frames and Riesz Bases*, Second ed., ser. Applied and Numerical Harmonic Analysis. Birkhäuser Basel, 2016.

[26] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, ser. LNCS. Springer International Publishing, 2014, pp. 419–442.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[28] D. Kingma and M. Welling, "Auto-encoding variational bayes." in *Proc. of ICLR*, 2014.

[29] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[32] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[33] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant q transform," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.

[34] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-q transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 775–785, 2013.

[35] N. Holighaus, G. Koliander, Z. Průša, and L. D. Abreu, "Characterization of analytic wavelet transforms and a new phaseless reconstruction algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 3894–3908, 2019.

[36] T. Necciari, N. Holighaus, P. Balazs, Z. Průša, P. Majdak, and O. Derrien, "Audlet filter banks: A versatile analysis/synthesis framework using auditory frequency scales," *Applied Sciences*, vol. 8, no. 1:96, 2018.