Generating Diverse Realistic Laughter for Interactive Art

M. Mehdi Afsar*
Mila & University of Calgary

Eric Park*
Mila & University of Waterloo

Étienne Paquette
Independent Artist

Gauthier Gidel†
Mila & University of Montreal

Kory W. Mathewson
DeepMind

Eilif Muller†
Mila & University of Montreal

Abstract

We propose an interactive art project to make those rendered invisible by the COVID-19 crisis and its concomitant solitude reappear through the welcome melody of laughter, and connections created and explored through advanced laughter synthesis approaches. However, the unconditional generation of the diversity of human emotional responses in high-quality auditory synthesis remains an open problem, with important implications for the application of these approaches in artistic settings. We developed LaughGANter, an approach to reproduce the diversity of human laughter using generative adversarial networks (GANs). When trained on a dataset of diverse laughter samples, LaughGANter generates diverse, high quality laughter samples, and learns a latent space suitable for emotional analysis and novel artistic applications such as latent mixing/interpolation and emotional transfer.[3]

1 Introduction

Modern society has refined the condition of solitude to the point where countless seniors, marginalized because of their age, have magically disappeared: left to their own devices, these individuals fade from social life and essentially live in a parallel world. The COVID-19 crisis and resulting lockdowns have both entrenched this phenomenon and helped to reveal how widespread it really is. Can artificial intelligence (AI) help to reconnect generations by making them part of a transgenerational art experience? At the crossroads of laughter—an act of communication between two individuals—and artificial intelligence—a purely functional entity—can we rediscover our humanity?

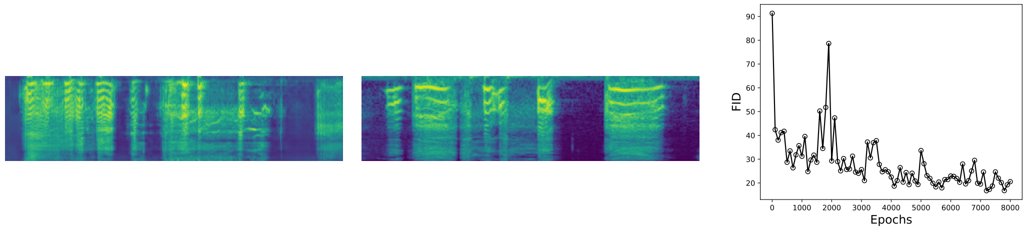
An interactive experience. The end-goal of this project is to connect people via an interactive web experience driven by synthetic laughter. Using our models, we will explore the phenomenon of empathy triggered by the sound of laughter, the relationship between individual memory and laughter, and how the sound of laughter evolves over a lifetime.

Laughter generation for advancing audio synthesis research. With stunning advancements in image synthesis [1, 2, 3, 4], Generative Adversarial Networks (GANs) [5] have gained the attention of researchers in the field of audio synthesis [6, 7, 8, 9]. Synthesizing audio opens new doors for musicians and artists and enables them to expand their repertoire of expression [6]. Despite significant progress by the ML community on methods for audio synthesis, there have been only a few attempts in the topic of laughter synthesis [10], and none leveraging modern approaches such as GANs.

Compared to speech, laughter is made challenging by its many context-dependent attributes, such as emotions [11], age, and gender. Moreover, compared to well-studied topics like speech synthesis,

*Equal contribution, corresponding author: mehdi.afsar@ucalgary.ca
†Canada CIFAR AI Chair
[3]Samples: https://bit.ly/2XQqcW0

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.



(a) Generated Mel spectrogram (b) Real Mel spectrogram (c) FID score for LaughGANter

Figure 1: (a)-(b) Log-magnitude spectrograms for real and generated samples show similar features on qualitative analysis and (c) FID score for LaughGANter decreases during training, as the diversity of the generated samples approaches that of the training data distribution.

there are not established evaluation methods for synthesized laughter. Thus laughter synthesis, has the potential to become a standard benchmark in unconditional audio synthesis.

Related work. Previous work in the field of laughter generation involves [12] the use of oscillatory system [13], formant synthesis [14], articulatory speech synthesis [15], and hidden Markov models (HMM) [16]. Recently, some researchers have also used deep learning [12, 17] methods for laughter synthesis. GANs are advantageous in learning of a compact latent space allowing for interpolation, mixing, and style transfer as well as emotional analysis. In this paper, we propose to use GANs for the purpose of unconditional laughter generation and manipulation (LaughGANter). Our aim is to enable a unique interactive art experience that surprises and connects through the primordial intimacy of our laughter interacting and juxtaposed with others.

2 Methodology

We adapt Multi-Scale Gradient GAN (MSG-GAN) [4] for laughter synthesis. Among other popular image synthesis methods, like DCGAN [18], ProgressiveGAN [1], and StyleGAN [2], LaughGANter employs multi-scale gradients on a DCGAN architecture to address the training instability prevalent in GANs. Progressive growing of network resolutions is avoided to limit the hyperparameters to be tuned (e.g. training schedule, learning rates for each resolution, etc) while the multi-scale discriminator penalizes intermediate and final layer outputs of the generator.

We refer the reader to [4] for an in-depth study of the MSG-GAN architecture. Concisely, the generator (G) samples a random vector z from a normal distribution and outputs $x = G(z)$. The generated samples are fed into the discriminator (D), along with real samples, in order to measure the divergence. We perform *pixel normalization* after every layer in G, and employ the *Relativistic Average Hinge* loss [19] in D. Moreover, inspired by [20, 21], we explored the impact of *induced receptive field expansion*, adding residual blocks with dilations after each upsampling layer in G, which exponentially increases the model's receptive field and can lead to better long range correlation in audio data.

Categorical Conditional Generation. A more directed data generation process is employed through a conditional adaptation of MSG-GAN [22], facilitating the laughter representation learning given additional context beyond unlabeled laughter (e.g. gender, age, humor style, etc). Here, categorical information augments the latent noise vector in G, and to each of the multi-scale vectors within D, through a concatenation with an embedding of context information.

3 Experiments

Setup. Our model is implemented in PyTorch. We use a laughter dataset containing 2145 laughter samples collected by the National Film Board of Canada. Samples are 1-8s long (22.05kHz mono), and were collected (and labeled) from subjects with different ages and genders (55% male, 45% female; 93% adult, 6% child, 1% teen). The audio data is augmented using a random combination of additive noise, shifting, and changing pitch and duration (using `pyrubberband`). Then, this data is converted to Mel spectograms and fed into the model. In addition to qualitative evaluation, i.e., listening to generated samples, we have used Fréchet inception distance (FID) [23] to assess the diversity of the generated samples compared to the training dataset. Instead of using Inception features used in the original FID score, we use features from a classifier (gender and age group) trained on the spectrograms of our laughter dataset.

References

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[4] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7799–7808, 2020.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[6] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[7] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

[8] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019.

[9] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*, 2019.

[10] Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J Griffin, Florian Lingenfelser, et al. Laugh when you're winning. In *International Summer Workshop on Multimodal Interfaces*, pages 50–79. Springer, 2013.

[11] Marc Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*. Citeseer, 2001.

[12] Hiroki Mori, Tomohiro Nagata, and Yoshiko Arimoto. Conversational and social laughter synthesis with wavenet. In *INTERSPEECH*, pages 520–523, 2019.

[13] Shiva Sundaram and Shrikanth Narayanan. Automatic acoustic synthesis of human-like laughter. *The Journal of the Acoustical Society of America*, 121(1):527–535, 2007.

[14] Jieun Oh and Ge Wang. Lolol: Laugh out loud on laptop. In *NIME*, pages 190–195, 2013.

[15] Eva Lasarcyk and Jürgen Trouvain. Imitating conversational laughter with an articulatory speech synthesizer. 2007.

[16] Jérôme Urbain, Hüseyin Çakmak, Aurélie Charlier, Maxime Denti, Thierry Dutoit, and Stéphane Dupont. Arousal-driven synthesis of laughter. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):273–284, 2014.

[17] Noé Tits, Kevin El Haddad, and Thierry Dutoit. Laughter synthesis: Combining seq2seq modeling with transfer learning. *arXiv preprint arXiv:2008.09483*, 2020.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[21] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Supplementary Materials

Ethical Implications

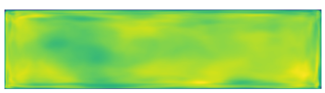
LaughGANter's facilitation in the rediscovery of our humanity and re-connectivity of generations through interactive art experiences allows for production of a diversity of human emotional responses. Such a system's capability in emotional transfer can be used to strengthen the interrelationship within human-and-machine interaction, however, is in tandem capable of coaxing accurate and precise emotional responses which could be used for downstream human manipulation tasks. This project is performed specifically for artistic purposes, so ethical considerations applied to all artistic projects also apply to our work.

Acknowledgements

We would like to thank Arnaud Roussel for his contributions to dataset processing and early MSG-GAN prototypes, as well as Isabelle Repelin, Isabelle Limoges, Martin Viau, Stephanie Quevillon, and Marie-Eve Babineau at the National Film Board of Canada for financial and project management support for this research.

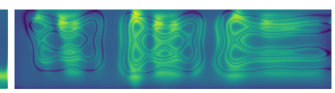
Experiments Results

Fig. 2 depicts how the performance of LaughGANter is improved in the course of training. In particular, Figs. 2(k) and (l) show that LaughGANter can generate laughter samples that are very close to real samples. Moreover, Fig. 3 depicts a linear interpolation between the generated laughter of a female and a male.

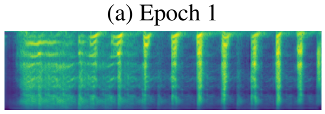


(a) Epoch 1

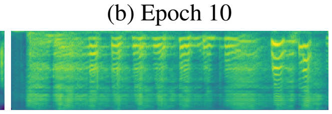
(b) Epoch 10



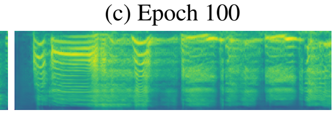
(c) Epoch 100



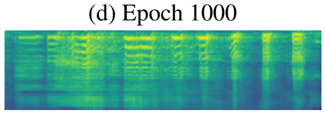
(d) Epoch 1000



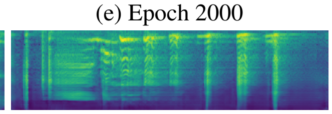
(e) Epoch 2000



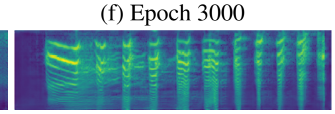
(f) Epoch 3000



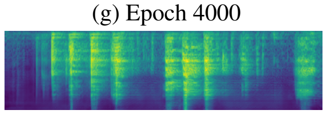
(g) Epoch 4000



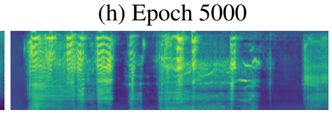
(h) Epoch 5000



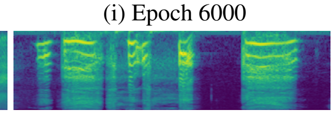
(i) Epoch 6000



(j) Epoch 7000



(k) Epoch 8000



(l) A real sample

Figure 2: (a)-(k) Log-magnitude spectrograms of generated laughter samples during training and (l) a real sample

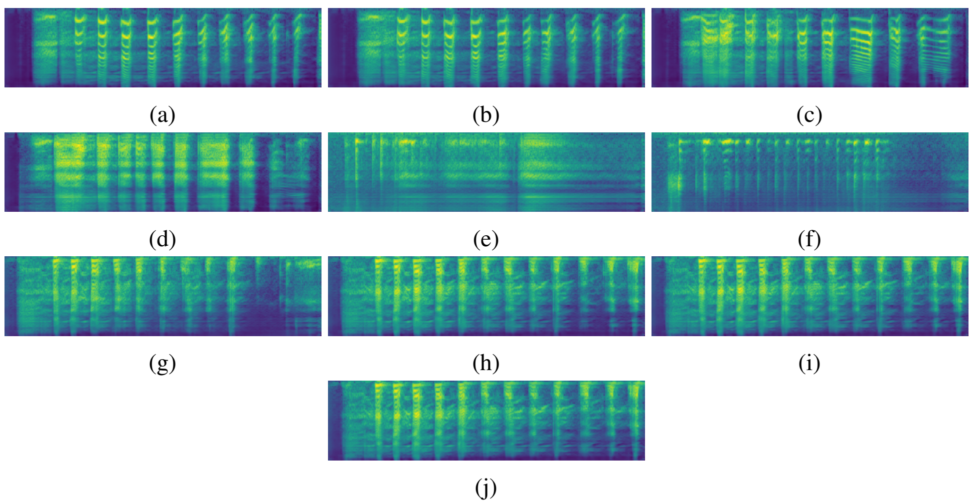


Figure 3: Interpolation between generated laughter respectively identified as female (a) and male (j)