An Ambient Audio Dataset for Sample-Level Audio Generation

Sean Hastings
sean_hastings@brown.edu

Abstract

Predicting and generating sample-level audio is hard, and so for comparison's sake most neural audio generation papers use the same or similar datasets. These datasets allow for cross-model performance comparisons, but all focus on very structured sound such as music or speech, where the high level temporal relationships (i.e. what phonemes tend to come after each other in speech) are relatively strict. I try to address this issue by compiling a dataset of ambient audio, which is effectively unstructured by the fact that the set of audible sources shifts unpredictably over time. I implement the WaveNet (Oord et al., 2016) neural architecture as a benchmark to test my dataset and compare results to existing datasets. The specific form of ambient audio I collected is recordings of busy rooms. People enter and leave, discussions begin and end, and occasional sounds of chairs moving or cutlery clinking can be heard.

1 Introduction

Audio generation is a growing problem space as virtual assistants like Apple's Siri and Microsoft's Cortana become increasingly common and people expect them to sound more human. At the same time, sample-level music generation is getting better as well. These primary use cases focus on structured audio, and as such the data used to benchmark the relevant models are composed of similarly structured audio. Unstructured audio is also important, however, as it is an inherently more difficult problem and more closely aligns with the reality of audio understanding, which is a related, if separate, problem. Predicting unstructured audio requires either an understanding of the latent structure of the scene where the audio was recorded or a probabilistic model of what sounds tend to follow what sounds at a high level (akin to, but notably more complex than, the same model over more structured audio data). Thus, it being a harder problem than predicting structured audio but still feasible by the same approaches, I argue that predicting unstructured audio would be a worthwhile benchmark to have alongside the more structured and directly applicable speech and music datasets.

There are many audio datasets currently in use by researchers for benchmarking and comparing models, but they tend to be either speech (King and Karaiskos, 2012; Kominek et al., 2003), short sounds[1], or music (Defferrard et al., 2017; Bertin-Mahieux et al., 2011). As mentioned above, these are all structured audio (or audio clips so short that structured vs unstructured in the sense used in this paper becomes a pointless discussion). While the temporal relationships within these datasets can be complex, they are far less so than those of unstructured audio such as ambient sound. I propose to address this issue by collecting a dataset of unstructured audio of a form commonly encountered, busy spaces. People navigating the space, communicating with each other, and doing audible actions such as closing doors and using cutlery make the problem of audio prediction much more difficult even at a high level because they happen in no particular order or pattern. In music, the tempo dictates when you can expect sounds to occur. With unconditional ambient sound, predicting when a door will shut, when a voice will be heard, etc is at a high level nearly if not fully infeasible. Under these conditions, questions such as how will audio generation models hold up? Will they fail to reproduce these unpredictable sounds, or will they produce audio that captures the distribution of these sounds from the dataset? can potentially become useful in comparing state of the art models. The answers to these questions likely vary

[1] Such as the Onomatopoeia set available by application from Ubisoft.

between types of and individual models, as the receptive field (how far into the past informs the next sample predicted) of a model determines how much high-level context it can use in predictions, and popular models range from as low as a quarter of a second (Oord et al., 2016) to effectively infinite (Mehri et al., 2016).

In order to evaluate audio synthesis datasets I implemented WaveNet, a convolutional model introduced by Oord et al. in 2016. I chose this model because it is relatively simple, both in theory and practice, as well as a commonly-used benchmark for other audio-synthesis models. The original publication omitted hyperparameters, so I used those reported by (Mehri et al., 2016) in their section on their reimplementation of WaveNet.

2 Related Work

There are tens if not hundreds of music and speech datasets, but there are also short-clip datasets and ambient datasets. Music and speech datasets have been discussed at length above, and short-clip datasets dont have the requisite length to fully express the unpredictability of unstructured audio. Ambient audio datasets do already exist (Gemmeke et al., 2017; Salamon et al., 2014), but each have their own drawbacks as unstructured audio benchmarks.

AudioSet (Gemmeke et al., 2017) features a library of two million 10-second audio clips pulled from YouTube. The biggest drawback of this is that they are extremely varied, with many of the examples having absolutely nothing in common, because currently common benchmarks such as Blizzard (King and Karaiskos, 2012) and Music[2] limit their variety by having one speaker and one instrument and composer, respectively. Thus, in theory a model expressive enough to capture the low-level audio of Blizzard and Music could fail to capture the low-level audio of AudioSet.

Urban Sound Dataset is a dataset of outdoor ambient audio featuring sounds such as car horns, dog barks, childrens voices, and idling engines. Similarly, the actual sounds within this dataset are far more varied than in the common benchmarks, but because the variation is limited to within a theme (urban ambient) it is more feasible than AudioSet as an unstructured benchmark dataset. The audio clips, however, are all less than 4 seconds. High-level temporal relationships can manifest in 4 seconds, but many of the clips are far shorter than that, limiting their relevance to my goal.

[2]Beethoven's 32 piano sonatas publicly available on https://archive.org/

3 Technical Approach

My goal was to build a dataset analogous to the Blizzard and Music datasets commonly used for benchmarking sample-level audio prediction and synthesis models. To achieve this, my approach was to focus on making the audio consistent and similar throughout the dataset, while maintaining just enough variation to make prediction non-trivial. For this reason I recorded ambient audio in busy indoor spaces, namely classrooms and a dining hall.

The classroom audio is primarily voices of students (discussing their work) and the dining hall audio is primarily voices of students (socializing) intermixed with the sounds of cutlery on plates in the background. The dual-context audio was used to add slightly more variation as the main difference between the two contexts was the silverware sound. The resulting dataset has two notable features: background voices of at least several people all throughout, and occasional cutlery sounds through half. Recording was done by placing a non-directional microphone on an unoccupied table in the room.

To evaluate this dataset and provide a benchmark to compare it to others, I implemented WaveNet. WaveNet is a 1d-convolutional model based on dilated convolution to exponentially increase the receptive field with depth. For a detailed explanation of the model read the paper, but the parts omitted from the paper, namely hyperparameters, will be provided here. I used 4 blocks of 10 layers, for a total of 40 layers. Each dilation filter has size 2, and outputs 128 channels (with 128 channels in the residual path as well). Skip connections output 1024 channels. Training was done on batches of 4 sequences with target sequence lengths of 1600. I used categorical cross entropy loss with Adam (Kingma and Ba, 2014) optimizer with a fixed learning rate of .0001, training for slightly over a week.

4 Evaluation

My goal, to build a dataset analogous to the Blizzard and Music datasets, is difficult to define quantitatively. My expectations going in were that unstructured audio, due to being less predictable at a

high level, would have a noticeably higher test loss than my structured baseline (Music) but likely not more than double. In preliminary tests with short training time and a smaller model this was observed as the model achieved a test loss of 2.46 on the Music dataset and 3.58 on my ambient dataset. This fits my expectations, but does not sufficiently show that I have achieved my goal because there are many reasons the loss could be what it is.

In order to properly evaluate results I would have to perform a qualitative assessment of samples generated from a model trained sufficiently on the ambient dataset, but as of writing this I do not have such a model, although one is currently training.

5 Conclusion

There is an unfilled niche for audio datasets to be used as unstructured benchmarks for sample-level audio prediction and generation. I propose to fill this niche with a dataset of recorded ambient sound. In order to evaluate this dataset and set a baseline, I implement the WaveNet neural architecture and train models on both my ambient dataset and the popular Music dataset. Preliminary results indicate that the ambient model may be able to learn the data well enough for meaningful qualitative evaluation, but training is ongoing so qualitative results are not yet available.

Sample-level audio synthesis is still a very open problem, especially conditional synthesis. Conditioning variables from text describing what is being spoken to video of the source of the sound can be used. One potentially interesting extension would be to train a model to read lips. Train on speech, conditioned on a cropped frontal view of the speakers mouth as they speak the text. To make it even more interesting, the model could be conditioned on speaker identity (arbitrary speaker-related variables or those discussed in the original conditional WaveNet paper). This could allow the identity of the speaker to be separated from the visual of their mouth, potentially encouraging the model to pull more specifically the phonemes being spoken than the speakers voice out of the video.

On a more architectural note, it would be interesting to explore the possibilities of combining sequential and non-sequential sample-level audio generation. GANs, for example, are currently being used to generate both images and audio from noise, fully parallel. To be able to harness the parallelization and therefore speed of this approach to an overall sequential model would allow much faster evaluation than fully sequential models. Something akin to a modified GAN, judging not whether an audio clip is real or generated but rather whether one audio clip follows another may allow sequential generation in chunks.

In the long term, the possibilities are nearly endless. The use cases for audio generation multiply as the technologies success does, in an ever-increasing cycle. It may not be feasible now, but what if decades down the line we could build a model capable of separating out the parts of a sound, pulling conversations, background sounds, and noise separately out of one audio sample. It is most definitely a problem worth pursuing.

References

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Simon King and Vasilis Karaiskos. 2012. The blizzard challenge 2012.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Kominek, Alan W Black, and Ver Ver. 2003. Cmu arctic databases for speech synthesis. Technical report.

Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*.

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

J. Salamon, C. Jacoby, and J. P. Bello. 2014. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.