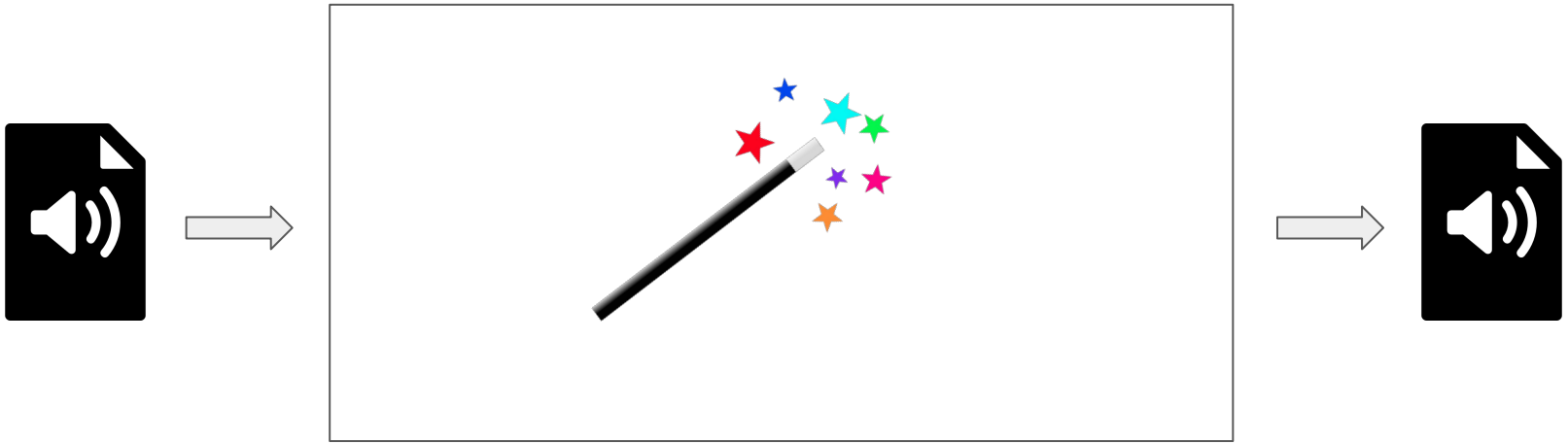
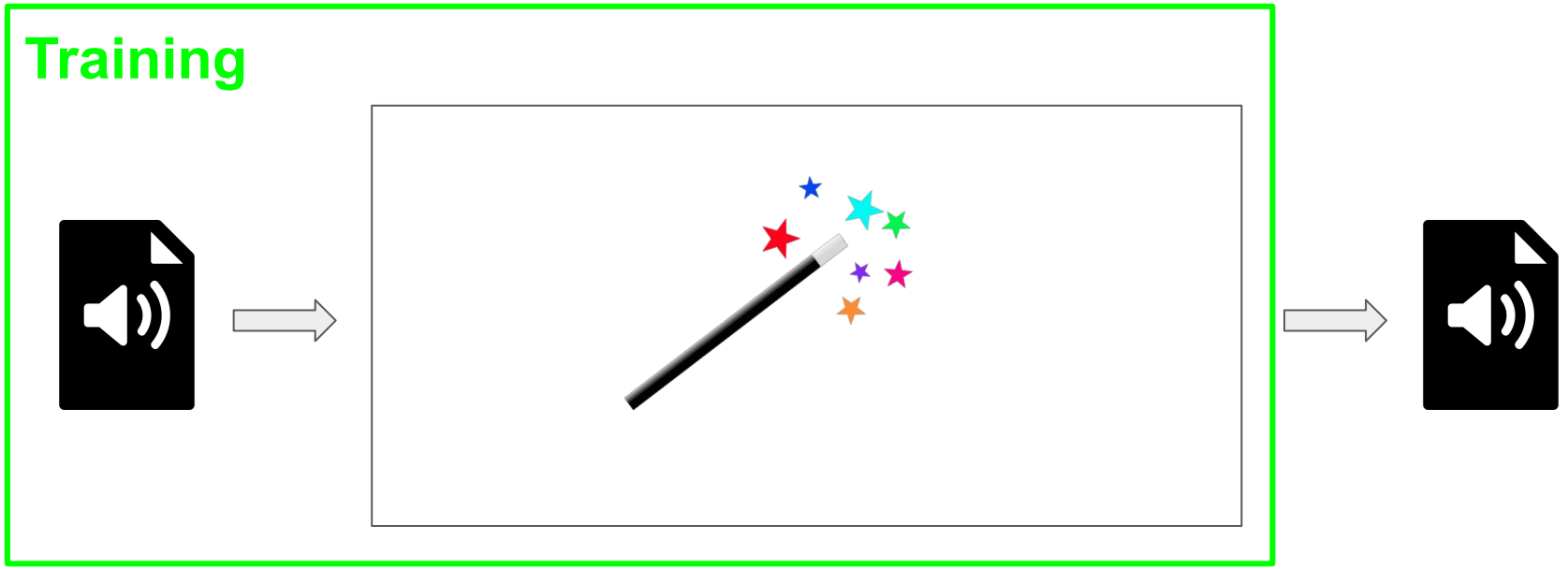
Sound Generation with Deep Learning: Approaches and Challenges

Valerio Velardo

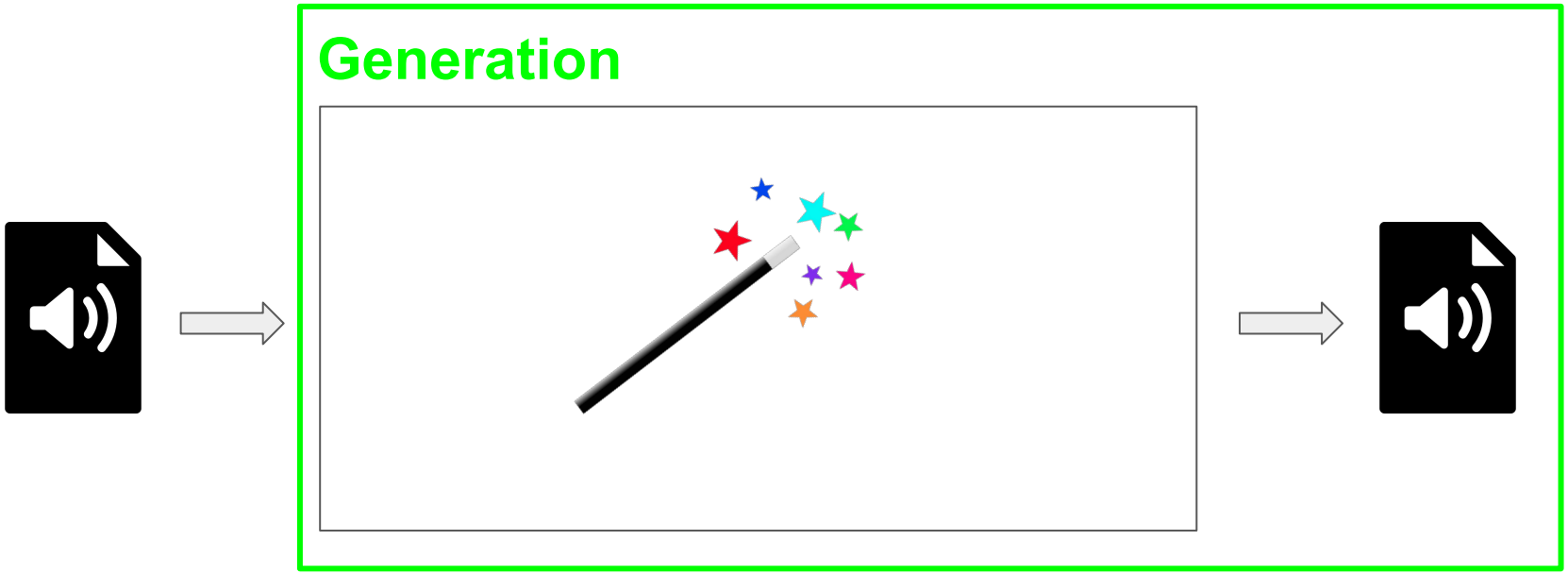
Sound generation task



Sound generation task



Sound generation task



Classification of sound generation systems

- What types of sounds are generated?
- What are the features used to train the system?
- What is the DL architecture employed?
- What are the inputs for generation?

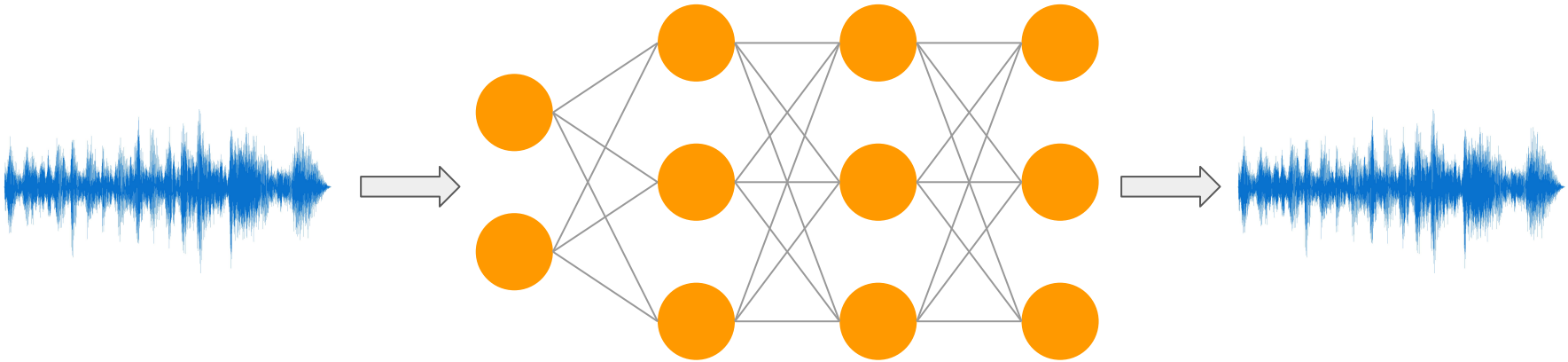
Types of generated sounds

- Speech (Text-to-Speech)
- Music
- Music notes (samples)
- Sound design
- ...

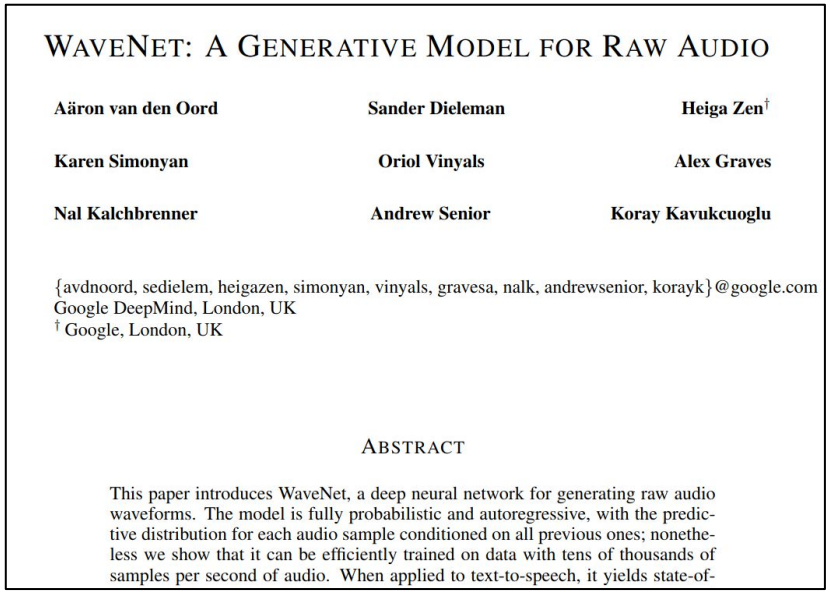
Sound representations

- Raw-audio
- Spectrograms

Generation from raw audio



Generation from raw audio



WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Aäron van den Oord **Sander Dieleman** **Heiga Zen**[†]

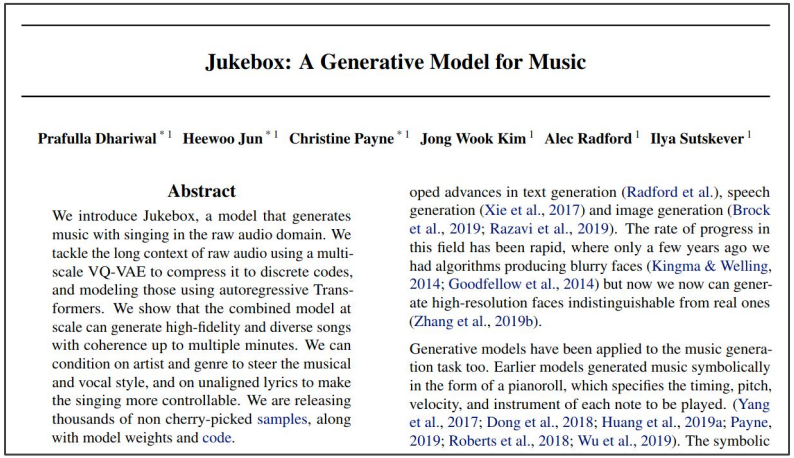
Karen Simonyan **Oriol Vinyals** **Alex Graves**

Nal Kalchbrenner **Andrew Senior** **Koray Kavukcuoglu**

{avdnoord, sedielem, heigazen, simonyan, vinyals, gravesa, nalk, andrewsenior, korayk}@google.com
Google DeepMind, London, UK
[†] Google, London, UK

ABSTRACT

This paper introduces WaveNet, a deep neural network for generating raw audio waveforms. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones; nonetheless we show that it can be efficiently trained on data with tens of thousands of samples per second of audio. When applied to text-to-speech, it yields state-of-



Jukebox: A Generative Model for Music

Prafulla Dhariwal [*1] **Heewoo Jun** [*1] **Christine Payne** [*1] **Jong Wook Kim** [1] **Alec Radford** [1] **Ilya Sutskever** [1]

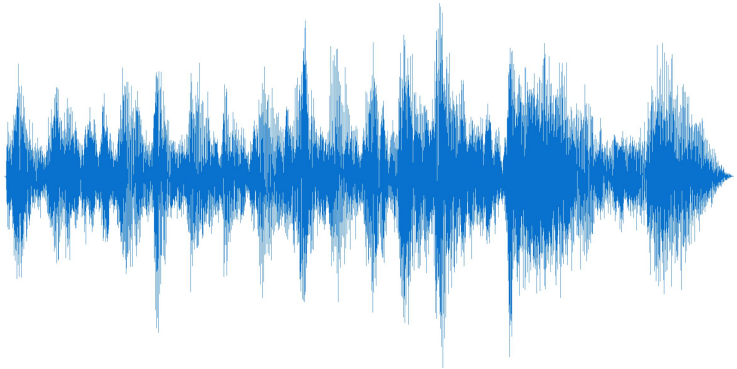
Abstract

We introduce Jukebox, a model that generates music with singing in the raw audio domain. We tackle the long context of raw audio using a multi-scale VQ-VAE to compress it to discrete codes, and modeling those using autoregressive Transformers. We show that the combined model at scale can generate high-fidelity and diverse songs with coherence up to multiple minutes. We can condition on artist and genre to steer the musical and vocal style, and on unaligned lyrics to make the singing more controllable. We are releasing thousands of non cherry-picked samples, along with model weights and code.

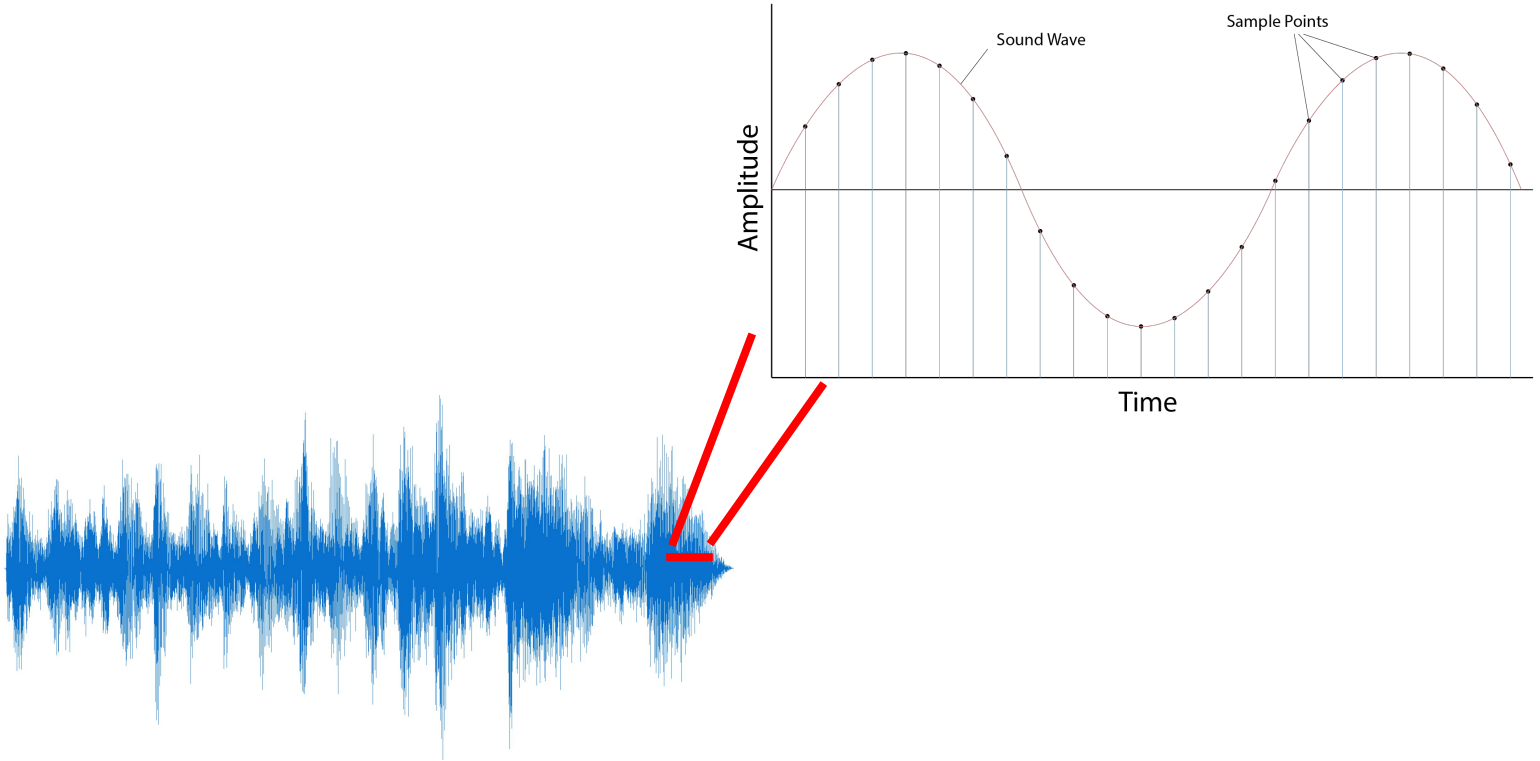
oped advances in text generation (Radford et al.), speech generation (Xie et al., 2017) and image generation (Brock et al., 2019; Razavi et al., 2019). The rate of progress in this field has been rapid, where only a few years ago we had algorithms producing blurry faces (Kingma & Welling, 2014; Goodfellow et al., 2014) but now we now can generate high-resolution faces indistinguishable from real ones (Zhang et al., 2019b).

Generative models have been applied to the music generation task too. Earlier models generated music symbolically in the form of a pianoroll, which specifies the timing, pitch, velocity, and instrument of each note to be played. (Yang et al., 2017; Dong et al., 2018; Huang et al., 2019a; Payne, 2019; Roberts et al., 2018; Wu et al., 2019). The symbolic

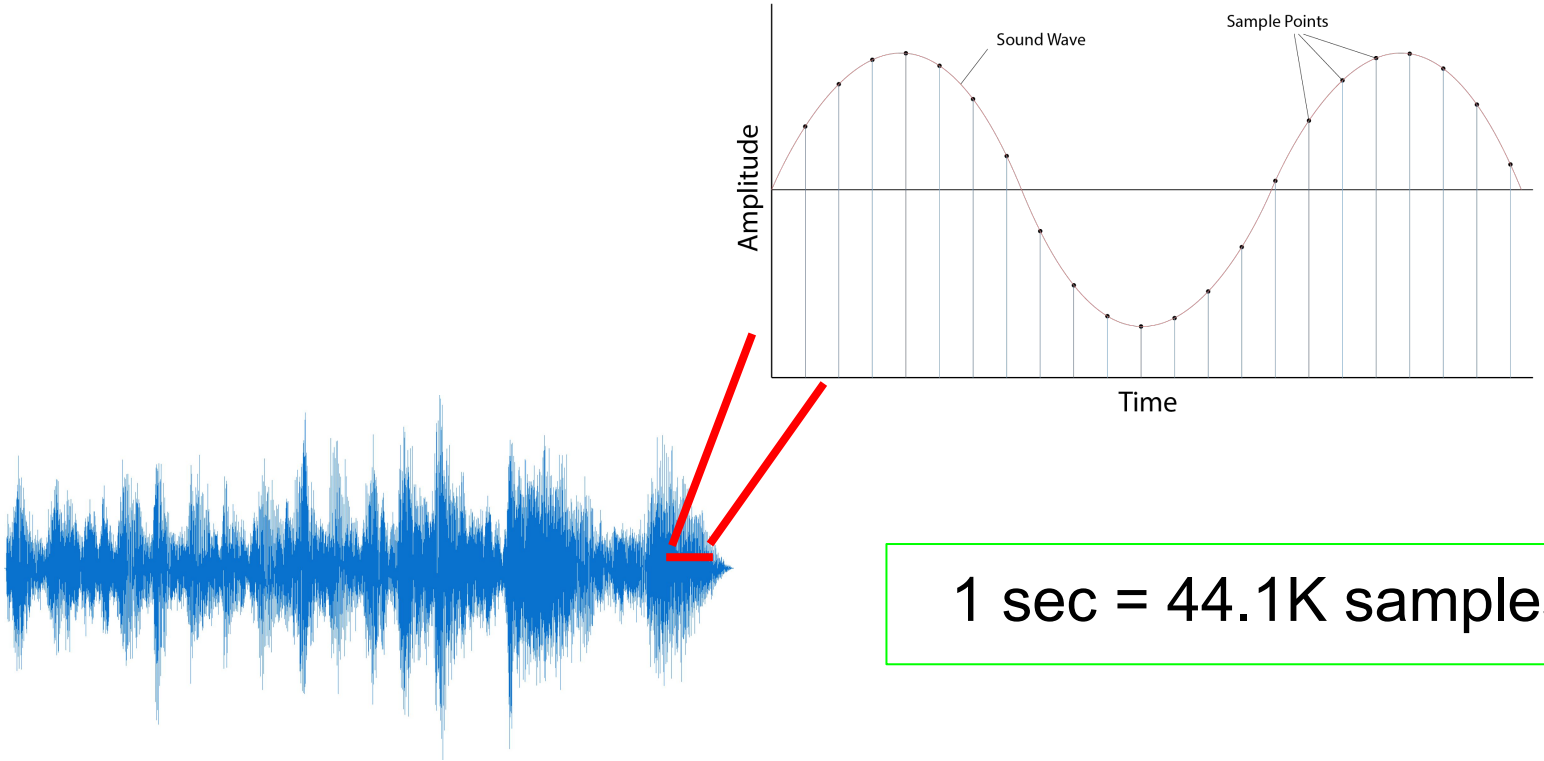
Generation from raw audio



Generation from raw audio



Generation from raw audio



1 sec = 44.1K samples


RAW AUDIO
YOUR DL MODEL
imgflip.com

Generation from raw audio: Challenges

- Difficult to capture long-range dependencies

Generation from raw audio: Challenges

- Difficult to capture long-range dependencies

Pitch

Melody

Structure

Timbre

Rhythm

Harmony

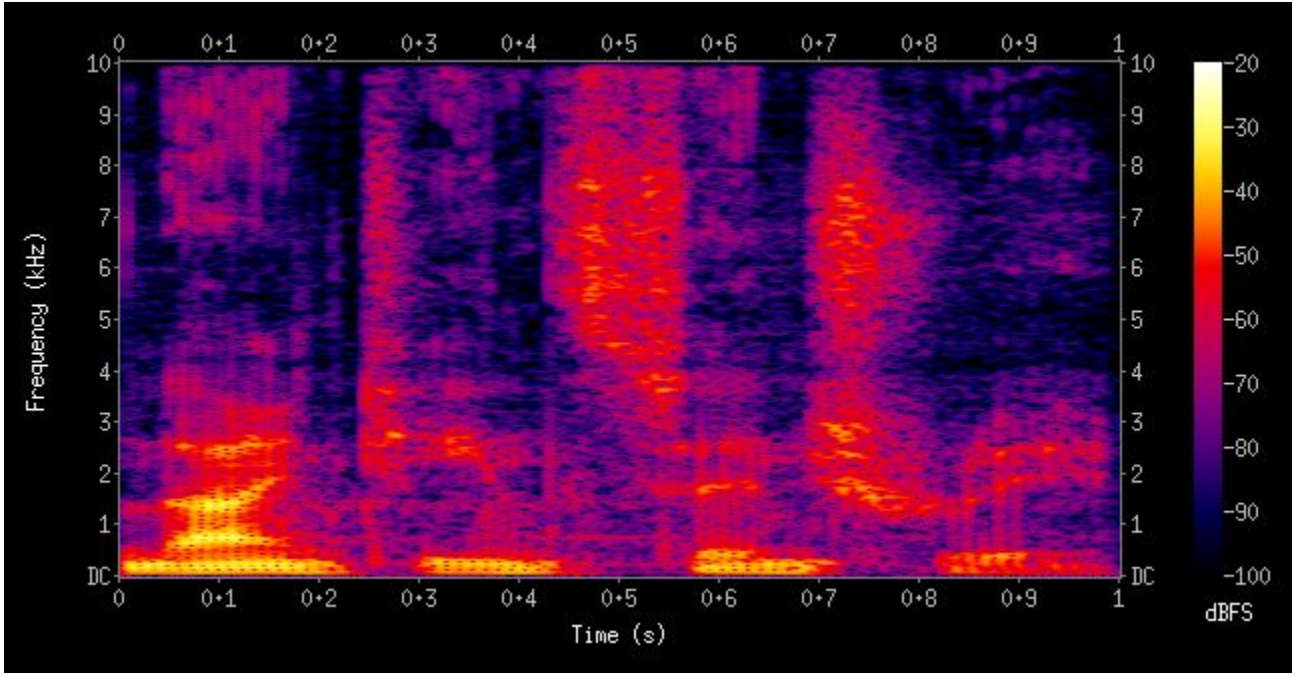
Generation from raw audio: Challenges

- Difficult to capture long-range dependencies
- Computationally expensive

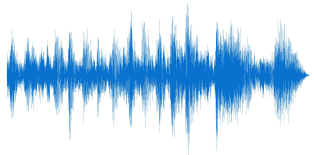
Generation from raw audio: Challenges

- Difficult to capture long-range dependencies
- Computationally expensive
- Generation is slow

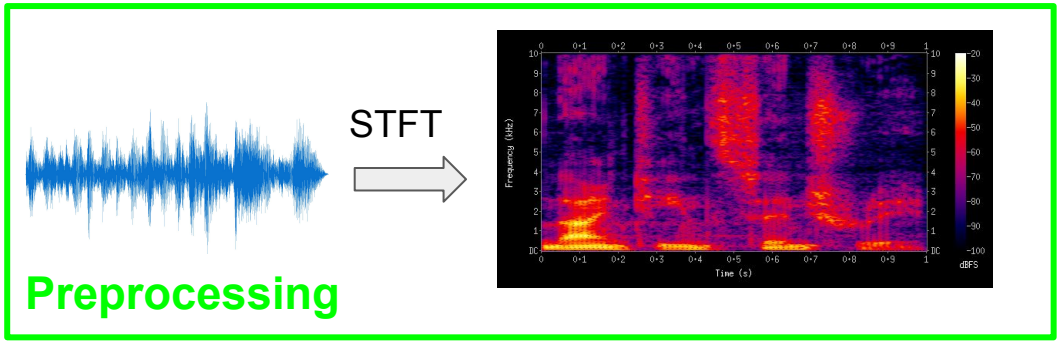
Use a more compact representation of sound


0
0·1
0·2
0·3
0·4
0·5
0·6
0·7
0·8
0·9
1
10
9
8
7
6
5
4
3
2
1
DC
Frequency (kHz)
Time (s)
-20
-30
-40
-50
-60
-70
-80
-90
-100
dBFS

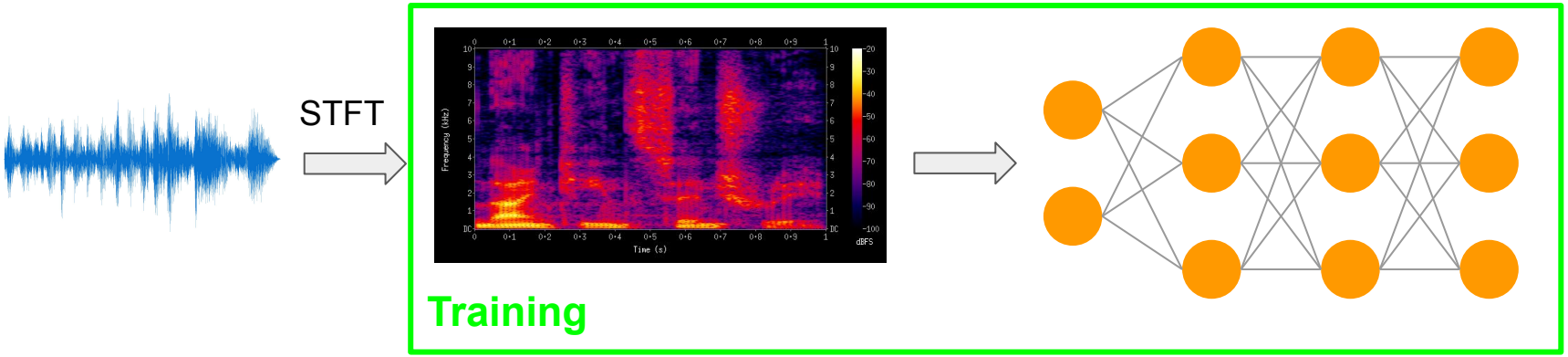
Generation from spectrograms



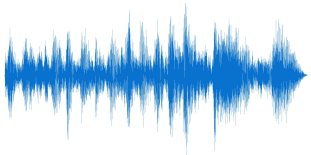
Generation from spectrograms



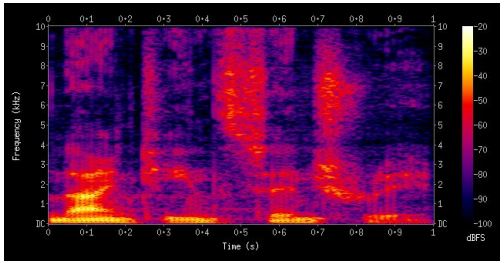
Generation from spectrograms

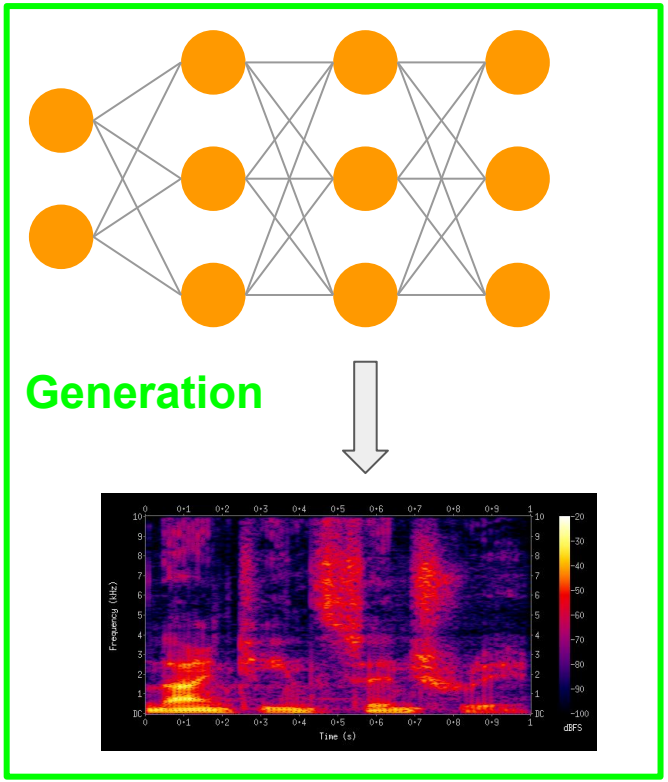


Generation from spectrograms

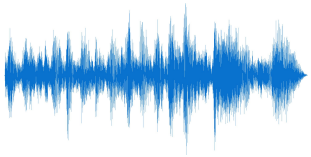


STFT

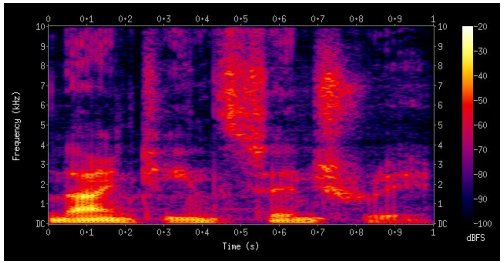


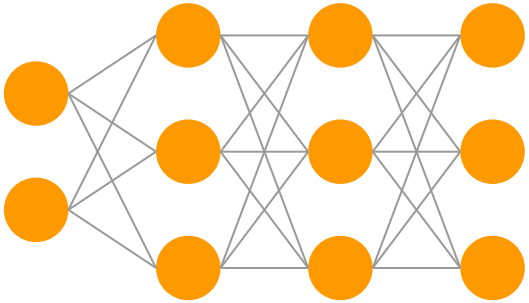


Generation from spectrograms

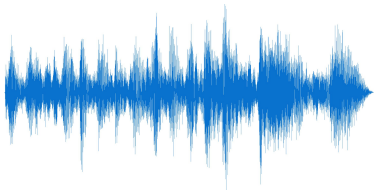


STFT

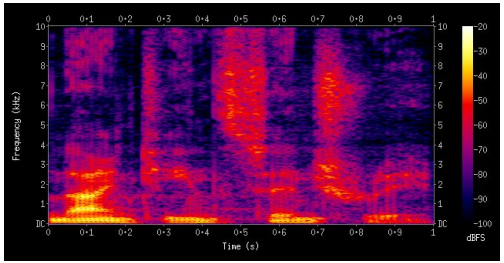




Postprocessing



ISTFT



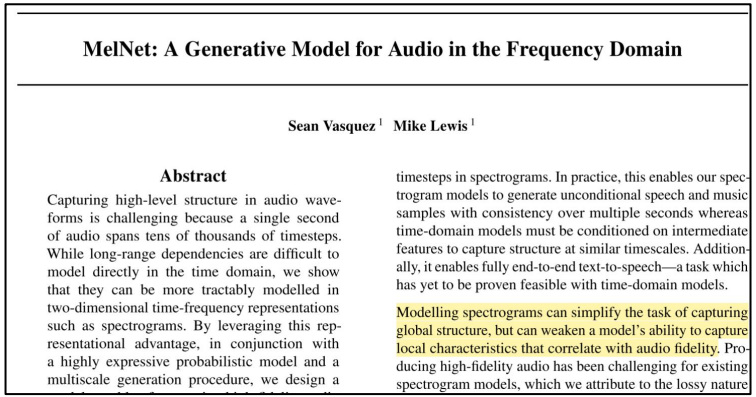
Spectrogram flavours for generation

- Vanilla spectrogram
- Log-amplitude spectrogram
- Log frequency-amplitude spectrogram
- Mel Spectrograms

Join the community!

thesoundofai.slack.com

Generation from spectrograms



MelNet: A Generative Model for Audio in the Frequency Domain

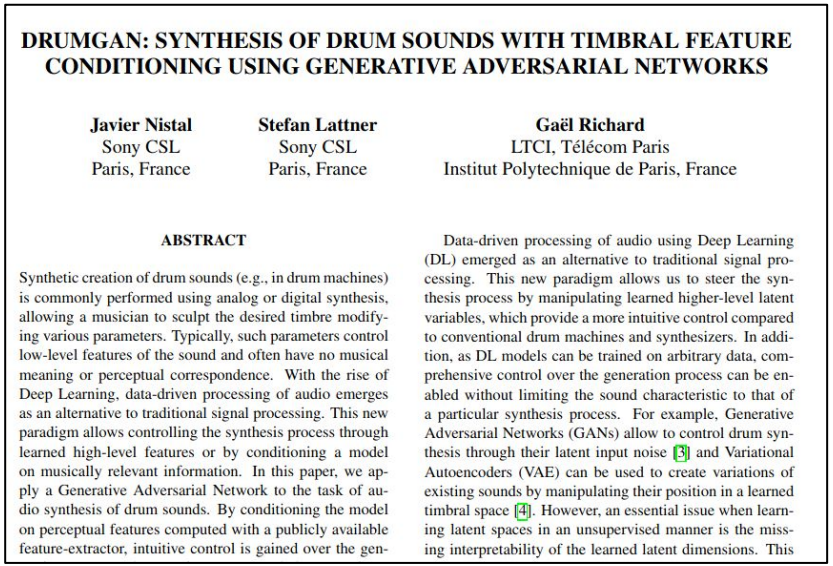
Sean Vasquez [1] Mike Lewis [1]

Abstract

Capturing high-level structure in audio waveforms is challenging because a single second of audio spans tens of thousands of timesteps. While long-range dependencies are difficult to model directly in the time domain, we show that they can be more tractably modelled in two-dimensional time-frequency representations such as spectrograms. By leveraging this representational advantage, in conjunction with a highly expressive probabilistic model and a multiscale generation procedure, we design a

timesteps in spectrograms. In practice, this enables our spectrogram models to generate unconditional speech and music samples with consistency over multiple seconds whereas time-domain models must be conditioned on intermediate features to capture structure at similar timescales. Additionally, it enables fully end-to-end text-to-speech—a task which has yet to be proven feasible with time-domain models.

Modelling spectrograms can simplify the task of capturing global structure, but can weaken a model's ability to capture local characteristics that correlate with audio fidelity. Producing high-fidelity audio has been challenging for existing spectrogram models, which we attribute to the lossy nature



DRUMGAN: SYNTHESIS OF DRUM SOUNDS WITH TIMBRAL FEATURE CONDITIONING USING GENERATIVE ADVERSARIAL NETWORKS

Javier Nistal
Sony CSL
Paris, France

Stefan Lattner
Sony CSL
Paris, France

Gaël Richard
LTCI, Télécom Paris
Institut Polytechnique de Paris, France

ABSTRACT

Synthetic creation of drum sounds (e.g., in drum machines) is commonly performed using analog or digital synthesis, allowing a musician to sculpt the desired timbre modifying various parameters. Typically, such parameters control low-level features of the sound and often have no musical meaning or perceptual correspondence. With the rise of Deep Learning, data-driven processing of audio emerges as an alternative to traditional signal processing. This new paradigm allows controlling the synthesis process through learned high-level features or by conditioning a model on musically relevant information. In this paper, we apply a Generative Adversarial Network to the task of audio synthesis of drum sounds. By conditioning the model on perceptual features computed with a publicly available feature-extractor, intuitive control is gained over the gen-

Data-driven processing of audio using Deep Learning (DL) emerged as an alternative to traditional signal processing. This new paradigm allows us to steer the synthesis process by manipulating learned higher-level latent variables, which provide a more intuitive control compared to conventional drum machines and synthesizers. In addition, as DL models can be trained on arbitrary data, comprehensive control over the generation process can be enabled without limiting the sound characteristic to that of a particular synthesis process. For example, Generative Adversarial Networks (GANs) allow to control drum synthesis through their latent input noise [3] and Variational Autoencoders (VAE) can be used to create variations of existing sounds by manipulating their position in a learned timbral space [4]. However, an essential issue when learning latent spaces in an unsupervised manner is the missing interpretability of the learned latent dimensions. This

Generation from spectrograms: Advantages

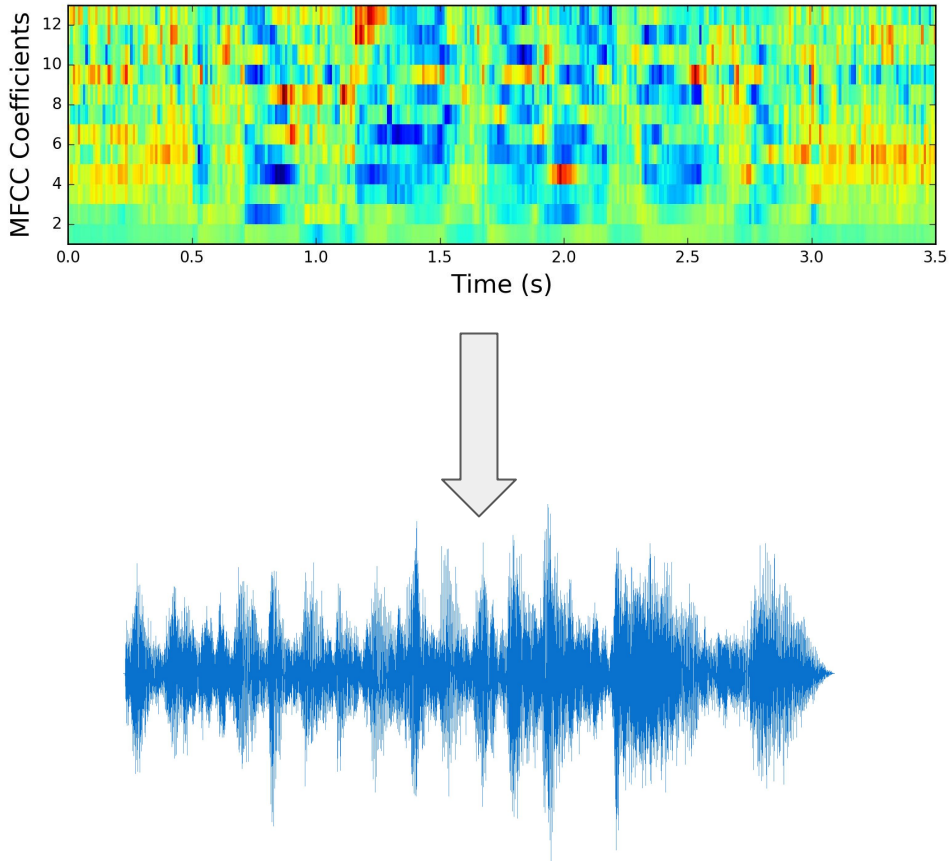
- Temporal axis of spectrogram is more compact than that of waveform
- Capture longer-time dependencies
- Computationally lighter than raw audio

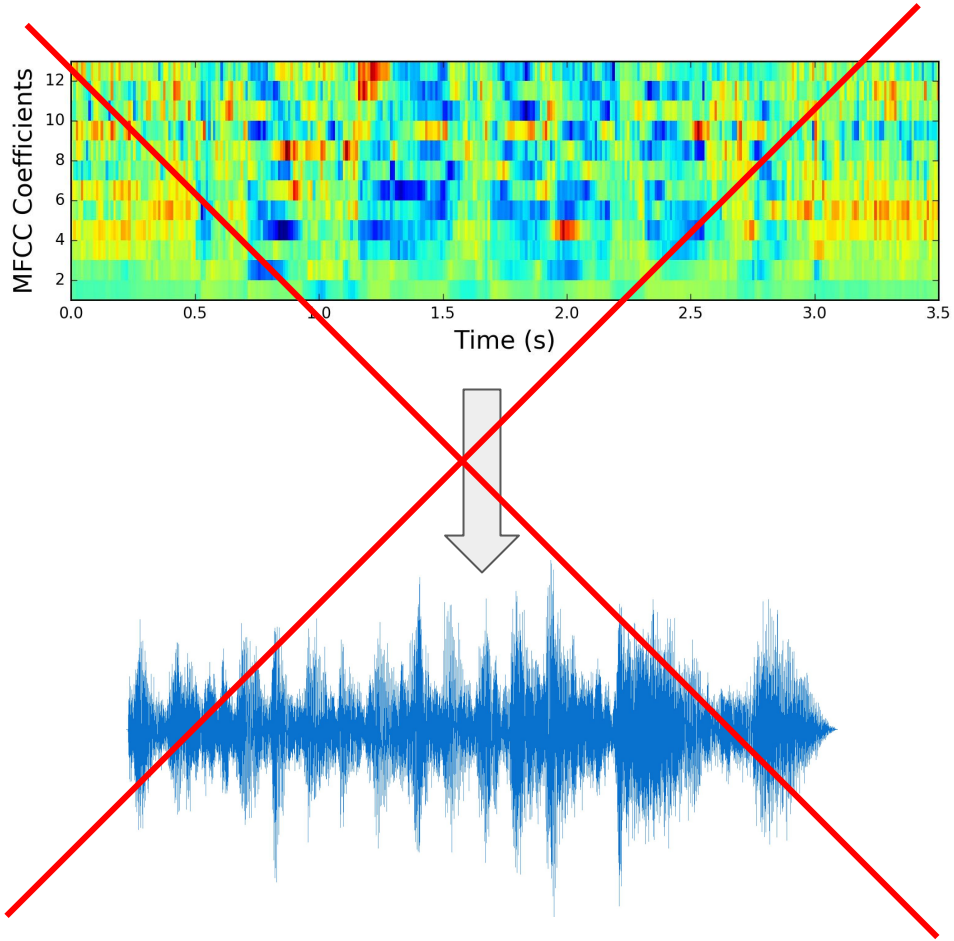
Generation from spectrograms: Challenges

- More difficult to capture local patterns (audio fidelity)
- Phase reconstruction can be problematic

Can we use Mel-Frequency Cepstral Coefficients for sound generation?

NO!


MFCC Coefficients
12
10
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Time (s)


MFCC Coefficients
12
10
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Time (s)

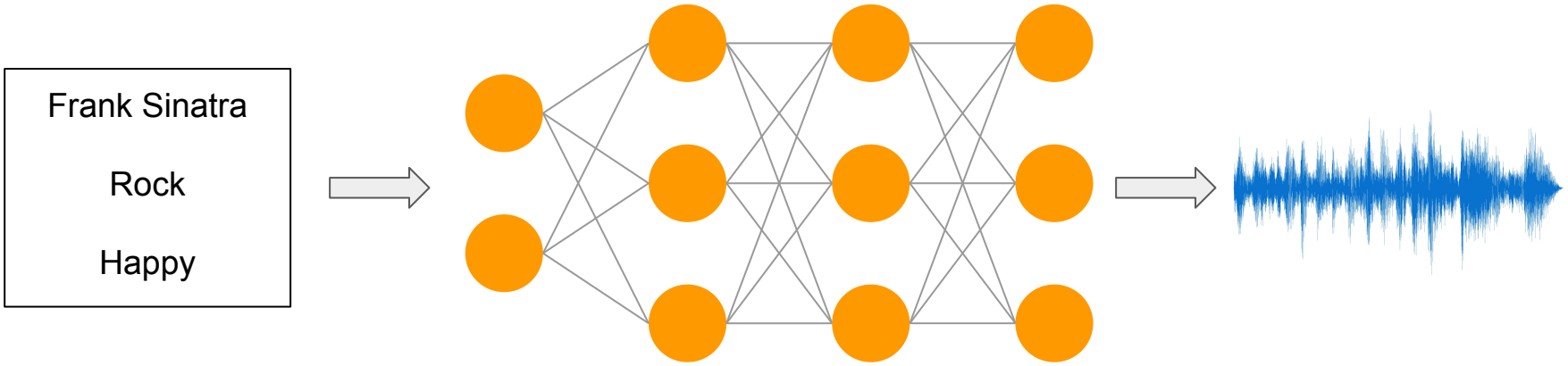
DL architectures for sound generation

- GAN
- Autoencoder
- Variational Autoencoder (VAE)
- VQ-VAE
- ...

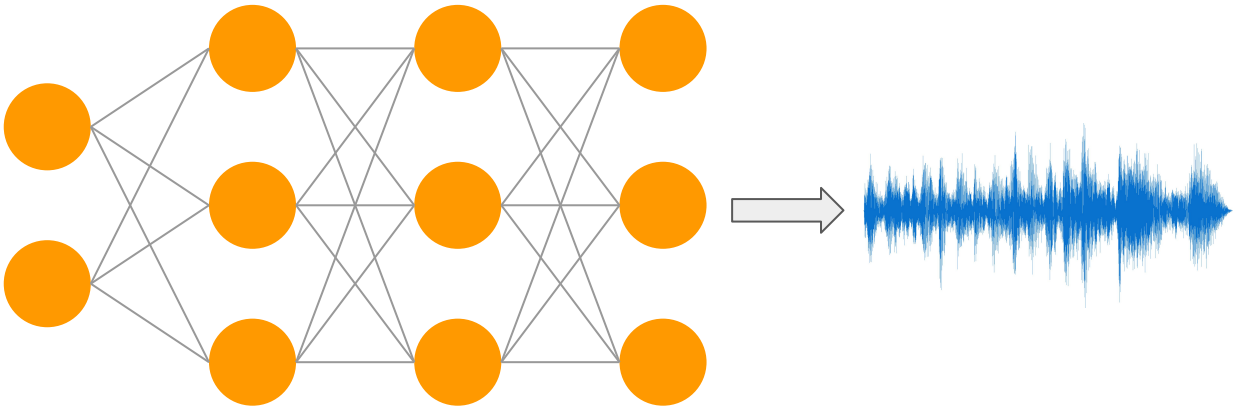
Inputs for generation

- Conditioning
- Autonomous
- Continuation

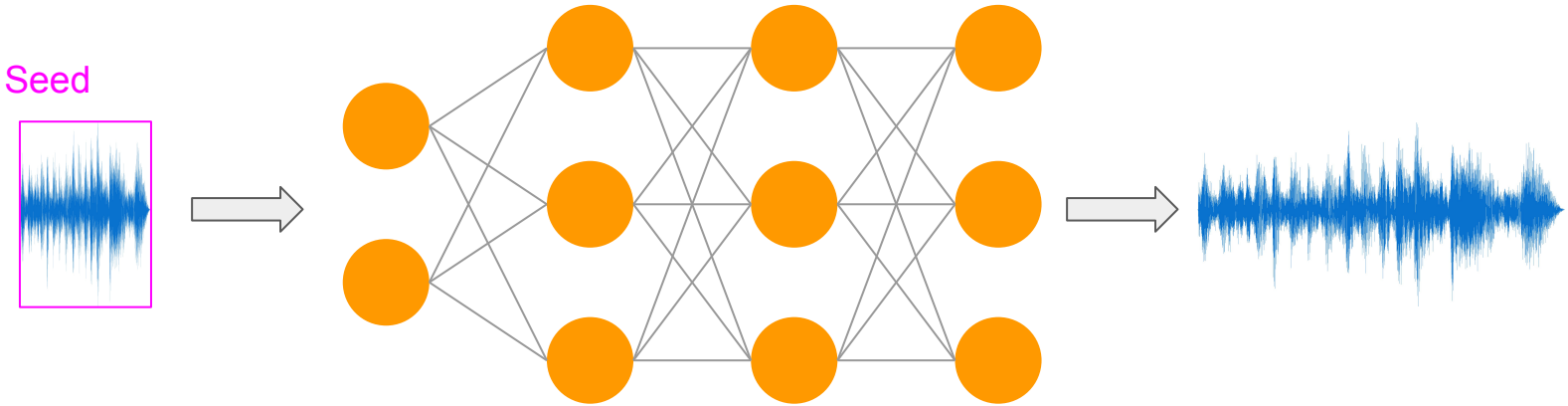
Conditioned generation



Autonomous generation



Continuation



Our generative sound system

- Music notes / music
- (Mel) Spectrograms
- Variational autoencoder
- Autonomous

What next?

- Intuition + theory behind autoencoders