Audio and cross-modal Generative AI

Gaël RICHARD

Professor, Telecom Paris, Institut polytechnique de Paris

Winter School on Generative AI – Feb 27th-28th, 2024, EMINES conference Center, Morocco

G. Richard

Content

- **A few examples of audio generative AI**
- **What is an audio signal ?**
- **Generative audio**
  - A short (and incomplete) historical perspective
  - Deep neural audio synthesis
    - *Autoregressive, VAE, VQ-VAE*
    - *Neural discrete representation (or tokenisation)*
    - *GANs, Diffusion models*
- **Cross modal audio generation : some examples**
- **Towards Hybrid Deep learning**
- **Conclusion**

G. Richard

A few examples of Audio Generative AI

Speech synthesis

- Several systems: Vall-E, VoiceBox (Meta), OpenAI, …
  - Vall-E(Microsoft)
    - Zero-shot TTS

Text	Speaker prompt	Ground Truth	VALL-E
They moved thereafter cautiously about the hut groping before and about them to find something to show that Warrenton had fulfilled his mission			

- … or keeping the speaker emotion

Text	Emotion	Speaker prompt	VALL-E
We have to reduce the number of plastic bags.	Anger		
	Sleepy		
	Amused		

VALL-E: https://www.microsoft.com/en-us/research/project/vall-e-x/vall-e/
VOICEBox: https://voicebox.metademolab.com/

A few examples of Audio Generative AI

Speech synthesis

G. Richard

- « Deepfake » voices by many actors : Resemble.ai, Speechify, Respeecher,…

- An example on my voice with Speechify (*OpenAI's API*)

Text	Training prompt (in French)	Generated voice
Hi, Gaël Richard! It's time to listen to your voice clone in action. Your voice clone opens up a world of possibilities.		

https://speechify.com/

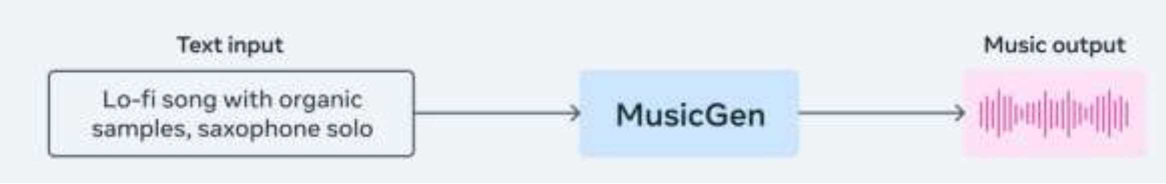
A few examples of Audio Generative AI

Audio/Music synthesis

- Several impressive models:


OpenAI (Jukebox,Musenet,) (unseen lyrics rendition, completion, ..)

- MusicGen/AudioGen (AudioCraft, Meta): text-to-music or text-to audio generation



An example with MusicLM

Text	Generated music
Slow tempo, bass-and-drums-led reggae song. Sustained electric guitar. High-pitched bongos with ringing tones. Vocals are relaxed with a laid-back feel, very expressive.	

https://openai.com/research/jukebox
https://audiocraft.metademolab.com/musicgen.html
https://google-research.github.io/seanet/musiclm/examples/

A few examples of Audio Generative AI

Cross modal audio synthesis

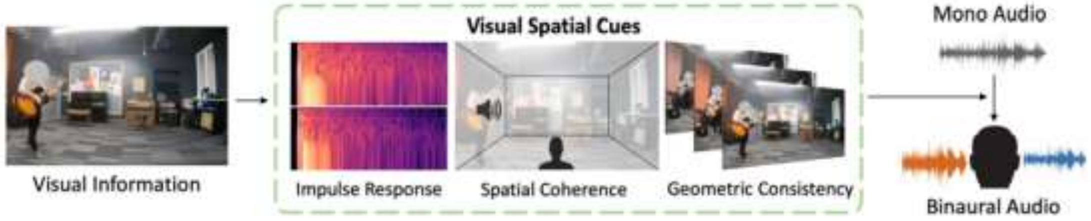
G. Richard

- Examples with MusicLM (text+ melody conditioning) generation,
  - *Painting Caption Conditioning:* An example with MusicLM

Guernica - Pablo Picasso

"The grey, black, and white painting, on a canvas 3.49 meters tall and 7.76 meters across, portrays the suffering wrought by violence and chaos. Prominent in the composition are a gored horse, a bull, screaming women, a dead baby, a dismembered soldier, and flames." By [wikipedia]

  - Other examples: visually guided audio spatialization + the sound of pixel (2018)



https://google-research.github.io/seanet/musiclm/examples/

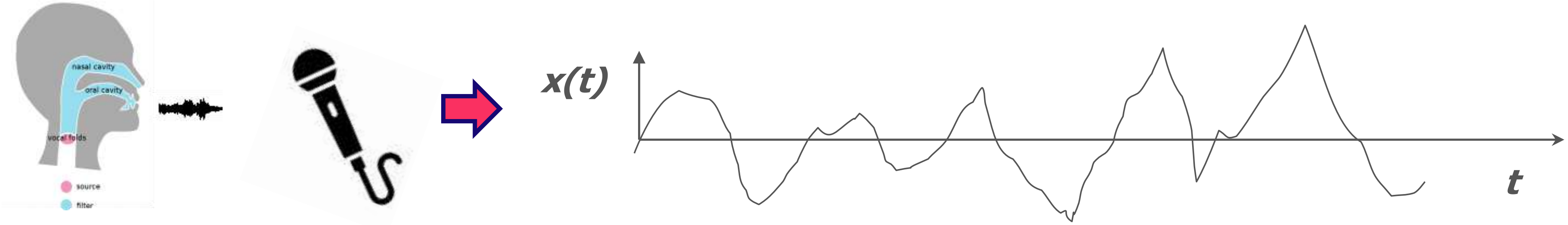
Rishabh Garg, Ruohan Gao, Kristen Grauman, Visually-Guided Audio Spatialization in Video with Geometry-Aware Multi-task Learning. International Journal of Computer Vision (IJCV). Vol 131. 2023. Special Issue for Best Papers of BMVC

… but what is an audio signal ?

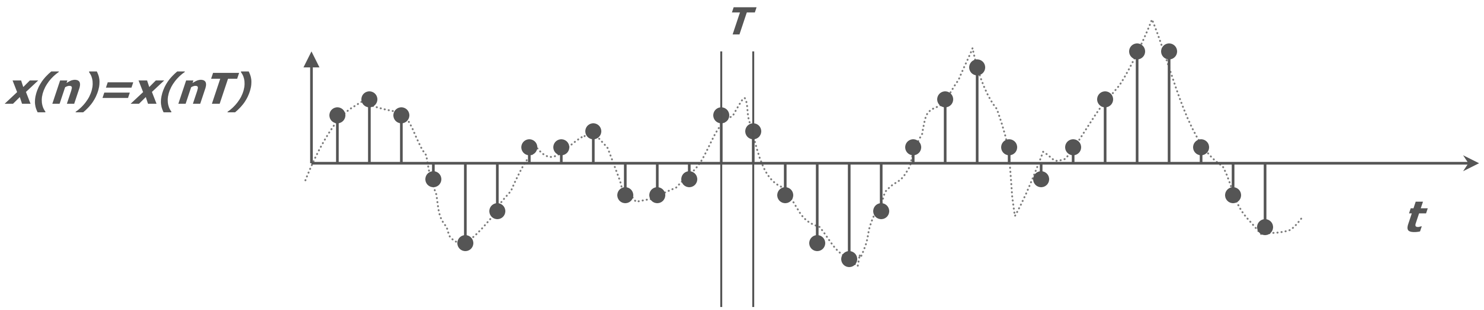
G. Richard

What is an audio signal ……

- The audio signal x(t) is an continuous acoustic signal



- Let x(nT) be the discrete signal sampled at time *t=nT*



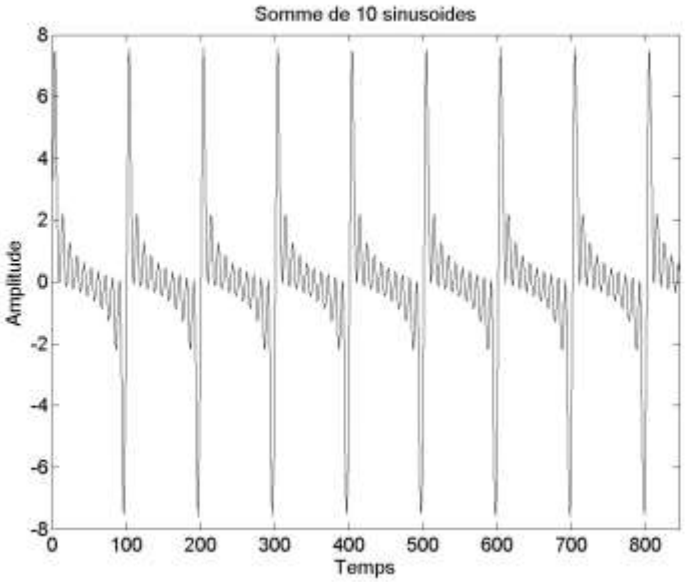
Time-Frequency representation

- Fourier Transform

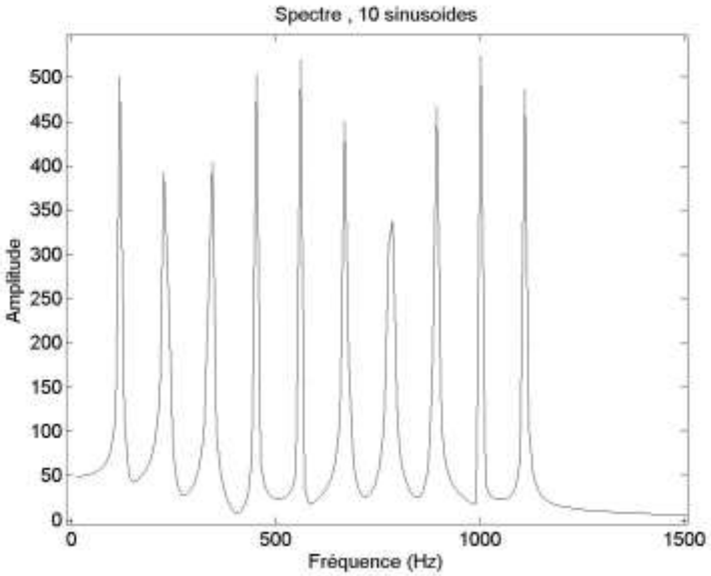
$$X_k = \sum_{n=0}^{N-1} x_n e^{-2j\pi nk/N}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{2j\pi nk/N}$$

x_n

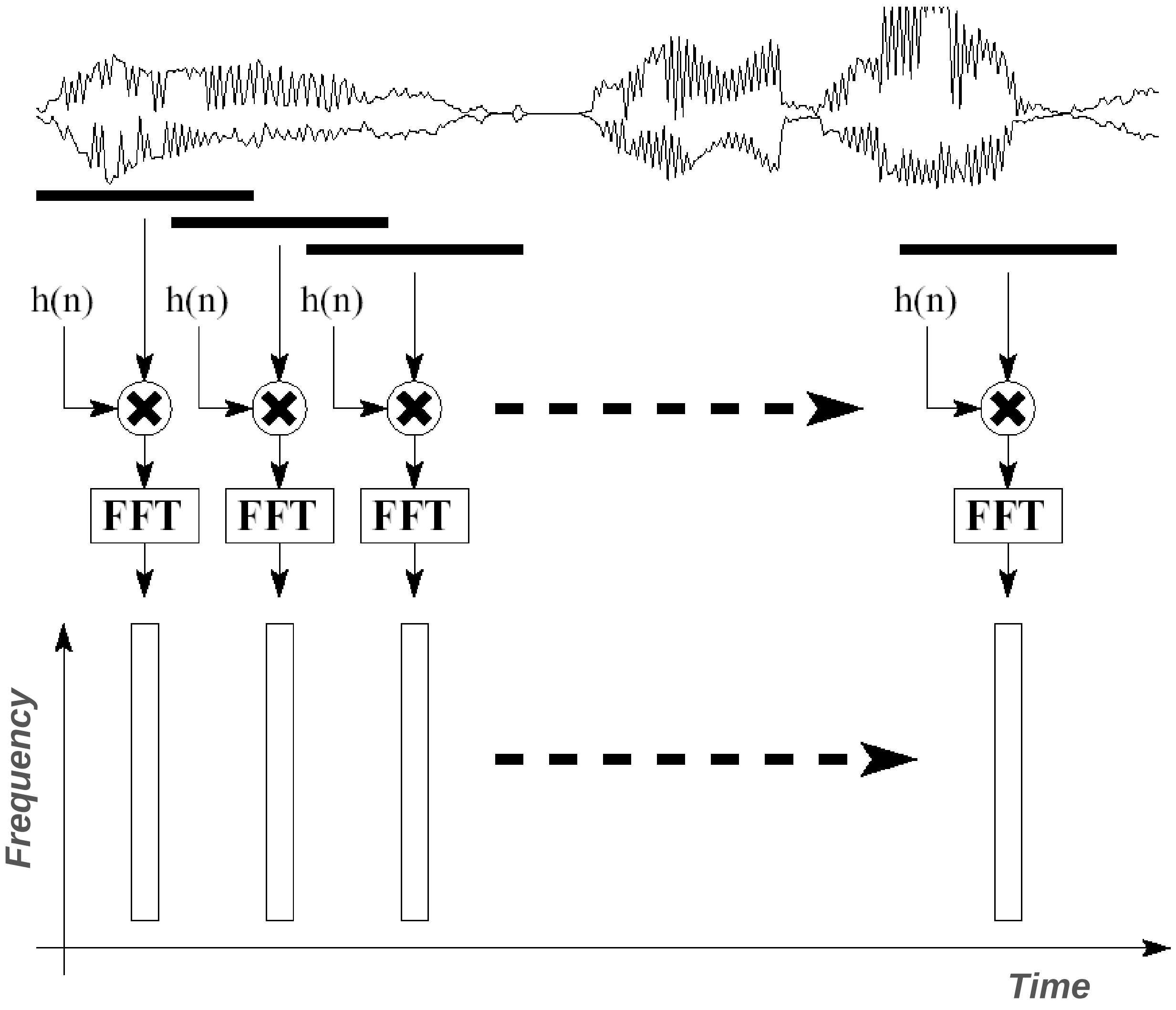


$|X_k|$



Spectral analysis of an audio signal (1)

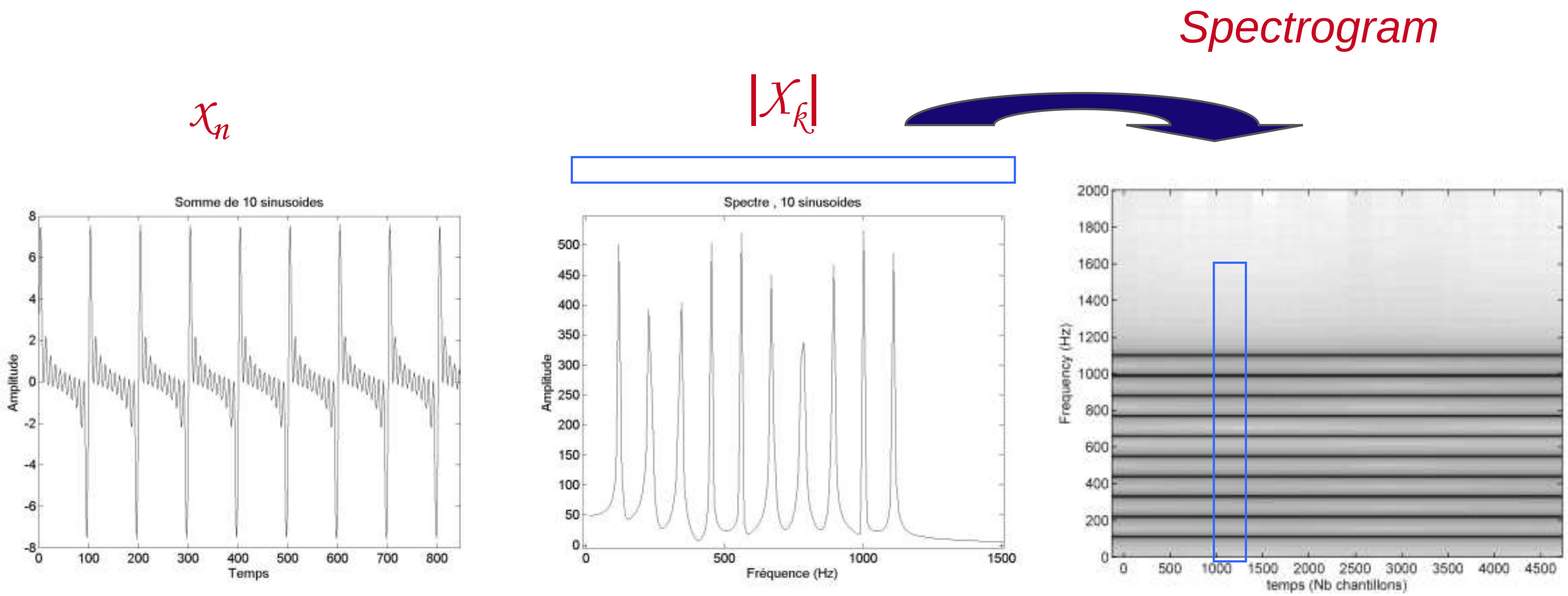
(drawing from J. Laroche)



G. Richard

Spectral analysis of an audio signal (2)

- Spectrogram of a sum of 10 stable sinusoids



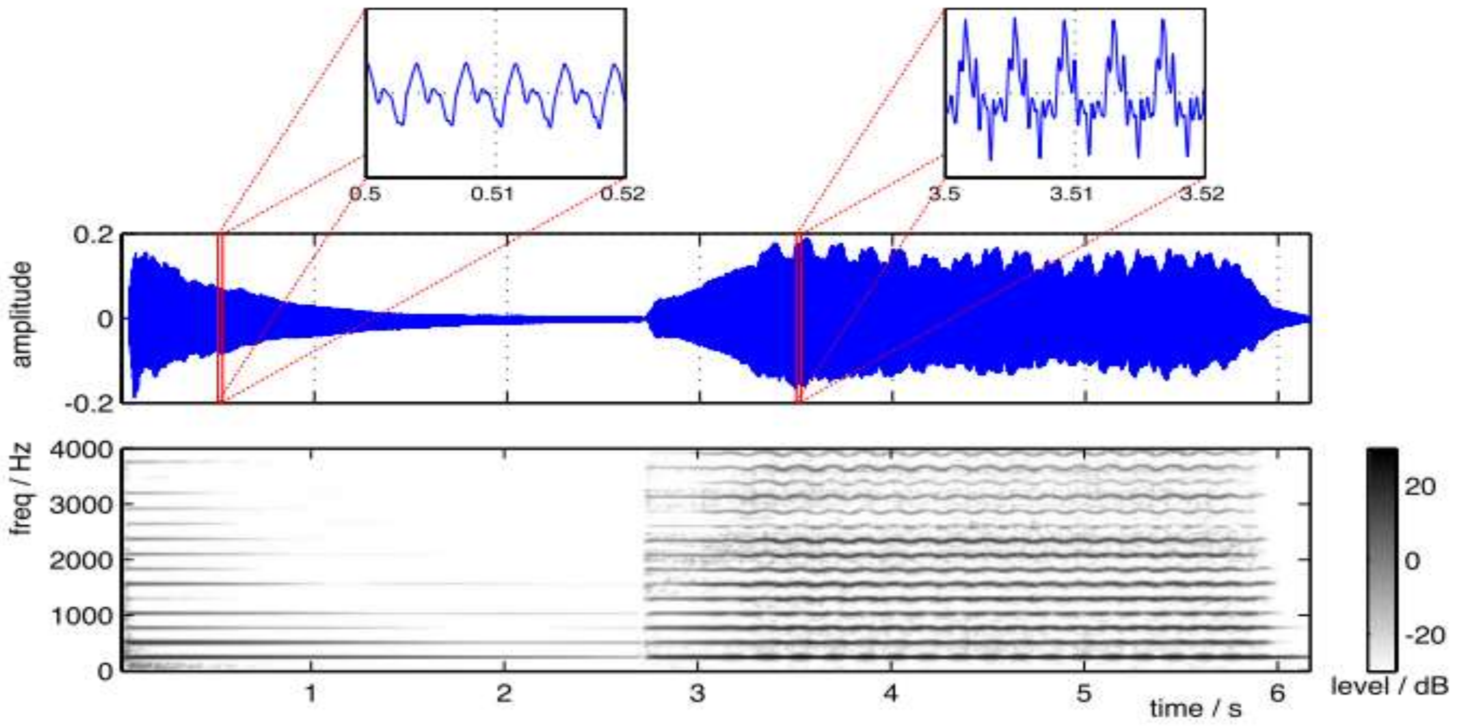
Audio signal representations

G. Richard

- Example on a music signal: note C (262 Hz) produced by a piano and a violin.

Temporal Signal

Spectrogram



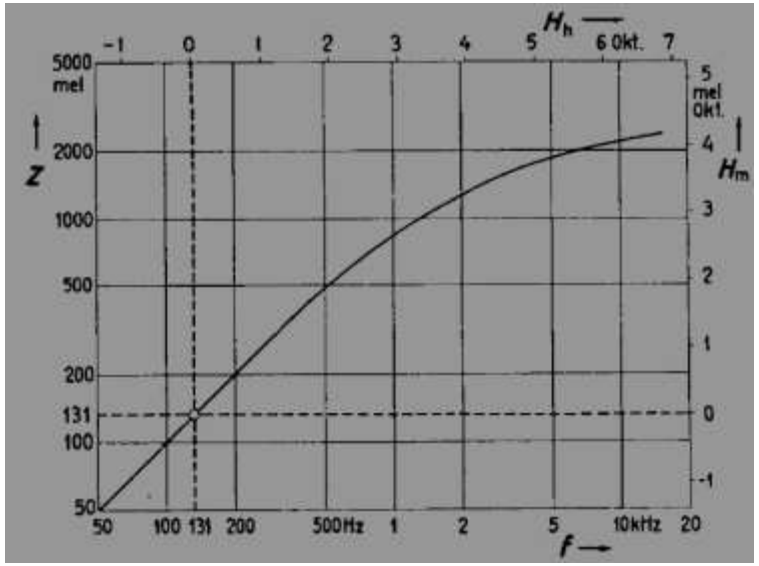
From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

G. Richard

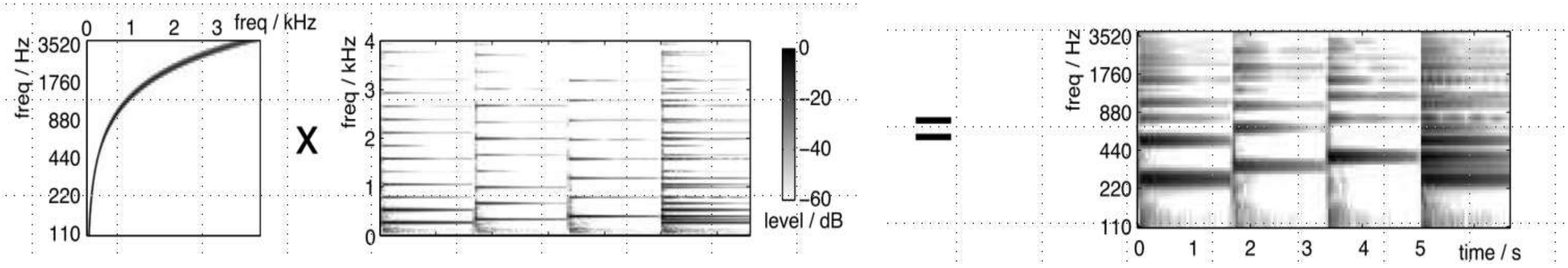
Towards a more specific representation

Mel-spectrogram

- Exploiting principles of sound perception
  - E.g. Tonal heights perception: Mel scale
    - From 0 à 500 Hz où 1 Mel = 1 Hz (linear)
    - Above 500 Hz, height perception (or « tonie ») growths logarithmically with frequency



- Example of analytical formula: $mel(f) = 1000 \log_2(1 + \frac{f}{1000})$



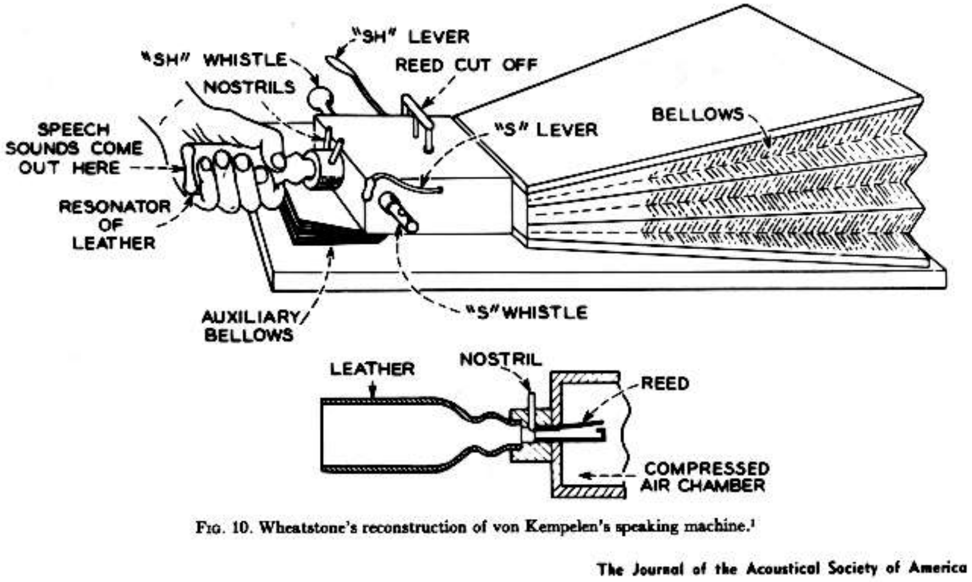
G. Richard

About Generative audio …

Generative audio … an old domain

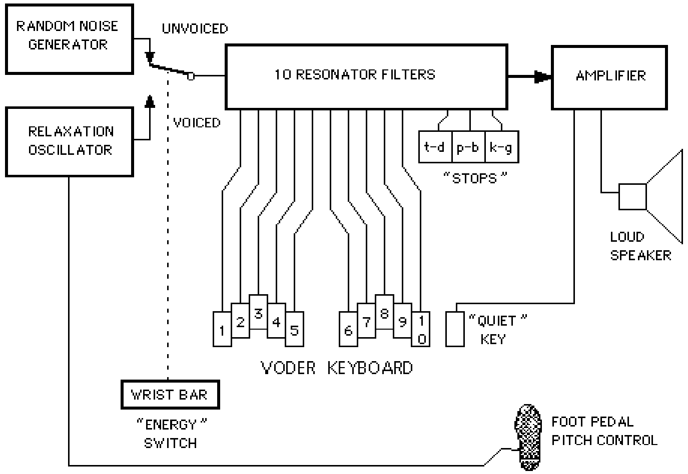
G. Richard

- …generating speech with an instrument or a machine

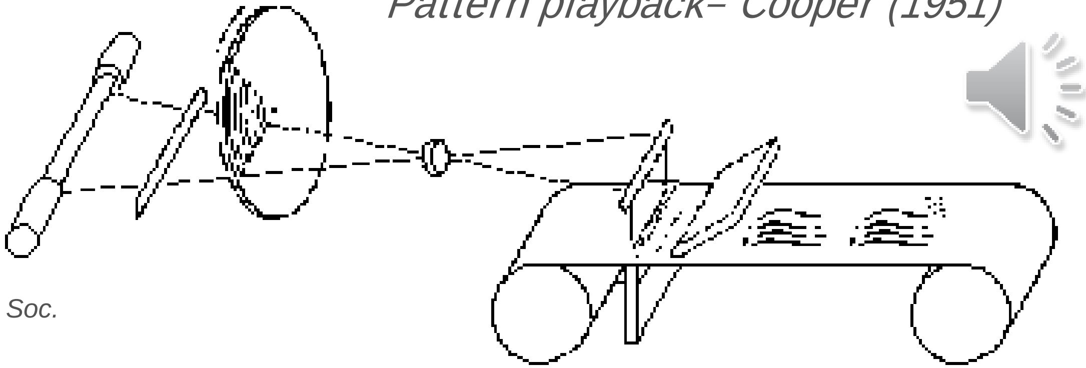


Van Kempelen machine (1791)

Voder – Dudley (1939)



Pattern playback– Cooper (1951)

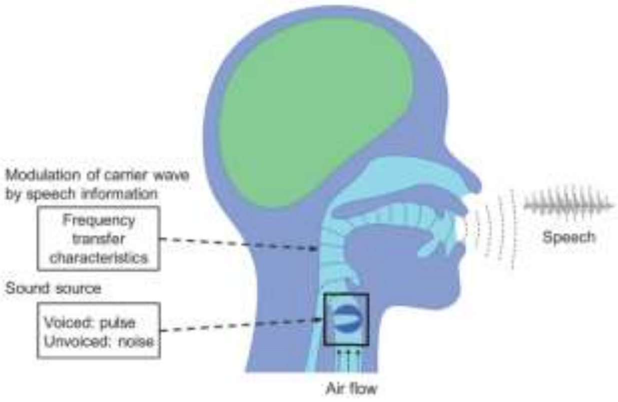


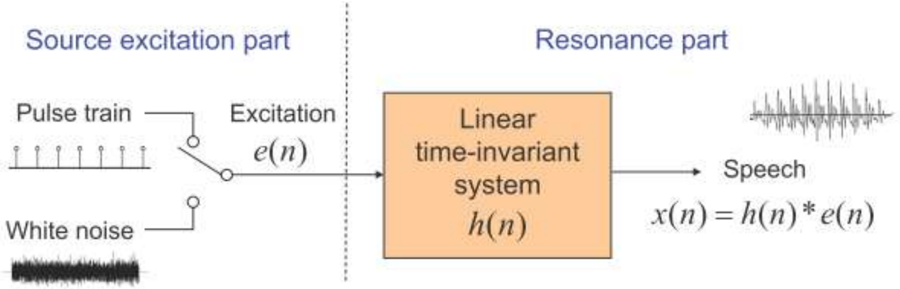
Dennis H. Klatt (1987), "Review of text-to-speech conversion for English" J. Acous. Soc. Amer. 82, 737-793

Generative audio … an old domain

G. Richard

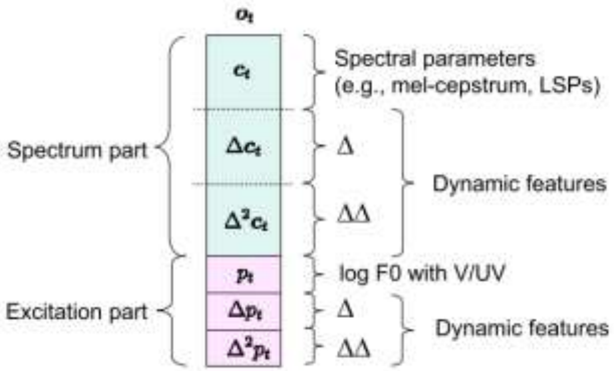
- …generating speech with a simplified « speech production » model

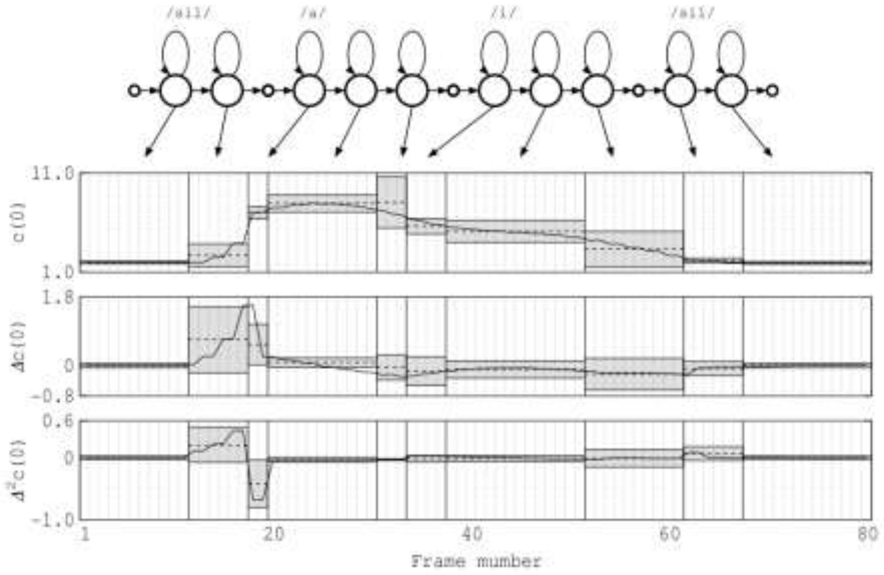




- ..Or using Hidden Markov Models (HMMs)

Observation vector



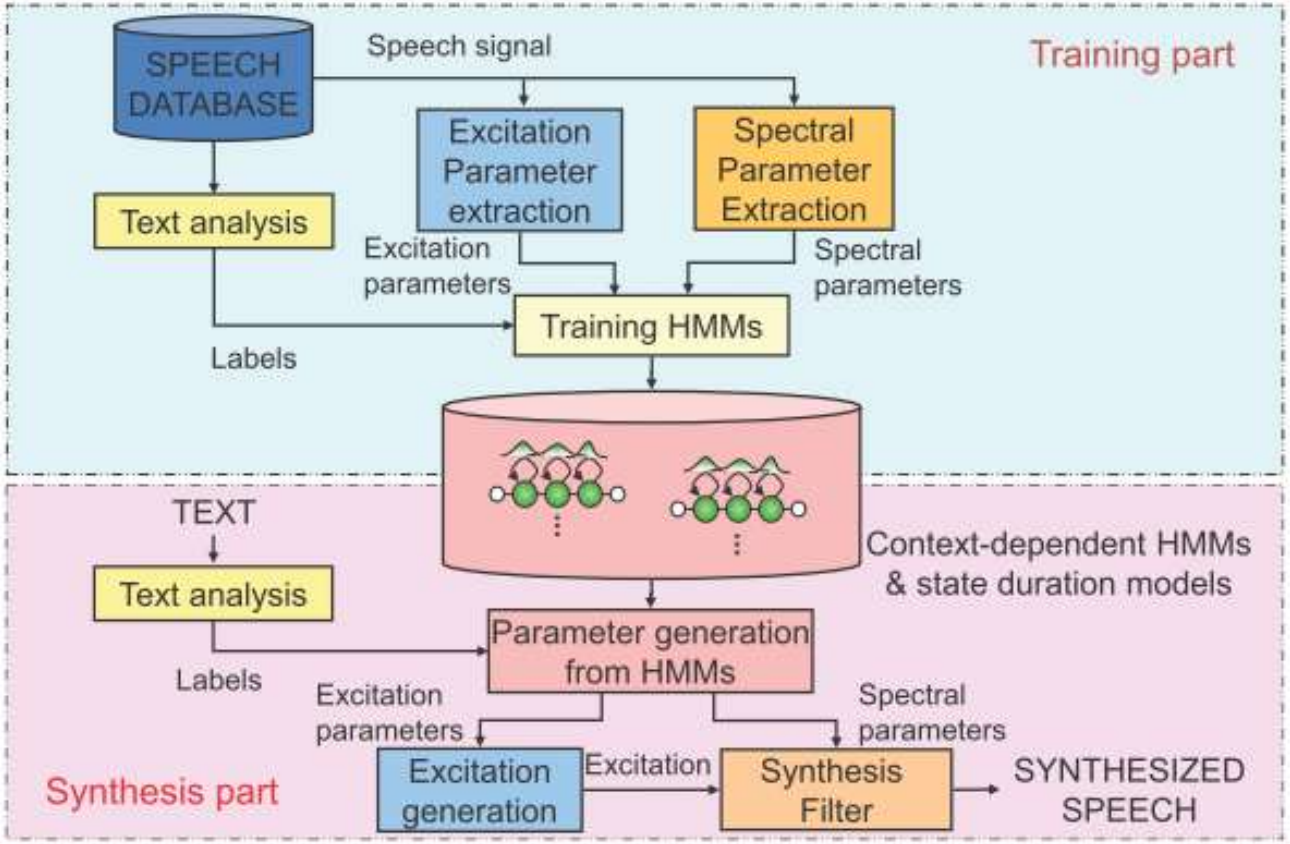


K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, "Speech Synthesis Based on Hidden Markov Models," in *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, May 2013

Generative audio … an old domain

G. Richard

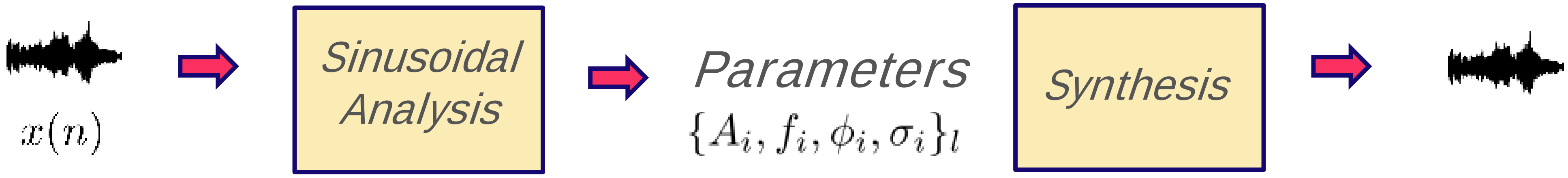
- …generating speech with Hidden Markov Models (HMMs)



K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, "Speech Synthesis Based on Hidden Markov Models," in *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, May 2013

Generative audio … an old domain

- …generating and transforming sound using an analysis/synthesis model



$$\hat{x}(n) = \sum_i A_i \sin(2\pi f_i n + \phi_i) + b_i(n)$$

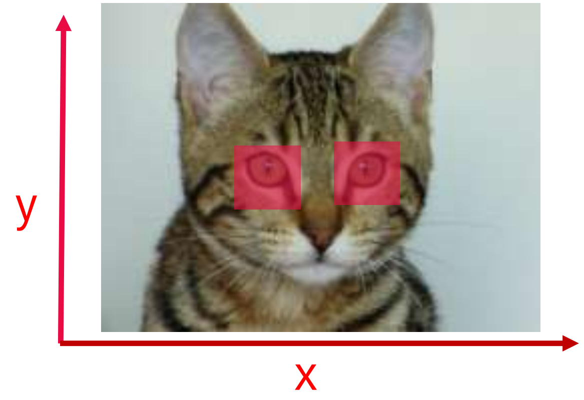
⇨Example on a piano signal

- Original signal:
- Transposed by a third:
- Signal S (« Sum of sinusoïds with vibrato effect»):
- Signal N (Noise):

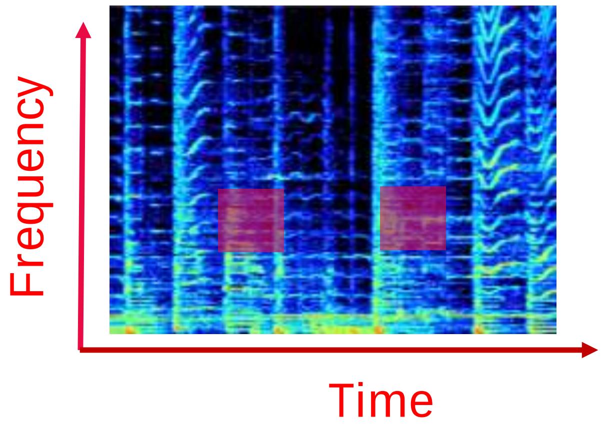
G. Richard

An « image of audio » (e.g spectrogram) is not the same as a natural image

- **Natural images**
  - the axes x and y represent the same concept (*spatial position*)
  - the elements of an image have the same meaning independently of their positions over x and y.
  - neighboring pixels:
    - usually highly correlated,
    - often belong to the same object



- **Time-frequency audio representations (for example a spectrogram)**
  - the axes x and y represent profoundly different concepts (time and frequency).
  - the elements of spectrogram (*such as the T/F area of a source*) have the same meaning independently of their position over time but not over frequency
  - no invariance over y, even in the case of log-frequencies
  - neighboring pixels:
    - are not necessarily correlated
    - a given sound source (such has an harmonic sound) can be distributed over the whole frequency in a sparse way (*the harmonics of a given sound can be spread over the whole frequency range*)



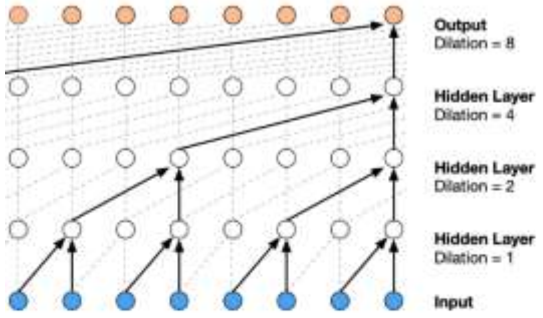
G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021 (to appear)

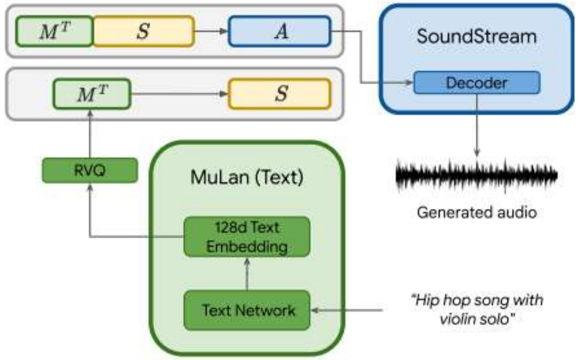
Deep neural audio synthesis

- Machine-learning based models "*uses large amount of data and machine learning to generate sounds"*
  - *A rapid growth and adoption of deep neural networks for audio synthesis*

From wavenet (2016) ….. MusicLM (2023)

(*autoregressive model*) (*Generating Music from Text*)

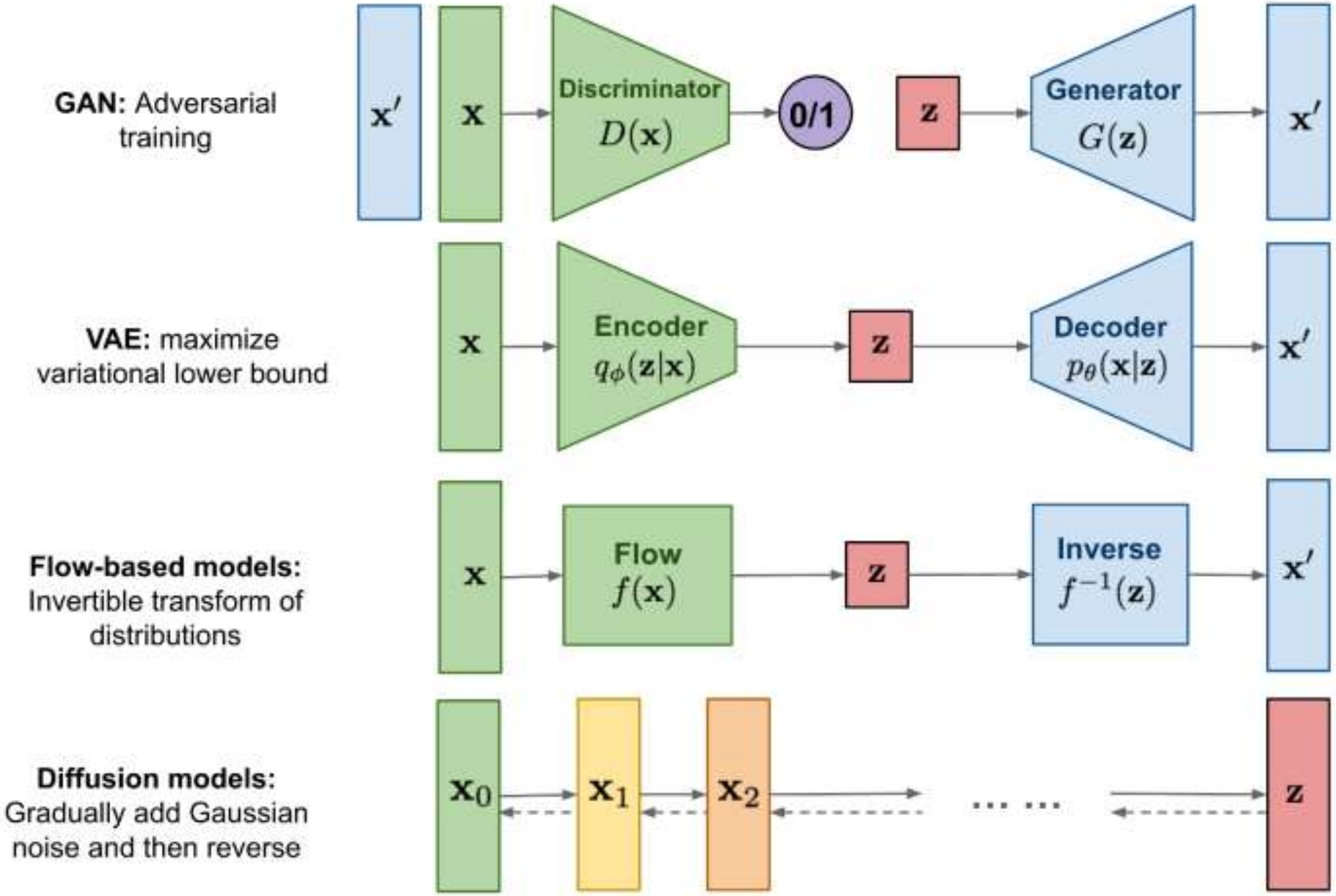




Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio (cite arxiv:1609.03499)
A. Agostinelli & al. MusicLM: Generating Music From Text, https://arxiv.org/abs/2301.11325, 2023.

A large variety of generative models …

G. Richard



L. Weng, What are diffusion models ?, 2021. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

G. Richard

Deep neural audio synthesis

Arch.	Name		Audio representation	Data	Conditioning
NAM	waveNet	[van den Oord et al., 2016a]	waveform	speech, piano	speaker ID, text
	Universal music Translation	[Mor et al., 2018]	waveform	classical music	-
	Hierarchical waveNet	[Dieleman et al., 2018]	waveform	piano music	-
	SampleRNN	[Mehri et al., 2017]	waveform	speech, piano music	-
	MelNet	[Vasquez and Lewis, 2019]	mag. spec.	speech, piano music	speaker ID text
	wavenetAE	[Engel et al., 2017]	waveform	tonal sounds	pitch
	sparse Transformer	[Child et al., 2019]	waveform	piano music	-
NFs	Parallel waveNet	[van den Oord et al., 2018a]	waveform	speech	text pitch
	ClariNet	[Ping et al., 2018]	waveform	speech	text
	FlowaveNet	[Kim et al., 2018]	waveform	speech	text Mel spec.
	waveGlow	[Prenger et al., 2018]	waveform	speech	text Mel spec.
	waveFlow	[Ping et al., 2020]	waveform	speech	text Mel spec.
	Blow	[Serrà et al., 2019]	waveform	speech	speaker ID
VAEs	Planet Drums	[Aouameur et al., 2019]	Mel-scaled mag. spec.	drums	instrument ID
	Jukebox	[Dhariwal et al., 2020]	waveform	music	artist & genre ID lyrics
	NOTONO	[Bazin et al., 2020]	mag. & IF	tonal instruments	pitch
	FlowSynth	[Esling et al., 2019]	mag.	synth. sounds	semantic tags
	Neural Granular Sound Synth.	[Bitton et al., 2020]	waveform	orchestral drums animals	pitch instrument ID
GANs	WaveGAN	[Donahue et al., 2019]	waveform	speech drums piano birds	-
	GANSynth	[Engel et al., 2019]	mag. & IF	tonal instruments	pitch ID
	MelGAN	[Kumar et al., 2019]	mag. spec.	speech music	Mel-scaled spec. text
	GAN-TTS	[Binkowski et al., 2020]	waveform	speech	pitch, text, speaker ID

J. Nistal, "Exploring Generative Adversarial Networks for Controllable Musical Audio Synthesis, PhD Thesis, IP Paris, 2022

Wavnet

a generative model, directly from the audio waveform

G. Richard

- The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$

is factorised as a product of conditional probabilities :

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

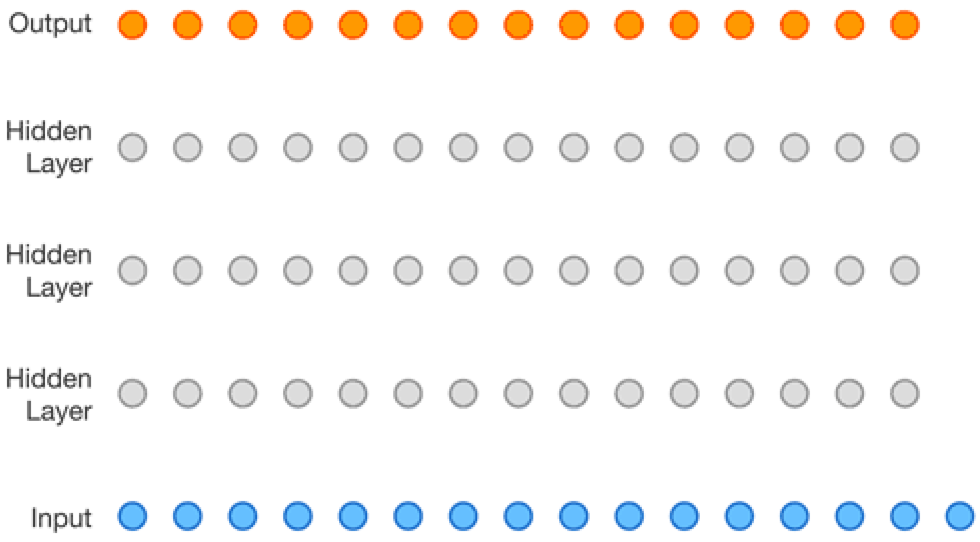
- the conditional probability distribution is modelled by a stack of convolutional layers;
- Output of the model: has the same time dimensionality as the input (no pooling)
- Output: a categorical distribution over the next value x_t with a softmax layer - optimized to maximize the log-likelihood of the data w.r.t. the parameters.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio (cite arxiv:1609.03499)

Wavnet

a generative model, directly from the audio waveform

- Dilated causal convolutions (the main ingredient!).
  - Classic causal convolutions needs many layers to increase the receptive fields (RF)
  - Dilated causal convolutions greatly increase RF



Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio (cite arxiv:1609.03499)

Wavnet

a generative model, directly from the audio waveform

G. Richard

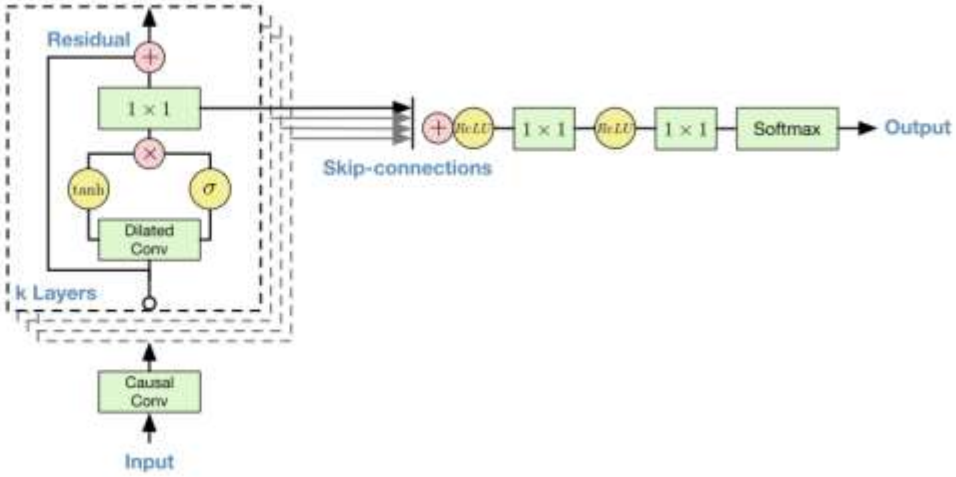
- Condition distributions $p(x_t|x_1,\ldots,x_{t-1})$ modelled using softmax distributions
- Use of mu-law to limit the number of “categories” (amplitude values):

$$f\left(x_t\right) = \mathrm{sign}(x_t)\frac{\ln\left(1+\mu\left|x_t\right|\right)}{\ln\left(1+\mu\right)}$$

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

- Use of Gated recurrent units

- .. and residual and skip connections



Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio (cite arxiv:1609.03499)

G. Richard

Wavnet : sound examples

(from https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio)

- Speech … but also music (no conditioning)

- But it is also possible to use conditions :

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^{T}\mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^{T}\mathbf{h}\right)$$

- Speech (with condition on the text)

G. Richard

Wavnet and other neural autoregressive models

- Wavnet remains complex (sample is generated one at a time)

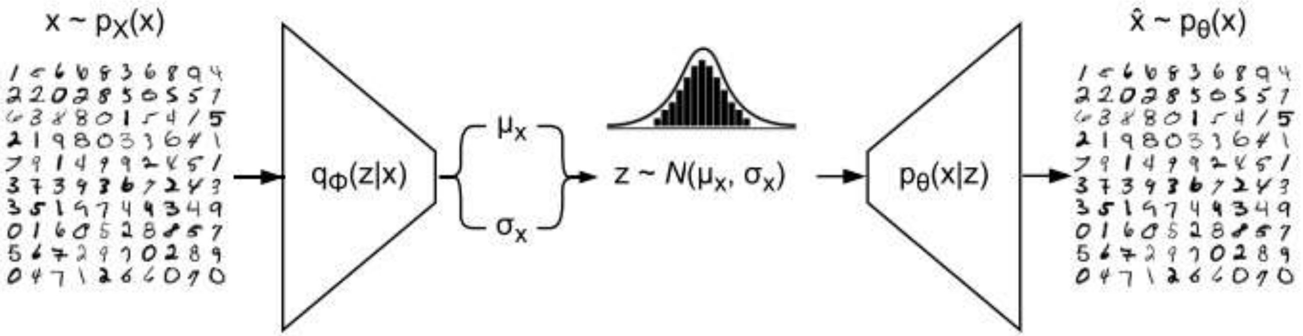
- Other neural autoregressive models

Arch.	Name		Audio representation	Data	Conditioning
NAM	waveNet	[van den Oord et al., 2016a]	waveform	speech, piano	speaker ID, text
	Universal music Translation	[Mor et al., 2018]	waveform	classical music	-
	Hierarchical waveNet	[Dieleman et al., 2018]	waveform	piano music	-
	SampleRNN	[Mehri et al., 2017]	waveform	speech, piano music	-
	MelNet	[Vasquez and Lewis, 2019]	mag. spec.	speech, piano music	speaker ID text
	wavenetAE	[Engel et al., 2017]	waveform	tonal sounds	pitch
	sparse Transformer	[Child et al., 2019]	waveform	piano music	-

Variational AutoEncoders

G. Richard

Schematic principle of Variational Autoencoders (VAEs)



The encoder $q_\phi(\mathbf{z}|\mathbf{x})$ approximates the true posterior distribution $q_\theta(z|x)$
The decoder $p_\theta(x|z)$ generates an approximation $\hat{x}$ from the encoding

Main idea of variational inference: :
- The complete model $p(\mathbf{x},\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, but the data follows complex distributions

- Exploit an approximate of the true posterior: $q_\phi(\mathbf{z}|\mathbf{x})$
- Variational inference: minimizing the difference between the approximation and the true density:

$$q^*_{\phi(\mathbf{z}|\mathbf{x})} = argmin_{q_\phi(\mathbf{z}|\mathbf{x})\in\mathcal{Q}} D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z}|\mathbf{x})]$$

G. Richard

Variational AutoEncoders

$$q^*_{\phi(\mathbf{z}|\mathbf{x})} = argmin_{q_\phi(\mathbf{z}|\mathbf{x}) \in \mathcal{Q}} D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{x})|p(\mathbf{z}|\mathbf{x})]$$

- This can be further expressed as :

$$\log p_\theta(\mathbf{x}) = D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\phi, \theta, \mathbf{x})$$

- It describes the quantity to model $\log p_\theta(\mathbf{x})$ minus the error we make by using an approximate q instead of the true p.

- We can maximize the **Evidenced Lower Bound (ELBO)**

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = -D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big(\log p_\theta(\mathbf{x}|\mathbf{z})\big)$$

Kingma et Welling, « An Introduction to Variational Autoencoders », Foundations and Trends in Machine Learning, vol. 12, nº 4, 2019, p. 307–392
K. Sachdeva: "Evidence Lower Bound (ELBO) - CLEARLY EXPLAINED!
https://www.youtube.com/watch?v=IXsA5Rpp25w

Variational AutoEncoders in Audio/music

Many examples

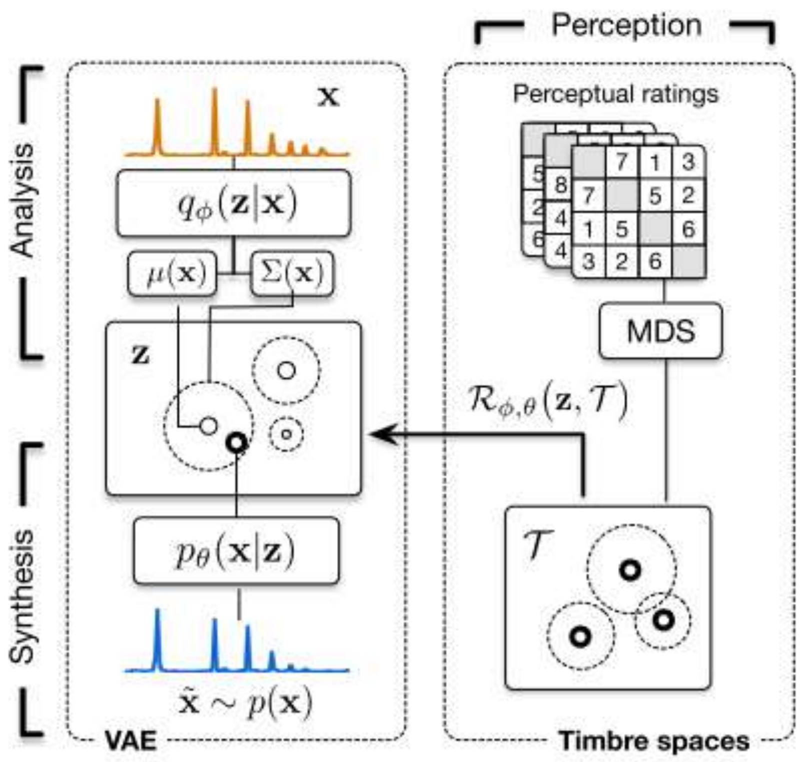
Arch.	Name		Audio representation	Data	Conditioning
VAEs	Planet Drums	[Aouameur et al., 2019]	Mel-scaled mag. spec.	drums	instrument ID
	Jukebox	[Dhariwal et al., 2020]	waveform	music	artist & genre ID lyrics
	NOTONO	[Bazin et al., 2020]	mag. & IF	tonal instruments	pitch
	FlowSynth	[Esling et al., 2019]	mag.	synth. sounds	semantic tags
	Neural Granular Sound Synth.	[Bitton et al., 2020]	waveform	orchestral drums animals	pitch instrument ID

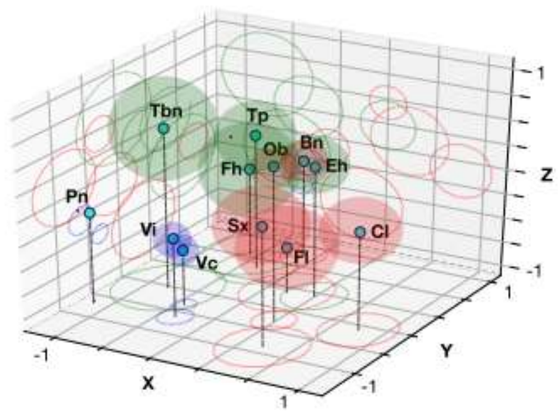
J. Nistal, "Exploring Generative Adversarial Networks for Controllable Musical Audio Synthesis, PhD Thesis, IP Paris, 2022

G. Richard

Variational AutoEncoders in Audio/music

Regularizing the latent space with timbre spaces (perception)





Multi-dimensional scaling (MDS)

P. Esling, A. Chemla-Romeu-Santos, A. Bitton, Bridging Audio Analysis, Perception and Synthesis with Perceptually-regularized Variational Timbre Spaces, in Proc. of ISMIR » 2018 »

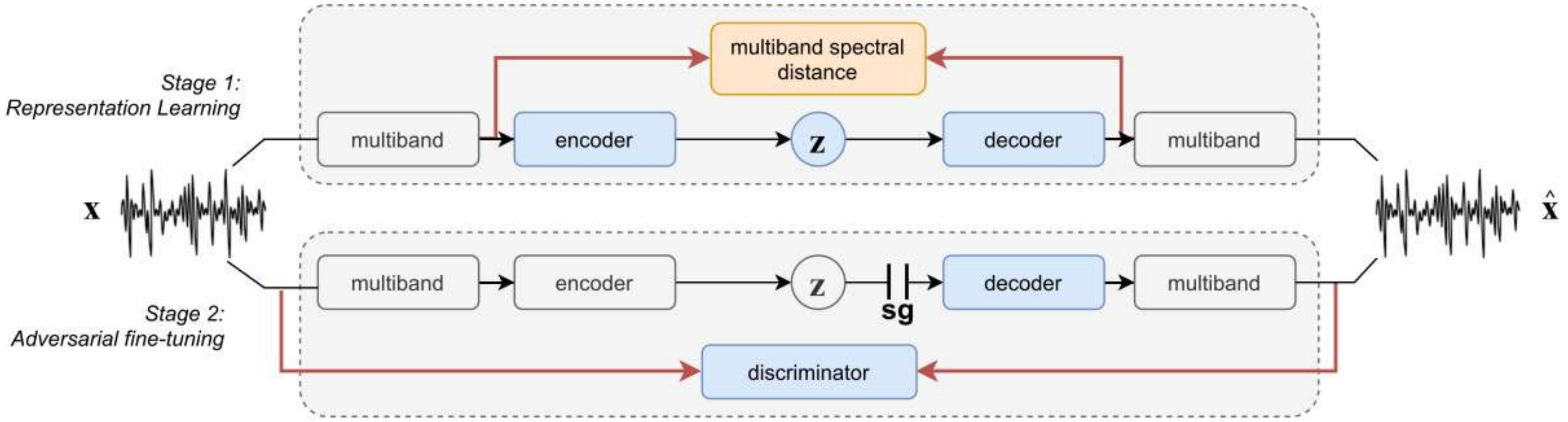
Variational AutoEncoders in Audio/music

Extensions

RAVE: Realtime Audio Variational autoEncoder

- Based on a two stage training:

1. representation learning with VAEs (stage 1)

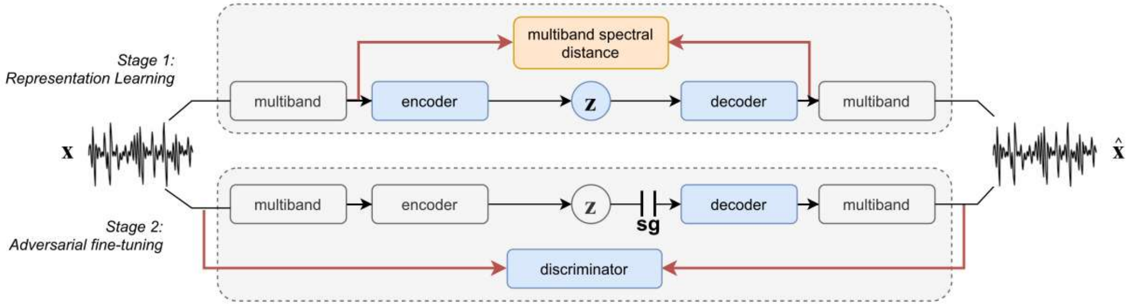


A. Caillon, Antoine; P. Esling. "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis." ArXiv abs/2111.05011 (2021)

G. Richard

Variational AutoEncoders in Audio/music

RAVE : some details



- The multispectral loss (from Engel2019 (DDSP))

$$S(\mathbf{x}, \mathbf{y}) = \sum_{n \in \mathcal{N}} \left[\frac{\|\mathrm{STFT}_n(\mathbf{x}) - \mathrm{STFT}_n(\mathbf{y})\|_F}{\|\mathrm{STFT}_n(\mathbf{x})\|_F} + \log\left(\|\mathrm{STFT}_n(\mathbf{x}) - \mathrm{STFT}_n(\mathbf{y})\|_1\right) \right]$$

- Latent representation compactness
  - To avoid *posterior* collapse (e.g situation where **the learned latent space is ignored)**
  - Based on variance normalisation, rank estimation (using SVD on the latent space)

A. Caillon, Antoine; P. Esling. "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis." ArXiv abs/2111.05011 (2021)
J. Engel & al., "DDSP: Differentiable Digital Signal Processing," in Int. Conf. on Learning Representations (ICLR), 2020.

Variational AutoEncoders in Audio/music

RAVE : some results

- Evaluation (in 2021)

Model	MOS	95% CI	Training time	Parameter count
Ground truth	4.21	±0.04	-	-
NSynth	2.68	±0.04	~ 13 days	64.7M
SING	1.15	±0.02	**~ 5 days**	80.8M
RAVE (Ours)	**3.01**	**±0.05**	~ 7 days	**17.6M**

- Synthesis examples:
  - Timbre transfer (model trained on speech, input :violin)

    Violin input

    Output (speech)

  - Darbouka synthesis:
    - Reconstruction

      Original

      Reconstructed

    - Unconditional generation

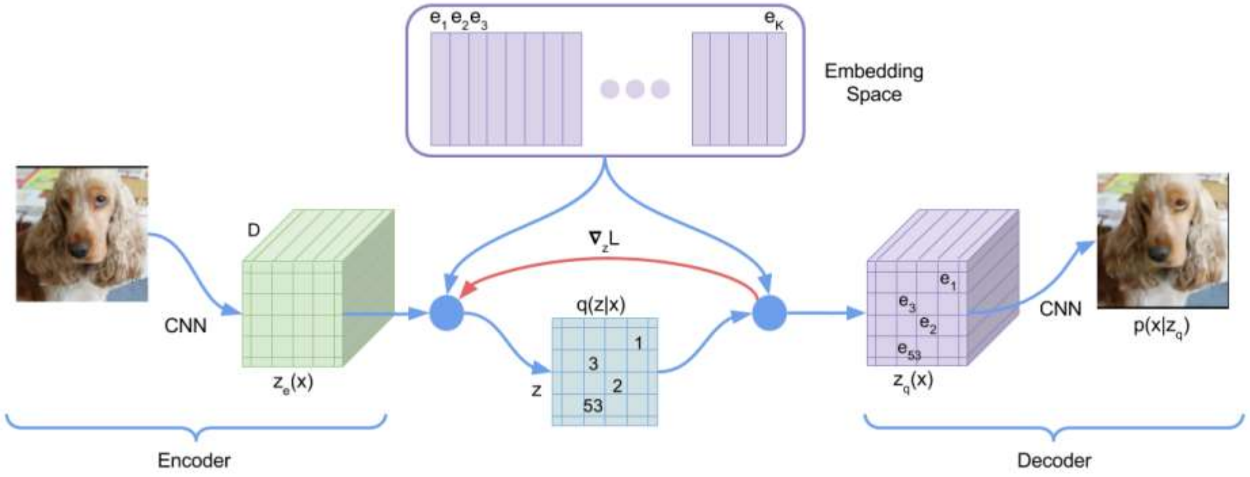
      unconditioned

A. Caillon, Antoine; P. Esling. "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis." ArXiv abs/2111.05011 (2021)

G. Richard

Vector-Quantized Variational AutoEncoders (VQ-VAEs)

- Combines VAEs with Vector quantization
- Helps to avoid *posterior* collapse of VAEs
- Offers the flexibility of a **discrete** neural representation
- Main principle



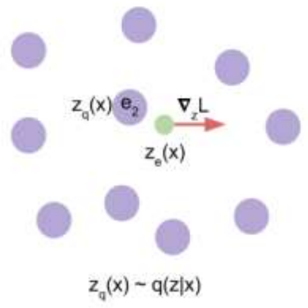
A.van den Oord, O. Vinyals, K. Kavukcuoglu.. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017

Vector-Quantized Variational AutoEncoders (VQ-VAEs)

- Discrete latent representation
  - The discrete latent variables are obtained by nearest neighbour look-up

$$q(z=k|x) = \begin{cases} 1 & \text{for } \mathrm{k} = \mathrm{argmin}_j \|z_e(x) - e_j\|_2, \\ 0 & \text{otherwise} \end{cases}$$

$$z_q(x) = e_k, \quad \text{where} \quad k = \mathrm{argmin}_j \|z_e(x) - e_j\|_2$$



A.van den Oord, O. Vinyals, K. Kavukcuoglu.. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017

G. Richard

Vector-Quantized Variational AutoEncoders (VQ-VAEs)

- Learning
  - A loss function with three components
    1. A reconstruction loss (or data term)
    2. A dictionary learning term (VQ):
    3. A commitment loss (to force a joint learning of encoder and dictionary)

$$L = \log p(x|z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2$$

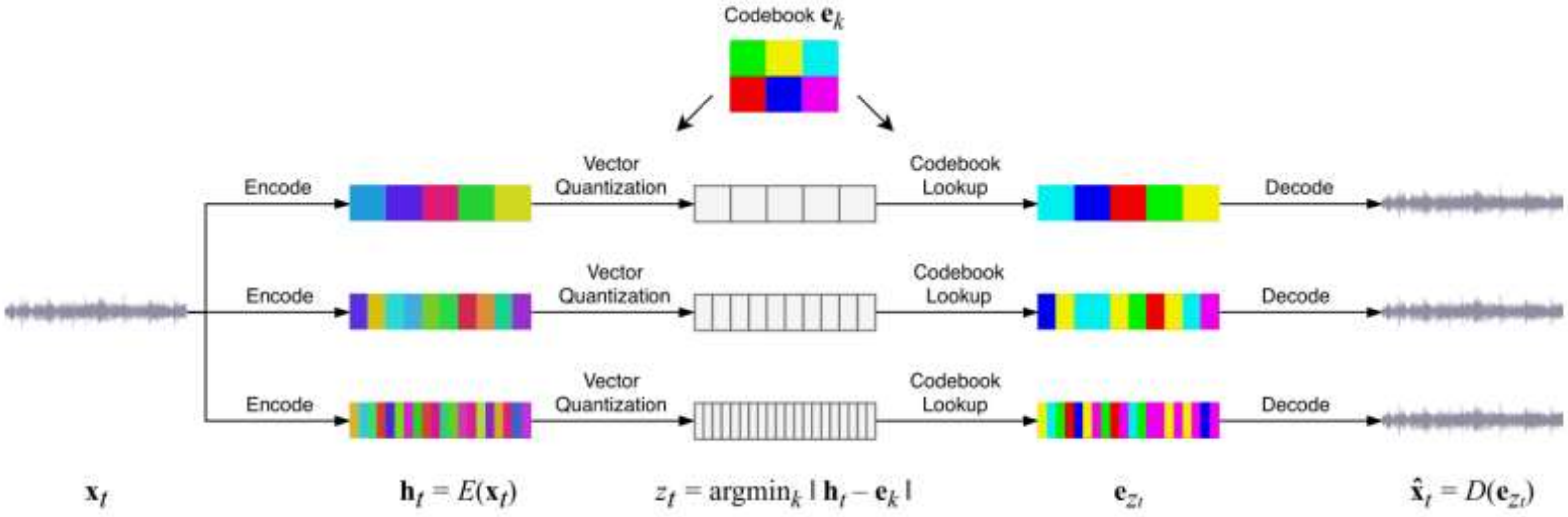
A.van den Oord, O. Vinyals, K. Kavukcuoglu.. Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017

VQ-VAEs in Audio and Music

An example with Jukebox

G. Richard

- Based on hierarchical VQ-VAE (VQ-VAE2), trained with an additional spectral loss
- Combined with sparse transformers for learning the latent prior for generation



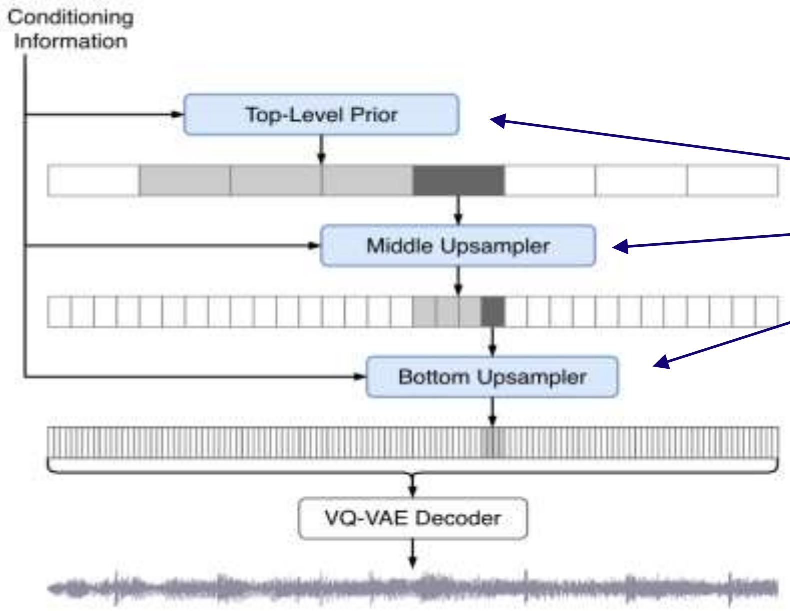
Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In Advances in Neural Information Processing Systems, 2019.
Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 , 2019.
P. Dhariwal & al. “Jukebox: A Generative Model for Music”, arXiv:2005.00341

VQ-VAEs in Audio and Music

An example with Jukebox

- Learning the latent prior once the separate VQ-VAEs are trained

$$p(\mathbf{z}) = p(\mathbf{z}^{\text{top}}, \mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{bottom}})$$
$$= p(\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{middle}}|\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{bottom}}|\mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{top}})$$



Sparse transformers

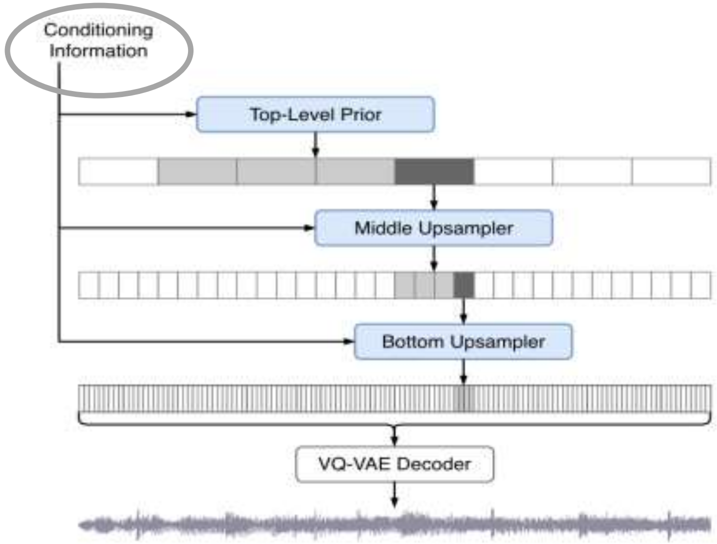
P. Dhariwal & al. "Jukebox: A Generative Model for Music", arXiv:2005.00341

VQ-VAEs in Audio and Music

An example with Jukebox

G. Richard

- Conditioning for controlling the synthesis



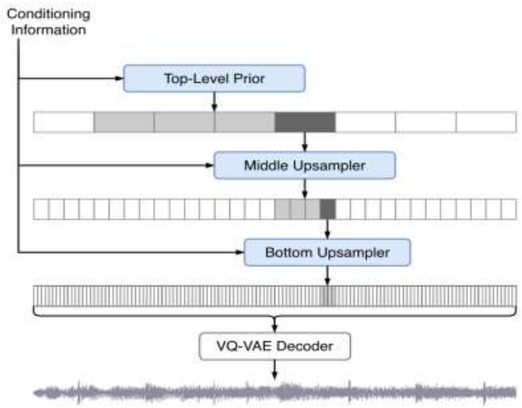
- **Artist, Genre, and Timing Conditioning** (to allow to learn patterns that depend on the structure… such as applause at the end)
- **Lyrics Conditioning (with necessity to learn lyrics/audio alignment)**

P. Dhariwal & al. "Jukebox: A Generative Model for Music", arXiv:2005.00341

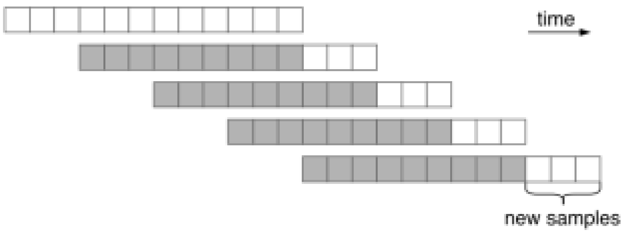
VQ-VAEs in Audio and Music

An example with Jukebox

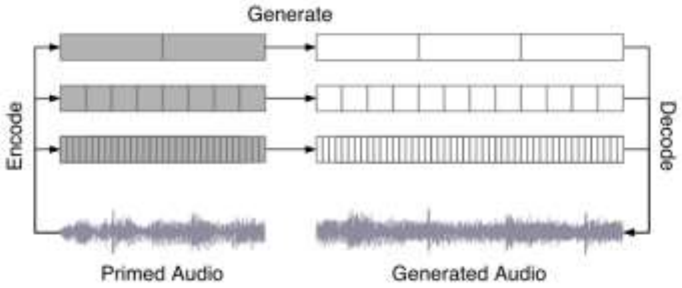
- Sampling methods for generating music



Windowed sampling for modelling sequences longer than initial context



Primed sampling: generate continuations by converting input into the VQ-VAE codes and sampling the subsequent codes in each level.



P. Dhariwal & al. "Jukebox: A Generative Model for Music", arXiv:2005.00341

VQ-VAEs in Audio and Music

An example with Jukebox

- Sound examples
  - Completion (with context of 12s of existing songs in the training)
  - Re-renditions (using pairs of lyrics-artist existing in the training)
  - Generation with novel lyrics (generated by GPT-2)
  - Generation with novel voices (by interpolating existing voice embeddings)

  - Many raw examples at https://jukebox.openai.com/

  - Some curated examples at https://openai.com/blog/jukebox/
    - One example of continuation with unknown lyrics: https://jukebox.openai.com/?song=795460096

- Original model is rather slow at sampling (9 hours to render 1’ of music)

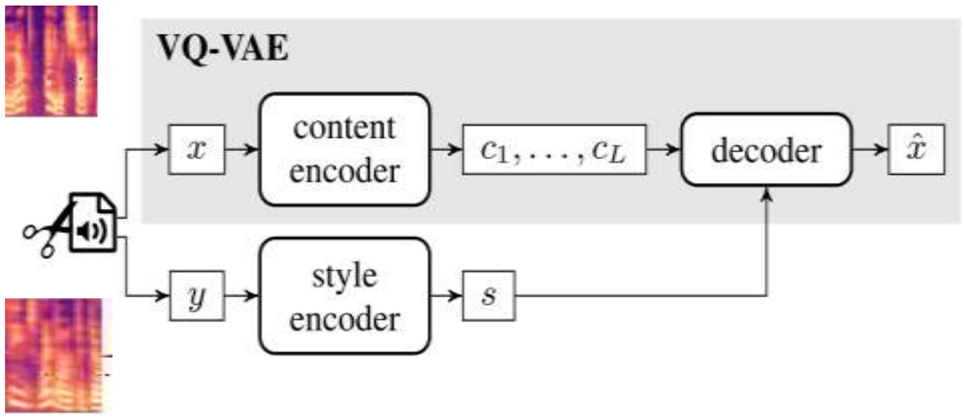
P. Dhariwal & al. “Jukebox: A Generative Model for Music”, arXiv:2005.00341

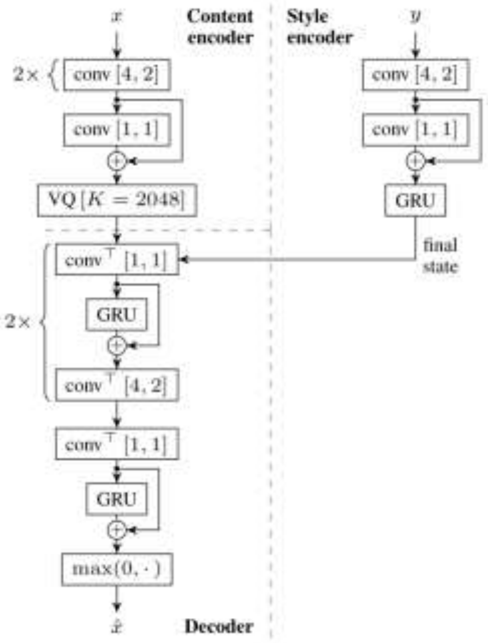
G. Richard

VQ-VAEs in Audio and Music

Another example for one-shot music style transfer

- Content is encoded using a VQ-VAE
- Style is encoded using a self-supervised strategy (*y is an audio-augmented version of a different segment than x, taken from the same recording*)





Ondřej Cífka, Alexey Ozerov, Umut Şimşekli and Gaël Richard. "Self-Supervised VQ-VAE for One-Shot Music Style Transfer." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

VQ-VAEs in Audio and Music

Another example for one-shot music style transfer

- Many sound examples at: https://adasp.telecom-paris.fr/rc/demos_companion-pages/cifka-ss-vq-vae/#examples

- Two examples
  1. Synthetic example

     Content input *Style input* *Target* *Output (VQ-VAE)*

  2. Real example

     Content input *Style input* *Output (VQ-VAE)*

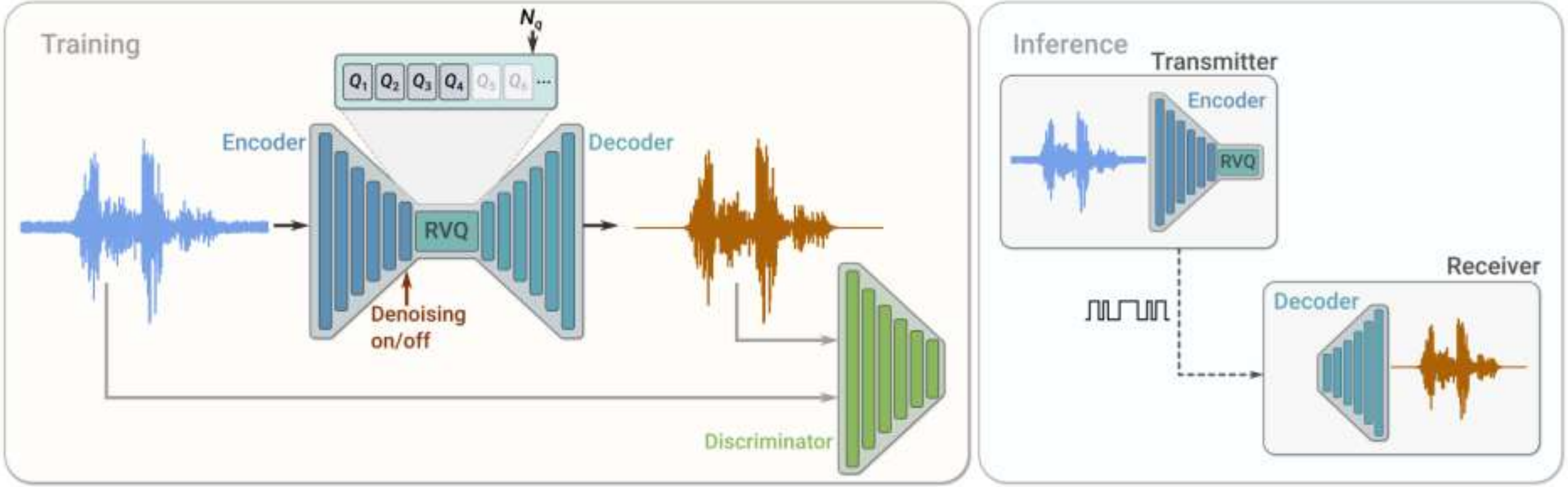
Ondřej Cífka, Alexey Ozerov, Umut Şimşekli and Gaël Richard. "Self-Supervised VQ-VAE for One-Shot Music Style Transfer." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

G. Richard

Discrete neural representation:

Soundstream: another powerfull generative model

- Designed for audio compression
- Exploits Residual Vector Quantification (RVQ)
- trained end-to-end together with a discriminator using the mix of adversarial and reconstruction losses



Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. IEEE ACM Trans. Audio Speech Lang. Process., 30, 2022

G. Richard

Discrete neural representation:

Soundstream: an other powerfull generative model

- Interest of RVQ

A concrete example with regular VQ :

- a codec with a target bitrate R = 6000 bps.
- For an audio at Fs = 24000 Hz (*striding factor of M = 32*), each second of audio is represented by S = 75 frames
- This leads to r = 6000/75 = 80 bits allocated to each frame.
- Using a plain vector quantizer, this requires storing a codebook with N= 2^80 vectors (this is Huge !!)

- RVQ = multi-stage Vector quantizer

- Cascade N_q layers of VQ
- Total rate budget is uniformly allocated to each VQ,
- $Ri = r/ N_q = \log_2 (N)$.
- Example: with $N_q = 8$, each quantizer uses a codebook of size $N = 2^{(r/ Nq)} = 2^{(80/8)} = 1024$.

Algorithm 1: Residual Vector Quantization

Input: $y = \mathrm{enc}(x)$ the output of the encoder, vector quantizers Q_i for $i = 1..N_q$
Output: the quantized $\hat{y}$
$\hat{y} \leftarrow 0.0$
residual $\leftarrow y$
for $i = 1$ *to* N_q **do**
  $\hat{y}$ += Q_i(residual)
  residual −= Q_i(residual)
return $\hat{y}$

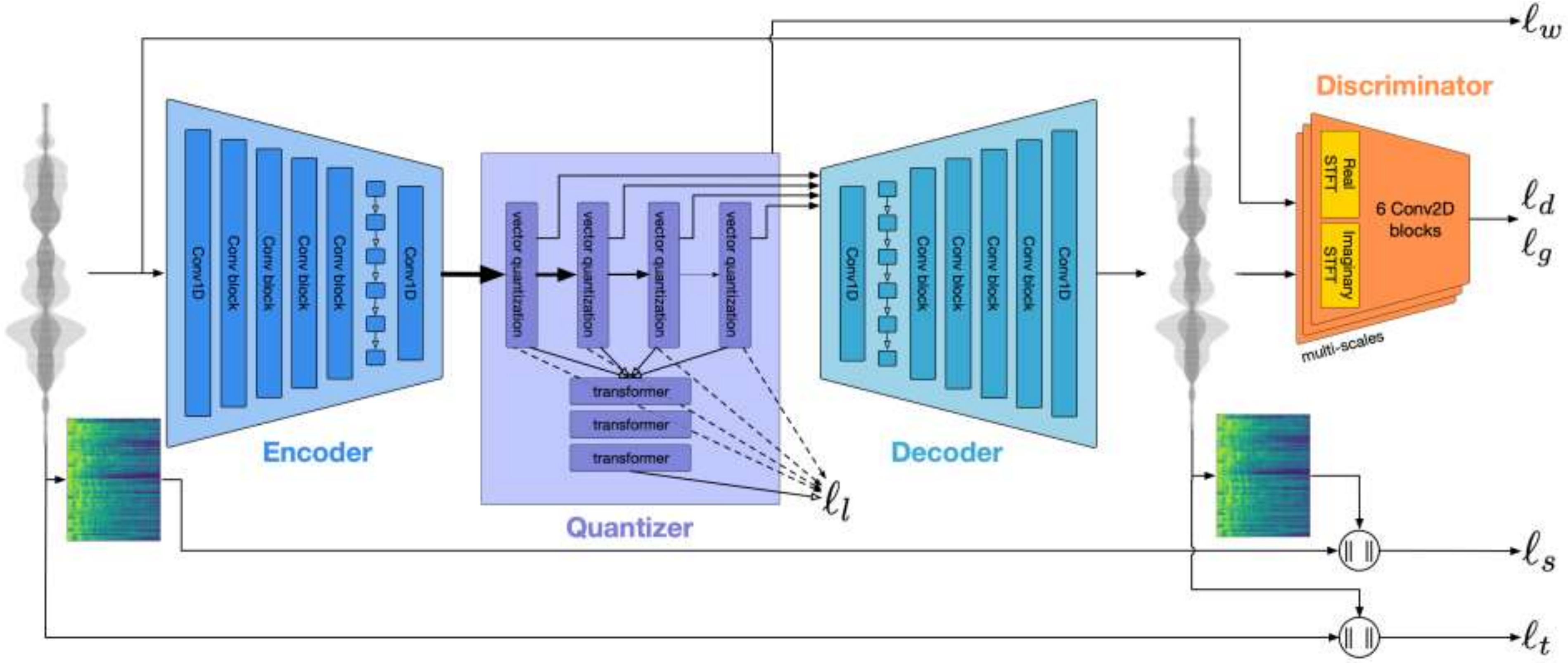
Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. IEEE ACM Trans. Audio Speech Lang. Process., 30, 2022

Discrete neural representation:

EncoDec: a slight extension of Soundstream

G. Richard

- E.g. Use of a small transformer model for better multi-stage VQ



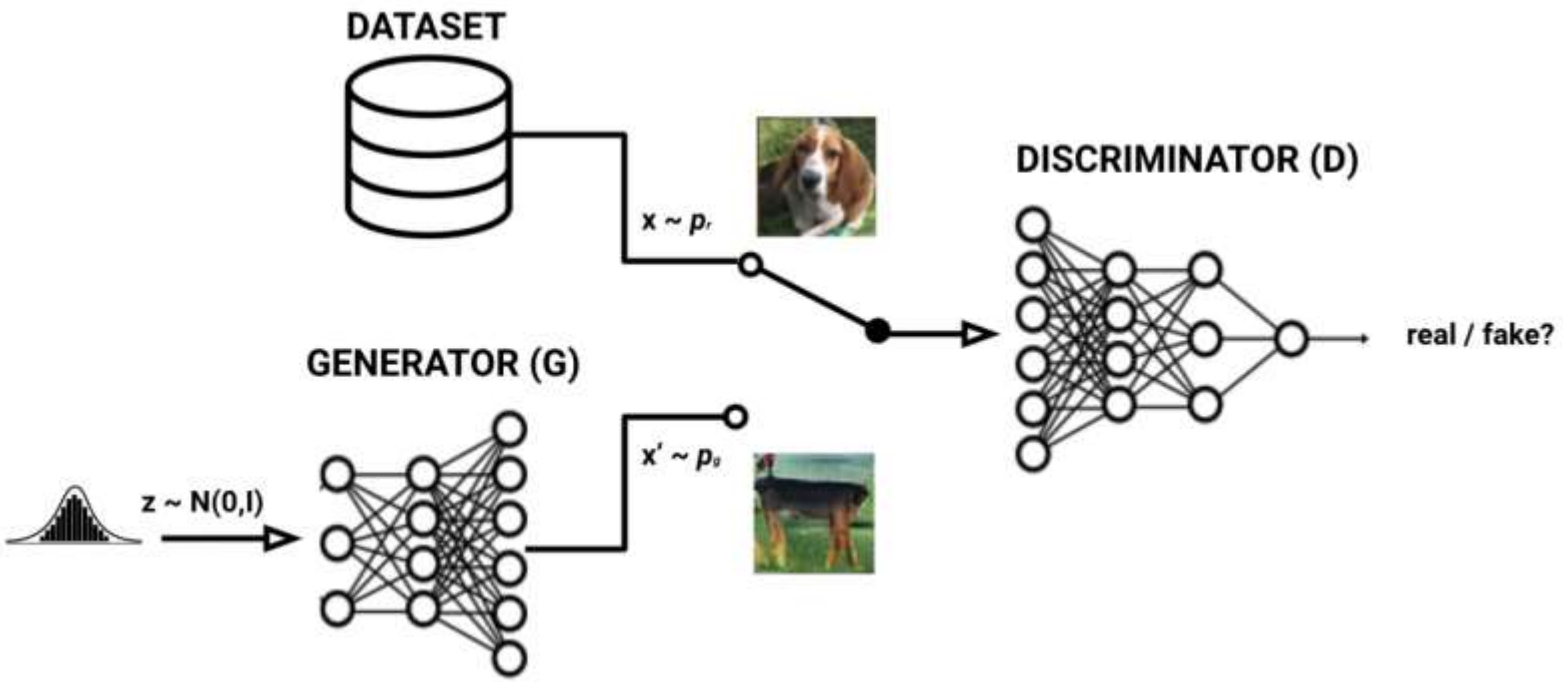
Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022.

GANS, Diffusion models for audio generation

Generative Adversarial Networks (GANs)

G. Richard

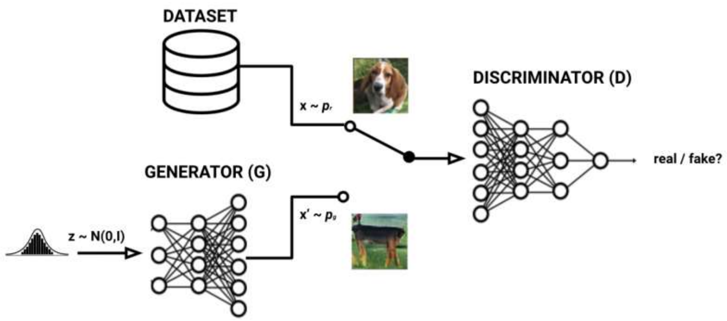
- Principle of GANs



Goodfellow, I. et al., 2014. Generative adversarial nets. In Advances in neural information processing systems.
Figure from J. Nistal, "Exploring generative Adversarial networks for controllable musical audio synthesis, PhD thesis, IP Paris, 2022

Generative Adversarial Networks (GANs)

G. Richard



- More formally
  - a generative network $G_\theta(\mathbf{z})$ that outputs $x_g \sim p_g$ from a random input $\mathbf{z}$
    After training, the output should follow the targeted probability distribution p_r
  - a discriminative network $D_\beta(\mathbf{x})$ trained to predict if the input comes from the real p_r or from the generated distribution p_g

  - Optimization problem: a competitive objective

$$\min_{G_{\boldsymbol{\theta}}} \max_{D_{\boldsymbol{\beta}}} V(D_{\boldsymbol{\beta}}, G_{\boldsymbol{\theta}}) = E_{\boldsymbol{x}\sim p_r}[\log D_{\boldsymbol{\beta}}(\boldsymbol{x})] + E_{\boldsymbol{x}\sim p_g}[1 - \log D_{\boldsymbol{\beta}}(G_{\boldsymbol{\theta}}(\boldsymbol{z}))]$$

Goodfellow, I. et al., 2014. Generative adversarial nets. In Advances in neural information processing systems.

G. Richard

Generative Adversarial Networks (GANs)

- Principle of conditional GANs for audio synthesis

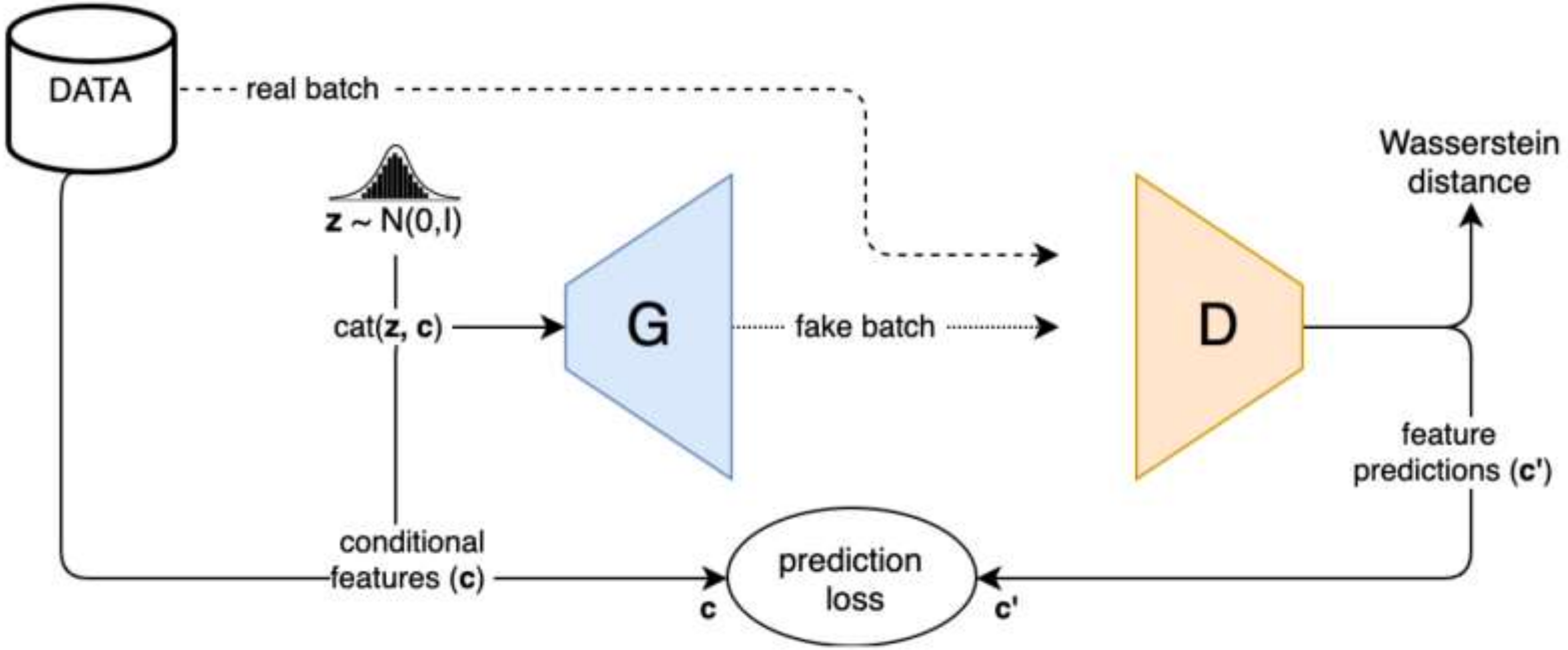
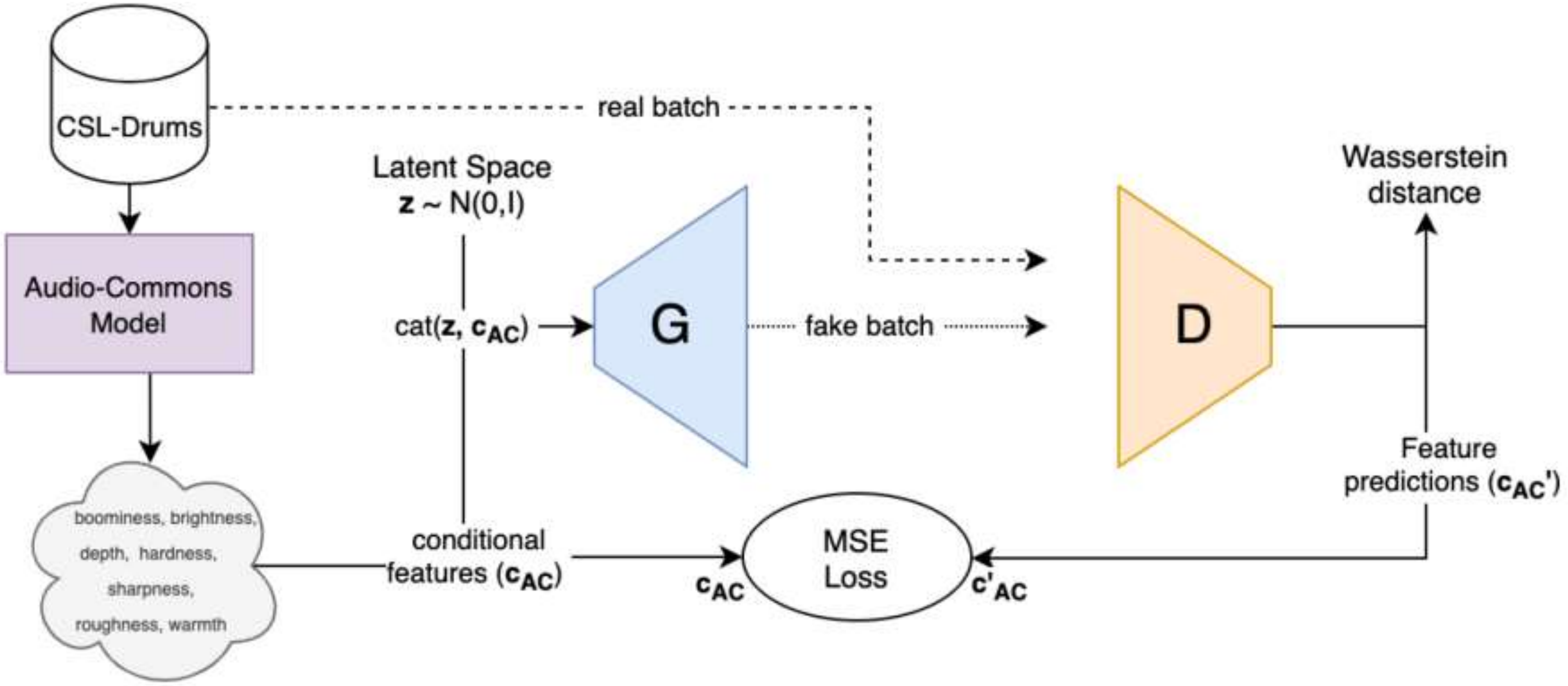


Figure from J. Nistal, "Exploring generative Adversarial networks for controllable musical audio synthesis, PhD thesis, IP Paris, 2022

Generative Adversarial Networks (GANs)

- DrumGAN: Synthesis of Drum sounds with timbral feature Conditioning using GANs synthesis



Nistal, J., Lattner, S., and, Richard, G. , "DrumGAN: Synthesis of Drum Sounds with Perceptual Feature Conditioning using GANs," in Proceedings of the 28th International Society for Music Information Retrieval, ISMIR , 2020.

Generative Adversarial Networks (GANs)

- DrumGAN: Demo
  - https://sites.google.com/view/drumgan?pli=1

- DrumGAN VST: A Plugin for Drum Sound Analysis/Synthesis with Autoencoding GANs
  - https://cslmusicteam.sony.fr/drumgan-vst/
  - Short demo on **Converting beatbox to drums**

Original

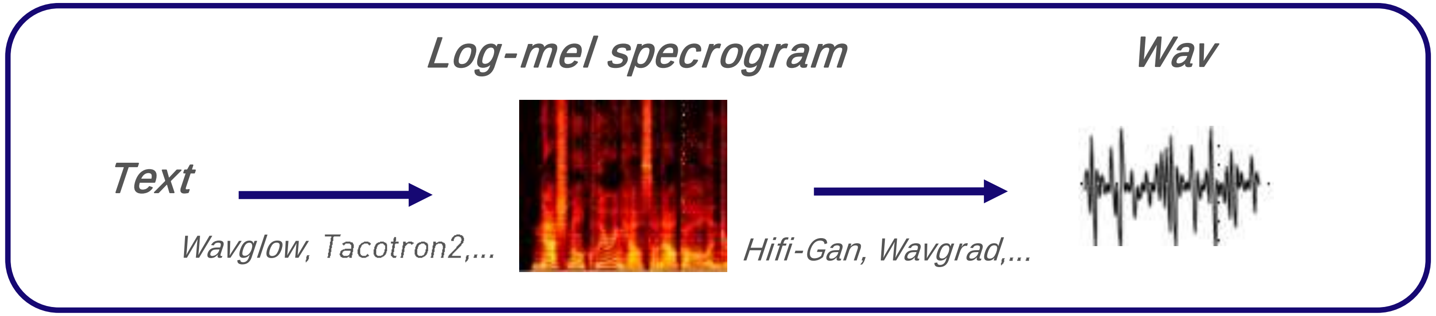
Original+decoded

Nistal, J., Lattner, S., and, Richard, G. , "DrumGAN: Synthesis of Drum Sounds with Perceptual Feature Conditioning using GANs," in Proceedings of the 28th International Society for Music Information Retrieval, ISMIR , 2020.

A classic pipeline for sound generation

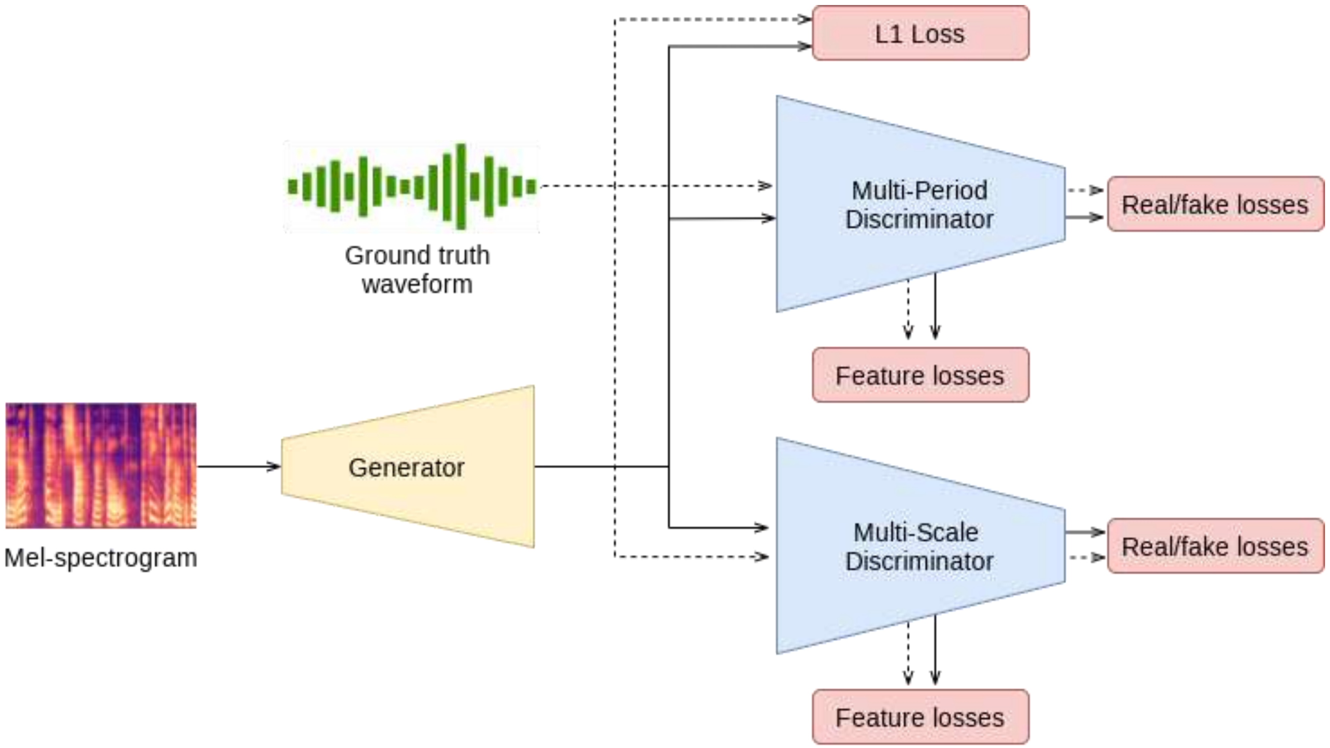
G. Richard

- For example, a classic pipeline in recent Text-to-speech



Hifi-Gan

- High computational efficiency and high sample quality
- 1 Generator (CNN) and 2 Discriminators

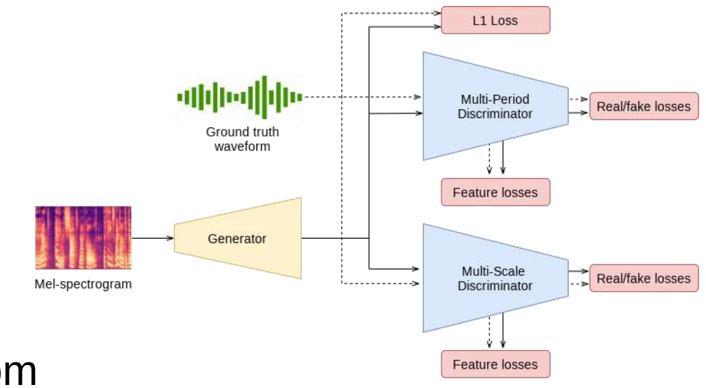


Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi -gan: Generative adversarial networks for efficient and high fi delity speech synthesis. Advances in Neural Information Processing Systems, 33:17022–17033, 2020
Demo at https://jik876.github.io/hifi-gan-demo/

Hifi-Gan

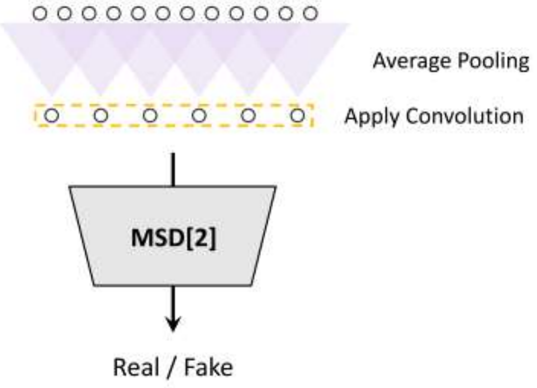
2 Discriminators: MSD and MPD

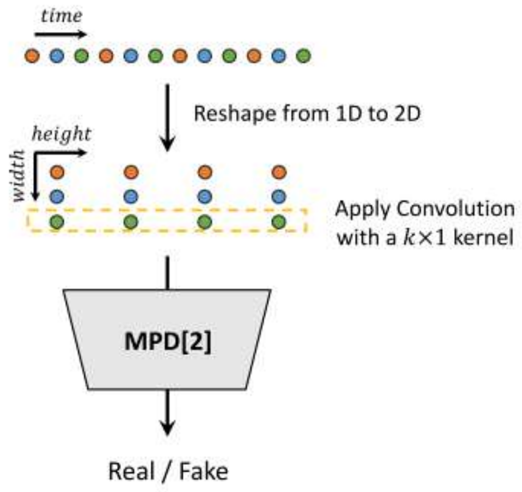
G. Richard



- **MPD = mixture of sub-discriminators**
- Sub-discriminators are designed to capture different implicit structures from each other by looking at different parts of an input audio
- each sub-discriminator only accepts equally spaced samples of an input audio
- the space (period) p is equal to [2, 3, 5, 7, 11]

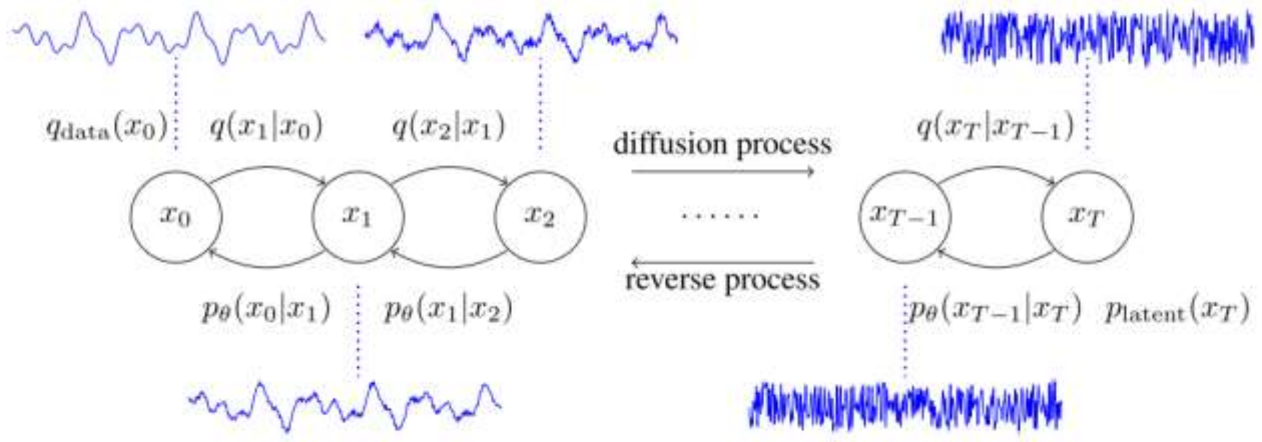
- **MSD to evaluate audio sequence at multiple scale**
- MSD is a mixture of three sub-discriminators operating on different input scales: raw audio, ×2 average-pooled audio, and ×4 average-pooled audio





Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In Advances in Neural Information Processing Systems 32, pages 14910–14921, 2019.
Mikołaj Binkowski, 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.

Diffusion models for audio synthesis …



- Based on two processes: the diffusion process, and the reverse process
  - The **diffusion process** is defined by a fixed Markov chain from data x_0 to the latent variable x_T

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

where each of $q(x_t|x_{t-1})$ is fixed to $\mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ for a small positive constant β_t

- The reverse process gradually converts the white noise signal into audio waveform through a Markov chain: .

$$p_{\text{latent}}(x_T) = \mathcal{N}(0, I), \text{ and } p_\theta(x_0, \cdots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t),$$

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). DiffWave: A Versatile Diffusion Model for Audio Synthesis. ArXiv, abs/2009.09761.

Diffusion models for audio synthesis …

- The models are often strongly conditioned
  - Example: wavgrad, specgrad conditioned on mel-spectrogram

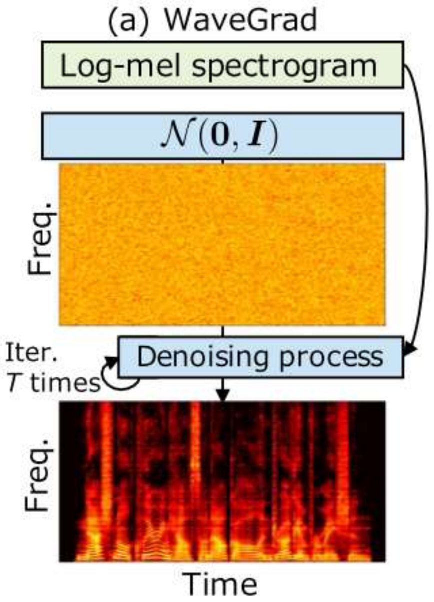
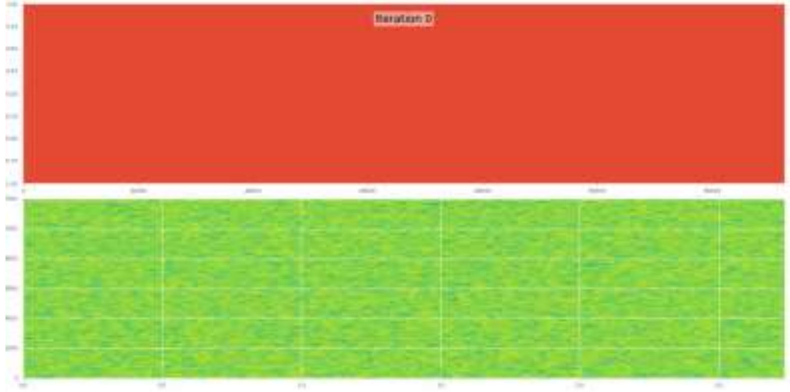


Illustration of the diffusion process (50 iterations)



Sound examples

reference

wavgrad

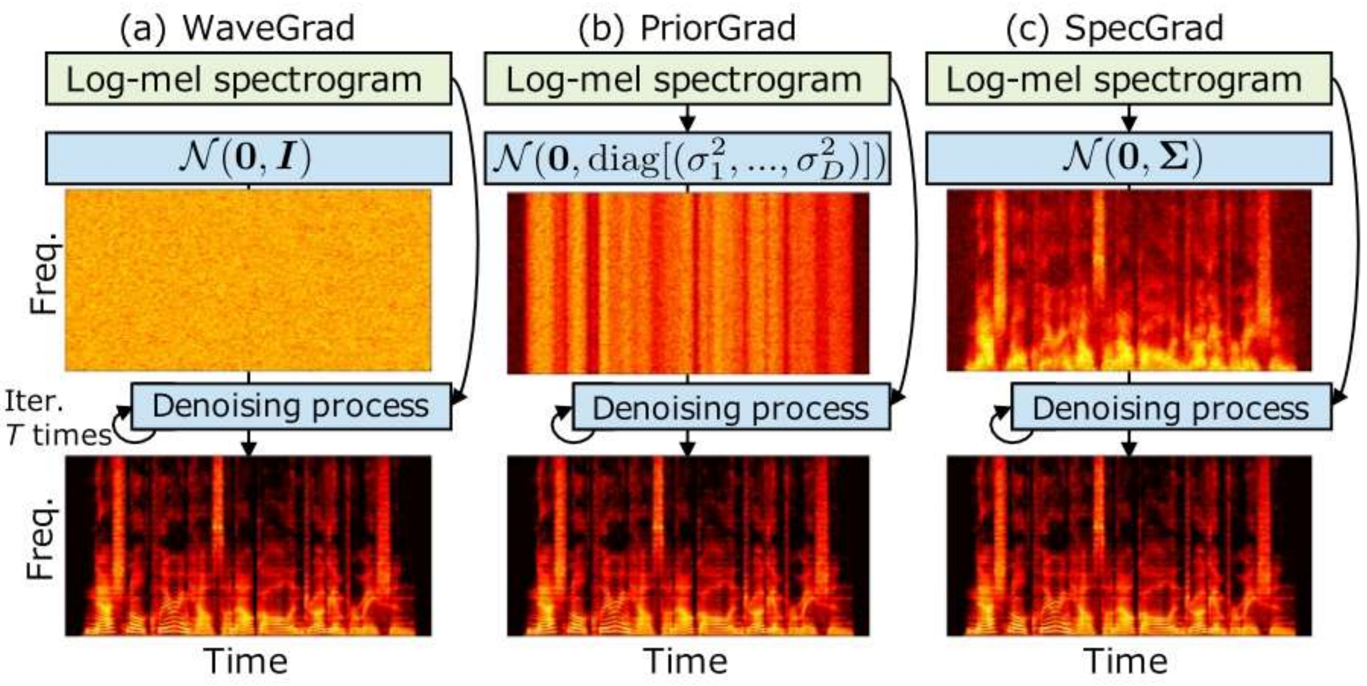
N. Chen & al.“WaveGrad: Estimating gradients for waveform generation,” in Proc. ICLR, 2021.
Koizumi, Yuma et al. “SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping.” Interspeech (2022).

Diffusion models for audio synthesis …

extensions of wavgrad

G. Richard

- The example of priorgrad, specgrad, …



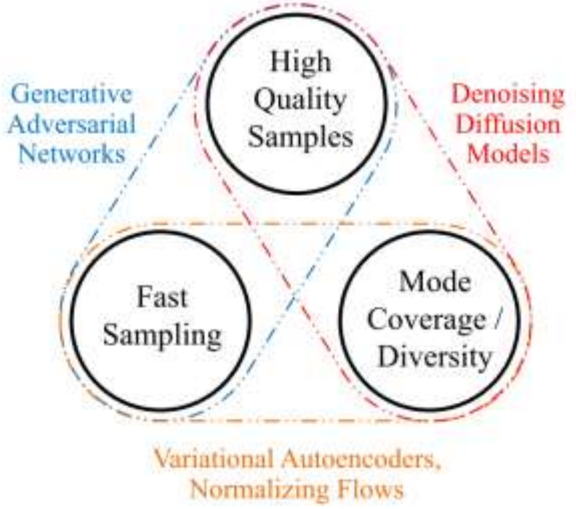
Koizumi, Yuma et al. "SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping." Interspeech (2022).

Diffusion models for audio synthesis …

Combining diffusion models with GANs

G. Richard

The Generative learning trilemma



- **Example of Denoising Diffusion Gan:**
    - Assumption: the slow sampling of diffusion models is due to the Gaussian assumption in the denoising distribution
    - Propose to employ complex, multimodal denoising distributions.
    - Propose denoising diffusion GANs, a diffusion model whose reverse process is parametrized by conditional GANs.
- Denoising diffusion GANs achieve several orders of magnitude speed-up compared to classic diffusion models for (image) generation

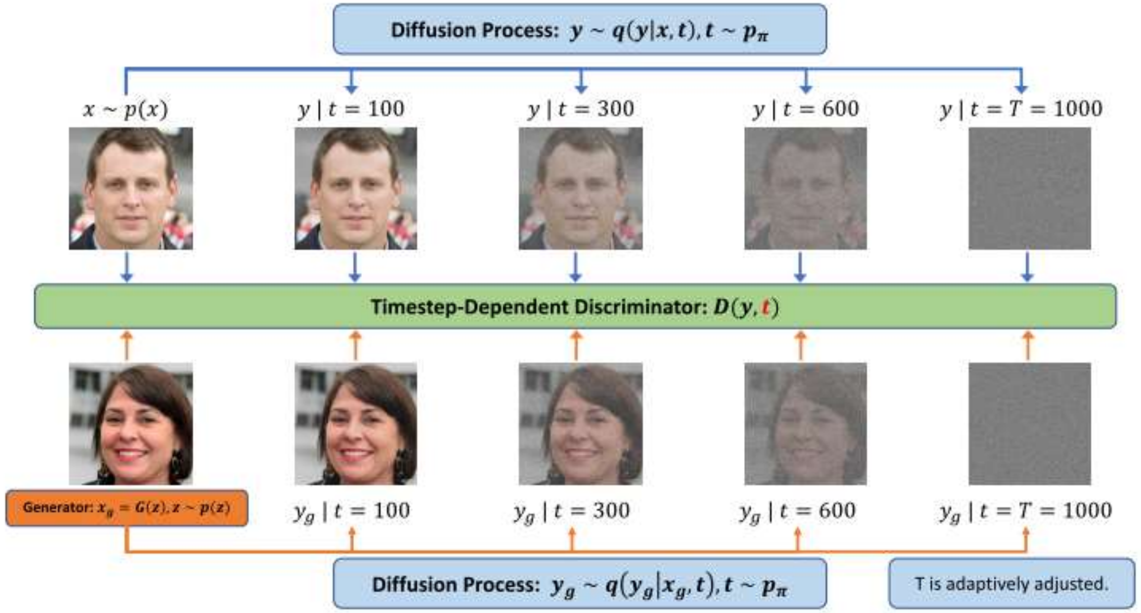
Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804, 2021

Diffusion models for audio synthesis …

Combining diffusion models with GANs

Diffusion-Gan: Training GANS with Diffusion

- The discriminator learns to distinguish a diffused real image from a diffused fake image at all diffusion steps.
  - Stabilizes the training of GANS ; Leads to improved performances (quality of images, complexity)



Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with diffusion," in Proc. ICLR, 2023

Diffusion models for audio synthesis …

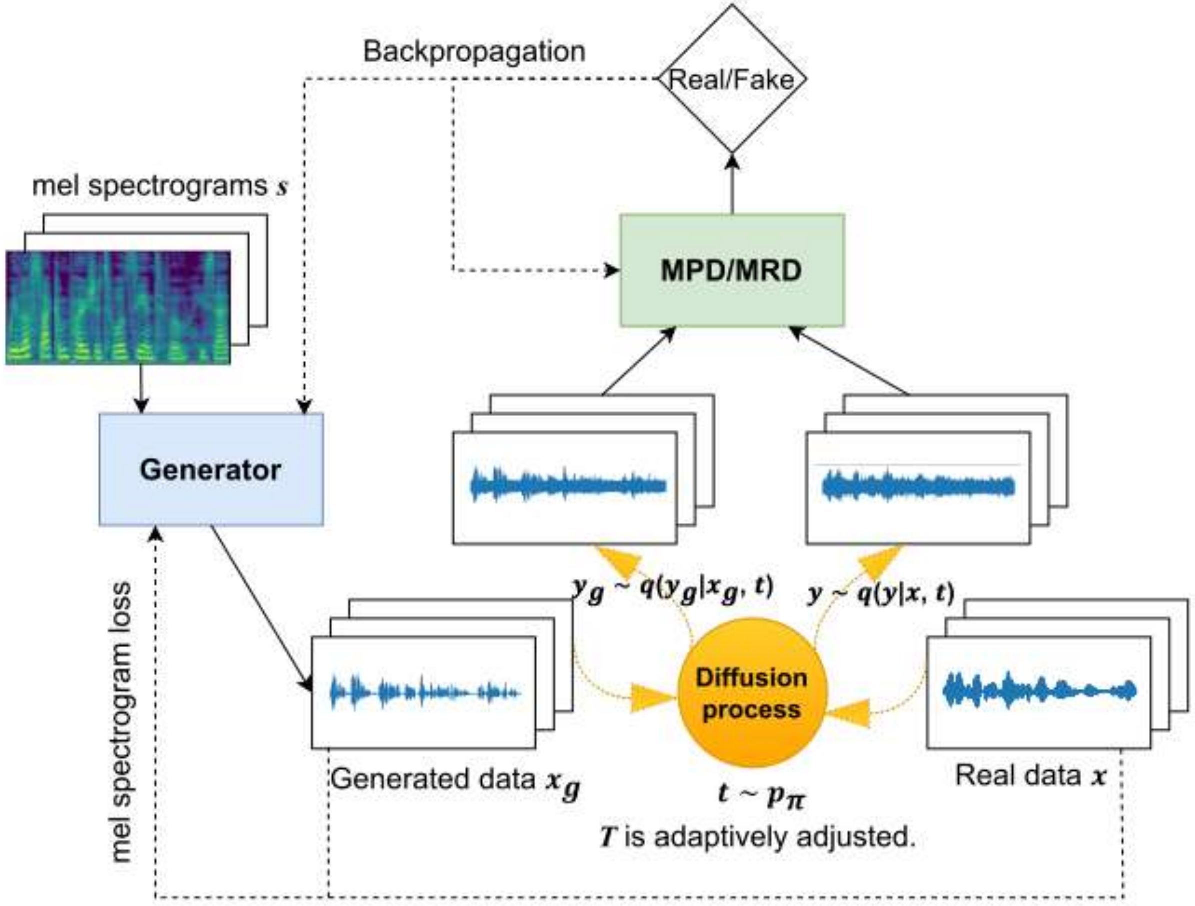
Combining diffusion models with GANs

G. Richard

SpecDiff-Gan

Combines principles of

- Diffusion-gans,
- Hifi-Gan
- and specgrad
- …for *speech and music*



Ground truth	SpecDiff-Gan
speech	
piano	
drums	

T. Baoueb, H. Liu, M. Fontaine, J. Le Roux, G. Richard , SpecDiff-GAN: A Spectrally-Shaped Noise Diffusion GAN for Speech and Music Synthesis, ICASSP 2024
Demo at: https://specdiff-gan.github.io/

Cross Model audio generation : some examples

Towards « text-prompt » to audio

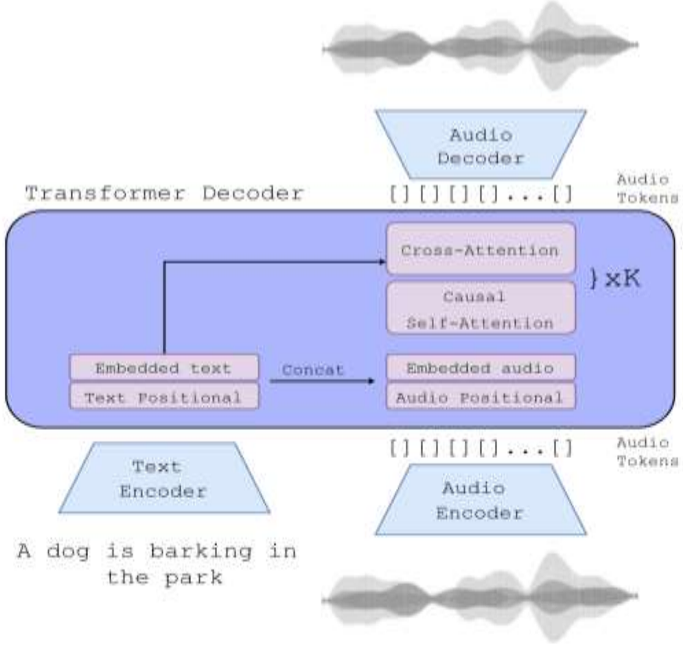
The exemple of AudioGen

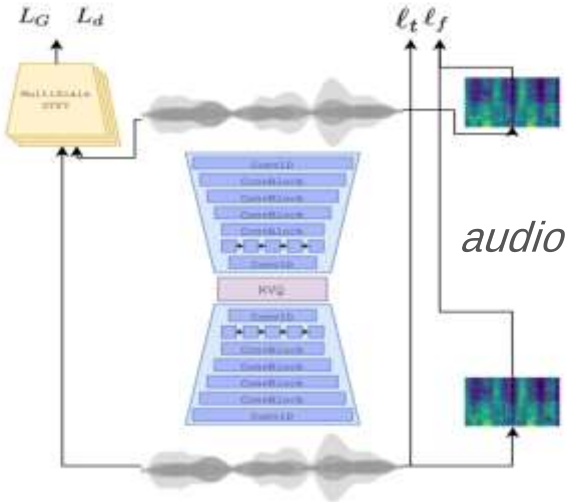
G. Richard

- AudioGen

- 2 main steps:
  - (i) an audio encoder-decoder to learn a discrete audio representation (RVQ)
  - (ii) training a Transformer language model over the learnt codes obtained from the audio encoder, conditioned on textual features.

- **Some specifities:**
  - Text representation obtained using a pretrained T5 text encoder
  - For text adherence: cross-attention between audio and text to each attention block of the transformer.
  - Augmentation method that fuses pairs of audio samples and their respective text captions, thus creating new concept compositions during training
  - Uses Classifier Free Guidance (CFG) to improve generation for Low resolution (e.g. randomly unconditional training)





audio encoder-decoder

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, DeviParikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. arXivpreprint arXiv:2209.15352, 2022a.

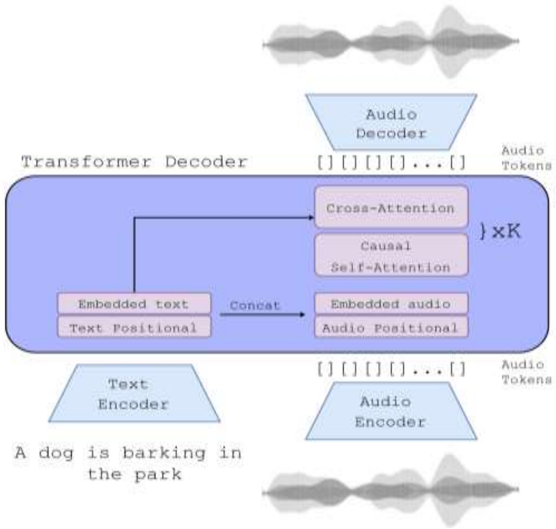
G. Richard

Towards « text-prompt » to audio

The exemple of AudioGen

- Demo

Text prompt
a man speaks as birds chirp and dogs bark
male speech with horns honking in the background
drums and music playing with a man speaking

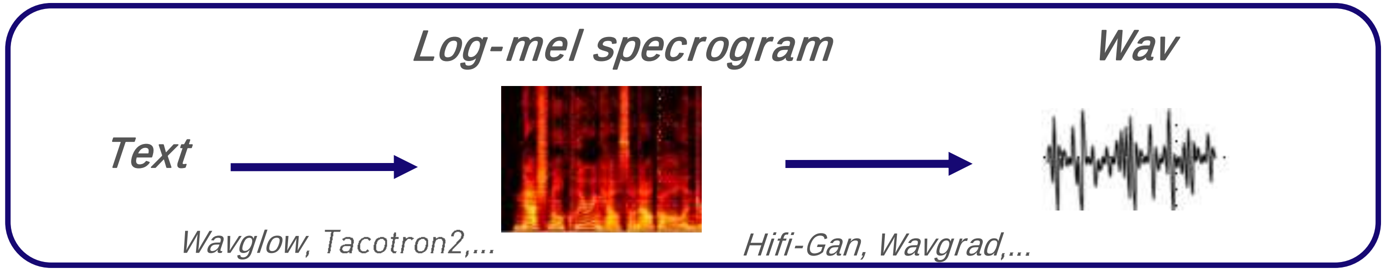


Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, DeviParikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. arXivpreprint arXiv:2209.15352, 2022a.
Demo at : https://felixkreuk.github.io/audiogen/

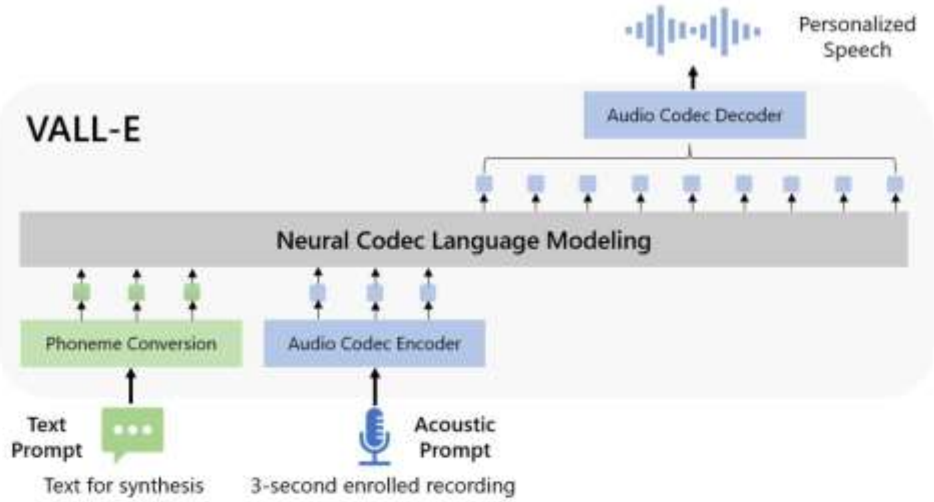
Vall-E: « text-to-speech (TTS) or *speech synthesis*

another possible model of discrete neural model

- A classic pipeline in recent Text-to-speech



- Vall-E: A different pipeline with *discrete codes* as intermediate representation

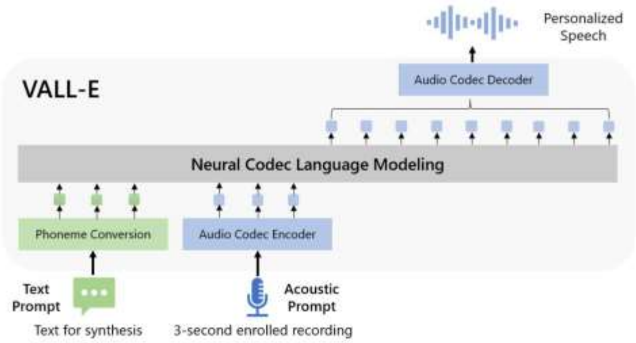


Chengyi Wang & al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers January 2023

G. Richard

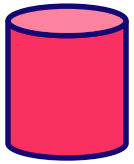
Vall-E: « text-to-speech (TTS) or *speech synthesis*

another possible model of discrete neural model



- **TTS as Conditional Codec Language Modeling**

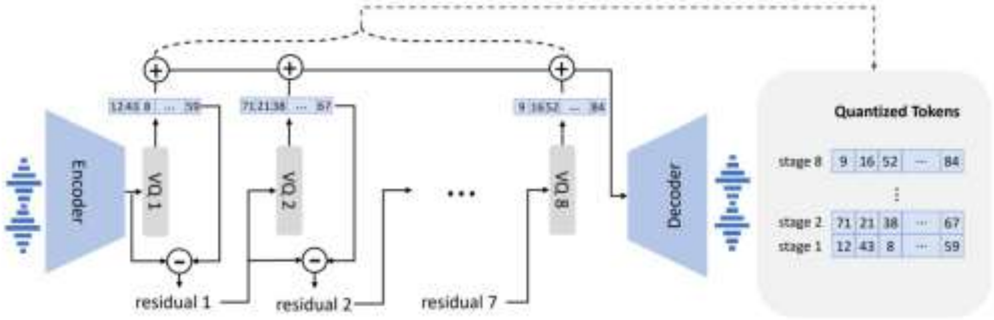
$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$$



$\mathbf{y}_i$: Audio signal

$\mathbf{x}_i$: corresponding phoneme transcription

- ***Use of an Tokenizer:* a pre-trained neural audio codec (Encodec)**



$$\text{Encodec}(\mathbf{y}) = \mathbf{C}^{T \times 8}$$

$\mathbf{c}_{t,:}$: 8 codes for frame t

$\mathbf{c}_{:,j}$: code sequence for codebook j

- **Train a neural LM to generate acoustic codes with an optimisation objective:**

$$\max\ (p(\mathbf{C}|\mathbf{x}, \tilde{\mathbf{C}})\ :\ \text{where } \tilde{\mathbf{C}} \text{ is the acoustic prompt}$$

Chengyi Wang & al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers January 2023

Vall-E: « text-to-speech (TTS) or *speech synthesis*

another possible model of discrete neural model

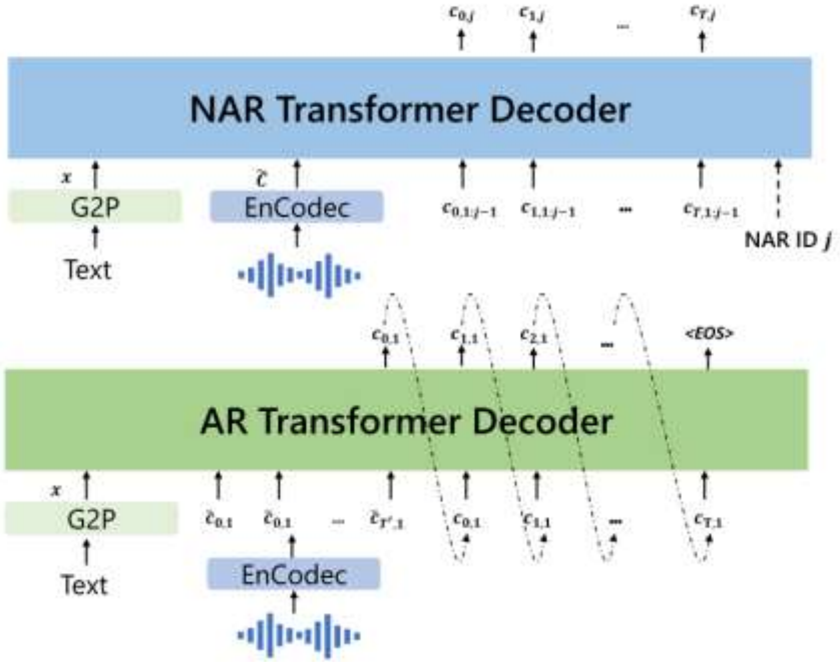
G. Richard

- The conditional codec language modelling
  - Association of 2 transformer models
  - An autoregressive model (AR) for the first codebook (*e.g. good quality*)

$$p(\mathbf{c}_{:,1}|\mathbf{x},\tilde{\mathbf{C}}_{:,1};\theta_{AR})) = \prod_{t=0}^{T} p(\mathbf{c}_{t,1}|\mathbf{c}_{<t,1},\tilde{\mathbf{c}}_{:,1},\mathbf{x};\theta_{AR})$$

  - An non-autoregressive (NAR) for the remaining ones (*e.g. less complex*)

$$p(\mathbf{C}_{:,2:8}|\mathbf{x},\tilde{\mathbf{C}};\theta_{NAR})) = \prod_{j=2}^{j=8} p(\mathbf{c}_{:,j}|\mathbf{C}_{:,<j},\mathbf{x},\tilde{\mathbf{C}};\theta_{NAR})$$



Chengyi Wang & al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers January 2023

G. Richard

Vall-E: « text-to-speech (TTS) or *speech synthesis*

another possible model of discrete neural model

- Inference: In-Context Learning via Prompting
  - Converts the text into a phoneme sequence and encodes the enrolled recording into an acoustic matrix, forming the phoneme prompt and acoustic prompt.
  - Both prompts are used in the AR and NAR models.
    - For the AR model, sampling-based decoding conditioned on the prompts is used
    - For the NAR model, greedy decoding is used to choose the token with the highest probability.
  - Finally, the neural codec decoder is used to generate the waveform conditioned on the eight code sequences.

Chengyi Wang & al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers January 2023

Vall-E demo
(replay)

- Vall-E(Microsoft)
  - Zero-shot TTS

Text	Speaker prompt	Ground Truth	VALL-E
They moved thereafter cautiously about the hut groping before and about them to find something to show that Warrenton had fulfilled his mission			

  - … or keeping the speaker emotion

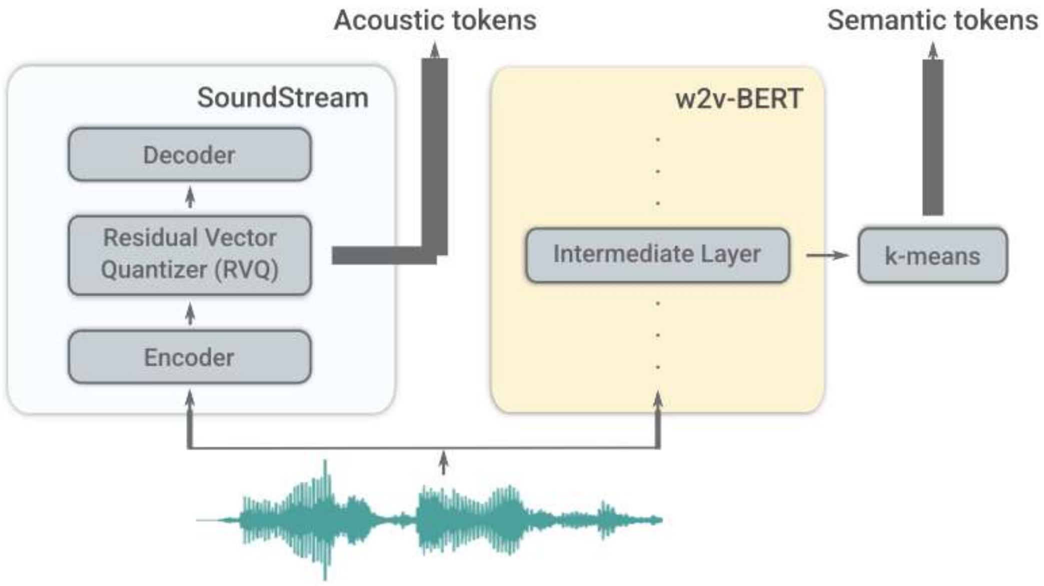
Text	Emotion	Speaker prompt	VALL-E
We have to reduce the number of plastic bags.	Anger		
	Sleepy		
	Amused		

VALL-E: https://www.microsoft.com/en-us/research/project/vall-e-x/vall-e/

AudioLM: using language models for audio generation

G. Richard

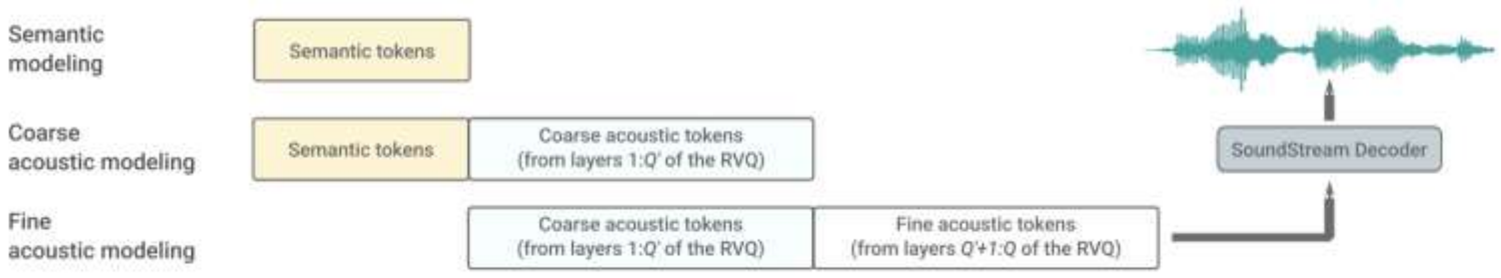
- 3 main components:
  - (i) A tokenizer model, which maps the input audio into a sequence of discrete tokens from a finite vocabulary
  - (ii) A decoder-only Transformer language model that operates on the discrete tokens. At inference time, the model predicts the token sequence autoregressively.
  - (iii) A detokenizer model, which maps the sequence of predicted tokens back to audio
- Motivation for the dual-token model (for speech signal):
  - **Acoustic tokens:** speaker identity and recording conditions (mostly)
  - **Semantic tokens:** capture the linguistic content (mostly)



Zalan Borsos et al. Audiolm:a language modeling approach to audio generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023a.
Y. Chung et al., "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in Proc. IEEE Autom. Speech Recognit. Understanding Workshop, 2021, pp. 244–250.

AudioLM: using language models for audio generation

- Hierarchical Modeling of Semantic and Acoustic Tokens
  - *In all stages: a separate (decoder) Transformer is trained for predicting next tokens given previous tokens*



- ***At inteference***
  - *Unconditional generation:* semantic tokens are sampled unconditionally and used as conditioning for acoustic modeling.
  - *Acoustic generation:* ground-truth semantic tokens are extracted from a test sequence as conditioning to generate the acoustic tokens.
  - *Generating continuations* **(**from a short prompt):

    1) generation of the continuation of semantic tokens autoregressively;

    2) concatenation of the entire semantic token sequence with the coarse acoustic tokens of the prompt and then feed as conditioning to the coarse acoustic model, which then samples the continuations of the corresponding acoustic tokens

    3) the coarse acoustic tokens are processed with the fine acoustic model.

    4) both the prompt and the sampled acoustic tokens are fed to the SoundStream decoder to reconstruct audio

Zalan Borsos et al. Audiolm:a language modeling approach to audio generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023a.

AudioLM: using language models for audio generation

Demo

- ***Speech (continuation)***

  Original (speech) *Prompt* *Continuation by AudioLM*

- ***Acoustic generation*** *« we sample the acoustic tokens given the semantic tokens extracted from the original samples from LibriSpeech test-clean»*

  Original (speech) *Generated (1)* *Generated (2)*

- **Generation without semantic tokens** : *« Continuations with a language model trained on the acoustic tokens only (without semantic tokens)"*

  Example 1 *Example 2*

- **An interesting example : piano continuation**

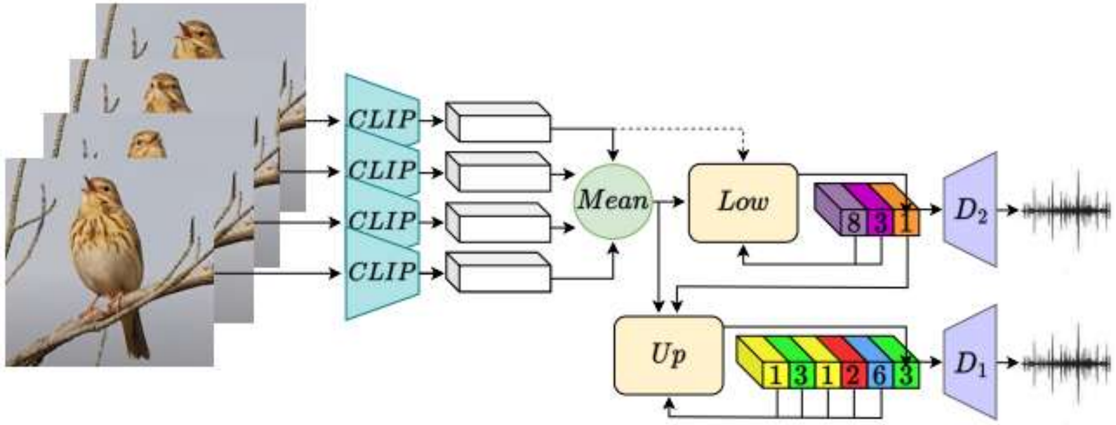
Original *Prompt* Continuation by acoustic-only model *Continuation by AudioLM*

A prompt of a known piano sonata example (Beethoven N° 18) is continued in … another known piano sonata (Beethoven – Moonlight sonata) !

More examples at https://google-research.github.io/seanet/audiolm/examples/

Towards image-to-audio: the model IM2WAV

- IM2WAV: a Transformer-based audio Language Model (LM) conditioned on image representation



- 3 main components:
  - (i) an audio encoder-decoder with a discrete internal representation (VQ-VAE)
  - (ii) a pre-trained image encoder (CLIP)
  - (iii) an audio language model which operates over the discrete audio tokens (autoregressive sparse transformer)
- Some specifities:
  - CLIP embeddings trained in a multimodal context
  - Use Classifier Free Guidance (CFG) to improve generation for Low resolution (e.g. randomly unconditional training)

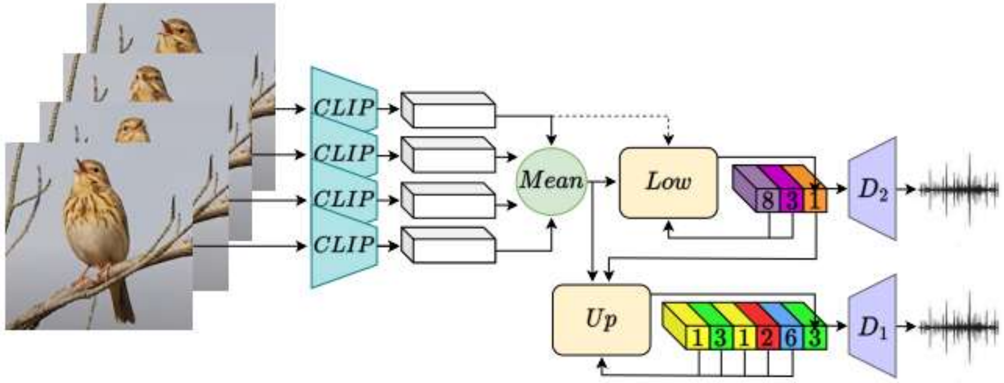
- **Parameters:**
  - 16 kHz Sampling frequency (4 s of sound)
  - 5 Conv. Layers for VQ-VAE (stride 2) Enc/dec.
  - 1st codebook after 3 layers (downsampling of 8)
  - 2nd codebook after 5 layers (downsampling of 32)
  - 2 k (resp. 5k) tokens in the UP (resp. LOW) model
  - Codebook: 2048 codes, embedding size of 128
  - Transformer: 48 layers, sparse attention

Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

G. Richard

The IM2WAV model

- Demo



DALL-E Image Guided
Audio Generation Example

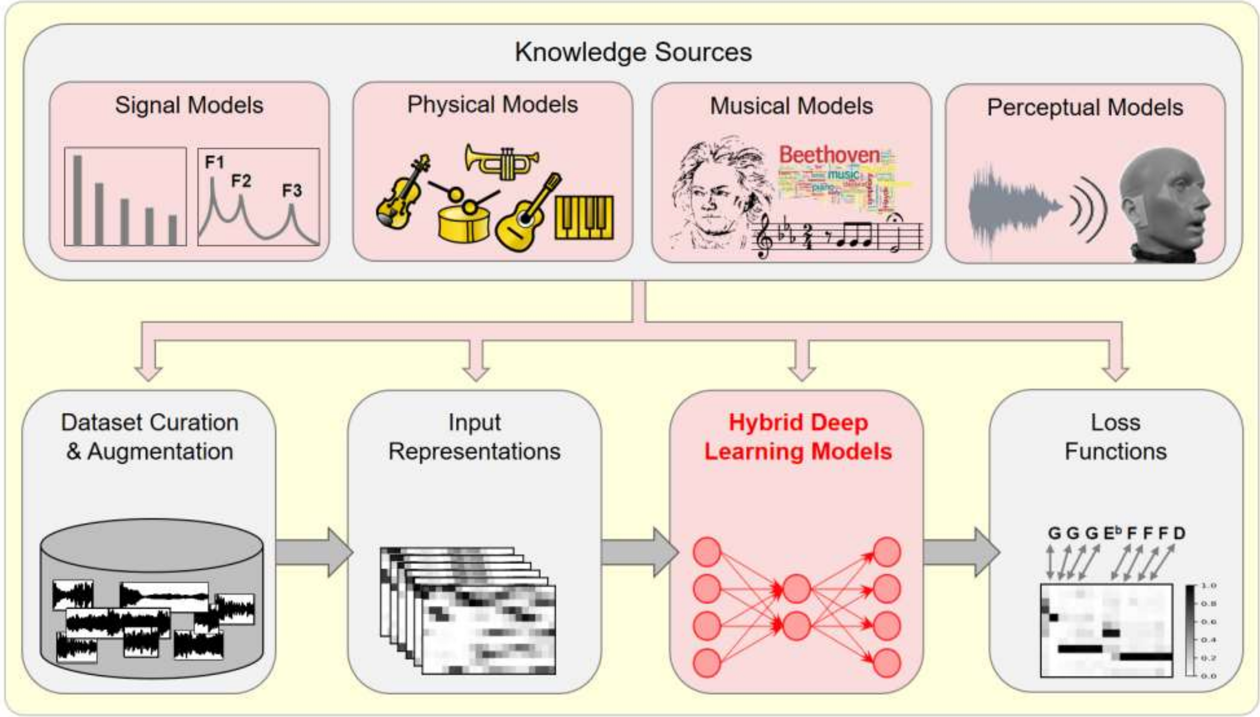
Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
Demo at :https://pages.cs.huji.ac.il/adiyoss-lab/im2wav/

Towards hybrid deep learning …

Towards Hybrid deep learning approaches

- Coupling model-based and deep learning:

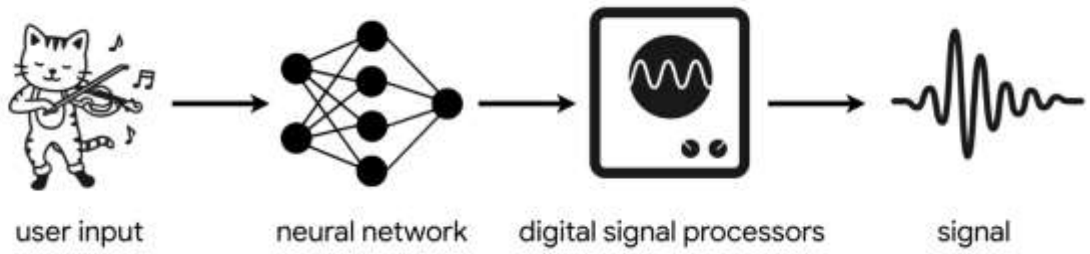
Example with Hybrid deep model for Music signals



G. Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Hybrid Deep Learning for Music Information Research", IEEE Signal Processing Magazine - Special Issue on Model-based and Data-Driven Audio Signal Processing, 2024 (under review)
Hi-Audio, Hybrid and Interpretable Deep neural audio machines, European Research Council "Advanced Grant" (AdG) project - https://hi-audio.imt.fr/

Towards Hybrid deep learning approaches

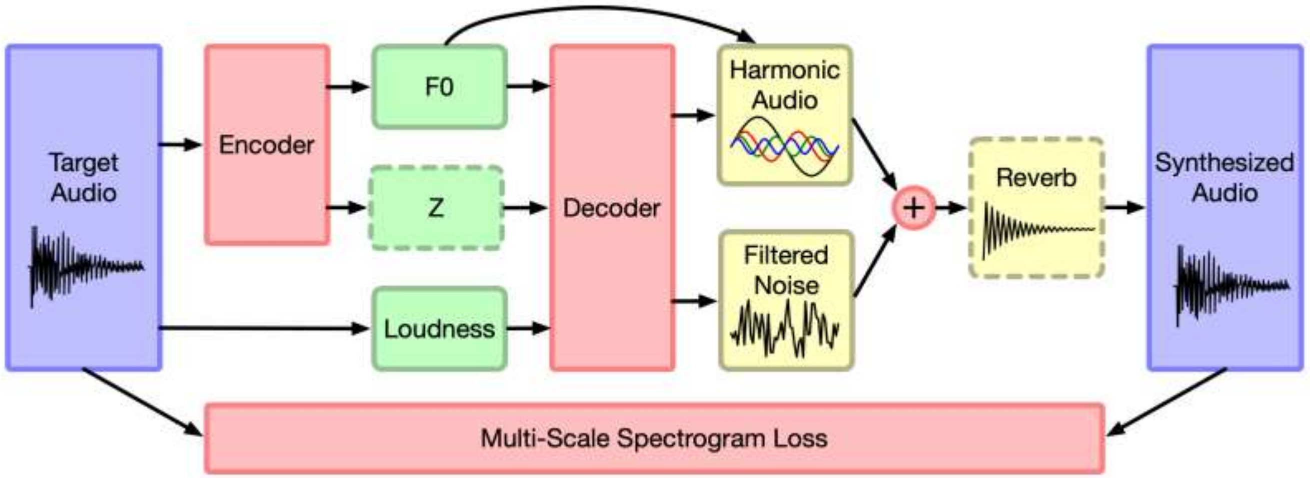
- Coupling model-based and deep learning
  - For example, using deep learning for learning the parameters of a signal processing model



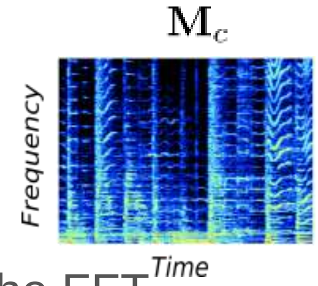
J. Engel & al., "DDSP: Differentiable Digital Signal Processing," in Int. Conf. on Learning Representations (ICLR), 2020.

Towards Hybrid deep learning approaches

- The example of DDSP



- A multi-scale spectral loss $\mathcal{L}_{rec} = \sum_c \mathcal{L}_c$



With $\mathcal{L}_c = \|\mathbf{M}_c - \tilde{\mathbf{M}}_c\|_1 + \|\log(\mathbf{M}_c) - \log(\tilde{\mathbf{M}}_c)\|_1$

and with c = [2048, 1024, 512, 256, 128, 64] indicates the FFT size used to compute the STFT.

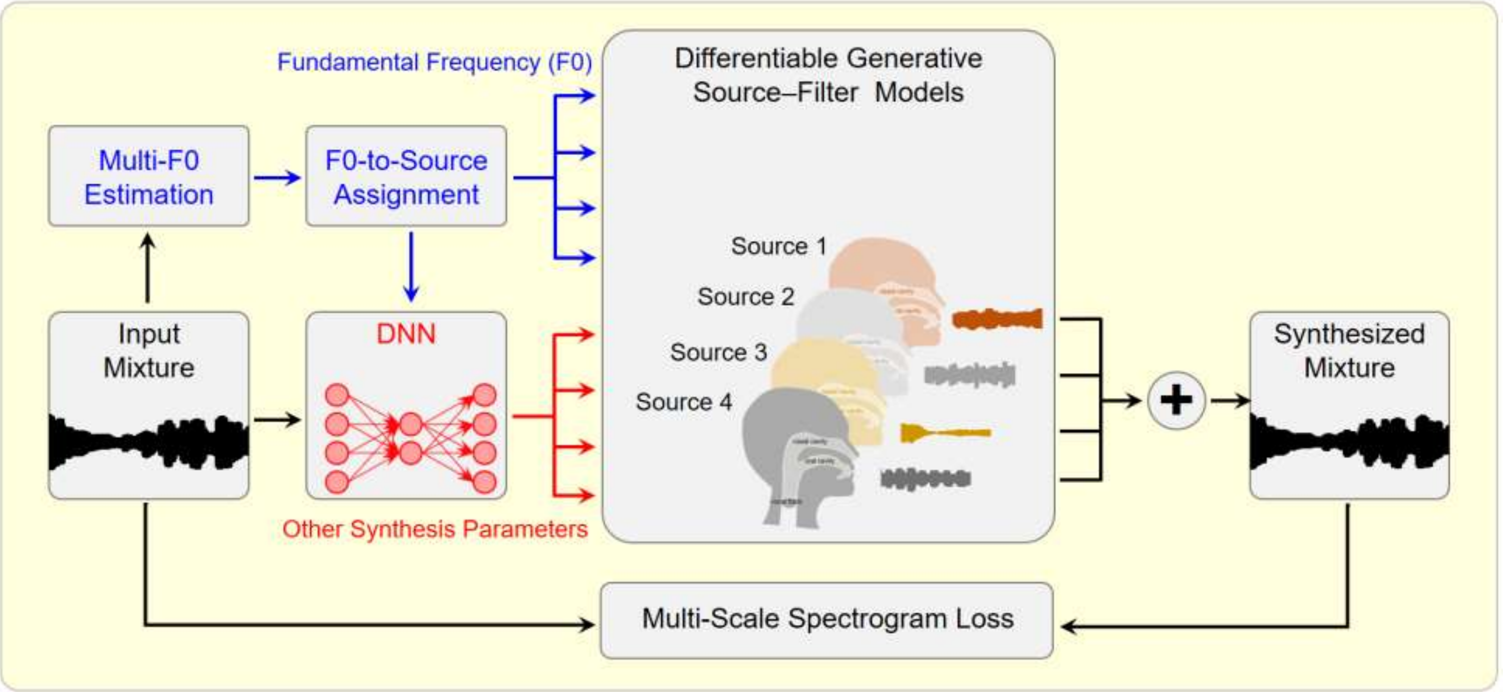
J. Engel & al., "DDSP: Differentiable Digital Signal Processing," in Int. Conf. on Learning Representations (ICLR), 2020.

G. Richard

Towards Hybrid deep learning approaches:

DDSP extensions and others…

- An example for unsupervised singing voice separation



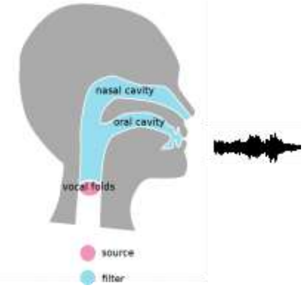
K Schulze-Forster, G. Richard, L. Kelley, C. Doire, R Badeau Unsupervised Music Source Separation Using Differentiable Parametric Source Models, IEEE Trans. On AASP, 2023
G. Richard, V. Lostanlen, Y.-H. Yang, M. Müller, "Hybrid Deep Learning for Music Information Research", IEEE Signal Processing Magazine - Special Issue on Model-based and Data-Driven Audio Signal Processing, 2024 (under review)

G. Richard

Towards Hybrid deep learning

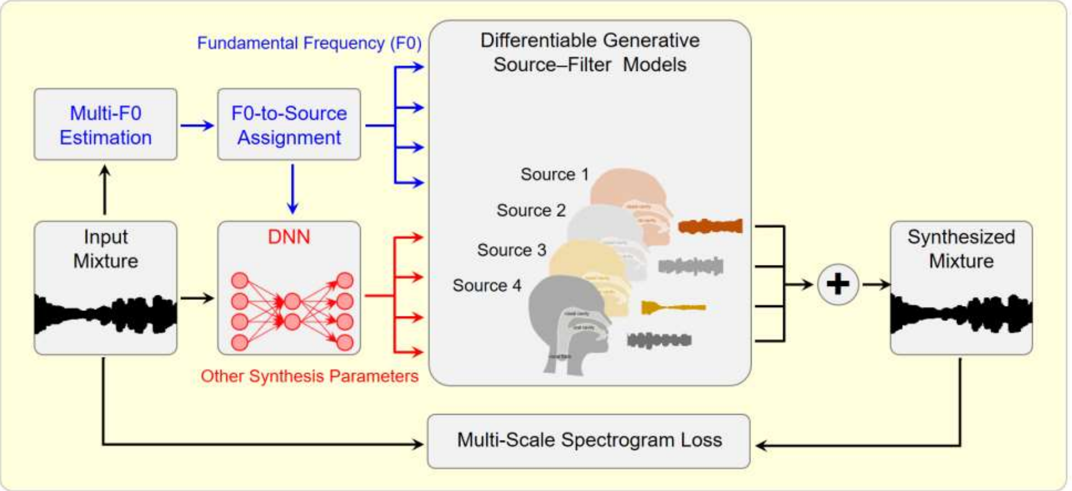
… by **integrating our prior knowledge** about the nature of the processed data.

Knowledge about « how the sound is produced « (e.g. sound production models)



Singing voice as a source / filter model :

- source = vibration of vocal folds
- Filter = resonances of vocal/nasal cavities



A new paradigm

- Model is at the « core » of neural architecture
- Source separation **by synthesis** (*no interference from other sources*)
- Learning only from the polyphonic recording (*no need of the true individual tracks*)

Novel sound transformation capabilities:

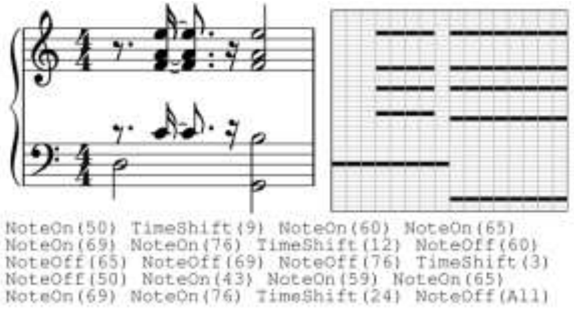
- Timbre/melody of the voice,
- Lyrics, translation
- Re-harmonization

Conclusion

G. Richard

- Generative AI goes beyond text generation…
- Generative Audio is gaining a strong interest and a variety of models and approaches are already proposed
- Note that I have not discussed the models for symbolic music (e.g. music scores as in MIDI)

Music score



MIDI representation (or piano roll)

Representation as sequence of tokens

  - … which includes transformer models for symbolic music, « theme » transformer, Groove2Groove (style transfer), long context modelling (with specific positional encoding….)

Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller, and Y.-H. Yang, "Theme Transformer: Symbolic music generation with theme-conditioned Transformer," IEEE Transactions on Multimedia, vol. 25, pp. 3495–3508, 2023.
O. Cıfka, U. Simsekli, and G. Richard, "Groove2groove: One-shot music style transfer with supervision from synthetic data," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2638–2650, 2020.
Manvi Agarwal, Changhong Wang, Gaël Richard, Structure-Informed Positional Encoding For Music Generation, Accepted for publication at ICASSP 2024.