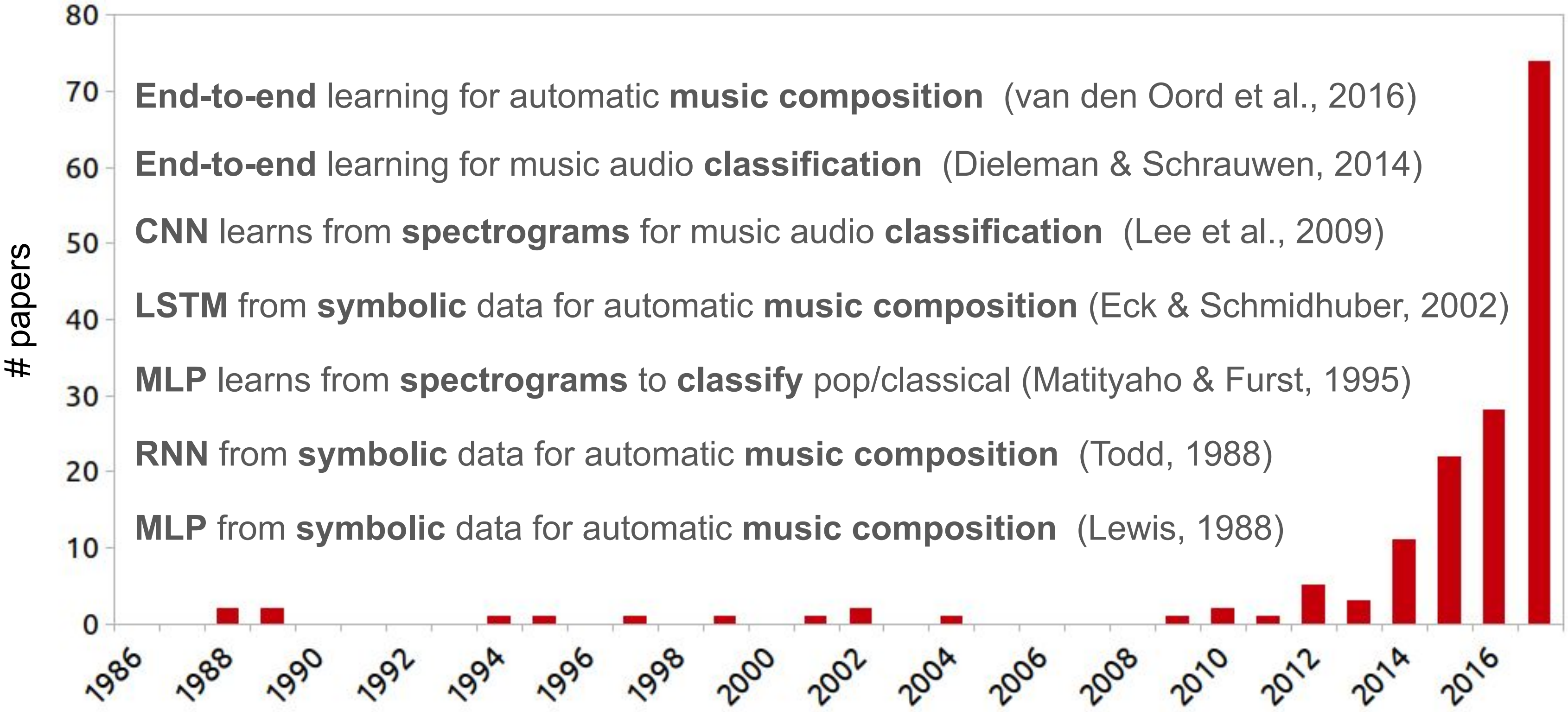
Waveform-based music processing with deep learning

Sander Dieleman, Jordi Pons, Jongpil Lee

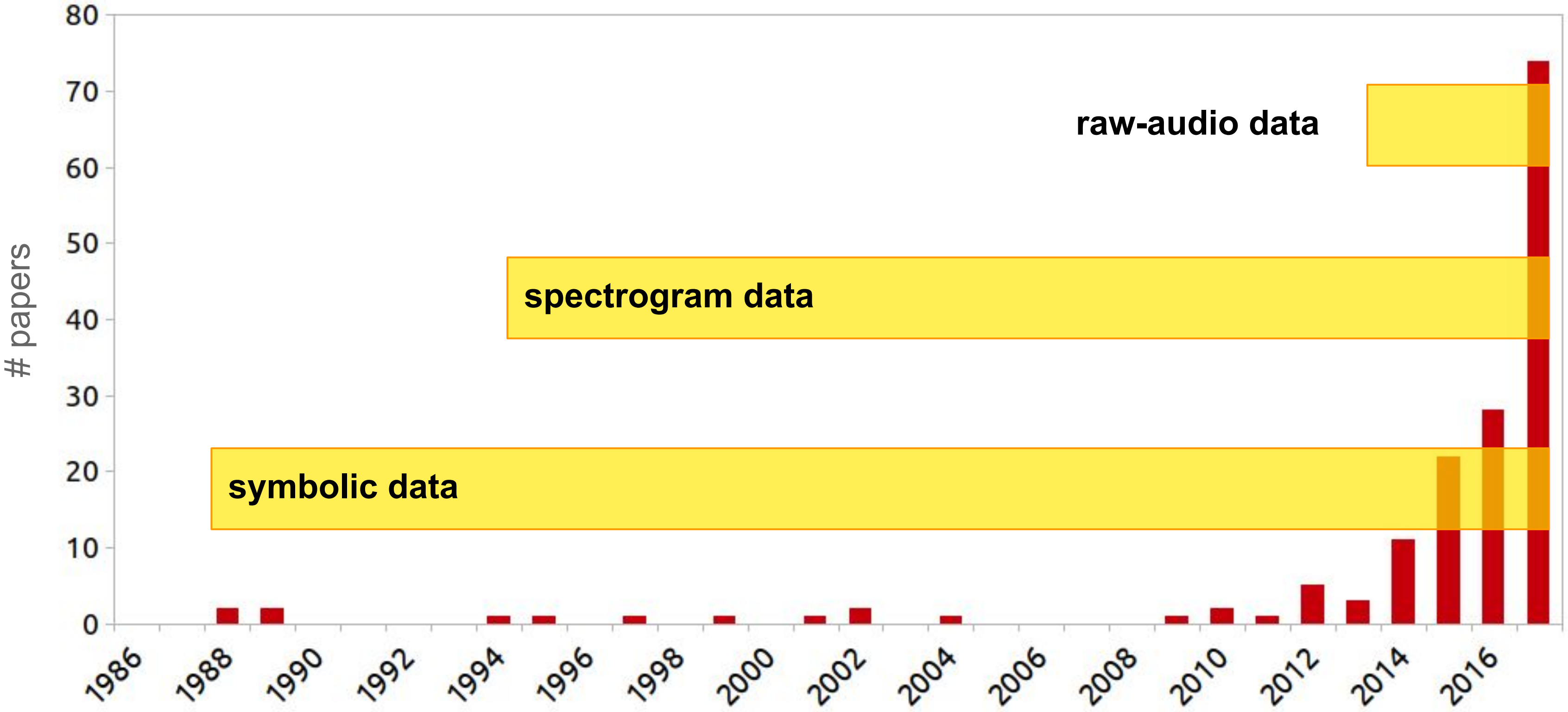
Historical perspective

Why now?

Papers on “neural networks & music”: milestones



Papers on “neural networks & music”: input data



References

Van den Oord et al., 2016. “Wavenet: a generative model for audio” in arXiv.

Dieleman and Schrauwen, 2014. “End-to-end learning for music audio”
in International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Lee et al., 2009. “Unsupervised feature learning for audio classification using convolutional deep belief networks”
in Advances in Neural Information Processing Systems (NIPS).

Matityaho and Furst, 1995. “Neural network based model for classification of music type”
in 18th Convention of Electrical and Electronics Engineers in Israel. IEEE, 4–3.

Eck and Schmidhuber, 2002. “Finding temporal structure in music: Blues improvisation with LSTM recurrent networks”
in Proceedings of the Workshop on Neural Networks for Signal Processing.

Todd, 1988. “A sequential network design for musical applications”
in Proceedings of the Connectionist Models Summer School.

Lewis, 1988. Creation by Refinement: A creativity paradigm for gradient descent learning networks”
in International Conference on Neural Networks.

Historical review from: Pons, 2019. “Deep neural networks for music and audio tagging”. PhD Thesis.

Waveform-based music processing with deep learning

Classification

by Jongpil Lee, PhD candidate at KAIST in Daejeon, South Korea.

Source Separation

by Jordi Pons, researcher at Dolby Laboratories in Barcelona.

Generation

by Sander Dieleman, research scientist at DeepMind in London, UK.

Music Classification

Learning specific properties of music

Outline

- Problem statement
- Audio classification models
    - *Engineered feature based model*
    - *Frame-level waveform based model*
    - *Sample-level waveform based model*
- Connect signal processing and deep learning
- Summary

Outline

- **Problem statement**
- Audio classification models
    - *Engineered feature based model*
    - *Frame-level waveform based model*
    - *Sample-level waveform based model*
- Connect signal processing and deep learning
- Summary

Music classification

Large music catalog

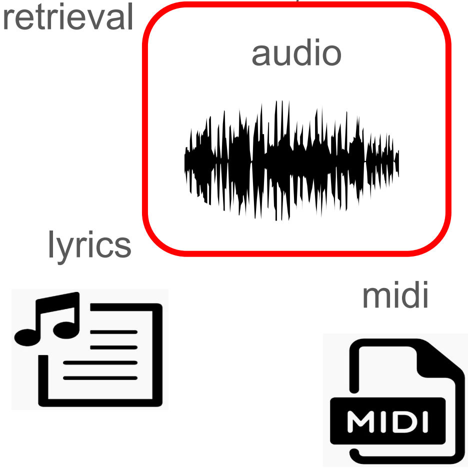
Retrieval / recommendation

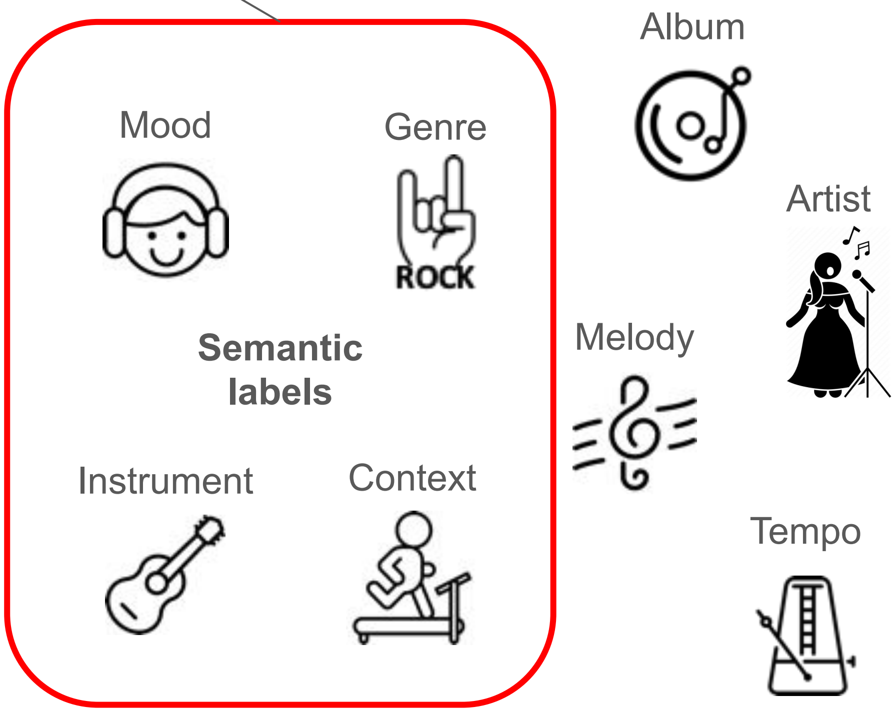
Collaborative filtering

Cold-start problem

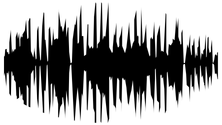
Tag-based music retrieval

Auto-tagging



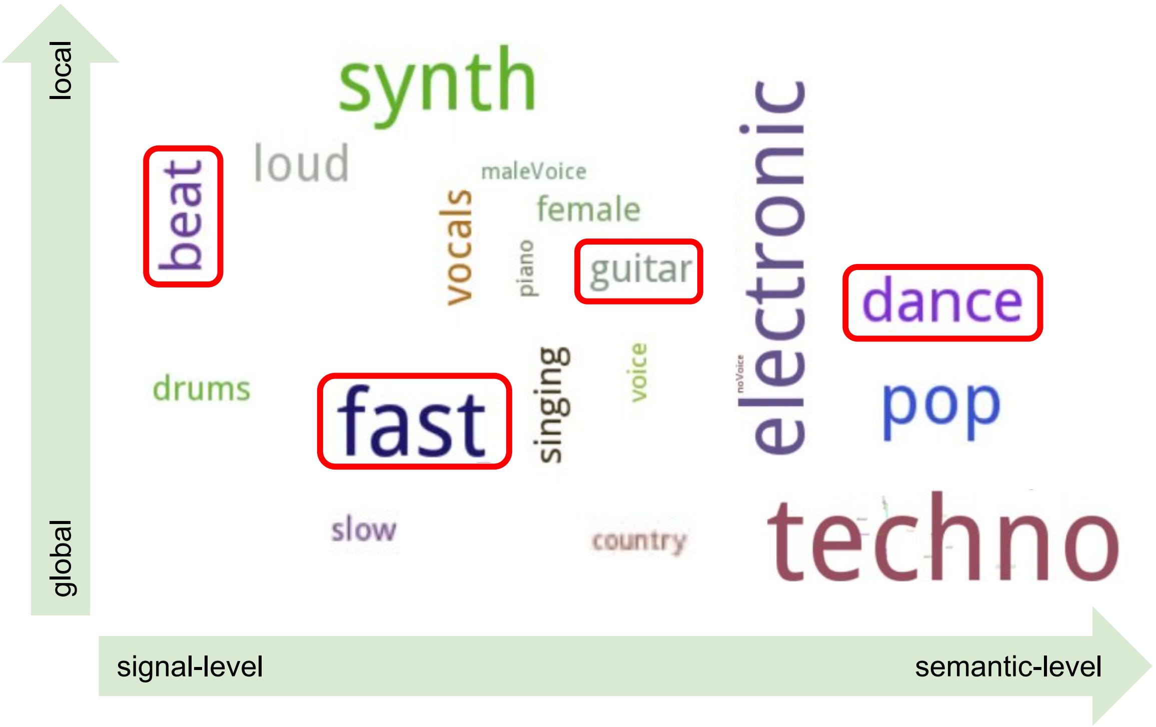


Celma, Oscar, 2010 "Music recommendation", *Music recommendation and discovery, Springer*.

Music 

	Scene	Music (artwork)	Speech (object)
Image			
Audio			
	Multi source **polyphonic**	Multi timbre (texture) **polyphonic**	Mostly **single** speaker
	Unstructured sound sources	**Structure** of sound sources	A **structured** sound source
		Dynamic control (mixing)	

Semantic labels

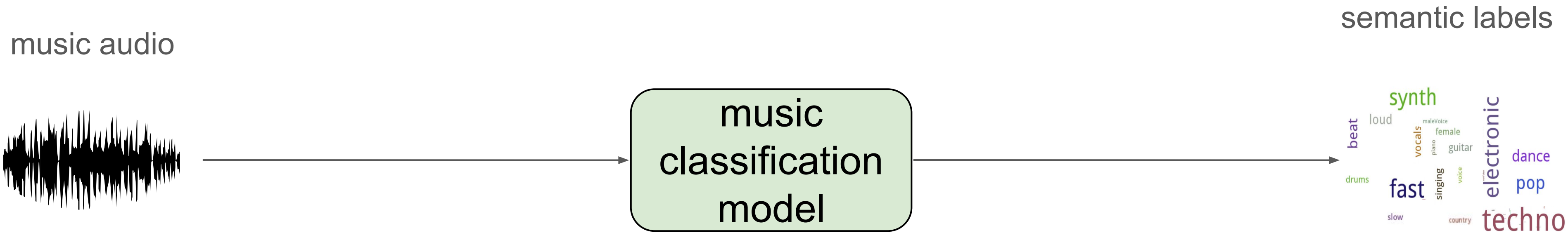


Semantic labels are highly diverse and have different levels of abstraction

Grosche et al., 2012 "Audio content-based music retrieval", *Dagstuhl Follow-Ups*.

Casey et al., 2008 "Content-based music information retrieval: Current directions and future challenges", *Proceedings of the IEEE*.

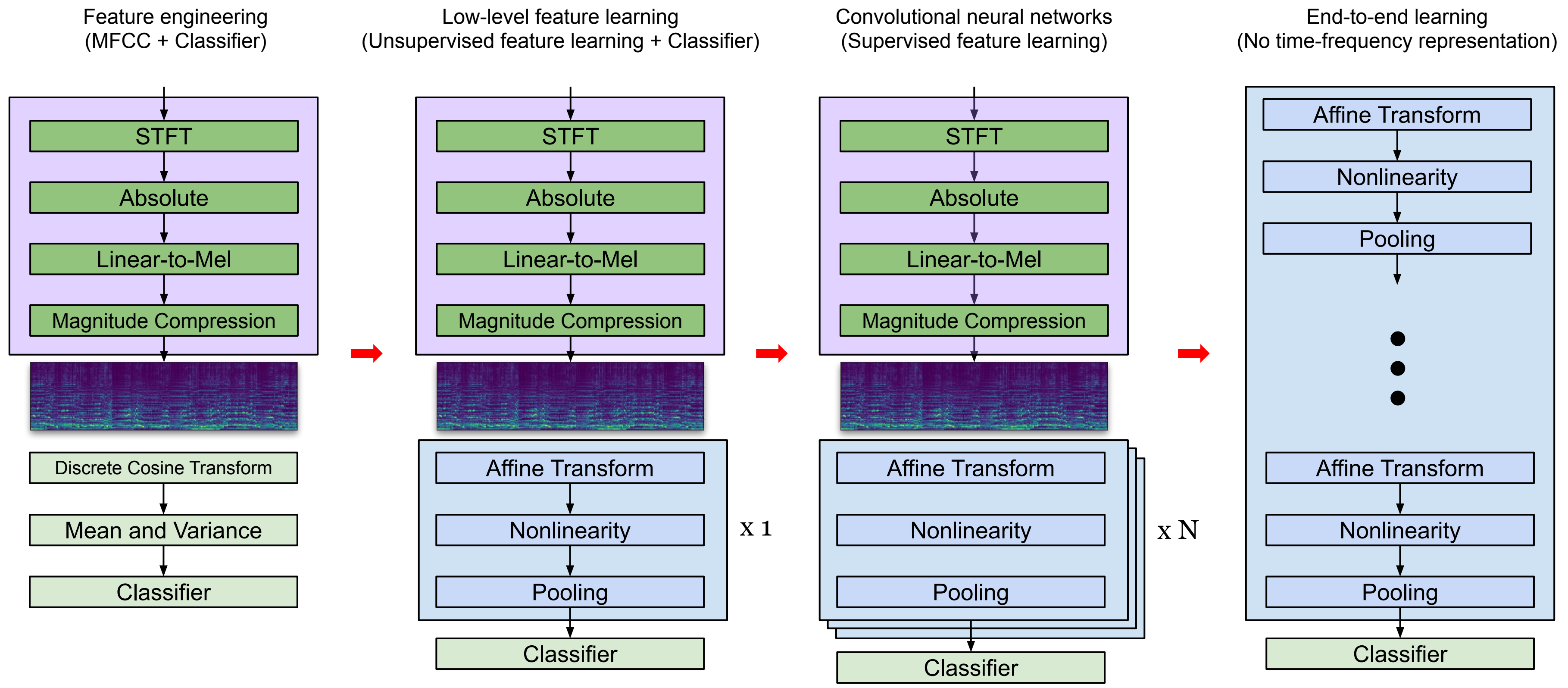
Problem definition



Outline

- Problem statement
- **Audio classification models**
  - *Engineered feature based model*
  - *Frame-level waveform based model*
  - *Sample-level waveform based model*
- Connect signal processing and deep learning
- Summary

From feature engineering to end-to-end learning



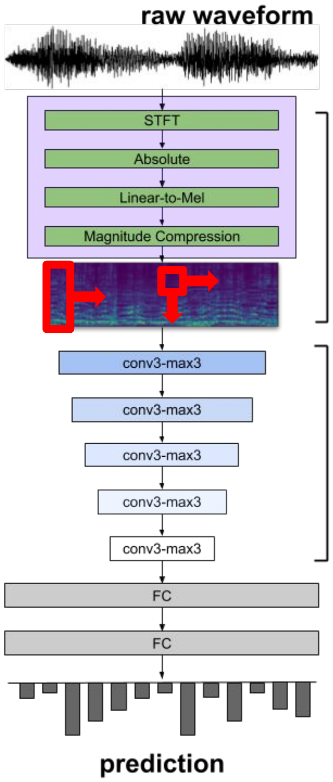
Nam et al., 2018. "Deep learning for audio-based music classification and tagging", *IEEE Signal Processing Magazine.*

Humphrey et al., 2013. "Feature learning and deep architectures: new directions for music informatics", *Journal of Intelligent Information Systems.*

Outline

- Problem statement
- Audio classification models
    - ***Engineered feature based model***
    - *Frame-level waveform based model*
    - *Sample-level waveform based model*
- Connect signal processing and deep learning
- Summary

Engineered feature based model



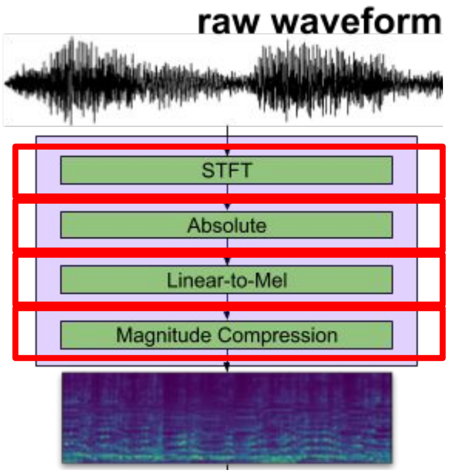
- MFCC, Chroma, Bag of low-level-features
- **Mel-spectrogram**
- Scattering transform
- **1D CNNs**
- **2D CNNs**
- RNNs
- Attention

Pons et al., 2018. "End-to-end learning for music audio tagging at scale", *ISMIR.*

Choi et al., 2016. "Automatic tagging using deep convolutional neural networks", *ISMIR.*

Purwins et al., 2019. "Deep Learning for Audio Signal Processing", *IEEE Journal of Selected Topics in Signal Processing.*

Fixed parameters in preprocessing stage



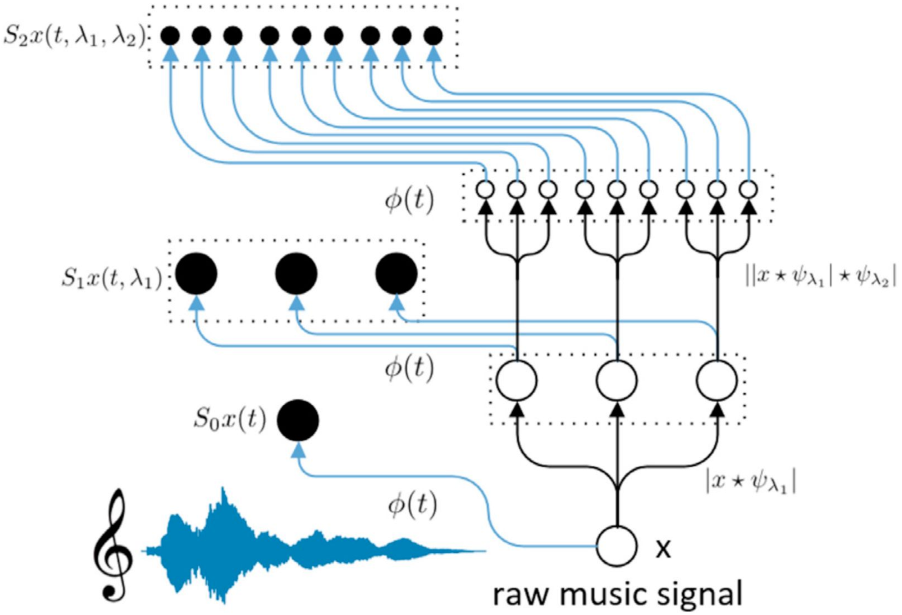
Window, hop sizes

Phase

Frequency scale, filter bandwidth

Dynamic range

Scattering transform



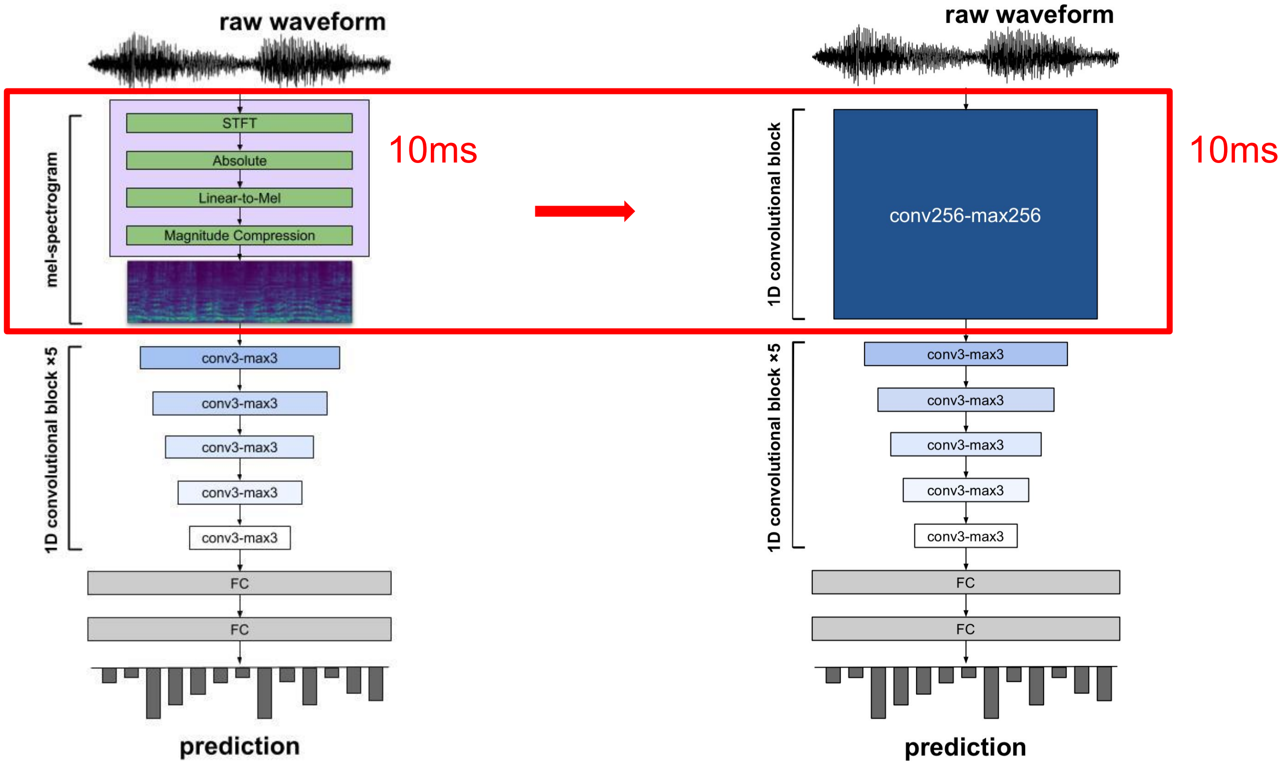
Andén et al., 2011. "Multiscale Scattering for Audio Classification", *ISMIR.*

Song et al., 2018. "Music auto-tagging using deep Recurrent Neural Networks", *Neurocomputing.*

Outline

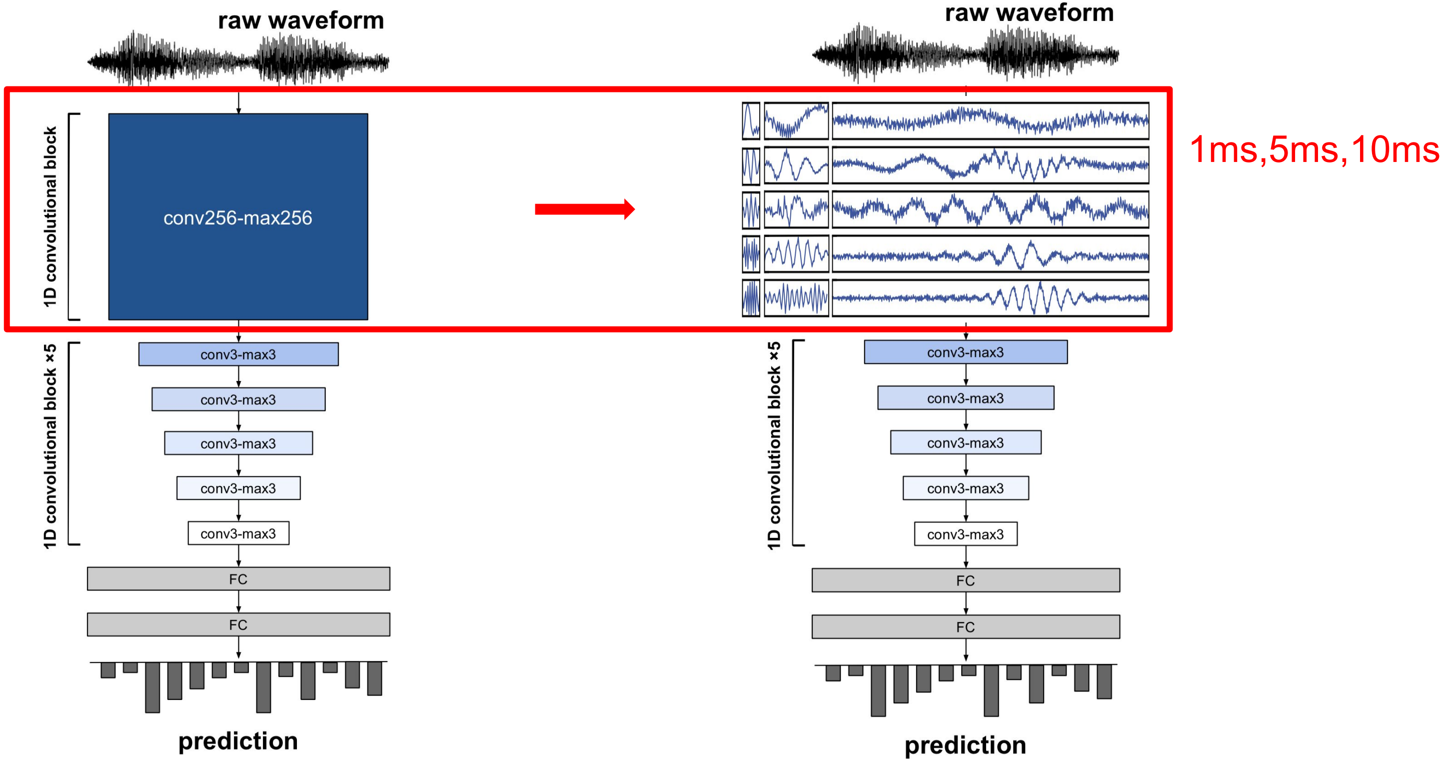
- Problem statement
- Audio classification models
  - *Engineered feature based model*
  - ***Frame-level waveform based model***
  - *Sample-level waveform based model*
- Connect signal processing and deep learning
- Summary

Frame-level waveform based model



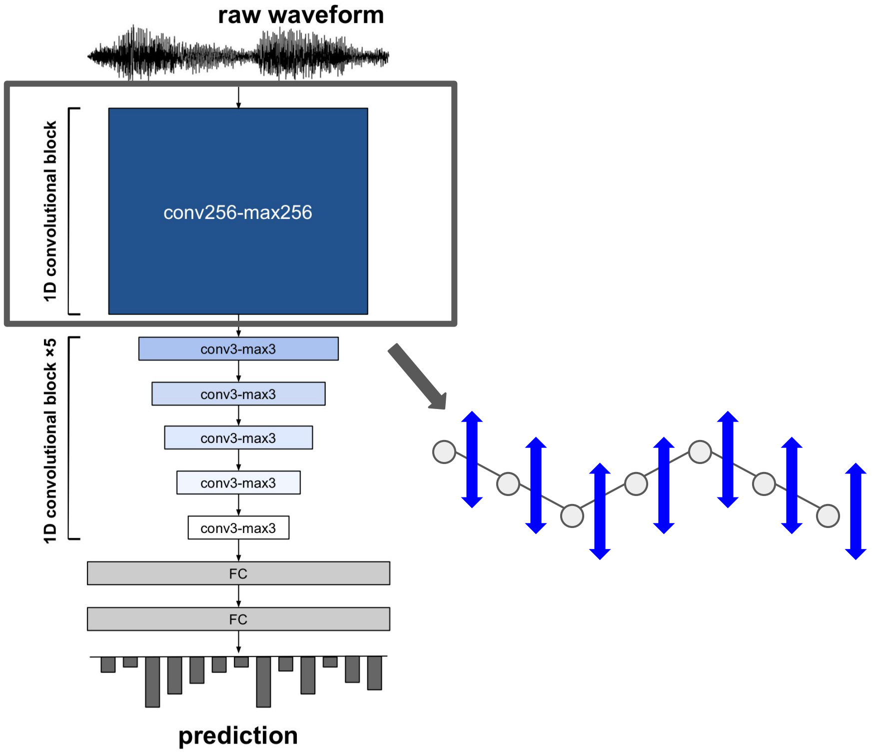
Dieleman and Schrauwen, 2014. "End-to-end learning for music audio", *ICASSP*.

Multi-scale approaches



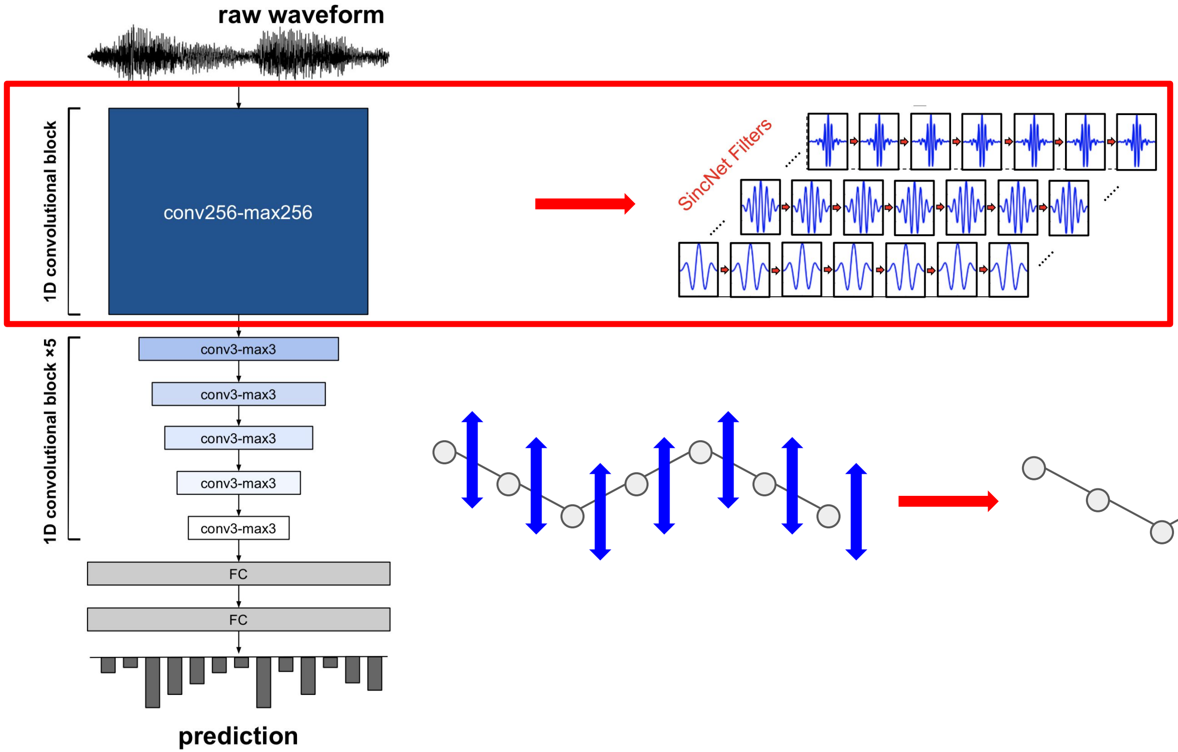
Zhu et al., 2016. "Learning multiscale features directly from waveforms", *Interspeech.*

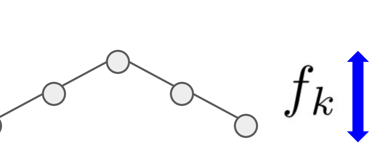
SincNet



Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", *IEEE Spoken Language Technology Workshop (SLT).*

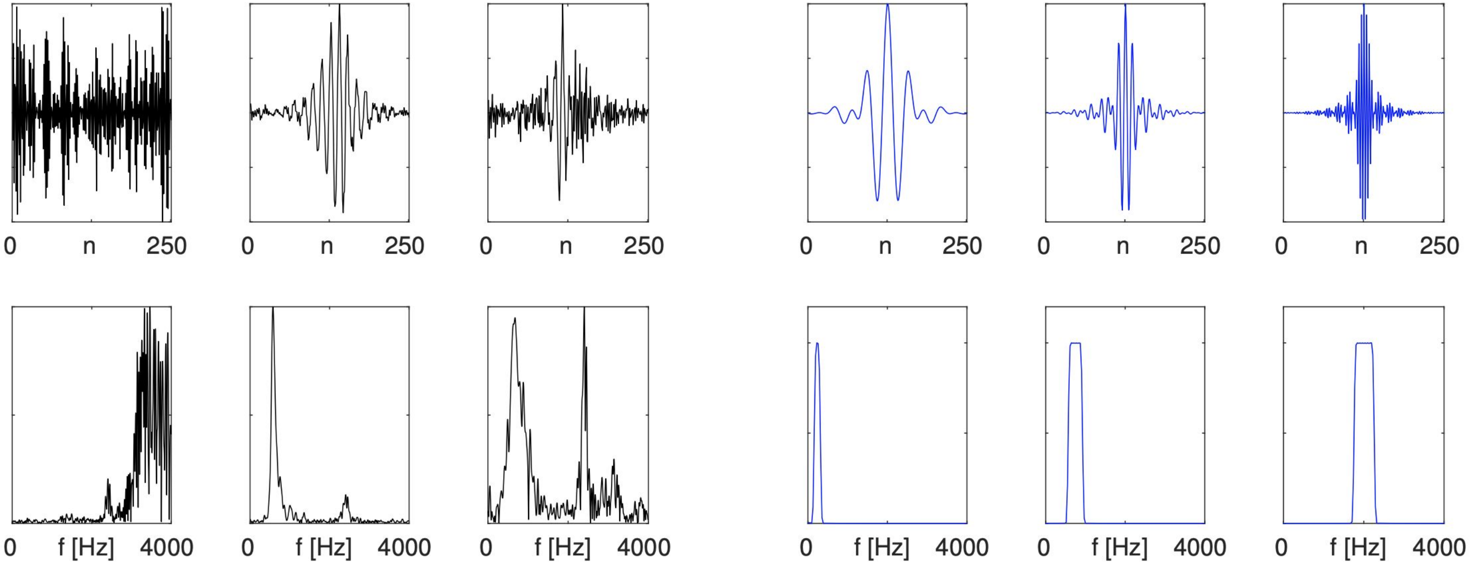
SincNet





Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", *IEEE Spoken Language Technology Workshop (SLT)*.

SincNet

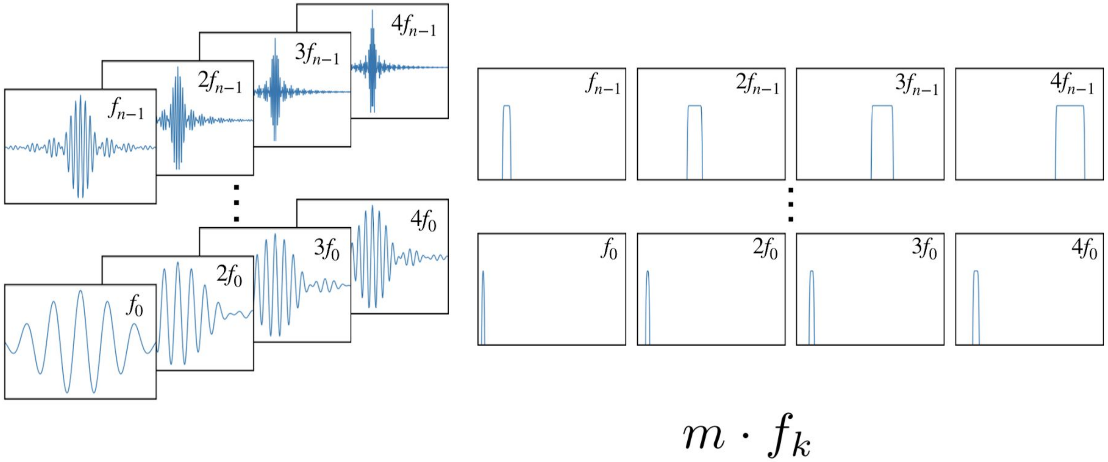


(a) CNN Filters

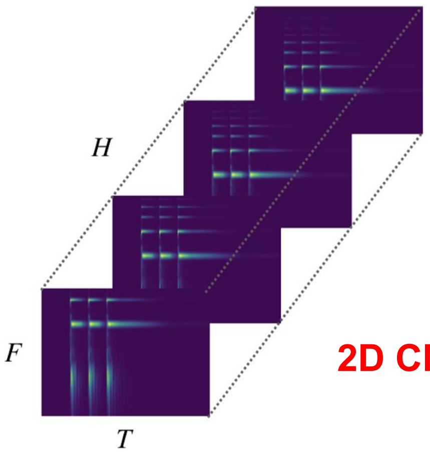
(b) SincNet Filters

Ravanelli and Bengio, 2018. "Speaker recognition from raw waveform with sincnet", *IEEE Spoken Language Technology Workshop (SLT).*

HarmonicCNN



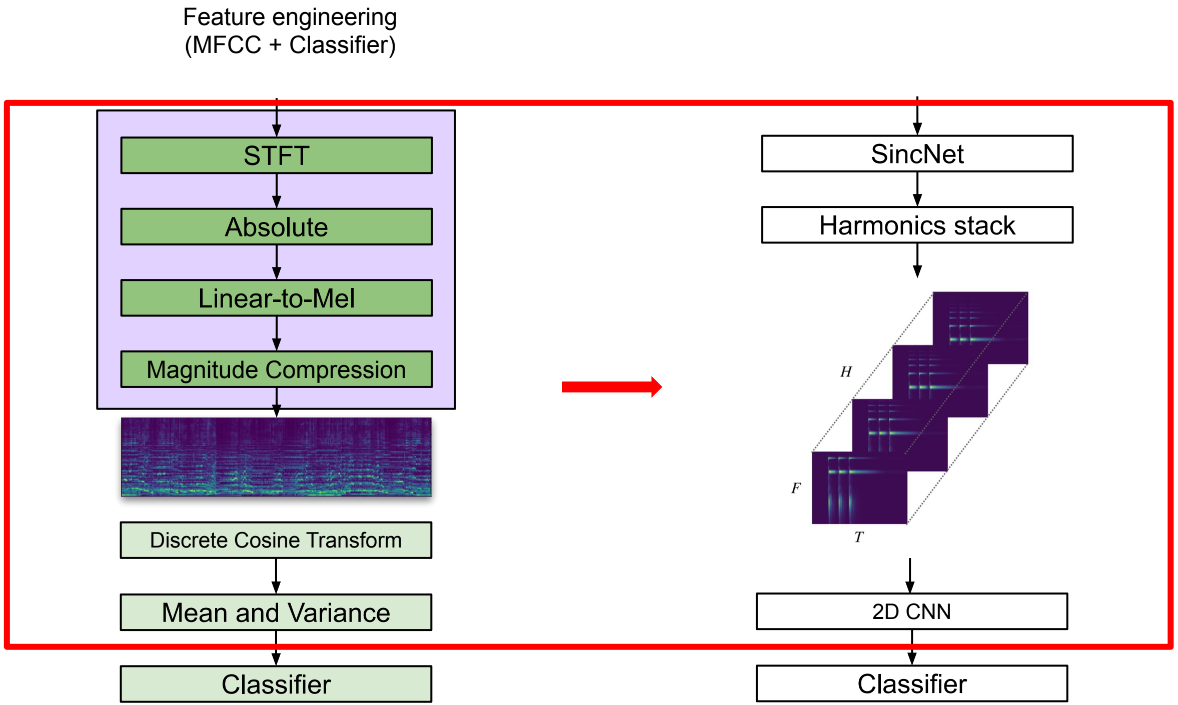
Harmonic embedding features are stacked at **channel dimension**



2D CNNs

Won et al., 2019. "Automatic music tagging with harmonic CNN", *ISMIR LBD*.

HarmonicCNN



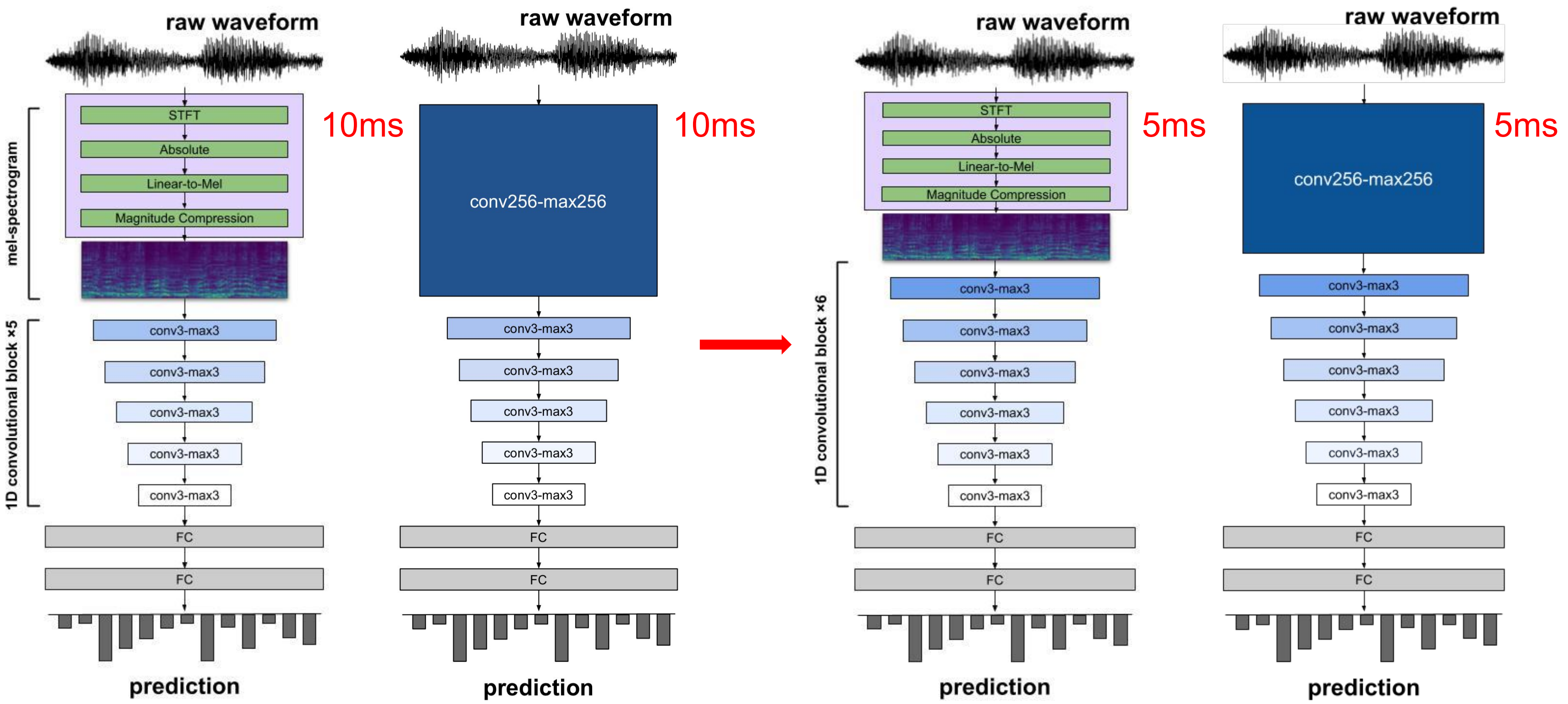
Won et al., 2019. "Automatic music tagging with harmonic CNN", *ISMIR LBD.*

Outline

- Problem statement
- Audio classification models
  - *Engineered feature based model*
  - *Frame-level waveform based model*
  - ***Sample-level waveform based model***
- Connect signal processing and deep learning
- Summary

Connect 1D CNN model to sample-level model

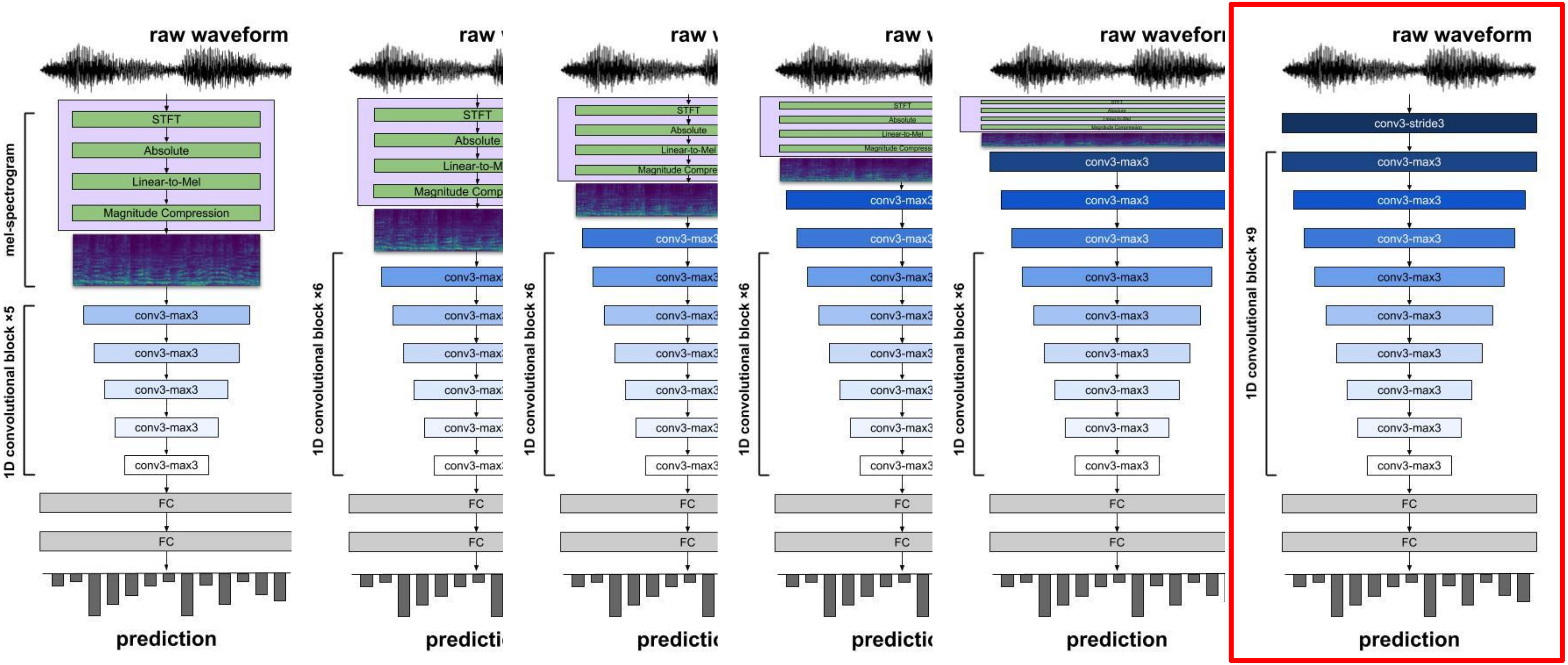
by reducing the sizes of window/filter and hop/stride , and, accordingly, increasing the number of convolutional blocks.



Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", *SMC*.

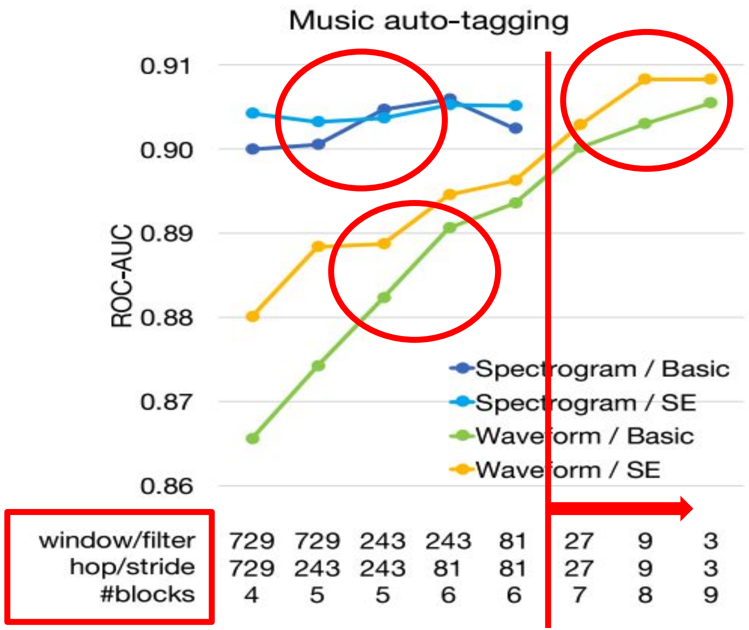
Connect 1D CNN model to sample-level model

by reducing the sizes of window/filter and hop/stride , and, accordingly, increasing the number of convolutional blocks.



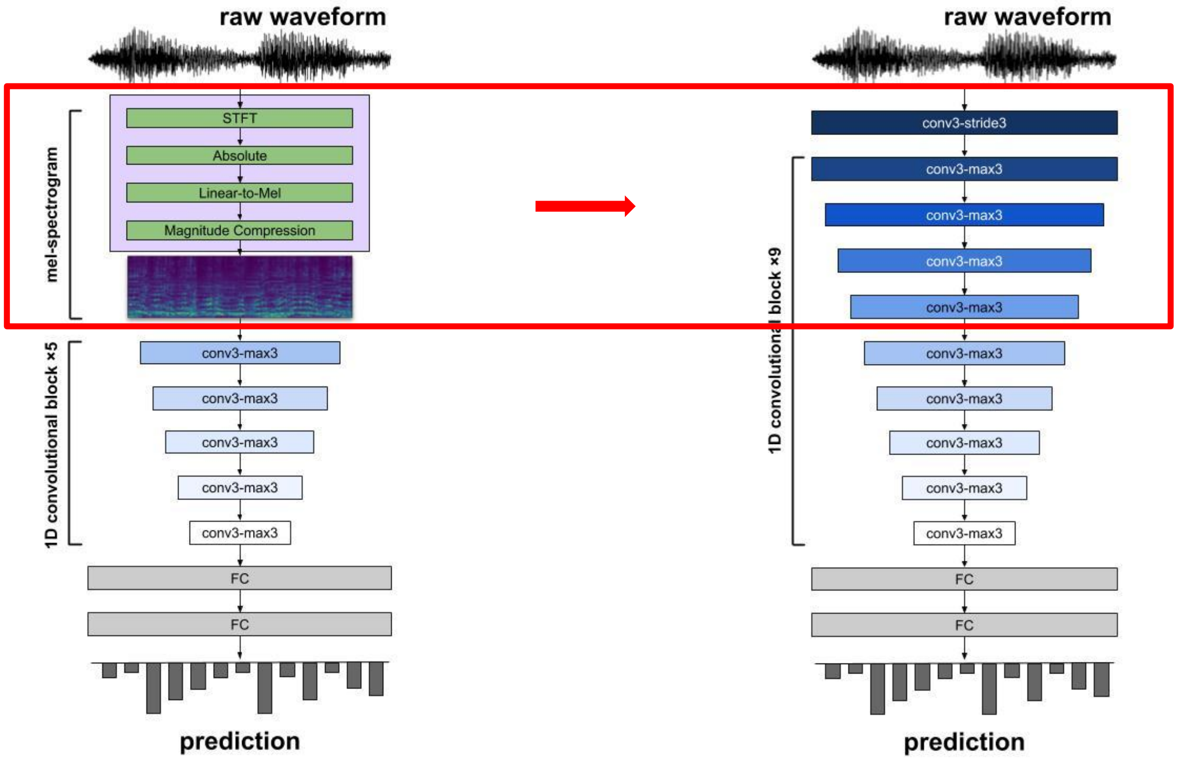
Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", *SMC*.

Connect 1D CNN model to sample-level model



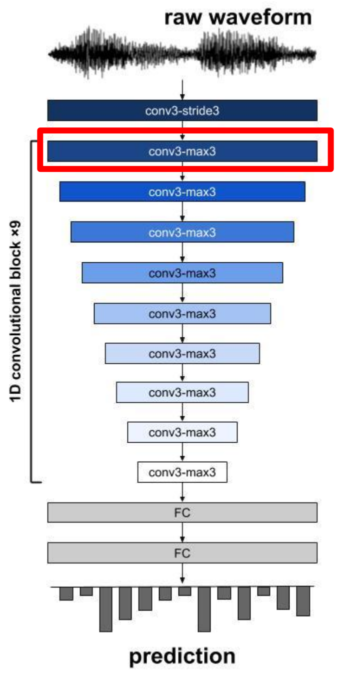
Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", *IEEE Journal of Selected Topics in Signal Processing*.

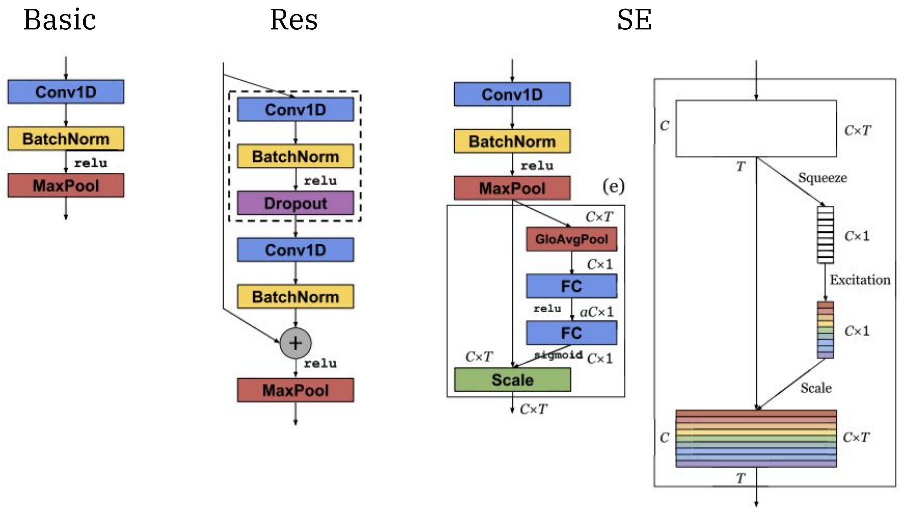
Sample-level waveform based model



Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", *SMC.*

More building blocks





Kim et al., 2018. "Sample-level CNN architectures for music auto-tagging using raw waveforms", *ICASSP*.

Outline

- Problem statement
- Audio classification models
  - ***Engineered feature*** *based model*
  - ***Frame-level waveform*** *based model*
  - ***Sample-level waveform*** *based model*
- Connect signal processing and deep learning
- Summary

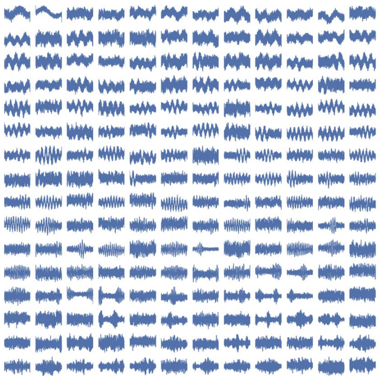
Pons et al., 2018. "End-to-end learning for music audio tagging at scale", *ISMIR.*

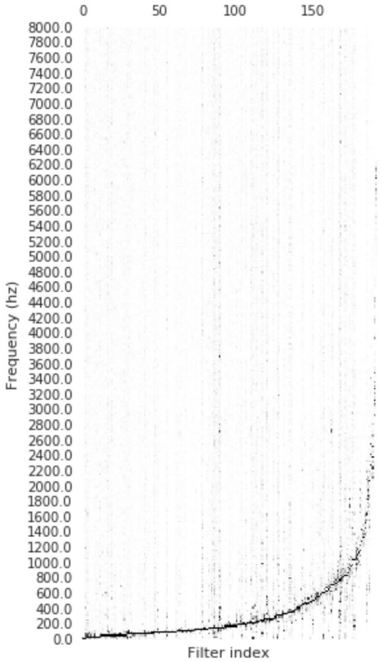
Outline

- Problem statement
- Audio classification models
    - *Engineered feature based model*
    - *Frame-level waveform based model*
    - *Sample-level waveform based model*
- **Connect signal processing and deep learning**
- Summary

Learned filter visualization

First convolutional layer filter of frame-level waveform model



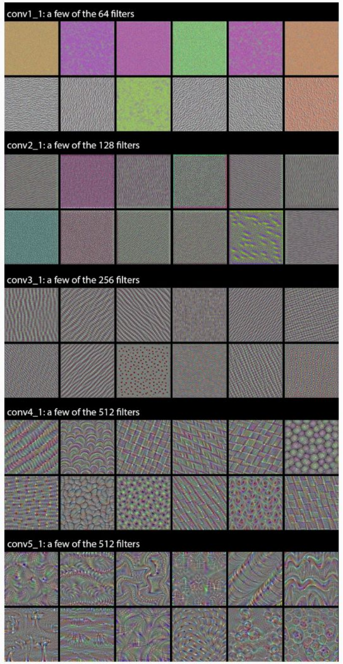


Dieleman and Schrauwen, 2014. "End-to-end learning for music audio", *ICASSP.*

Ardila et al., 2016. "Audio deepdream: Optimizing raw audio with convolutional networks", *ISMIR LBD.*

Learned filter visualization

Activation Maximization

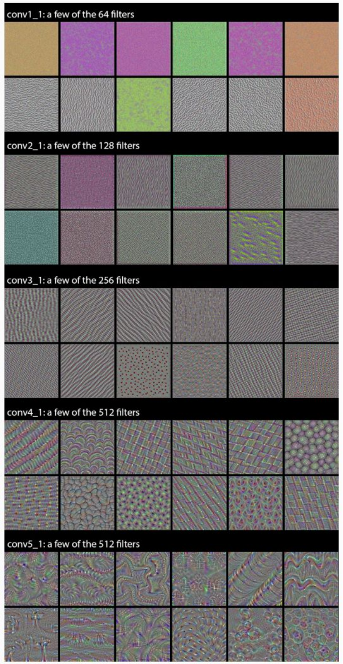


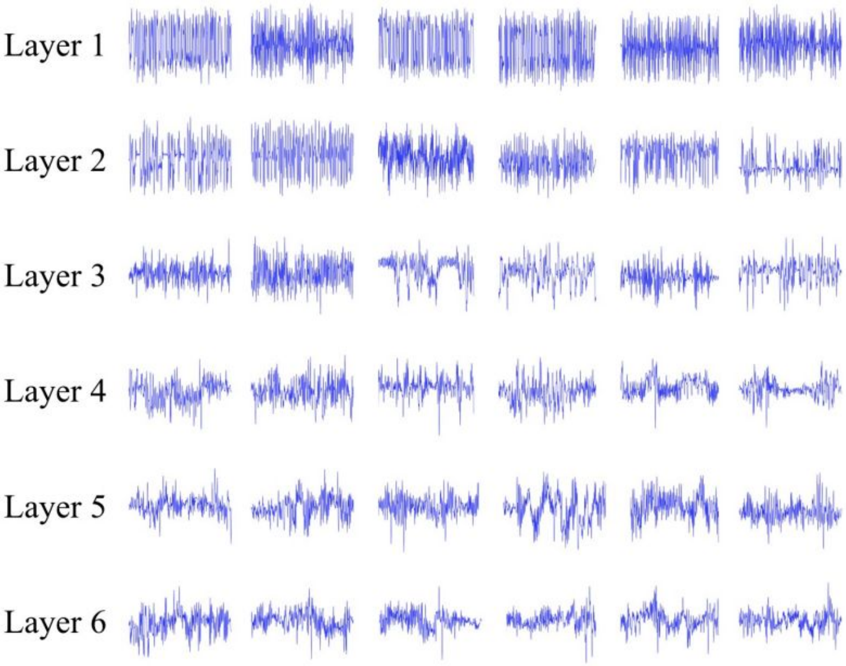
1. Generate random waveform
2. Feed-forward to the target filter
3. Obtain the gradient of the input layer (waveform)
4. Update waveform
5. Repeat

Image borrowed from “https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html”

Erhan et al., 2009. "Visualizing higher-layer features of a deep network", *University of Montreal.*

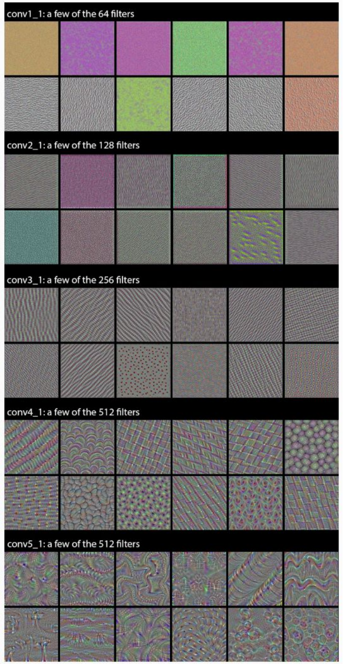
Learned filter visualization

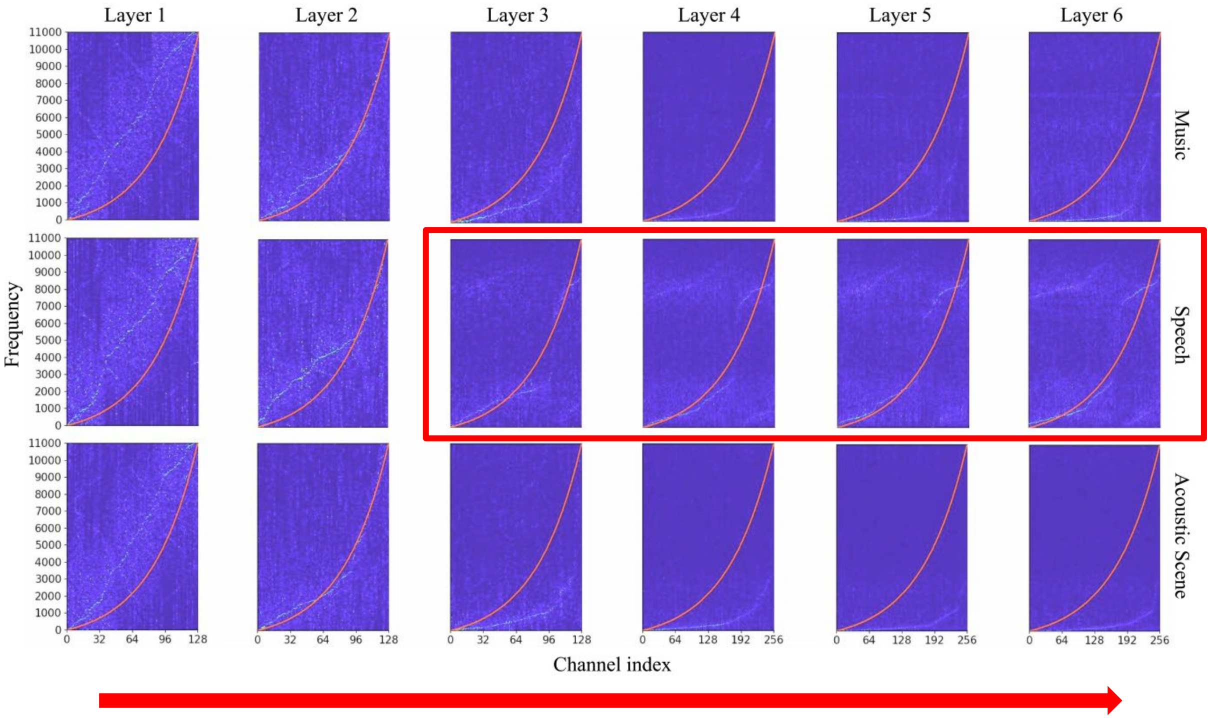




Lee et al., 2017. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms", *SMC*.

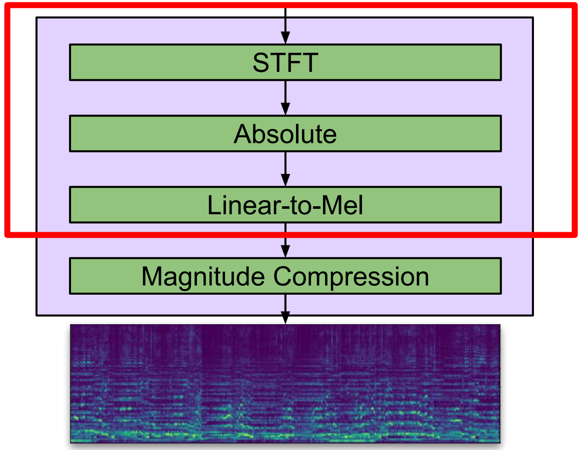
Learned filter visualization

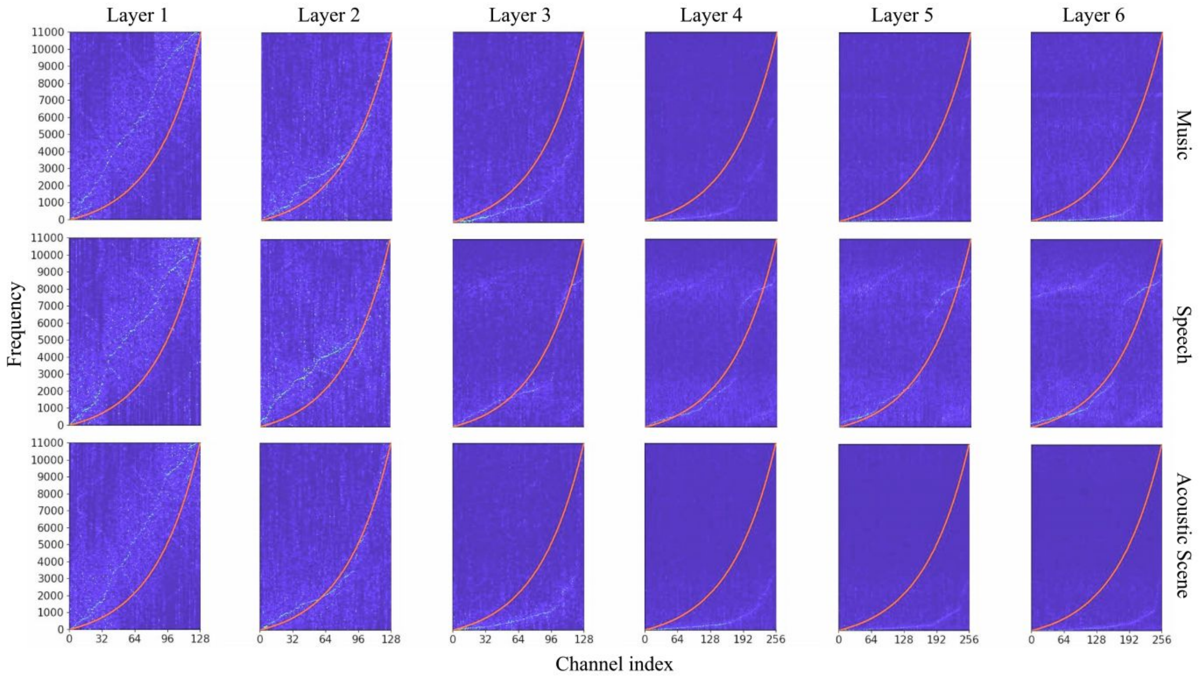




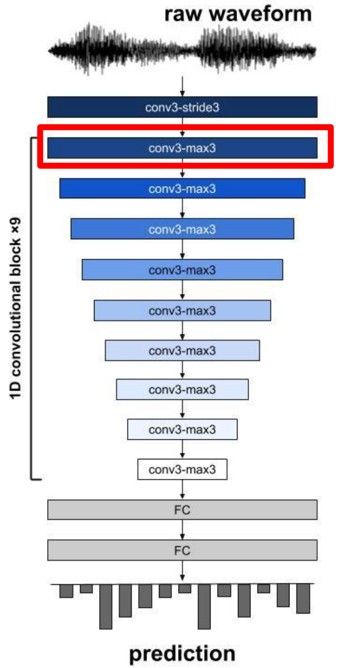
Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", *IEEE Journal of Selected Topics in Signal Processing.*

Learned filter visualization

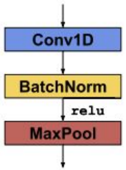




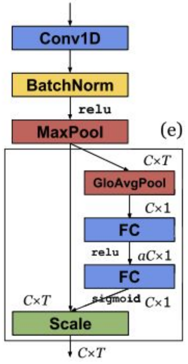
Extended Architecture

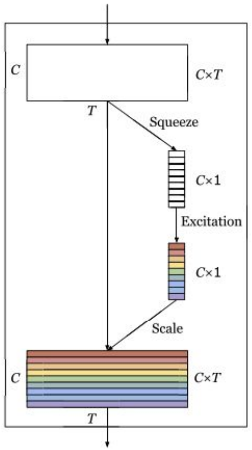


Basic



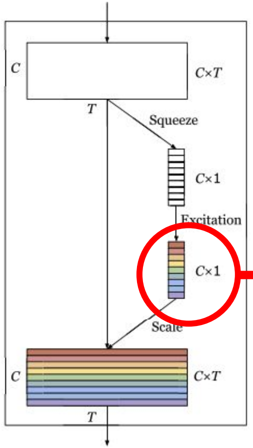
SE





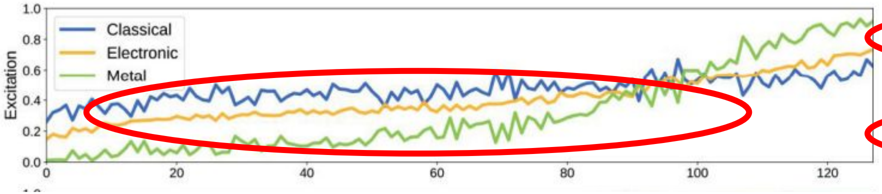
Hu et al., 2018. "Squeeze-and-excitation networks", *CVPR.*

Excitation Analysis



Averaged excitation value of the

first SE block by class.



(x-axis: channel index)

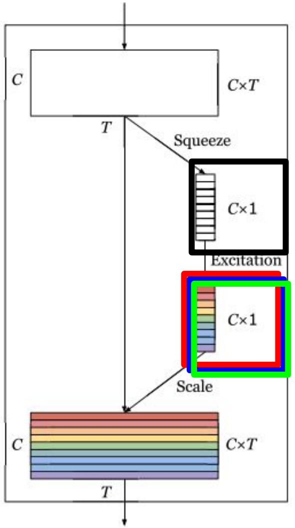
Averaged relative loudness by class.

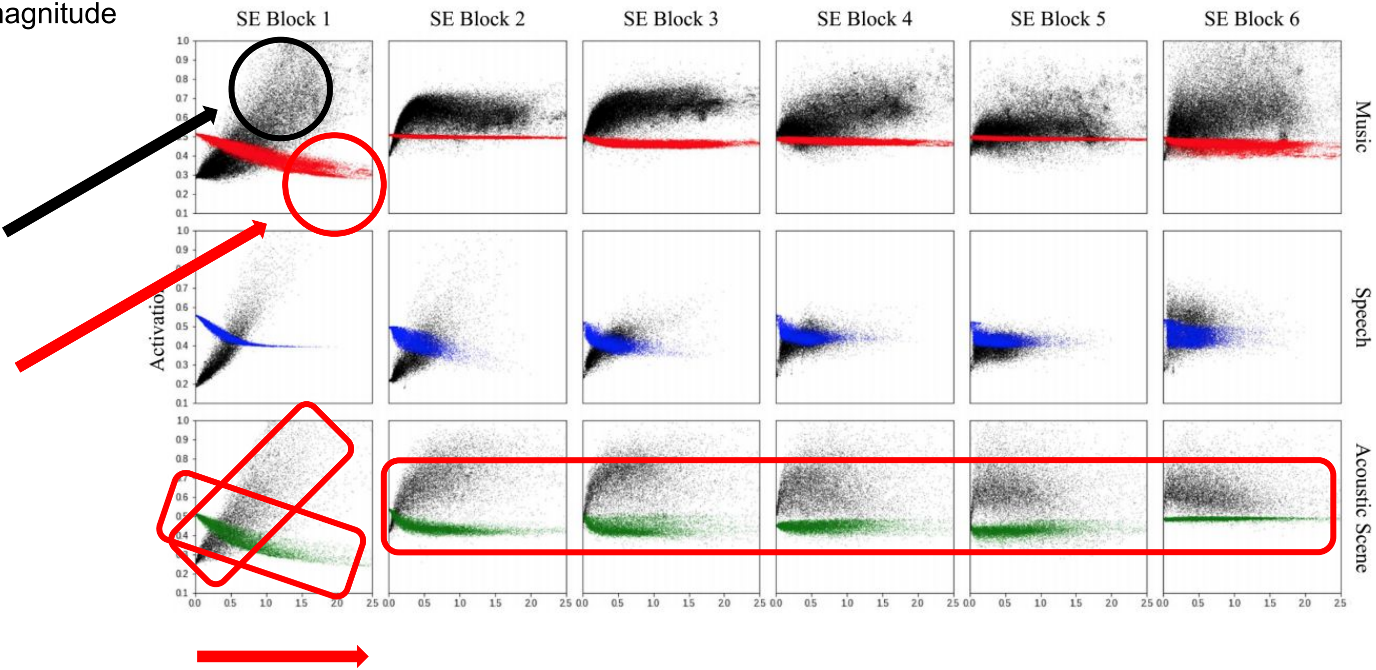
Kim et al., 2019. "Comparison and Analysis of SampleCNN Architectures for Audio Classification", *IEEE Journal of Selected Topics in Signal Processing.*

Excitation Analysis

Colored : Excitation value

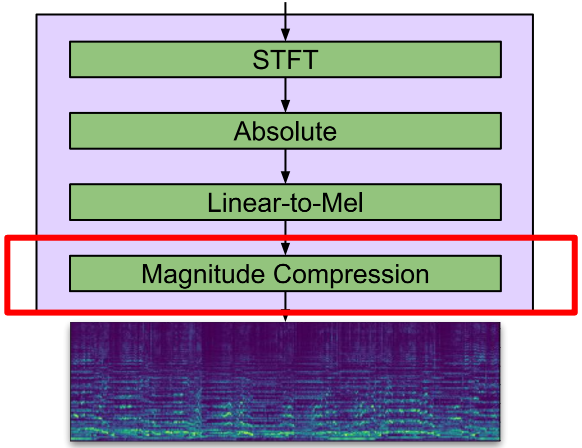
Black : Averaged channel magnitude

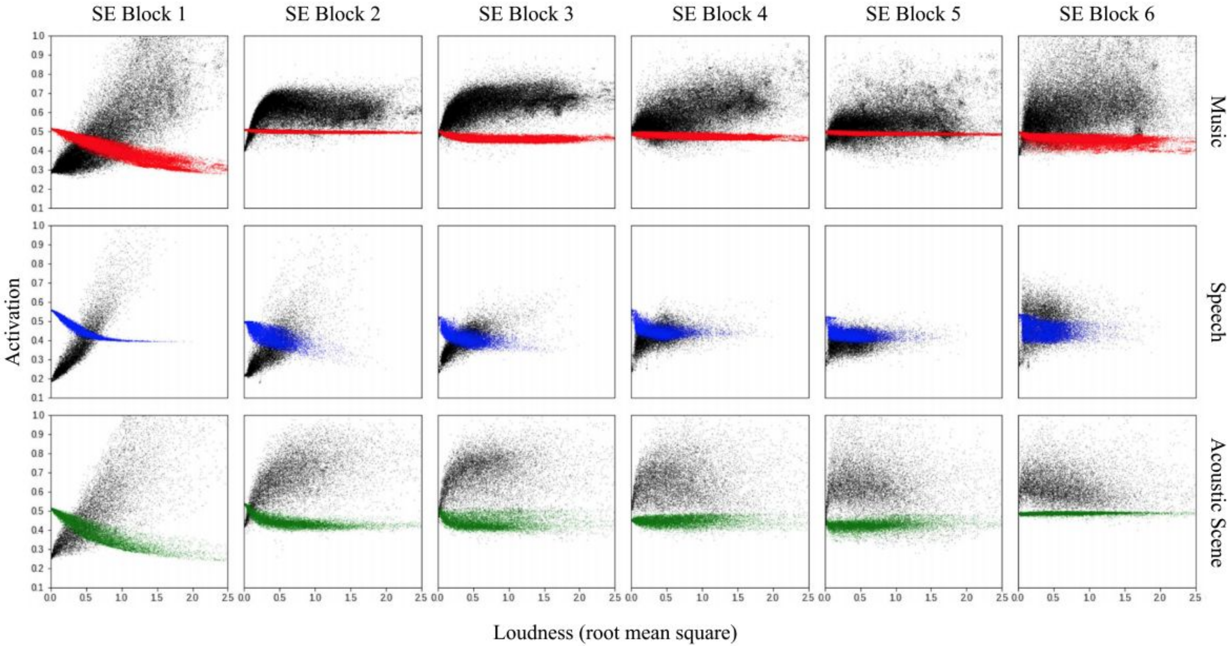




x-axis : loudness (root mean square)

Excitation Analysis



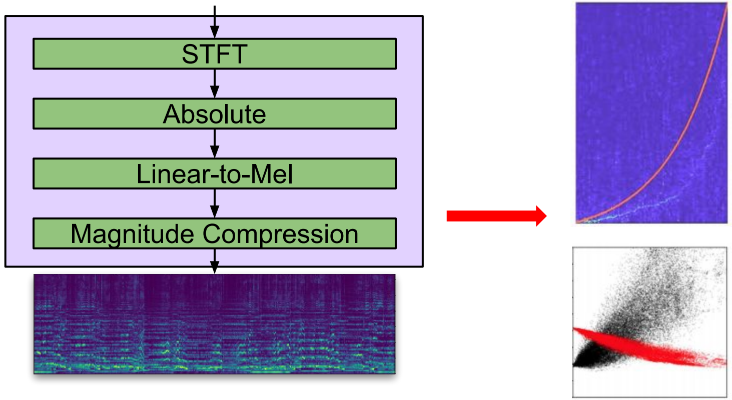


Outline

- Problem statement
- Audio classification models
    - *Engineered feature based model*
    - *Frame-level waveform based model*
    - *Sample-level waveform based model*
- Connect signal processing and deep learning
- **Summary**

Summary

- Music classification
- Audio classification models
- Waveform based model analysis



References

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 13, NO. 2, MAY 2019

Deep Learning for Audio Signal Processing

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath

MUSIC SIGNAL PROCESSING

Juhan Nam, Keunwoo Choi, Jongpil Lee, Szu-Yu Chou, and Yi-Hsuan Yang

Deep Learning for Audio-Based Music Classification and Tagging

Teaching computers to distinguish rock from Bach

END-TO-END LEARNING FOR MUSIC AUDIO TAGGING AT SCALE

Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona

Pandora Media Inc., Oakland, CA

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 13, NO. 2, MAY 2019

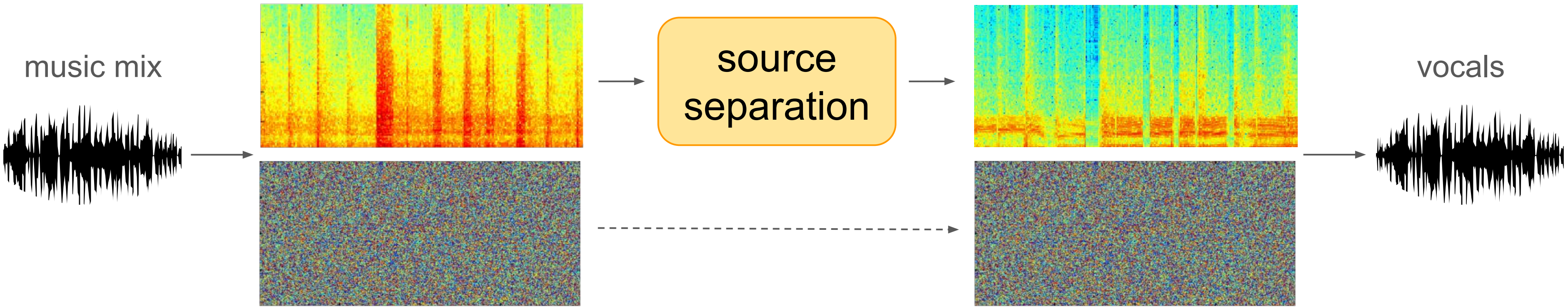
Comparison and Analysis of SampleCNN Architectures for Audio Classification

Taejun Kim, Jongpil Lee, and Juhan Nam, Member, IEEE

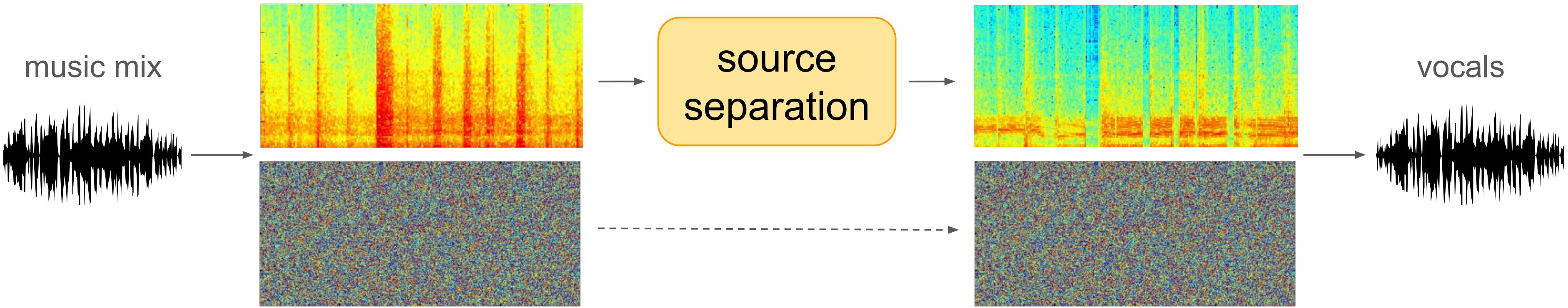
Music
Source Separation

Time to explore end-to-end learning

Why end-to-end music source separation?



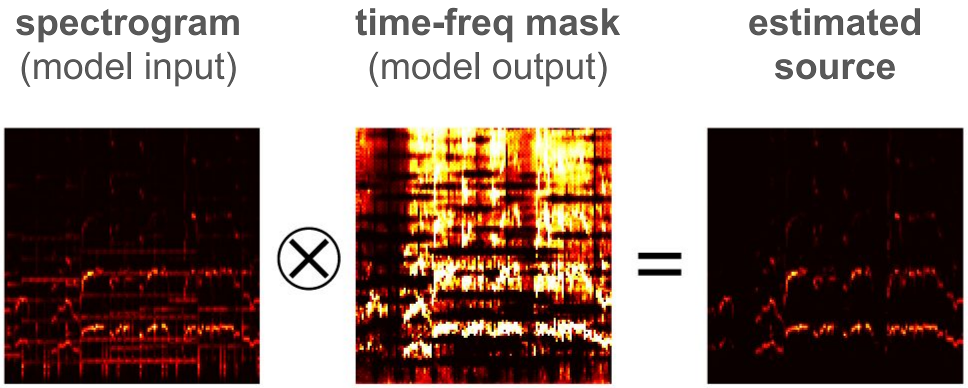
Why end-to-end music source separation?



I) Are we missing crucial information when **discarding the phase**?

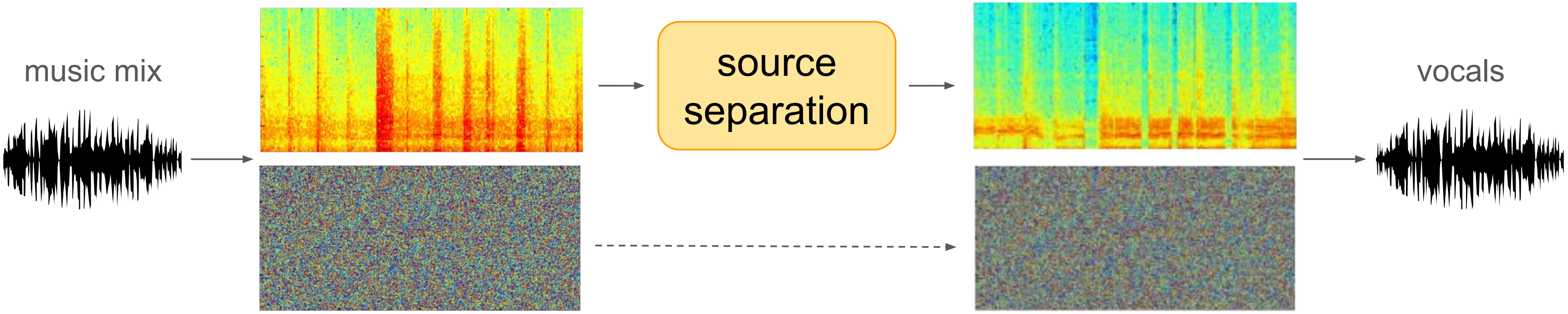
II) When using the **phase of the mixture at synthesis time**,are we introducing artifacts that are limiting our model's performance?

Why filtering spectrograms with masks?



III) **Filtering** spectrograms does **not allow recovering masked (perceptually hidden sound) signals**

Figure from: Jansson et al., 2017. "Singing voice separation with deep U-Net convolutional networks" in ISMIR.

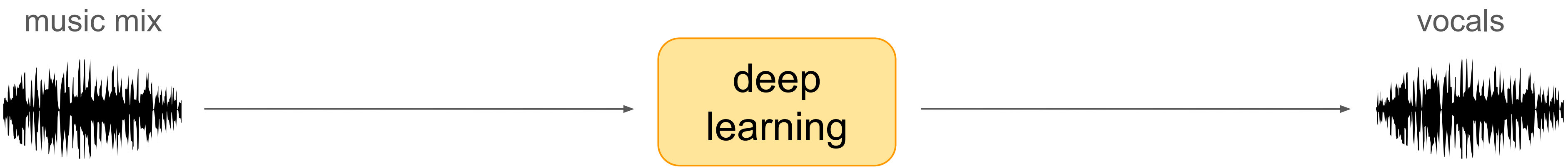


I) Are we missing crucial information when **discarding the phase**?

II) When using the **phase of the mixture at synthesis time**,are we introducing artifacts that are limiting our model's performance?

III) **Filtering** spectrograms does **not allow recovering masked signals**

End-to-end music source separation



Other (active) research directions:

Use the complex STFT as i/o interface?

Kameoka et al., 2009. “ComplexNMF: A new sparse representation for acoustic signals” in ICASSP.

Dubey et al., 2017. “Does phase matter for monaural source separation?” in arXiv.

Le Roux et al., 2019. “Phasebook and friends: Leveraging discrete representations for source separation” in IEEE Journal of Selected Topics in Signal Processing.

Tan et al., 2019. “Complex Spectral Mapping with a CRNN for Monaural Speech Enhancement” in ICASSP.

Liu et al., 2019. “Supervised Speech Enhancement with Real Spectrum Approximation” in ICASSP.

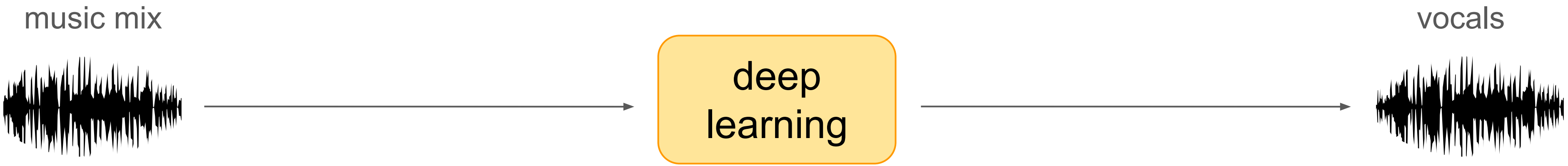
Other (active) research directions:

Alternative models at synthesis time?

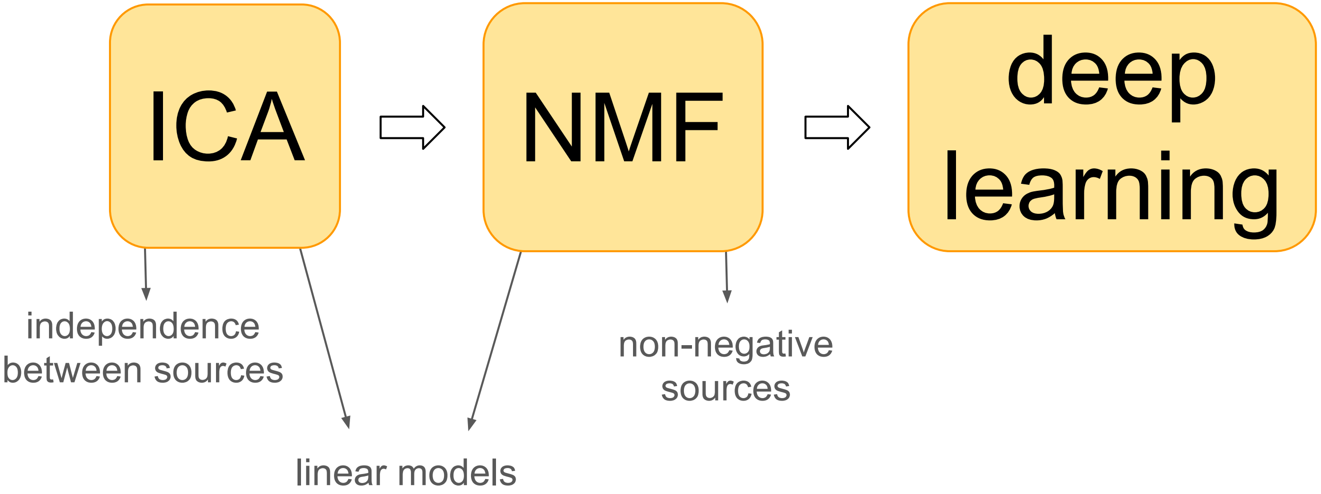
Virtanen and Klapuri, 2000. “Separation of harmonic sound sources using **sinusoidal modeling**,” in ICASSP.

Chandna et al., 2019. “A **vocoder** based method for singing voice extraction” in ICASSP.

End-to-end music source separation



Historical perspective: unsupervised & linear models



Linear model example

linear approximation

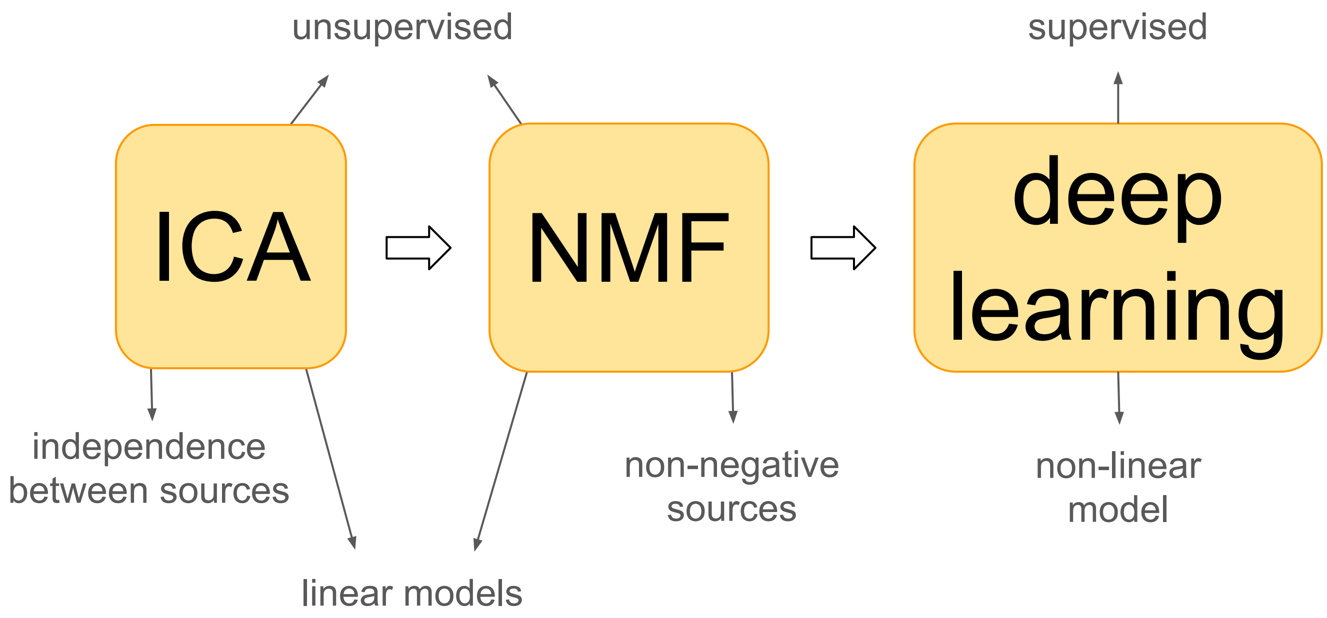
activations

$$X \approx \hat{X} = \sum_k w_k h_k = WH$$

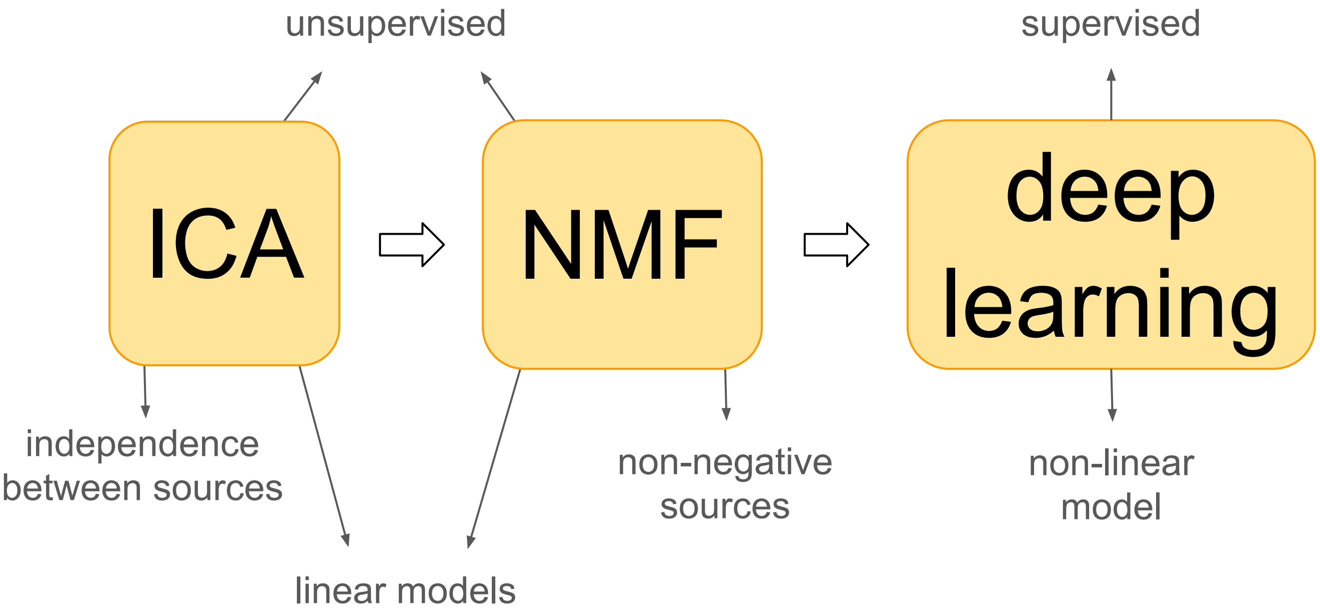
bases

Unsupervised factorization of the mixture
into **bases** (w) and **activations** (h)

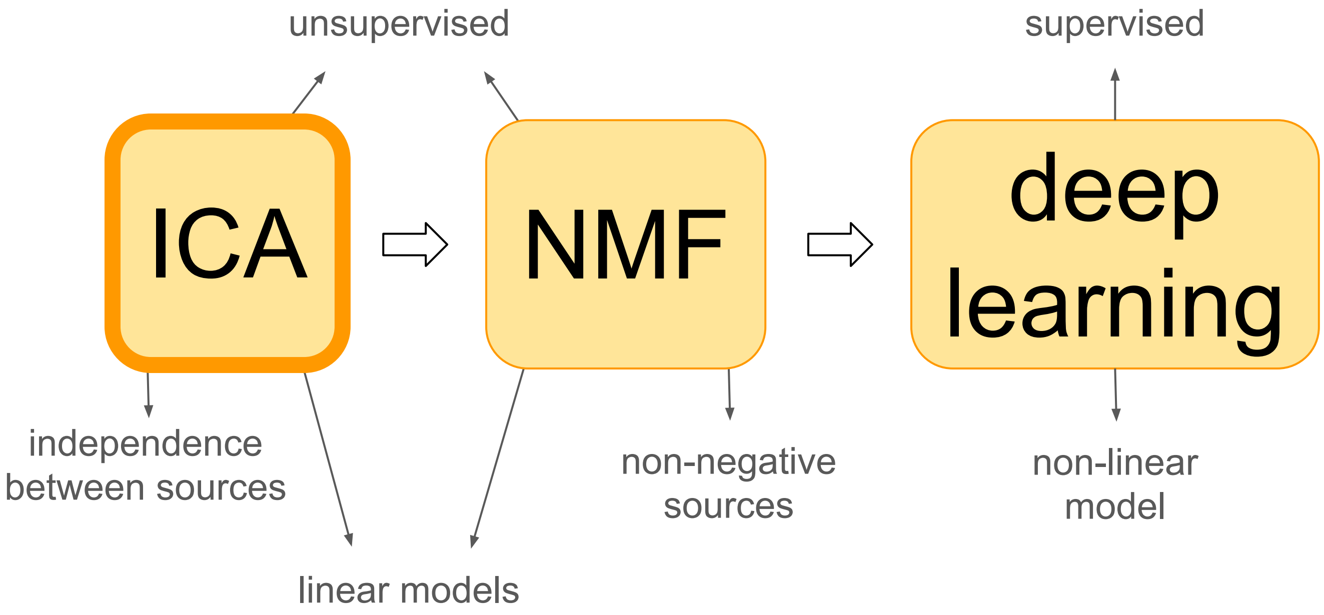
Historical perspective: unsupervised & linear models



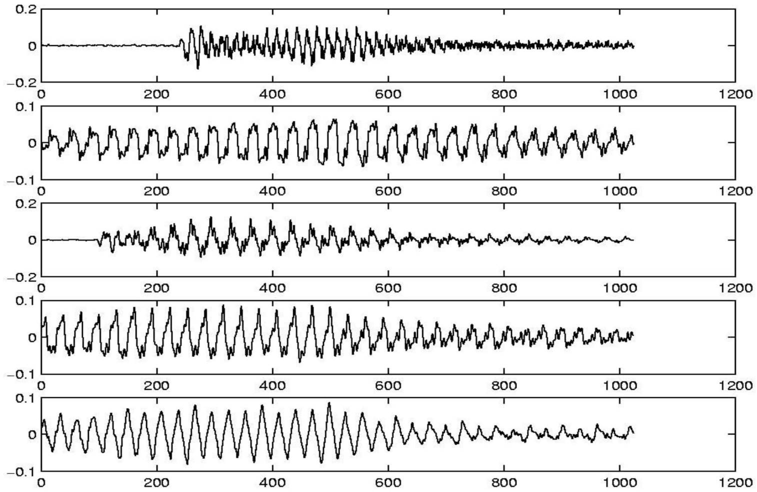
Historical perspective: waveform-based models?



Historical perspective: waveform-based models?



waveform-based ICA



$$X \approx \hat{X} = \sum_k w_k h_k = WH$$

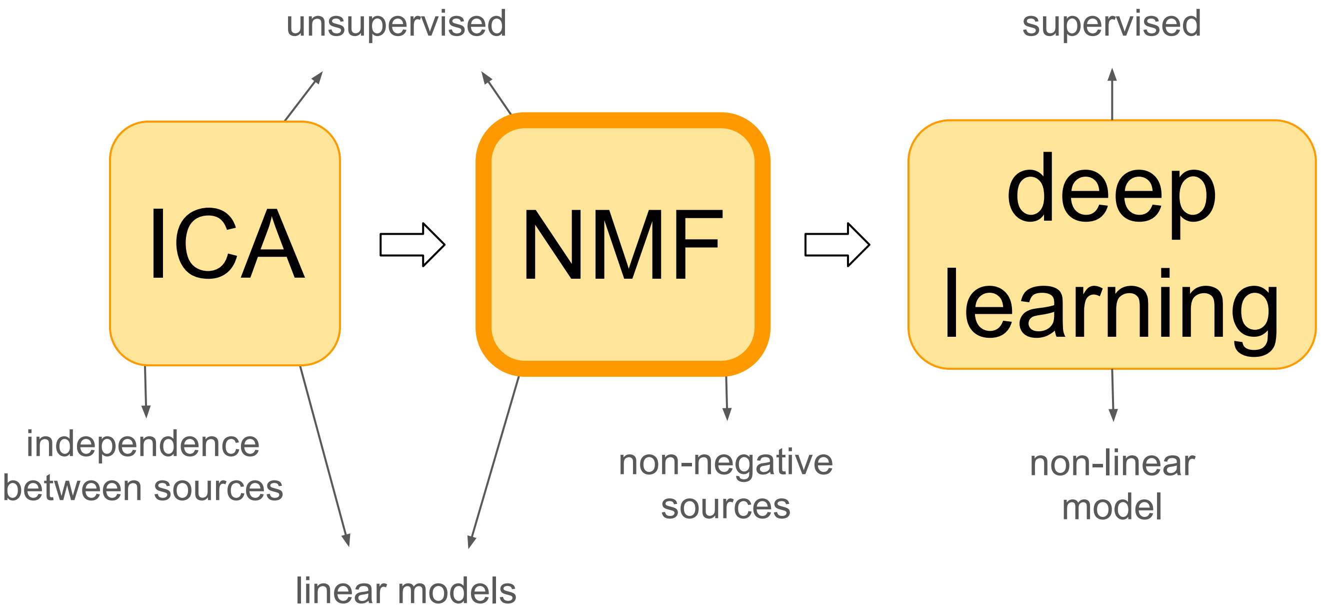
bases

activations

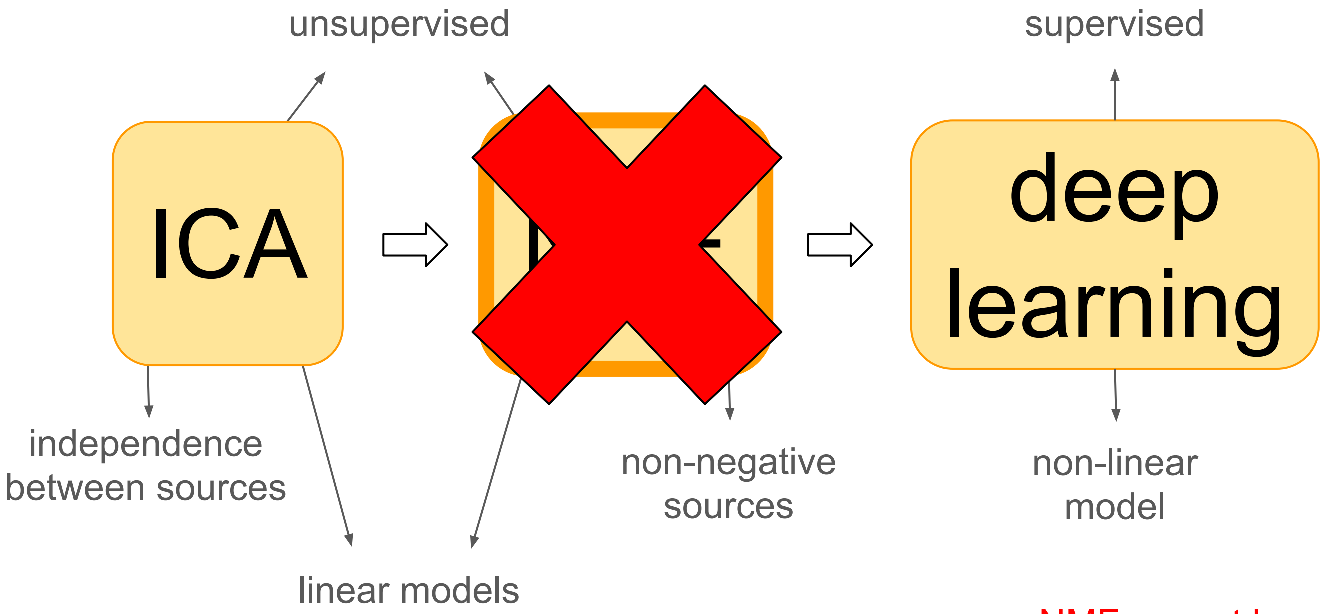
Problem 1: phase sensitive basis
Problem 2: simplicity of the linear model

Figure from: Blumensath and Davies, 2004. “Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music,” in ICASSP.

Historical perspective: waveform-based models?

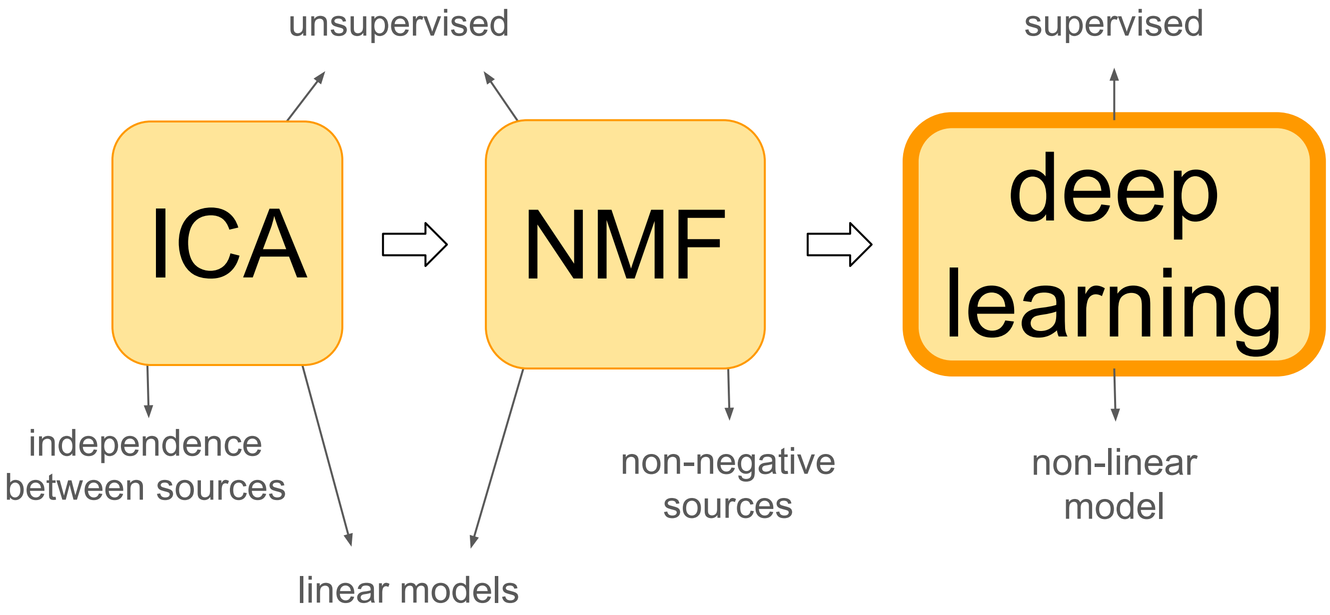


Historical perspective: waveform-based models?



NMF cannot be used with waveforms
due to its non-negative constraint!
(waveforms range from -1 to 1)

Historical perspective: waveform-based models?



A widely-used set of tools:

filtering spectrograms

linear models

unsupervised learning

audio domain knowledge

..maybe we could try another toolset?

~~filtering~~ → synthesis?

~~linear models~~ → non-linear models?

~~unsupervised learning~~ → supervised learning?

~~audio domain knowledge~~ → data driven?

End-to-end music source separation: 9 publications

Stoller et al., 2018. “Wave-u-net: A multi-scale neural network for end-to-end audio source separation” in arXiv.

Grais et al., 2018. “Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders” in EUSIPCO.

Lluis, et al., 2018. “End-to-end music source separation: is it possible in the waveform domain?” in arXiv.

Slizovskaia et al., 2018. “End-to-end Sound Source Separation Conditioned on Instrument Labels” in arXiv.

Cohen-Hadria et al., 2019. “Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation” in arXiv.

Kaspersen, 2019. “HydraNet: A Network For Singing Voice Separation”. Master Thesis.

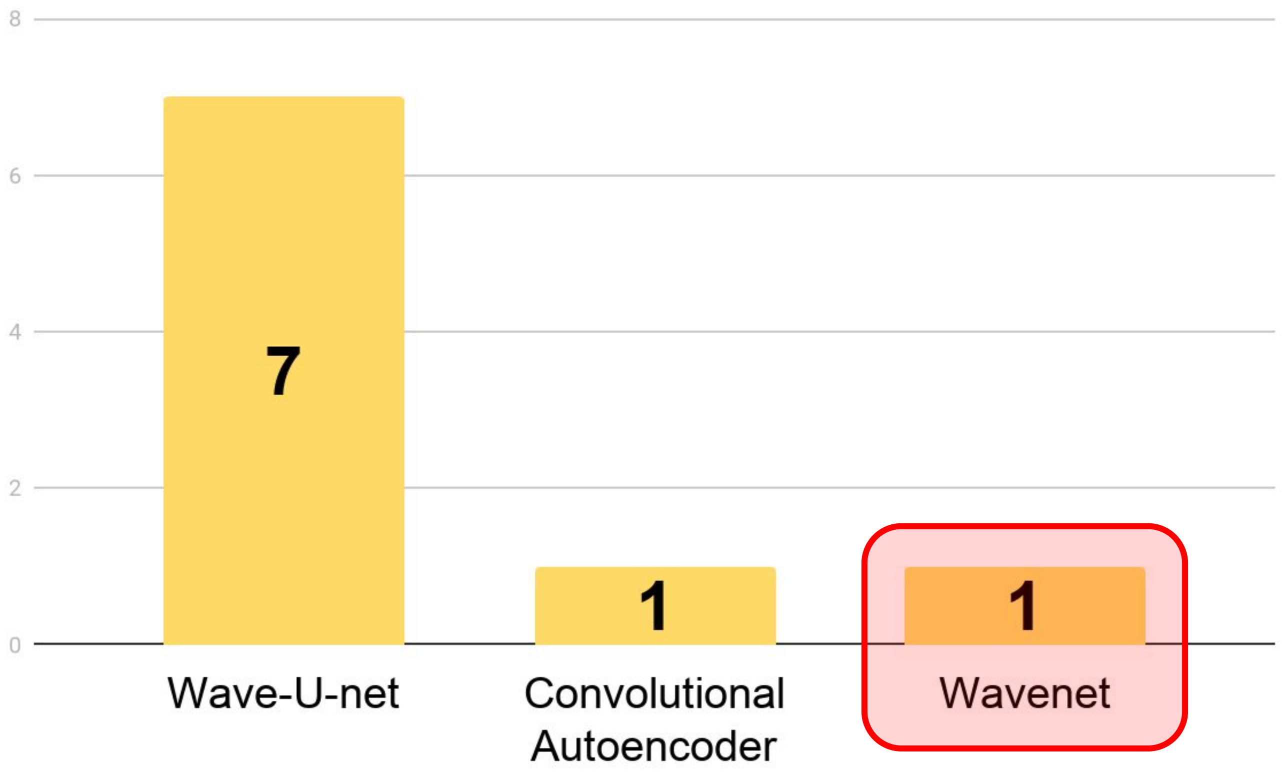
Akhmetov et al., 2019. “Time Domain Source Separation with Spectral Penalties”. Technical Report.

Défossez et al., 2019. “Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed” in arXiv.

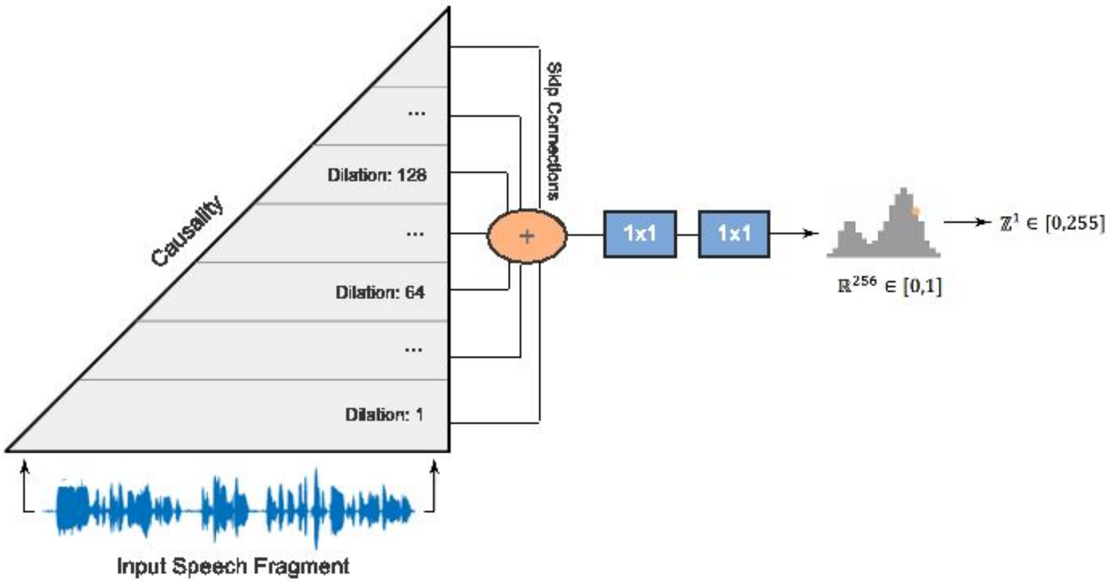
Narayanaswamy et al., 2019. “Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets” in arXiv.

ALL THE PUBLICATIONS IN CHRONOLOGICAL ORDER AS OF OCTOBER 2019

End-to-end music source separation: architectures



Introduction: the “generative” Wavenet

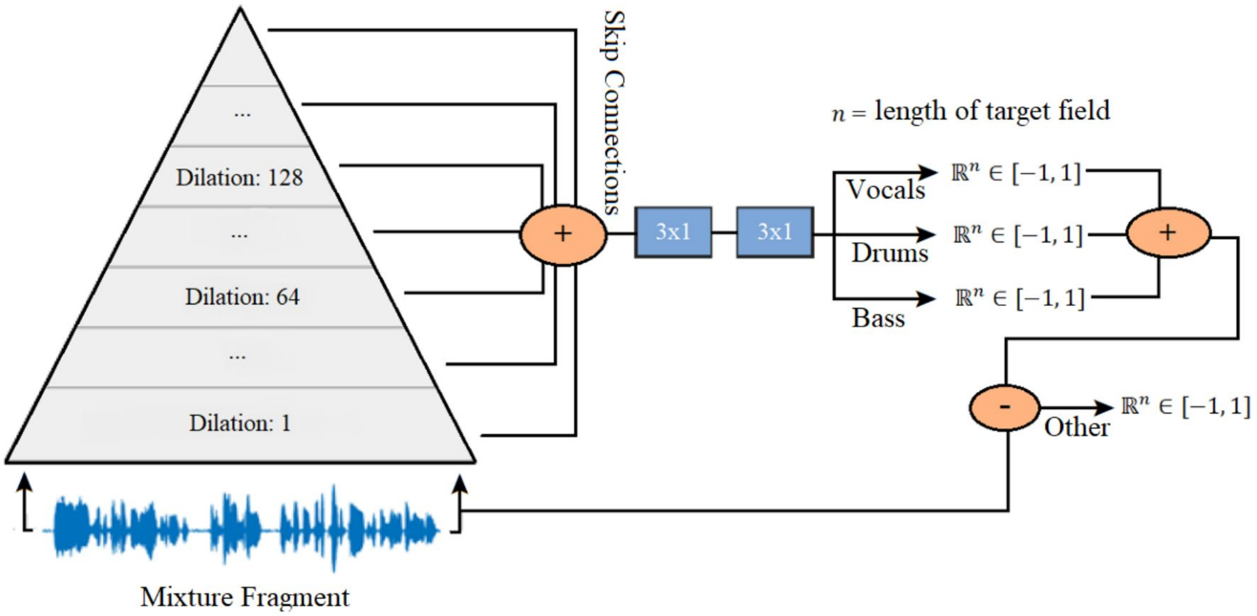


Causal

Softmax-output distribution *modeling p(x)*

Van den Oord et al., 2016. “Wavenet: a generative model for audio” in arXiv.

A “regression” Wavenet for music source separation

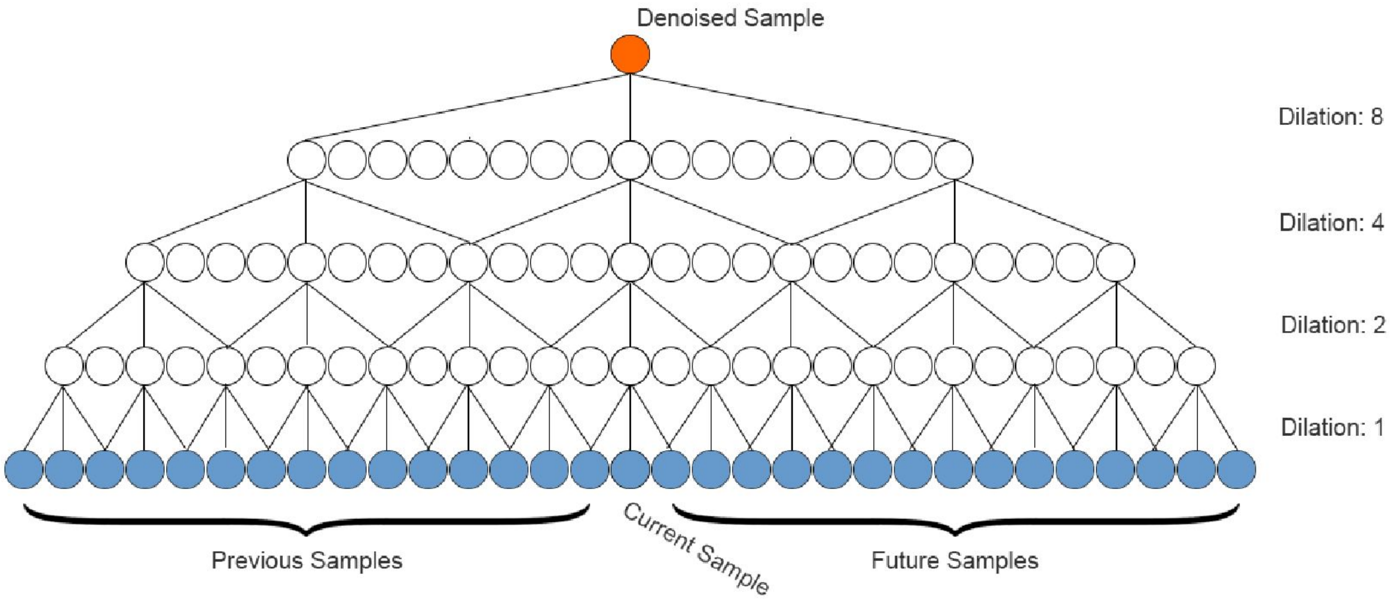


Non-causal

Regression output
p(y|x) discriminative
post-processing!

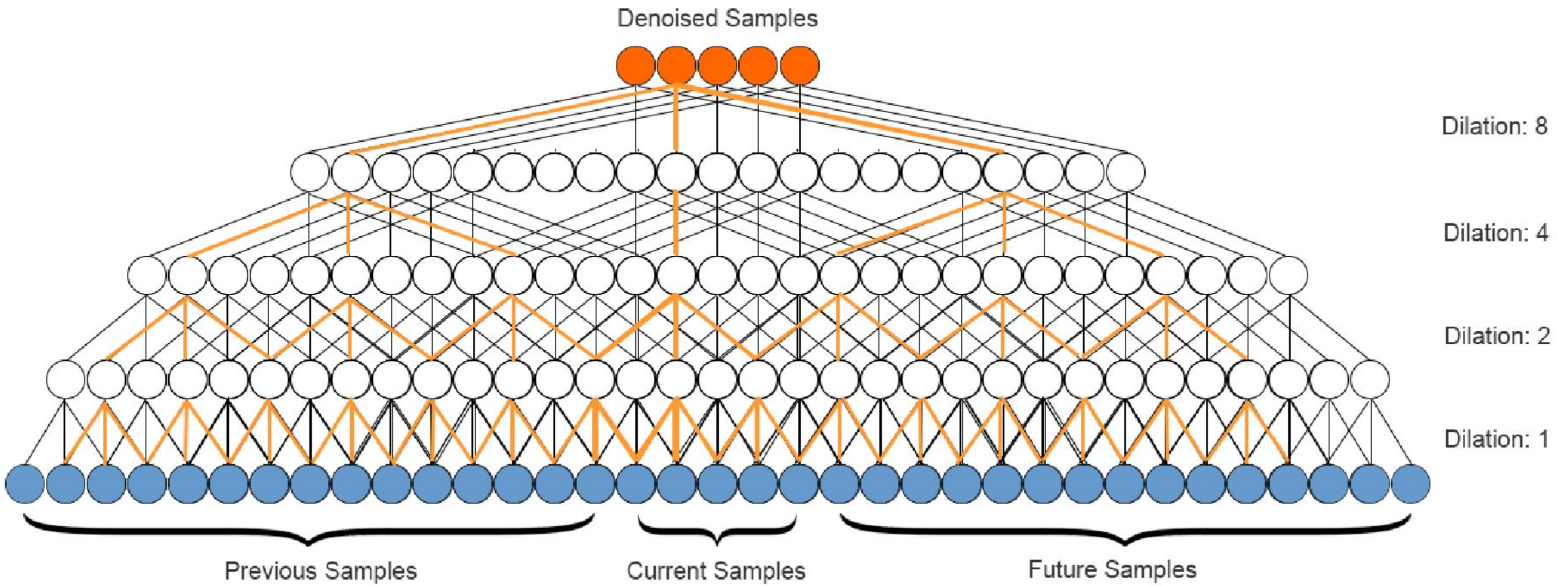
Lluis, et al., 2019. “End-to-end music source separation: is it possible in the waveform domain?” in Interspeech.

Fully convolutional & deterministic



Lluis, et al., 2019. “End-to-end music source separation: is it possible in the waveform domain?” in Interspeech.

Fully convolutional & deterministic

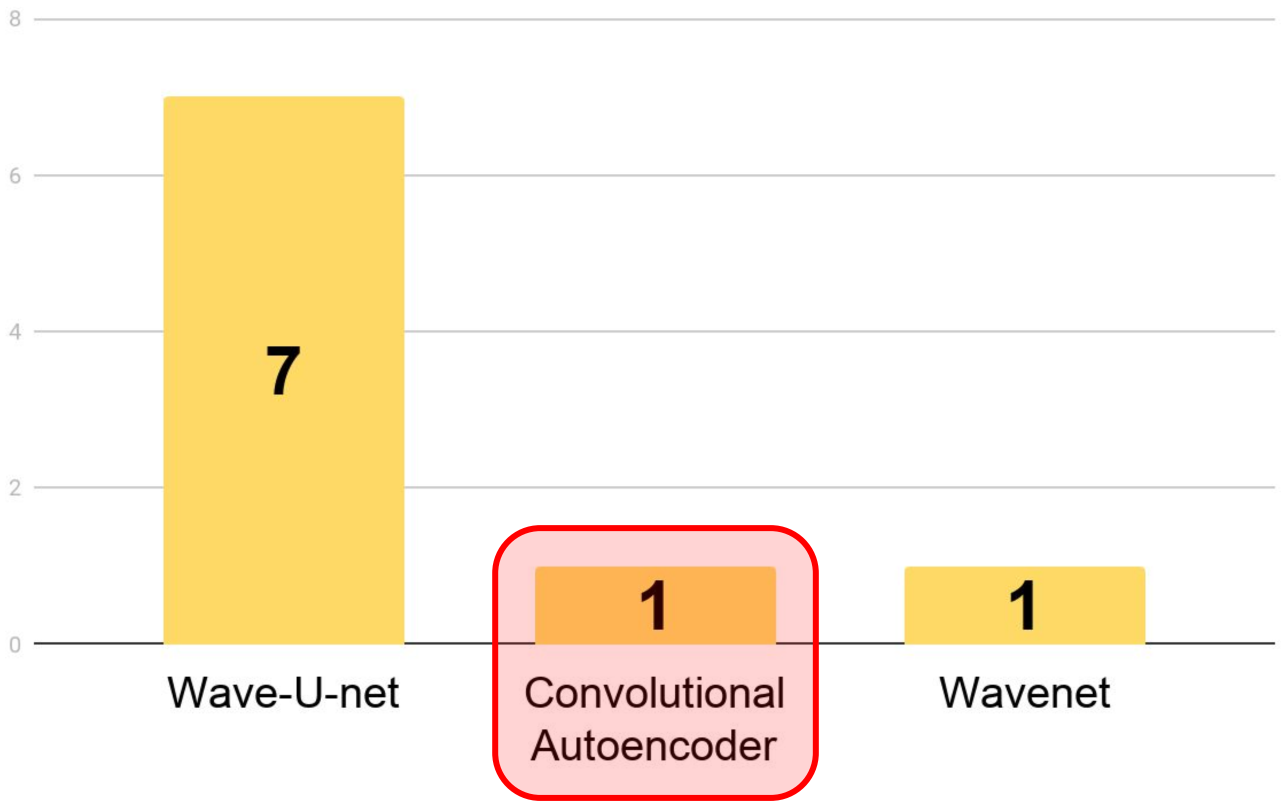


Real time inference!

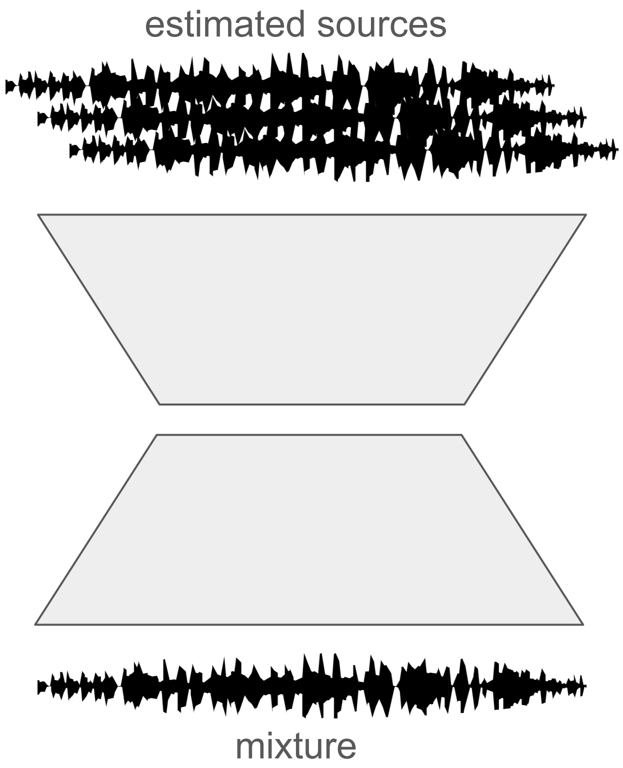
1601 samples input ⟶ denoising time: ≈ 0.56 sec per second of music on GPU!

Lluis, et al., 2019. “End-to-end music source separation: is it possible in the waveform domain?” in Interspeech.

End-to-end music source separation: architectures

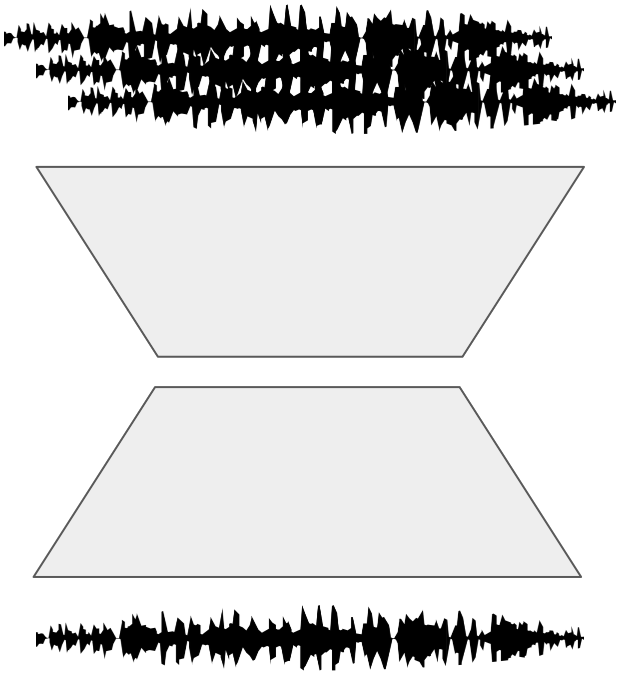


Autoencoders

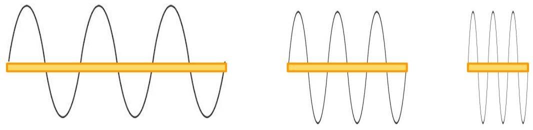


Multi-resolution & Convolutional autoencoder

estimated sources



mixture

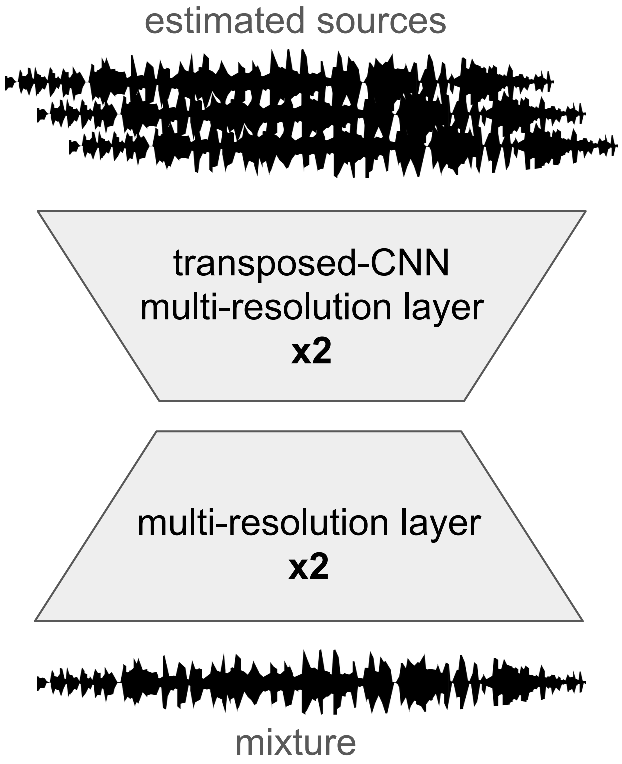


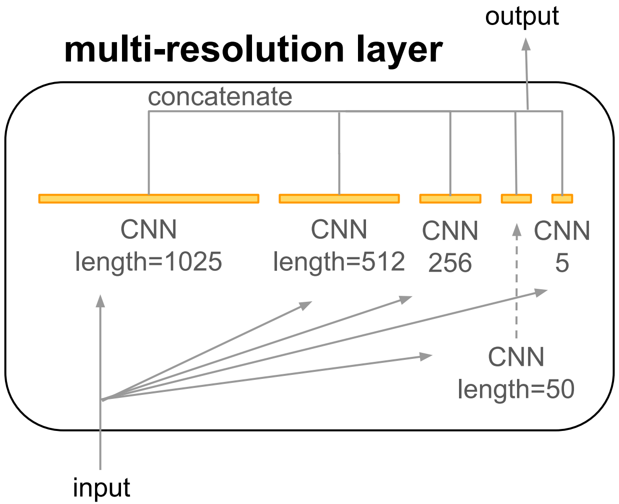
Multi-resolution CNN: efficient way to represent 3 periods!

Multi-resolution CNN = Inception CNN (different filter shapes in the same CNN layer)

Grais et al., 2018. “Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders” in EUSIPCO.

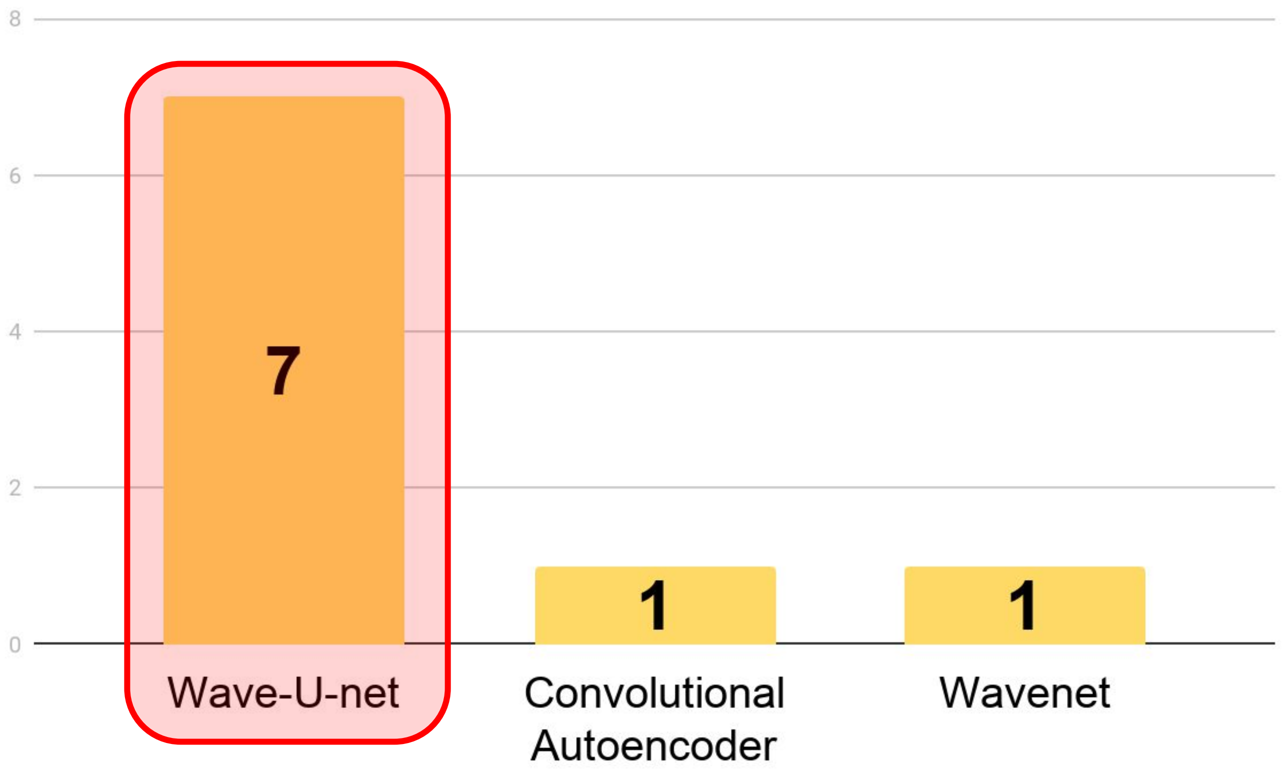
Multi-resolution & Convolutional autoencoder



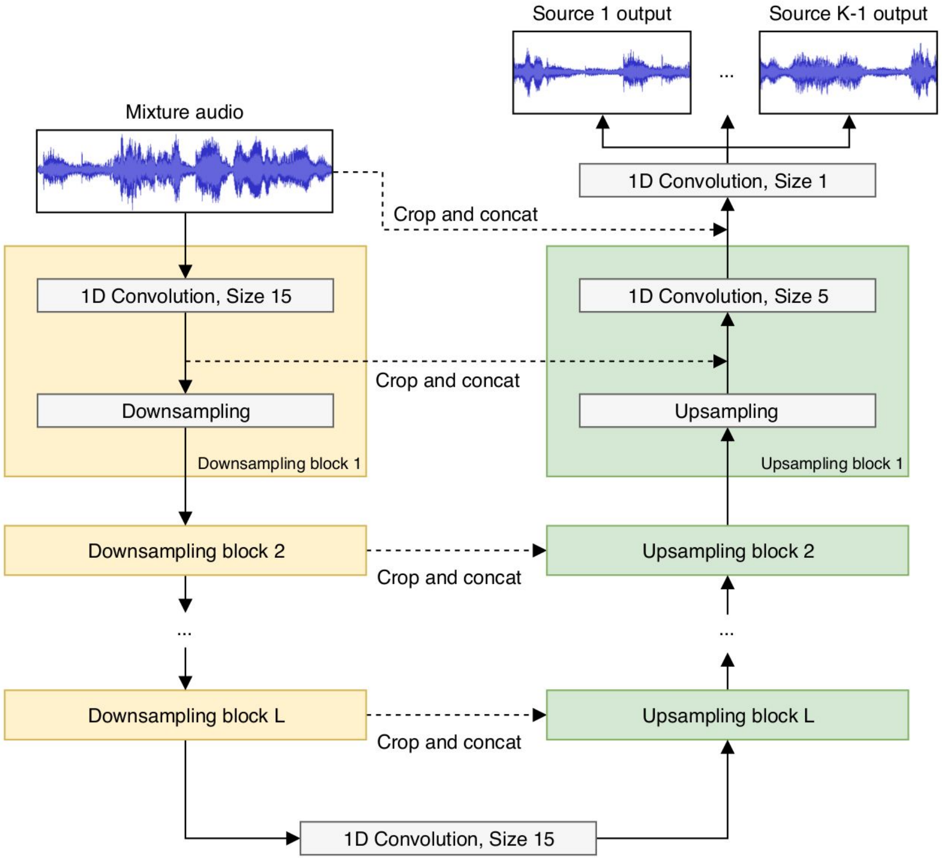


Grais et al., 2018. "Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders" in EUSIPCO.

End-to-end music source separation: architectures



Wave-U-net



Stoller et al., 2018. “Wave-u-net: A multi-scale neural network for end-to-end audio source separation” in arXiv.

Wave-u-net extensions

- Multiplicative conditioning using instrument labels at the bottleneck.

  Slizovskaia et al., 2019. “End-to-end Sound Source Separation Conditioned on Instrument Labels” in ICASSP.

- Data augmentation: ≈ 1 dB SDR improvement.

  Cohen-Hadria et al., 2019. “Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation” in arXiv.

Data augmentation strategies

It is used to **artificially expand the size of a training dataset** by creating modified versions of it.

- Random swapping left/right channel for each source
- Random scaling sources
- Random mixing of sources from different songs
- Pitch-shifting
- Time-stretching

Uhlich et al, 2017. “Improving music source separation based on deep neural networks through data augmentation and network blending” in ICASSP.

Cohen-Hadria et al., 2019. “Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation” in arXiv.

Wave-u-net extensions

- Multiplicative conditioning using instrument labels at the bottleneck.

  Slizovskaia et al., 2019. “End-to-end Sound Source Separation Conditioned on Instrument Labels” in ICASSP.

- Data augmentation: ≈ 1 dB SDR improvement.

  Cohen-Hadria et al., 2019. “Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation” in arXiv.

- Add BiLSTMs at the bottleneck: ≈ 1 dB SDR improvement.

  Kaspersen, 2019. “HydraNet: A Network For Singing Voice Separation”. Master Thesis.

- Loss function in the spectral domain.

  Akhmetov et al., 2019. “Time Domain Source Separation with Spectral Penalties”. Technical Report.

- Use dilated convolutions and dense CNNs.

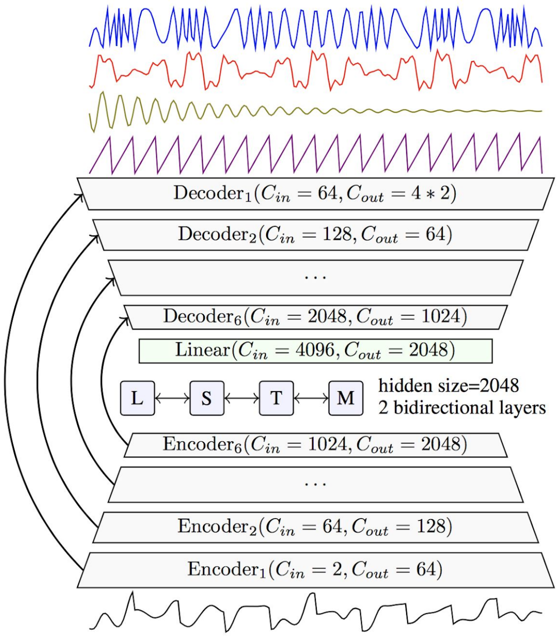
  Narayanaswamy et al., 2019. “Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets” in arXiv.

- Achieve comparable results to a spectrogram-based model: Demucs.

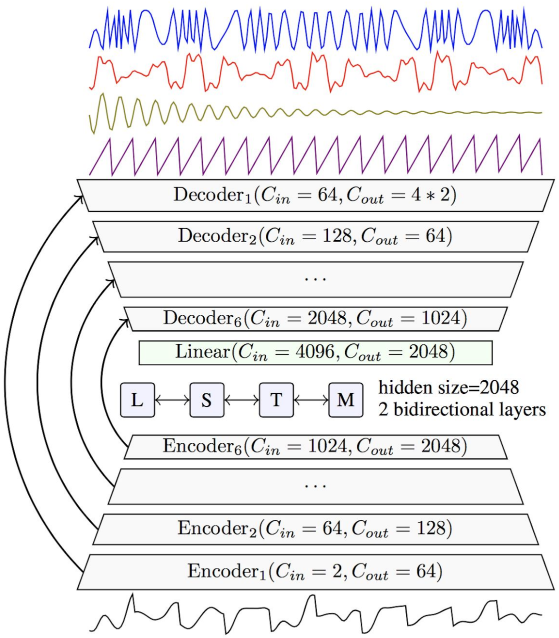
  w/ BiLSTMs at the bottleneck, data augmentation, and some additional architectural changes: ≈ 1.6 dB SDR improvement.

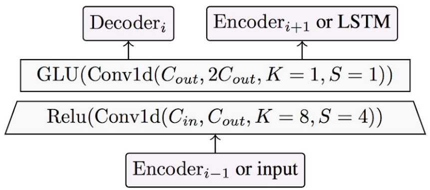
  Défossez et al., 2019. “Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed” in arXiv.

Wave-u-net extensions: Demucs

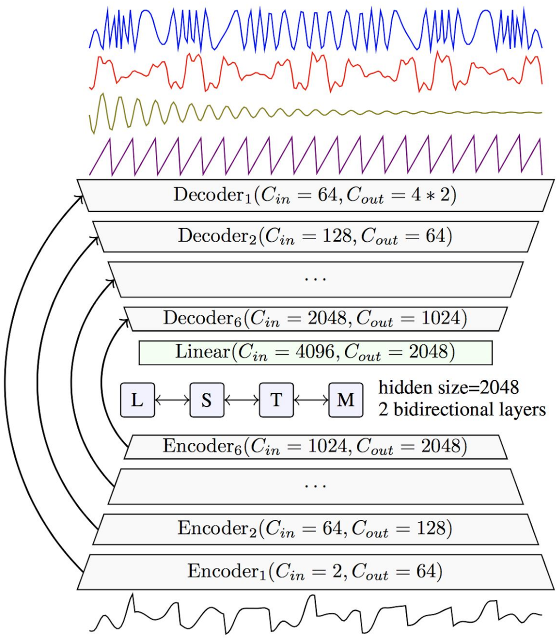


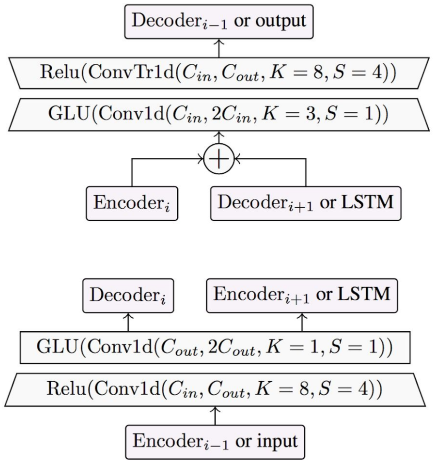
Wave-u-net extensions: Demucs





Wave-u-net extensions: Demucs



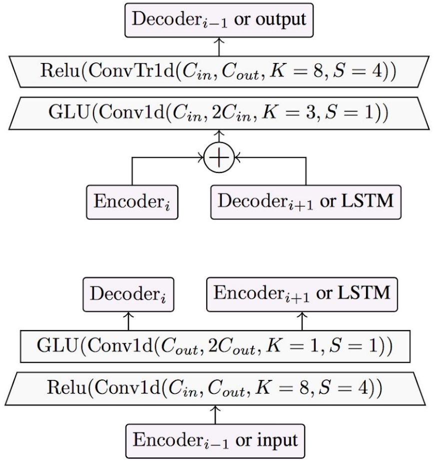


Wave-u-net extensions: Wave-U-net vs. Demucs

Block	Operation	Shape
	Input	$(16384, 1)$
DS, repeated for $i = 1, \ldots, L$	`Conv1D`$(F_c \cdot i, f_d)$ `Decimate`	 $(4, 288)$
	`Conv1D`$(F_c \cdot (L+1), f_d)$	$(4, 312)$
US, repeated for $i = L, \ldots, 1$	`Upsample` `Concat`(DS block i) `Conv1D`$(F_c \cdot i, f_u)$	 $(16834, 24)$
	`Concat`(Input) `Conv1D`$(K, 1)$	$(16834, 25)$ $(16834, 2)$

Wave-U-net: building blocks

ENCODER



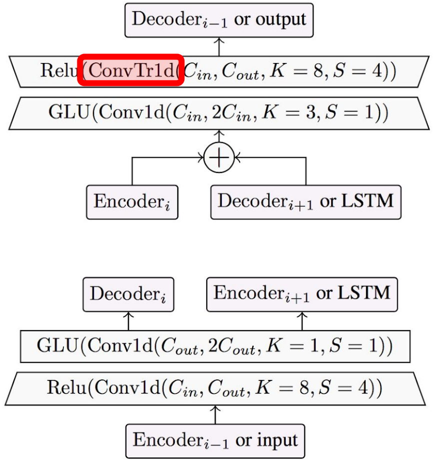
Demucs: building blocks

Wave-u-net extensions: Wave-U-net vs. Demucs

Block	Operation	Shape
	Input	$(16384, 1)$
DS, repeated for $i = 1, \ldots, L$	Conv1D($F_c \cdot i$, f_d) Decimate	 $(4, 288)$
	Conv1D($F_c \cdot (L+1)$, f_d)	$(4, 312)$
US, repeated for $i = L, \ldots, 1$	Upsample Concat(DS block i) Conv1D($F_c \cdot i$, f_u)	 $(16834, 24)$
	Concat(Input) Conv1D(K, 1)	$(16834, 25)$ $(16834, 2)$

DECODER

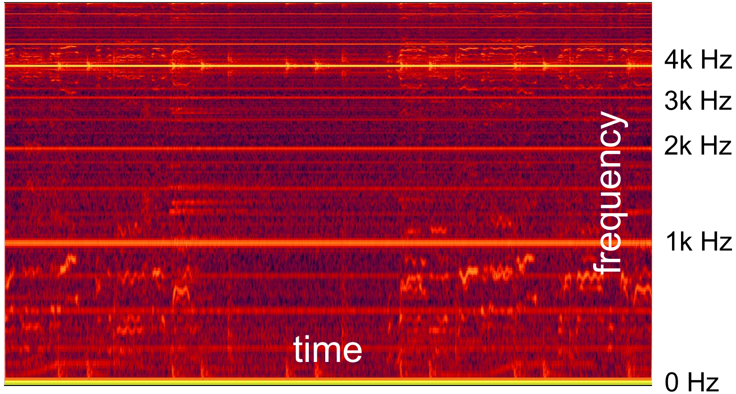
Wave-U-net: building blocks



Demucs: building blocks

Deconvolutions and high-frequency artifacts

Checkerboard artifacts in images



High-frequency buzz in audio

Odena et al., 2016. "Deconvolution and Checkerboard Artifacts" in Distill.

Comparison: a perceptual study

Lluis, et al., 2019. “End-to-end music source separation: is it possible in the waveform domain?” in Interspeech.

Mixture

Spectrogram-based (DeepConvSep)

Waveform-based (Wavenet)

Waveform-based (Wave-U-net)

Evaluation metrics: SDR, SIR, SAR

$$\text{SDR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}$$

"overall performance"

$$\text{SIR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}$$

"interference from other sources"

$$\text{SAR} := 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}$$

"algorithmic artifacts"

http://craffel.github.io/mir_eval/

https://github.com/sigsep/sigsep-mus-eval/

Vincent et al., 2006. "Performance measurement in blind audio source separation" in IEEE TASLP.

Comparison: a perceptual study

Lluis, et al., 2019. “End-to-end music source separation: is it possible in the waveform domain?” in Interspeech.

Mixture

Spectrogram-based (DeepConvSep)

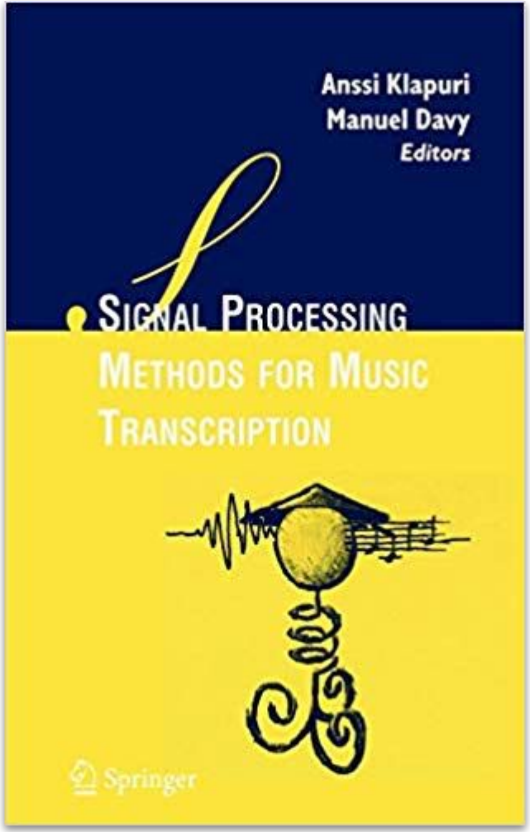
Waveform-based (Wavenet)

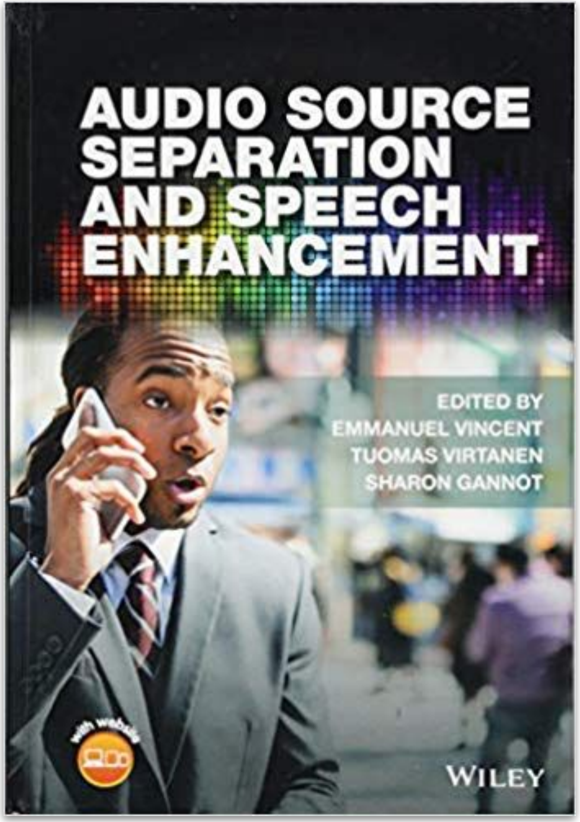
Waveform-based (Wave-U-net)

MOS	*Wavenet-based*	*Wave-U-Net*
Vocals	3.0 ± 1.0	3.3 ± 0.85

(t-test: p-value=0.049, 15 participants)

Additional references





S. Dubnov, 2002. “Extracting sound objects by independent subspace analysis” in AES Conference.

Blumensath and Davies, 2004. “Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music,” in ICASSP.

Jang and Lee, 2003. “A maximum likelihood approach to single-channel source separation” in Journal of Machine Learning Research.

Music Generation

Expressive music modeling

Overview

Why audio? Why raw audio?

Generative models

Likelihood-based models of raw audio

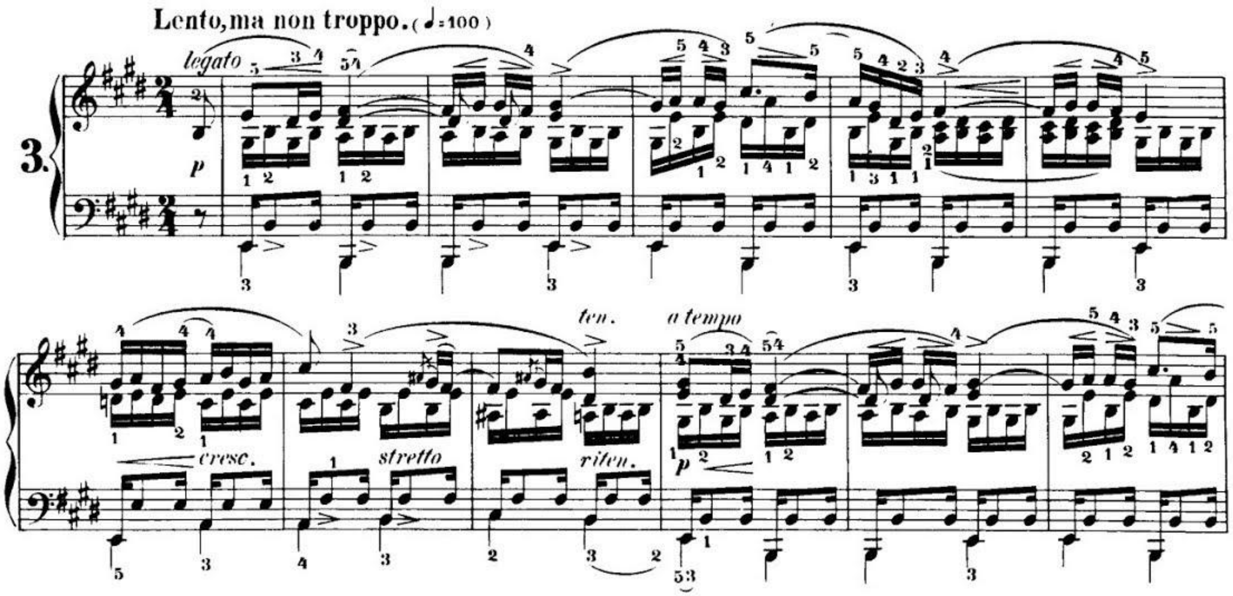
Adversarial models of raw audio

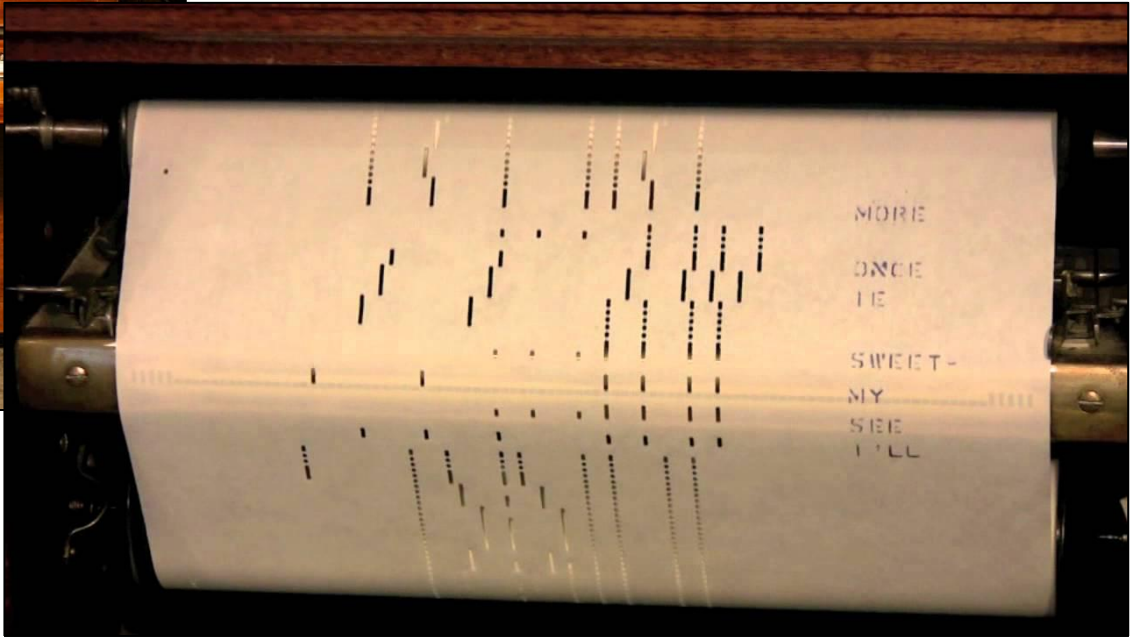
Summary

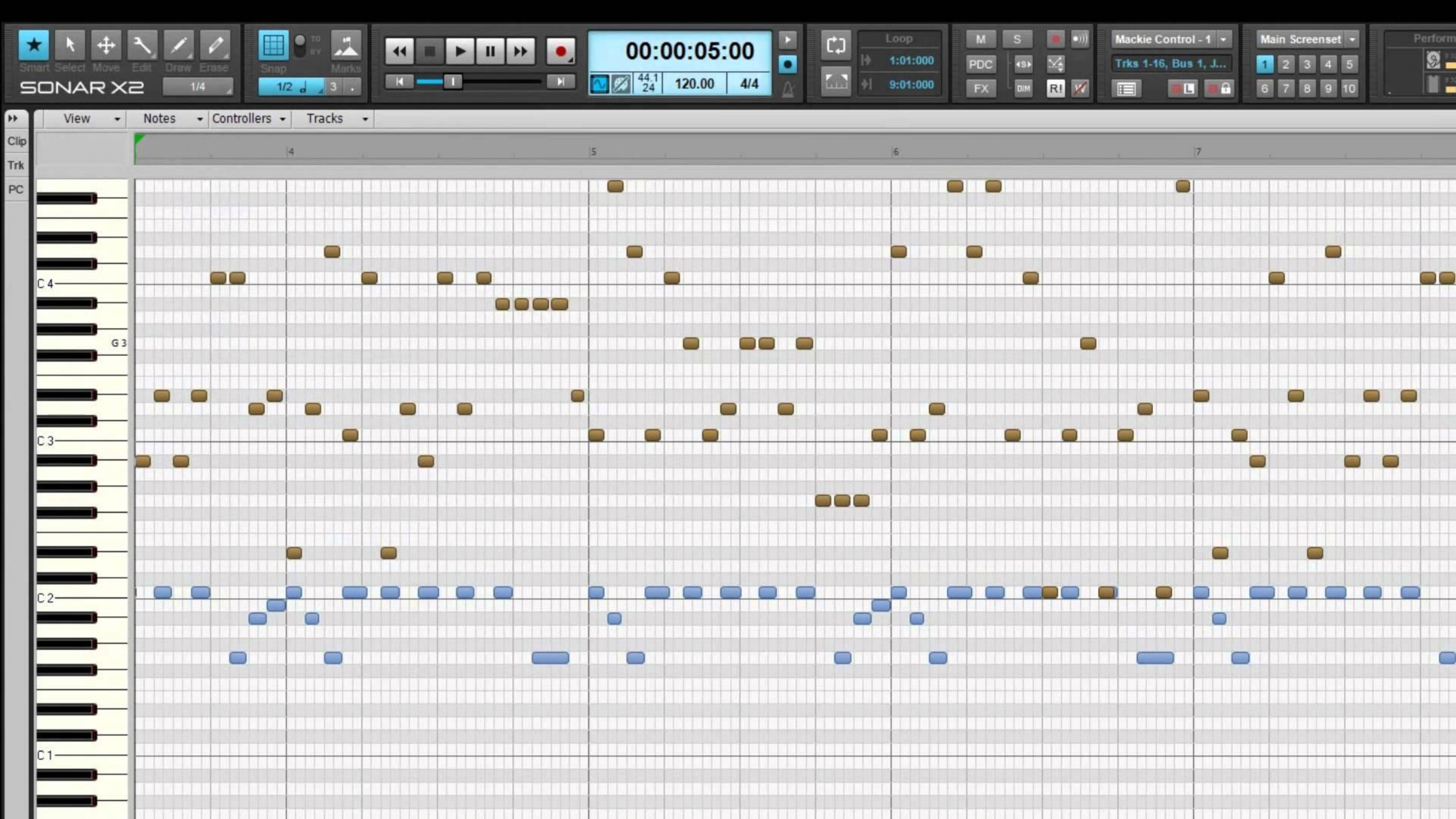
Why audio? Why raw audio?

Why audio?

Music generation is typically studied in the symbolic domain




MORE
ONCE
SWEET-
MY
SEE


Smart Select Move Edit Draw Erase
SONAR X2
1/4
Snap
Marks
1/2
3
00:00:05:00
44.1
24
120.00
4/4
Loop
1:01:000
9:01:000
M
S
PDC
FX
DIM
R!
Mackie Control - 1
Trks 1-16, Bus 1, J...
Main Screenset
1 2 3 4 5
6 7 8 9 10
Perform
View
Notes
Controllers
Tracks
Clip
Trk
PC
4
5
6
7
C 4
G 3
C 3
C 2
C 1

Why audio?

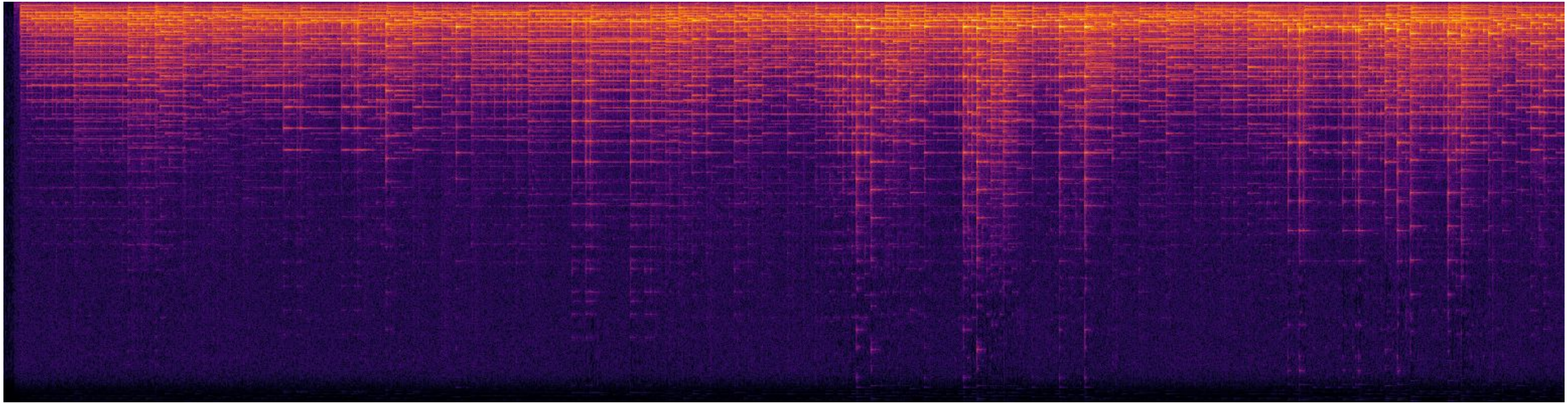
Many instruments have complex action spaces

Rich palette of sounds and timbral variations

Guitar

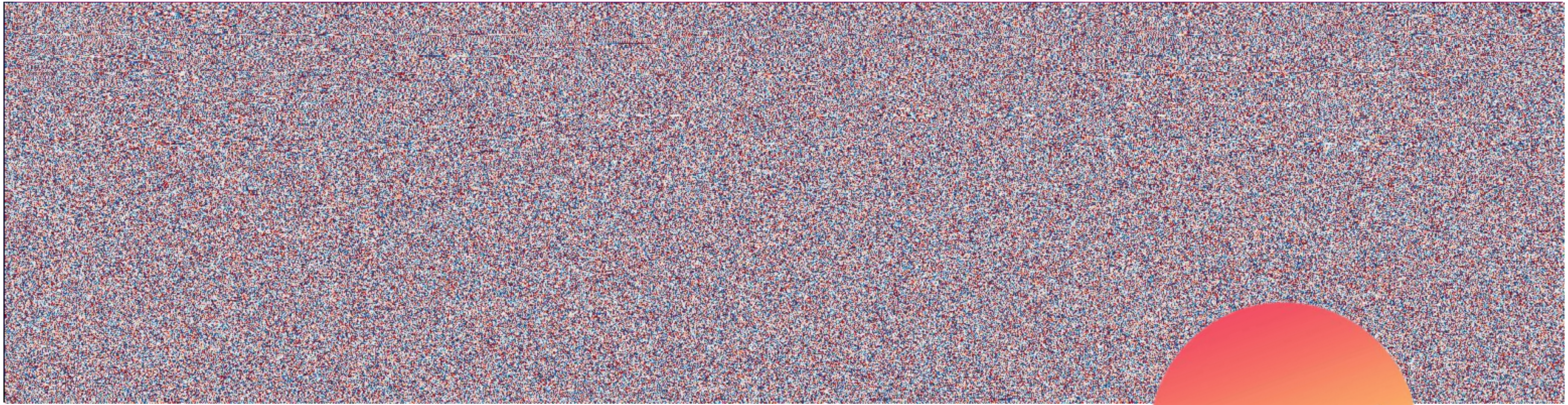
- pick vs. finger
- picking position
- frets
- harmonics
- ...

Why raw audio?



Magnitude spectrogram

Why raw audio?



Phase spectrogram

Why raw audio?

Phase is often unimportant in discriminative settings,
but is very important perceptually!

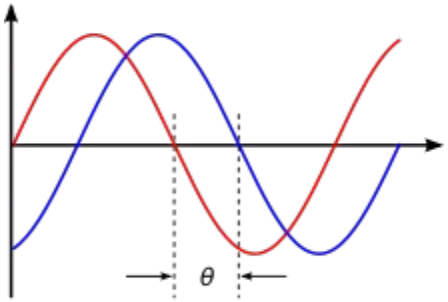
original phase

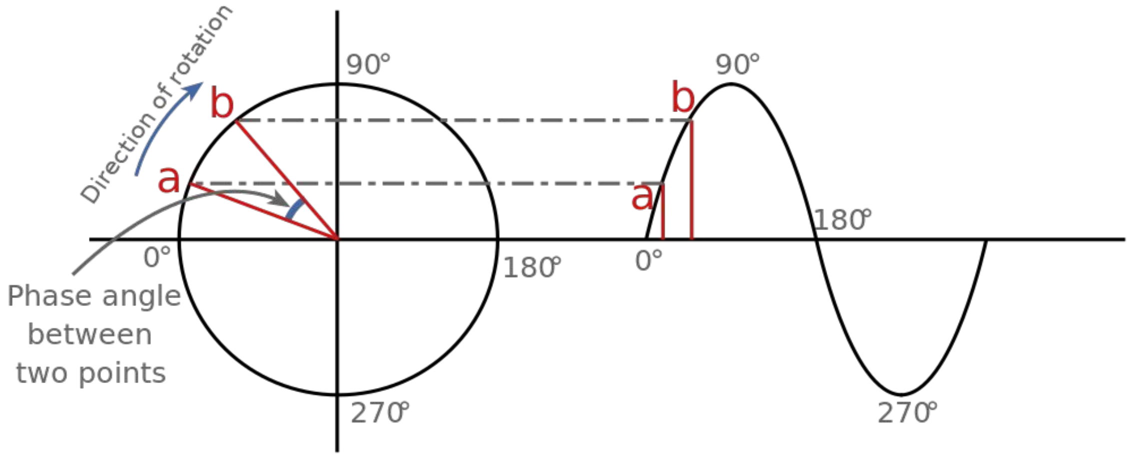
random phase

Why raw audio?

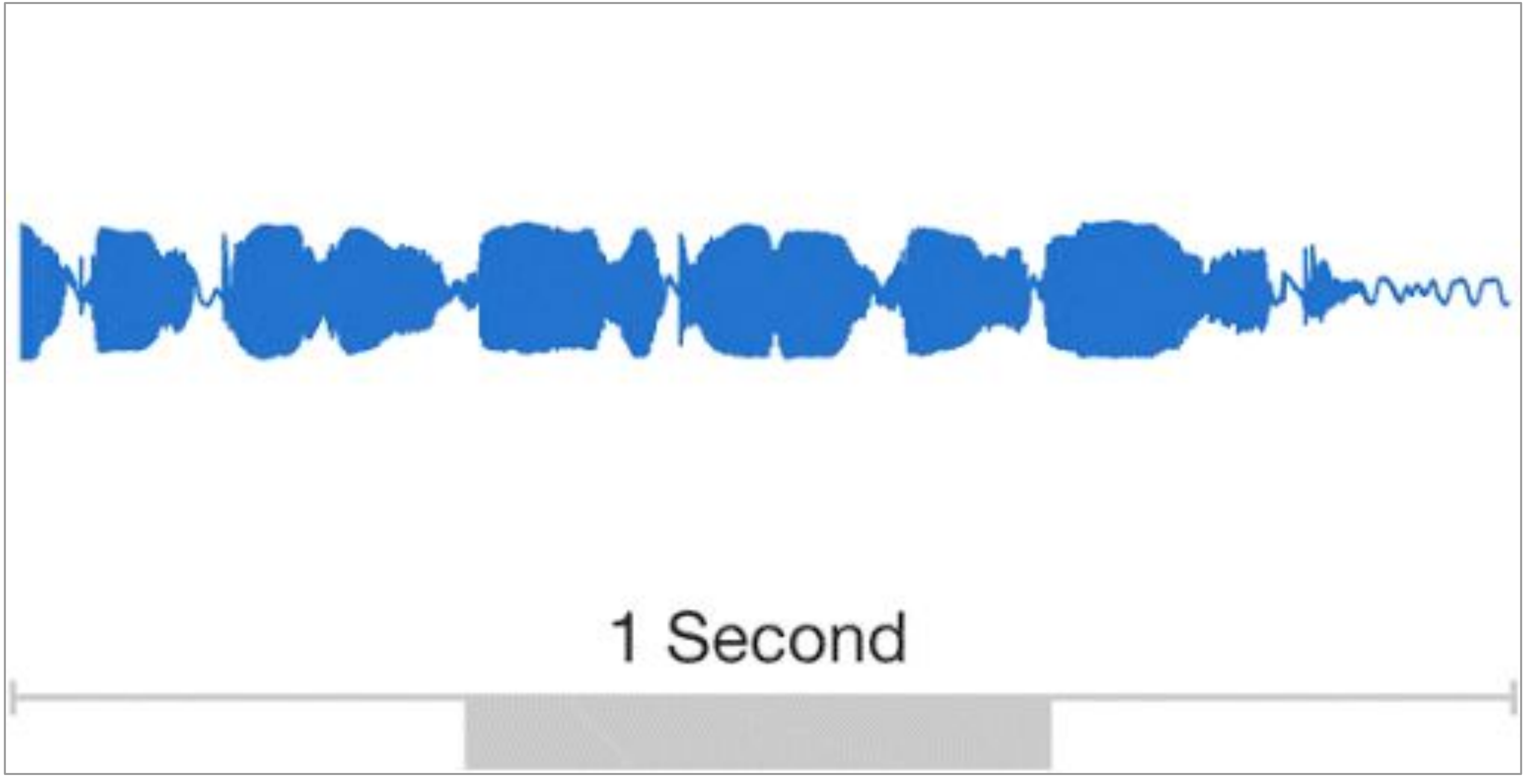
Phase is hard to model:

- it is an angle, so it wraps around
- it becomes random as the magnitude tends to 0
- absolute phase is less meaningful, but relative phase differences matter



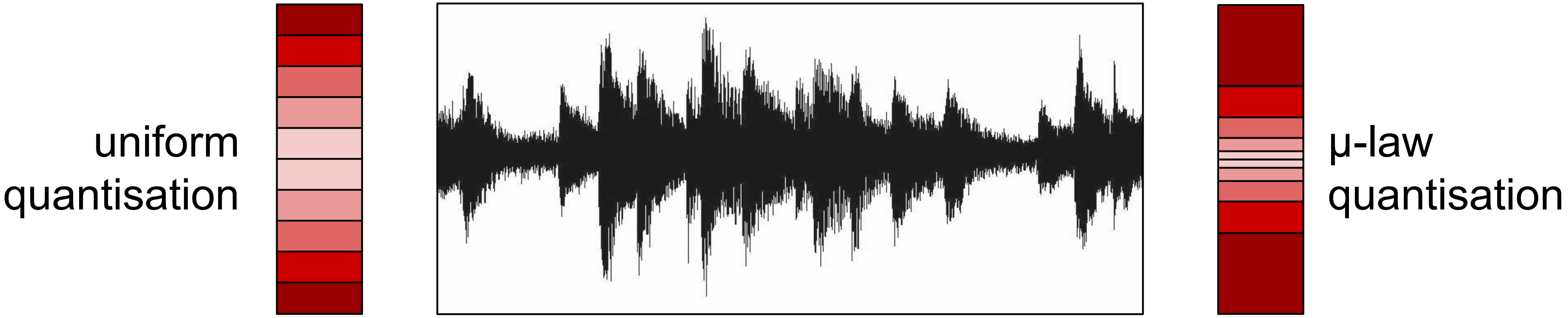


What is “raw audio” anyway?



Discretising audio

- Time
- Amplitude



Generative models

Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**

Generative models

Given a dataset of examples **X** drawn from **p(X)**:

a generative model estimates **p(X)**

Explicit: given $\mathbf{x} \in \mathbf{X}$, model can infer **p(x)**

Implicit: model can produce new samples $\mathbf{x} \sim \mathbf{p(X)}$

Generative models

Brock et al., 2019. “Large Scale GAN Training for High Fidelity Natural Image Synthesis”, ICLR.

Likelihood-based models

Likelihood-based models parameterise **p(X)** directly

Objective function: maximise $\sum_{\mathbf{x} \in \mathbf{X}} \mathbf{\log p(x)}$

Autoregressive models

Autoregressive models factorise **p(X)** into simpler (scalar) distributions

$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n)$

$\mathbf{p}(\mathbf{x}) = \prod_i \mathbf{p}(\mathbf{x}_i|\mathbf{x}_{<i})$ chain rule of probability

We can use the same model $\mathbf{p}(\mathbf{x}_i|\mathbf{x}_{<i})$ for all i!

Flow-based models

Flow-based models transform **p(X)** to a simple (factorised) distribution with an invertible mapping

$\mathbf{p(x) = p(z) \cdot |det\ J|^{-1}}$ change of variables theorem

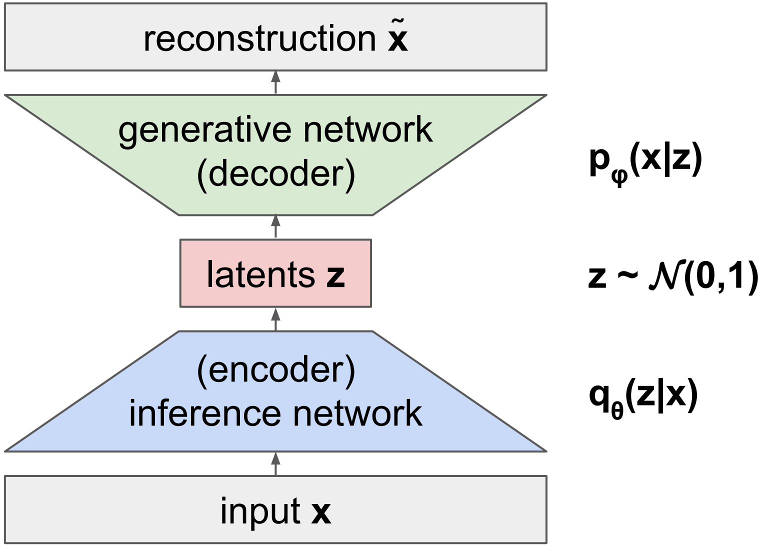
$\mathbf{J = d(g(z))/dz} \quad \mathbf{x = g(z)}$

Important constraints:

g(z) must be invertible **det J** must be tractable

Dinh et al., 2014. "NICE: Non-linear Independent Components Estimation", arXiv.
Dinh et al., 2016. "Density estimation using Real NVP", arXiv.

Variational autoencoders (VAEs)



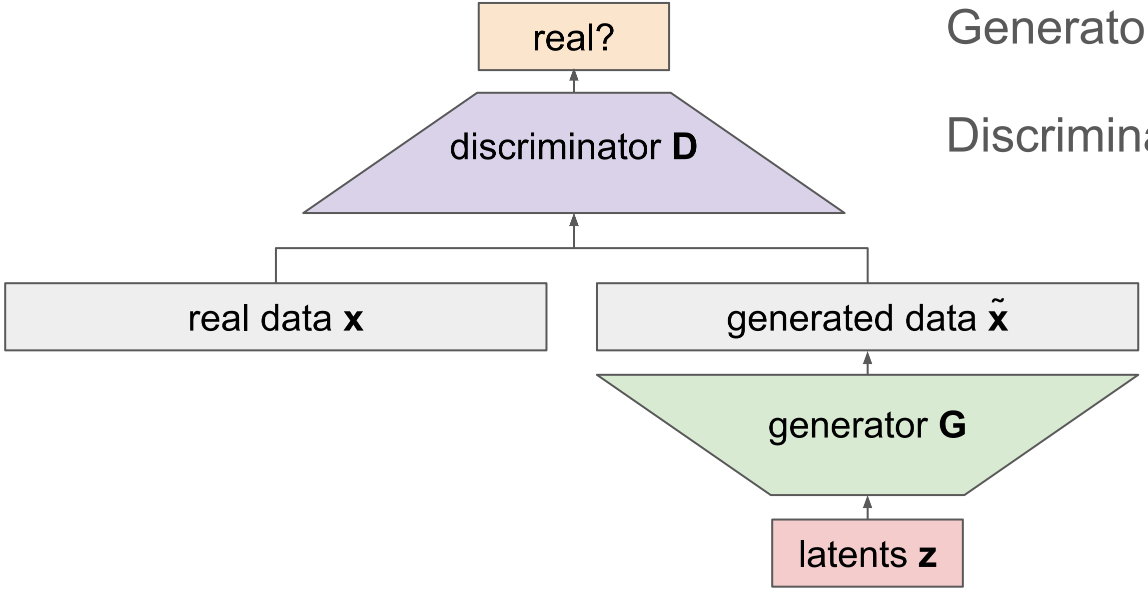
VAE maps latents **z** from a simple distribution to **x** with a (non-invertible) generative network

The inference network approximates the inverse operation

p(x) cannot be computed exactly, the Evidence Lower BOund (ELBO) is maximised instead

Kingma & Welling, 2014. “Auto-Encoding Variational Bayes”, ICLR 2014.
Rezende et al., 2014. “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”, ICML 2014.

Adversarial models



$$\mathcal{L} = \mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1 - D(G(z)))]$$

Generator minimises $\mathcal{L}$

Discriminator maximises $\mathcal{L}$

Goodfellow et al., 2014. “Generative Adversarial Nets”, NeurIPS.

More exotic flavours

- Implicit quantile networks
- Energy-based models
- Optimal transport (e.g. Wasserstein autoencoders)
- Score-based generative modelling
- ...

Conditional generative models

Conditioning is “side information” which allows for control over the model output

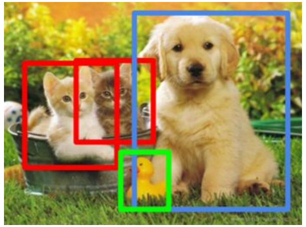
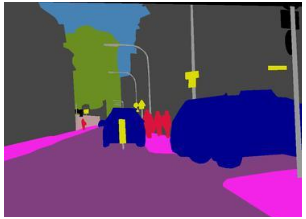
p(x|c) vs. **p(x)**

Conditional generative models

Conditioning is “side information” which allows for control over the model output

p(x|c) vs. **p(x)**

sparsely conditioned → densely conditioned

image models	class labels	bounding boxes	segmentation	grayscale image (colorisation)
	y = “cat”			

Conditional generative models

Conditioning is “side information” which allows for control over the model output

p(x|c) vs. **p(x)**

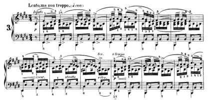
sparsely conditioned → densely conditioned

music audio models

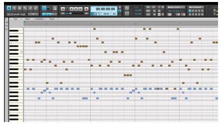
composer
instrument(s)
tempo
timbre
...

note density
musical form
…

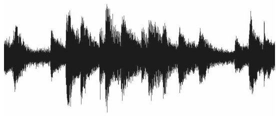
score



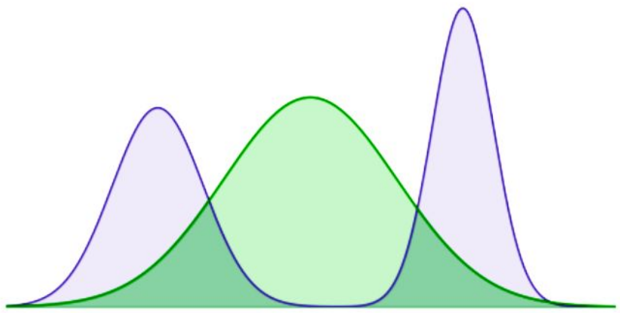
MIDI



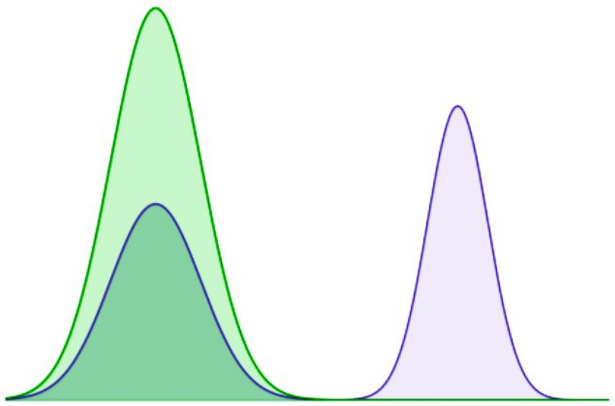
other audio signals



Mode-covering vs. mode-seeking behaviour



mode-covering

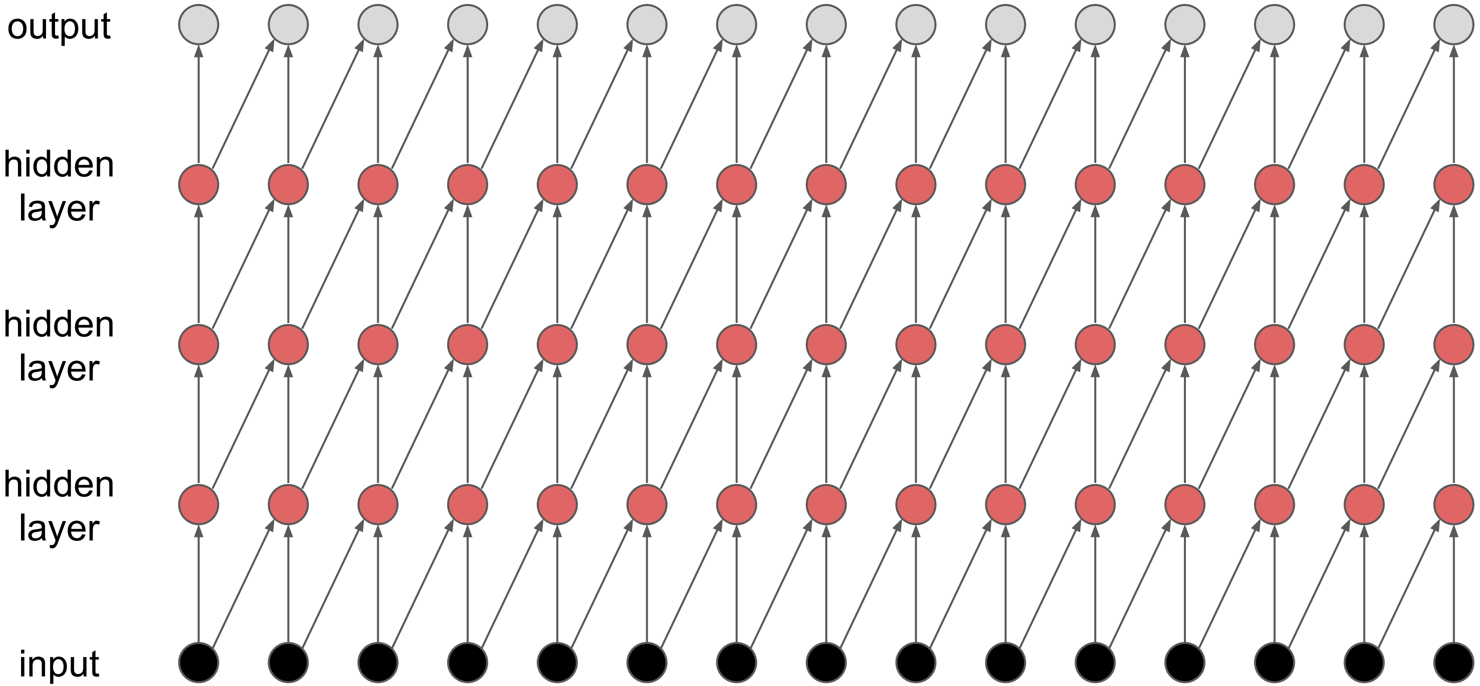


mode-seeking

- Likelihood-based models are mode-covering
- Adversarial models are (typically) mode-seeking
- In more densely conditioned settings, we tend to care less about covering all the modes

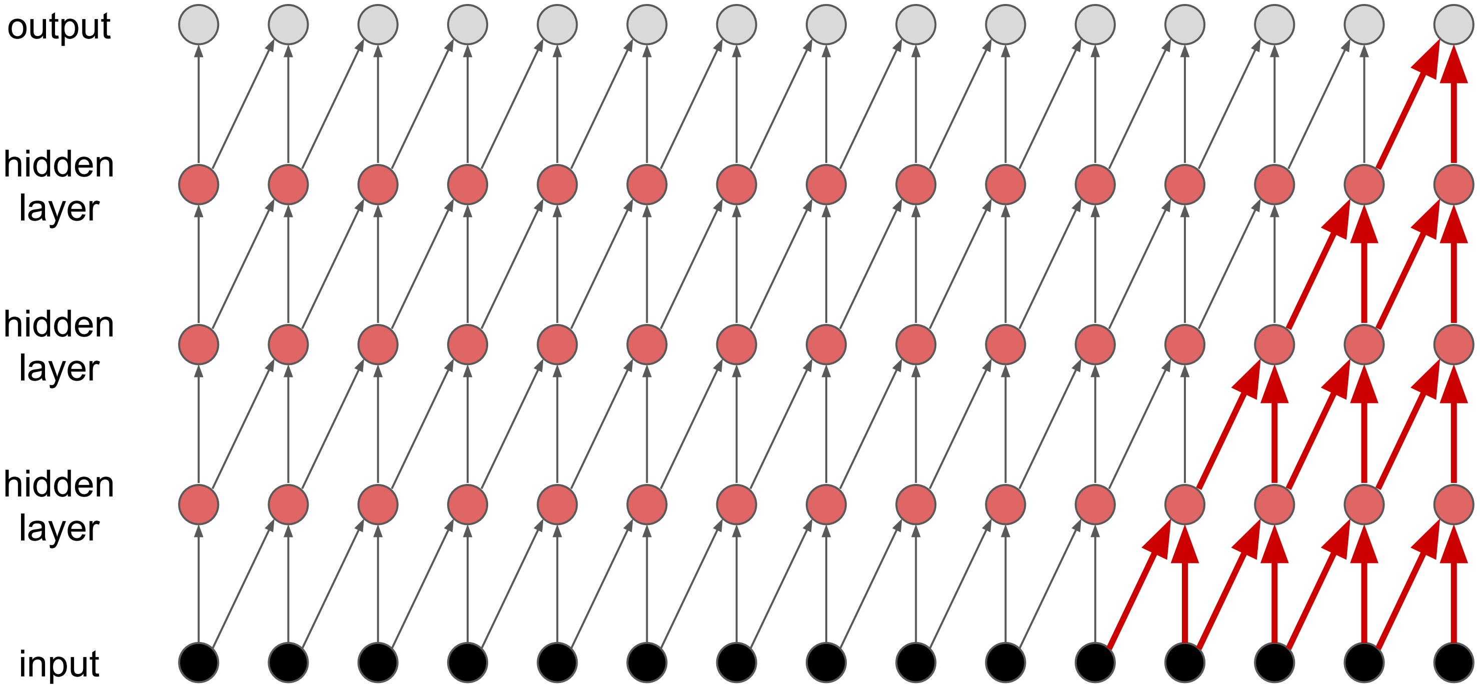
Likelihood-based models

WaveNet



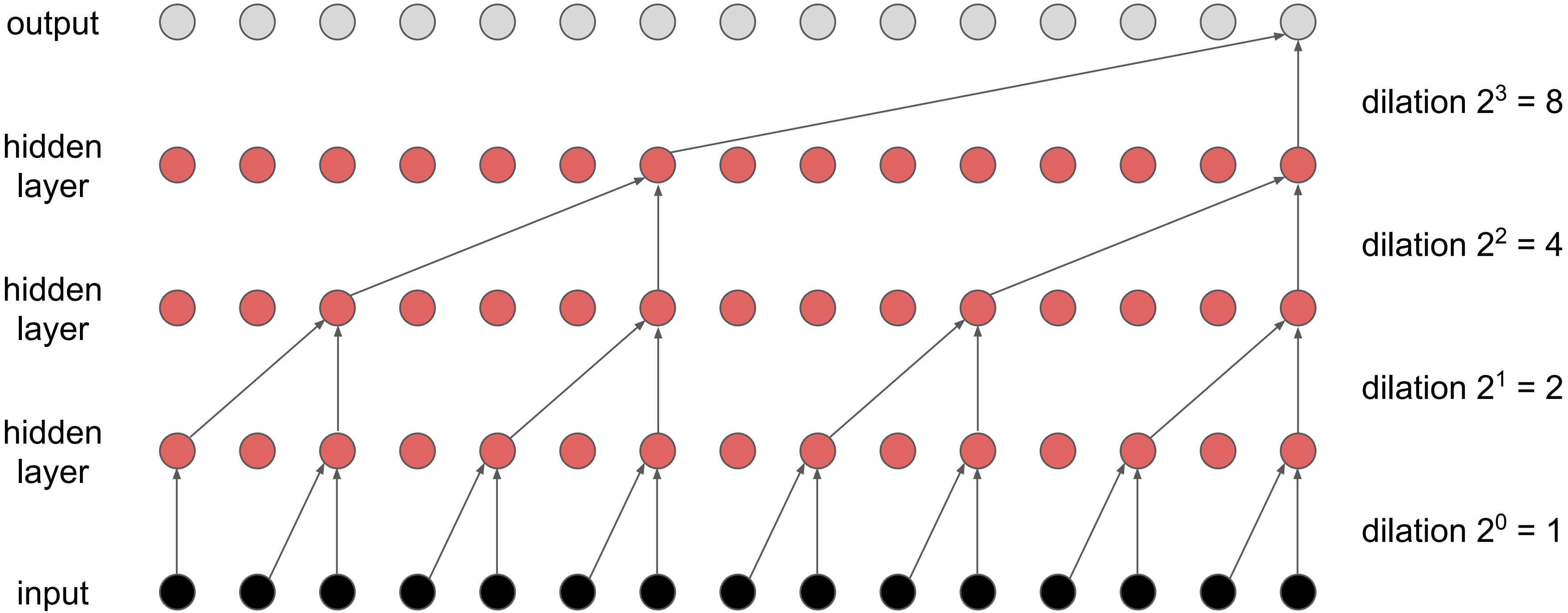
van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

WaveNet



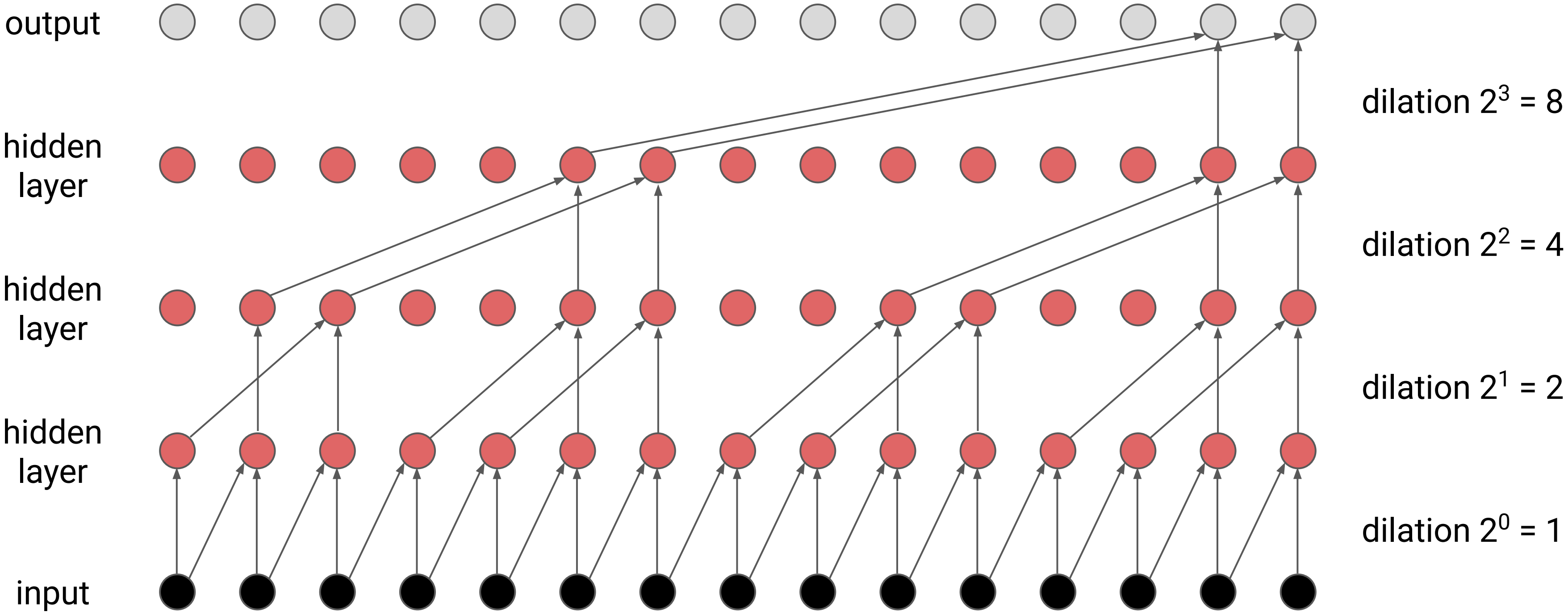
van den Oord et al., 2016. "WaveNet: A Generative Model for Raw Audio", arXiv.

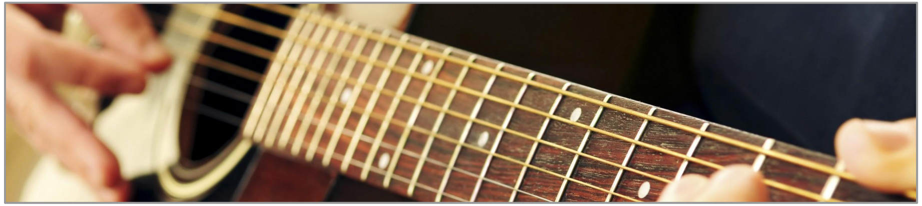
WaveNet: dilated convolutions

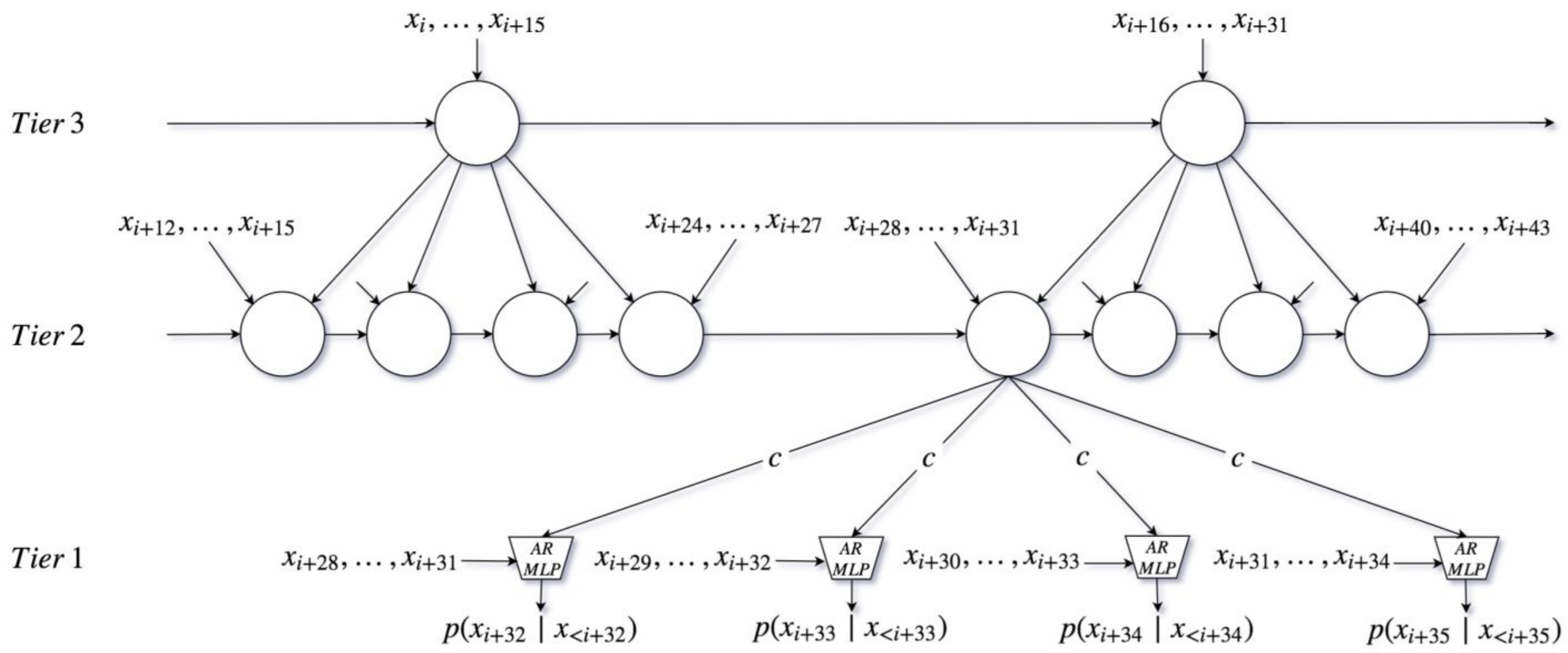


van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

WaveNet: dilated convolutions

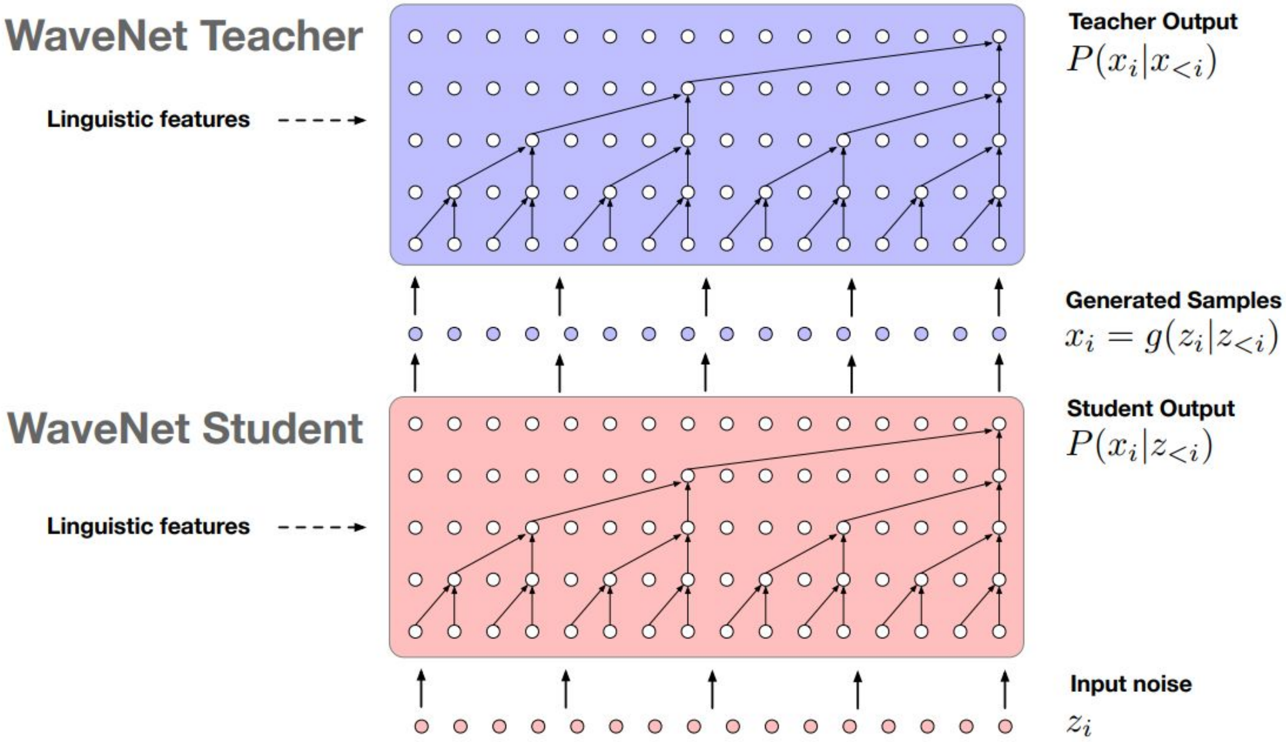


van den Oord et al., 2016. “WaveNet: A Generative Model for Raw Audio”, arXiv.

SampleRNN



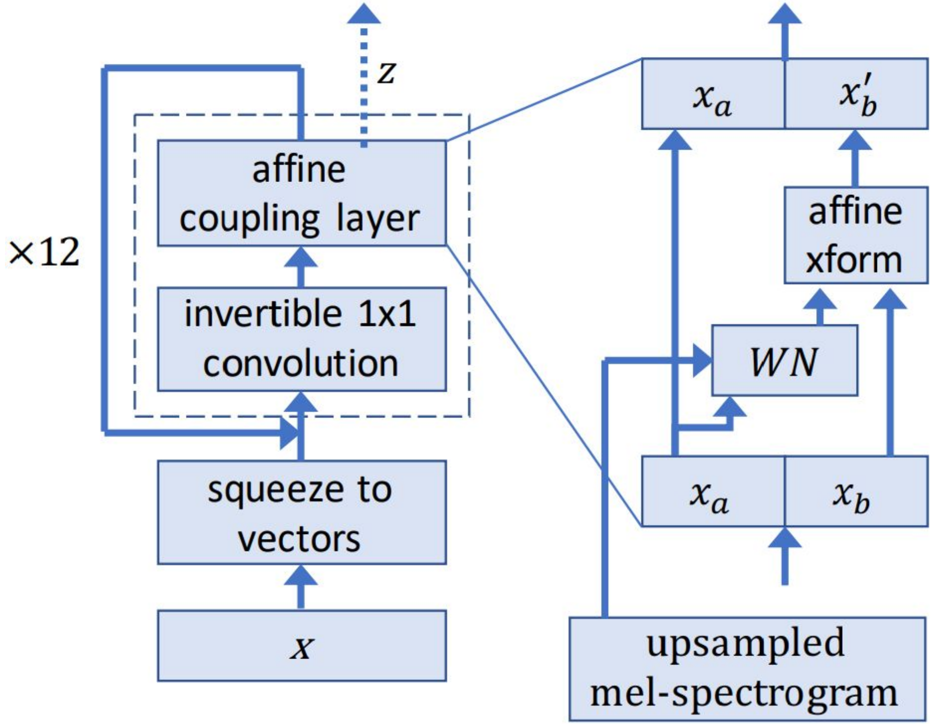
Mehri et al., 2017. “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model”, ICLR.

Parallel WaveNet, ClariNet



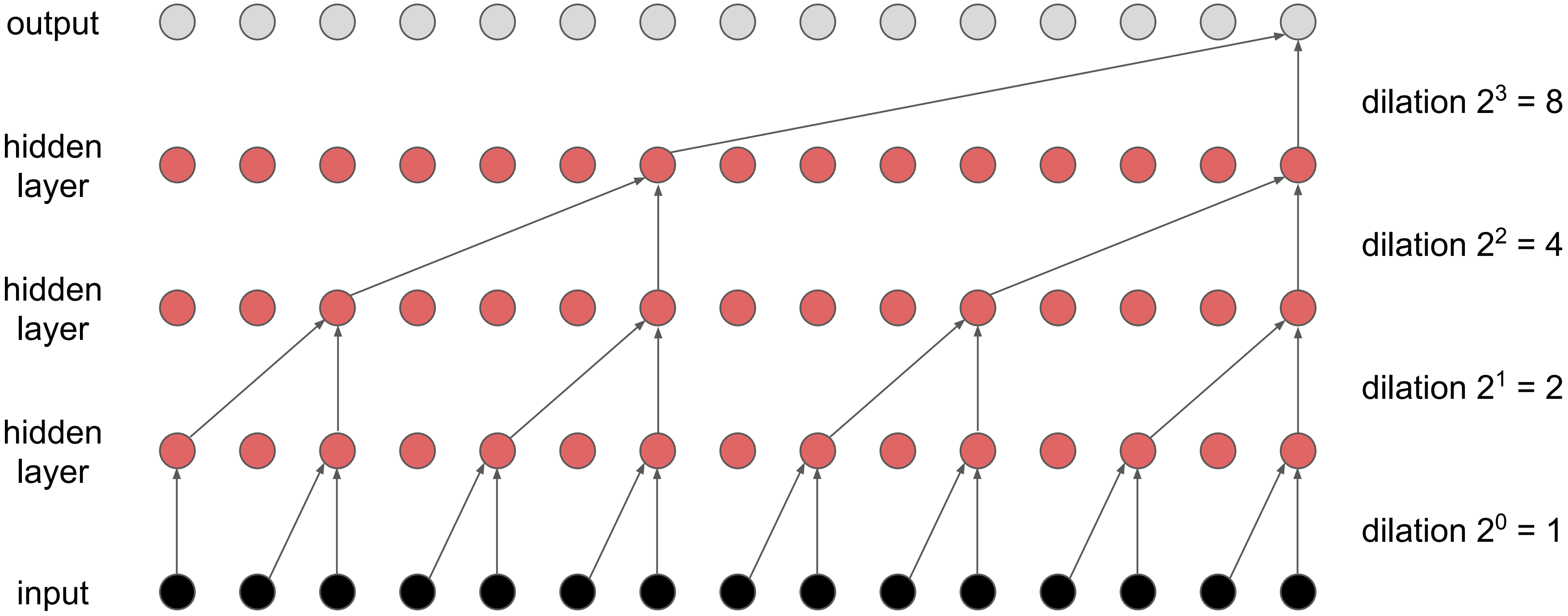
van den Oord et al., 2019. “Parallel WaveNet: Fast High-Fidelity Speech Synthesis”, ICML.

WaveGlow, FloWaveNet



Prenger et al., 2019. “Waveglow: A flow-based generative network for speech synthesis”, ICASSP.
Kim et al., 2019. “FloWaveNet: A generative flow for raw audio”, ICML.

WaveNet: #layers ~ log(receptive field length)



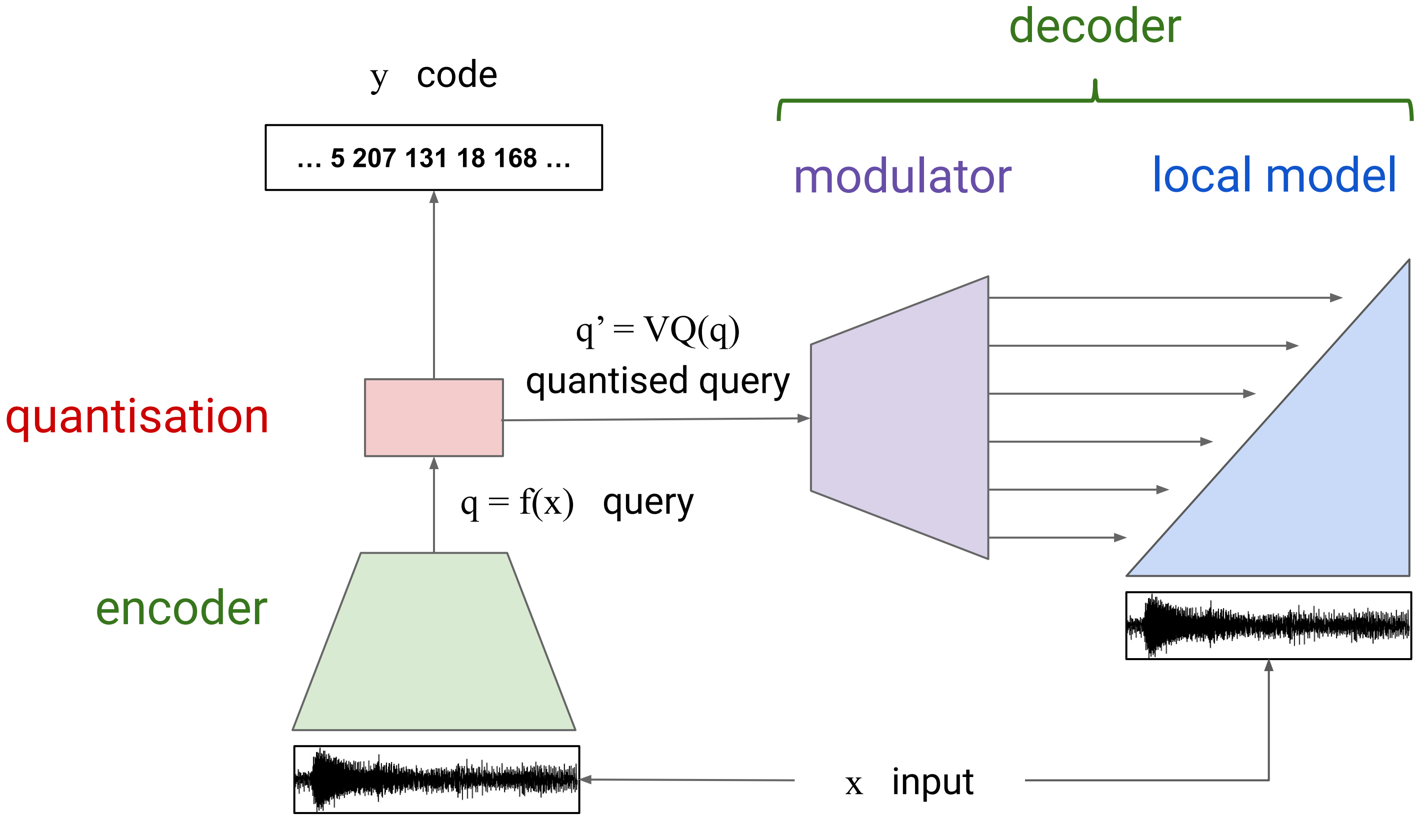
… but memory usage ~ receptive field length

Required model depth is **logarithmic** in the desired receptive field length

Required memory usage during training is still **linear** in the desired receptive field length!

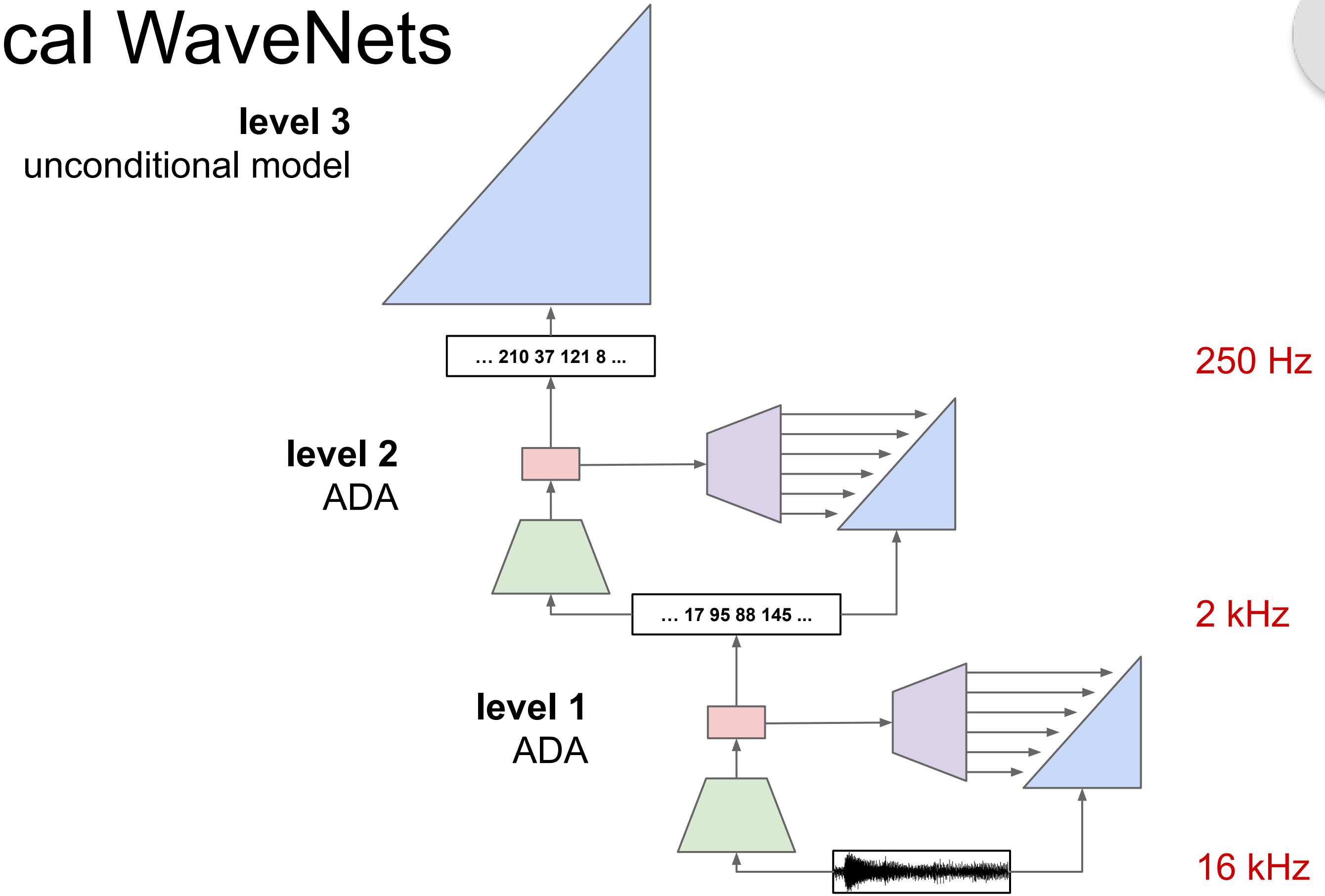
⇒ **We cannot scale indefinitely using dilation**

Autoregressive discrete autoencoders



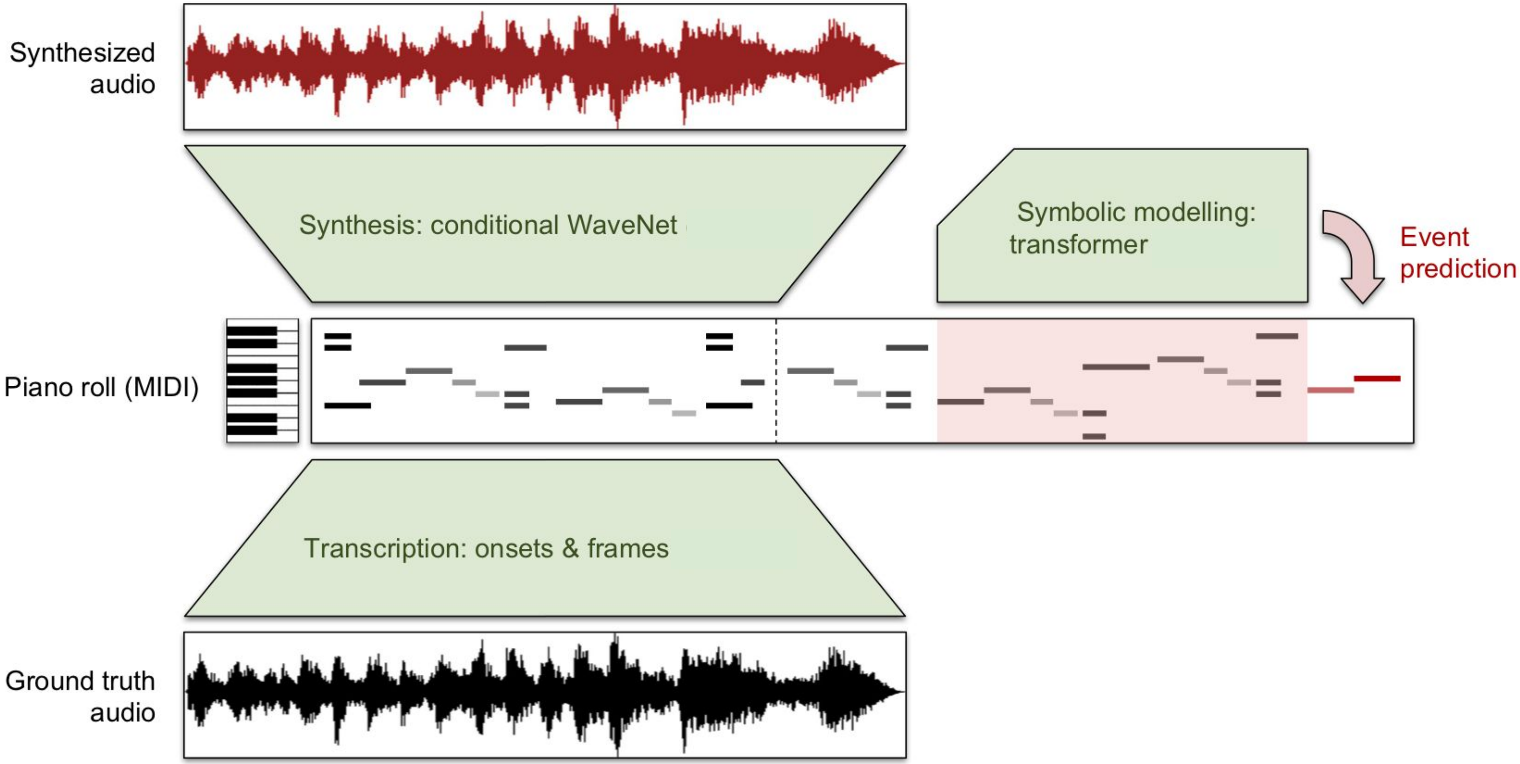
van den Oord et al., 2017. “Neural discrete representation learning”, NeurIPS.

Hierarchical WaveNets



Dieleman et al., 2018. “The challenge of realistic music generation: modelling raw audio at scale”, NeurIPS.

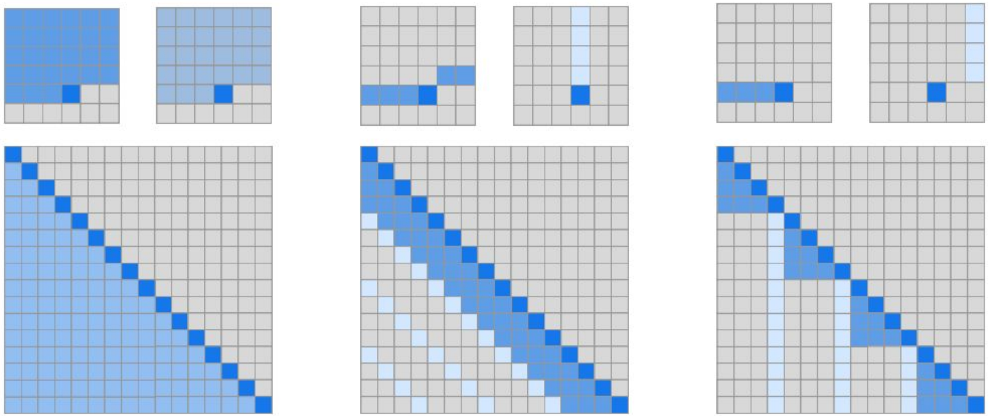
Wave2Midi2Wave and the MAESTRO dataset



Hawthorne et al., 2019. “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset”, ICLR.

Sparse transformers

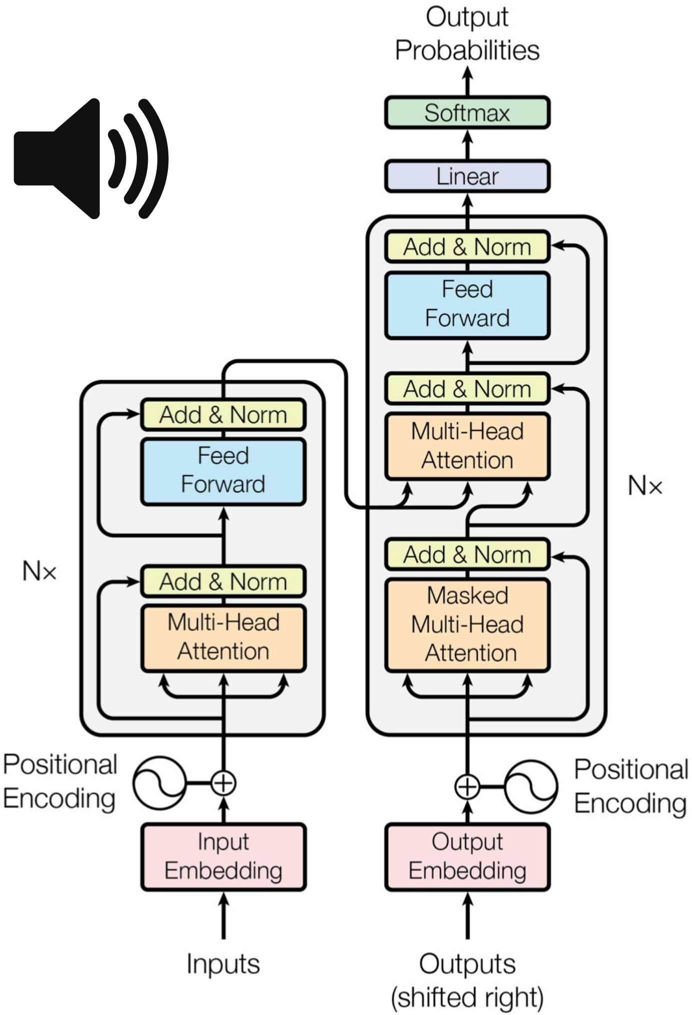
Attention (with sparse masks) instead of recurrence / convolutions.



(a) Transformer

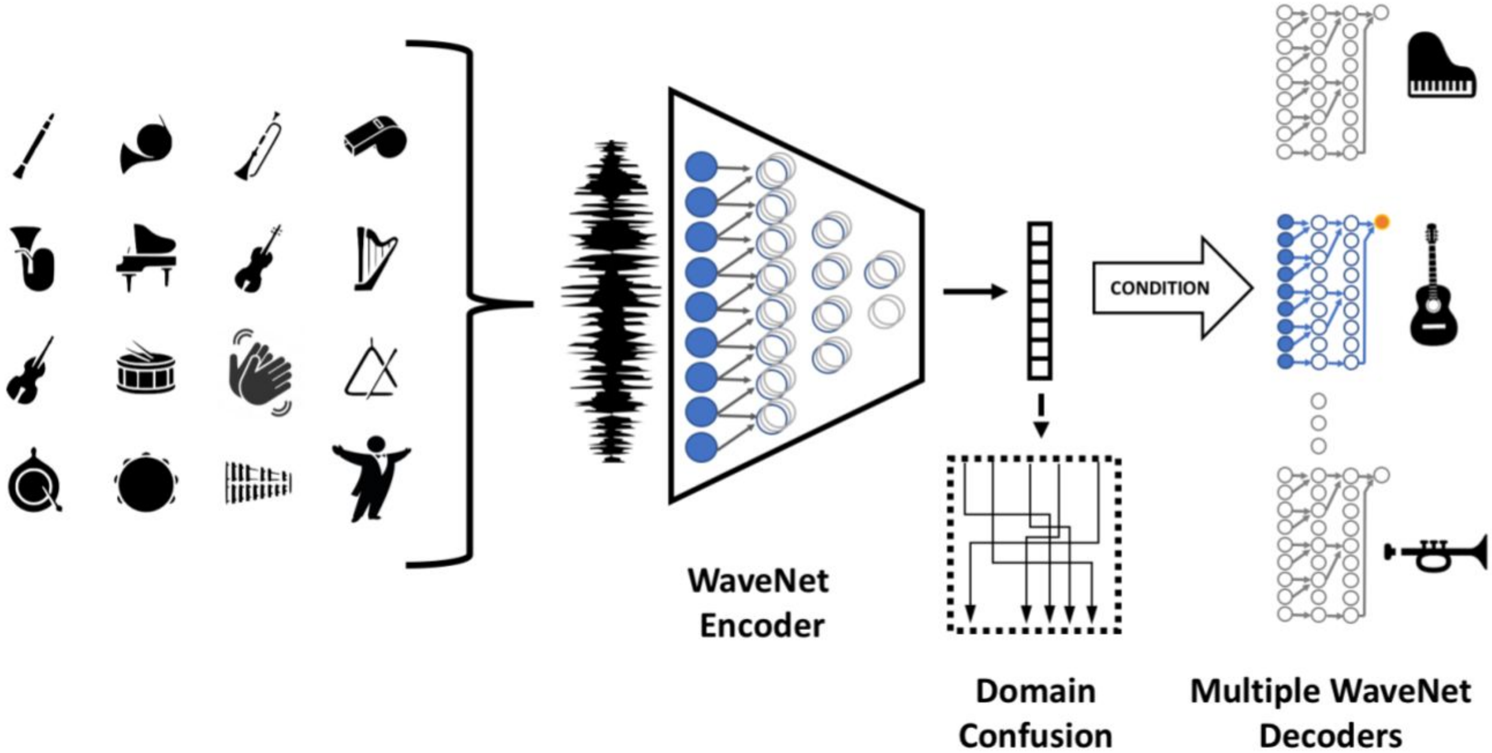
(b) Sparse Transformer (strided)

(c) Sparse Transformer (fixed)



Child et al., 2019. “Generating long sequences with sparse transformers”, arXiv.

Universal music translation network



Mor et al., 2019. “A Universal Music Translation Network”, ICLR.

http://dadabots.com/

Neural networks

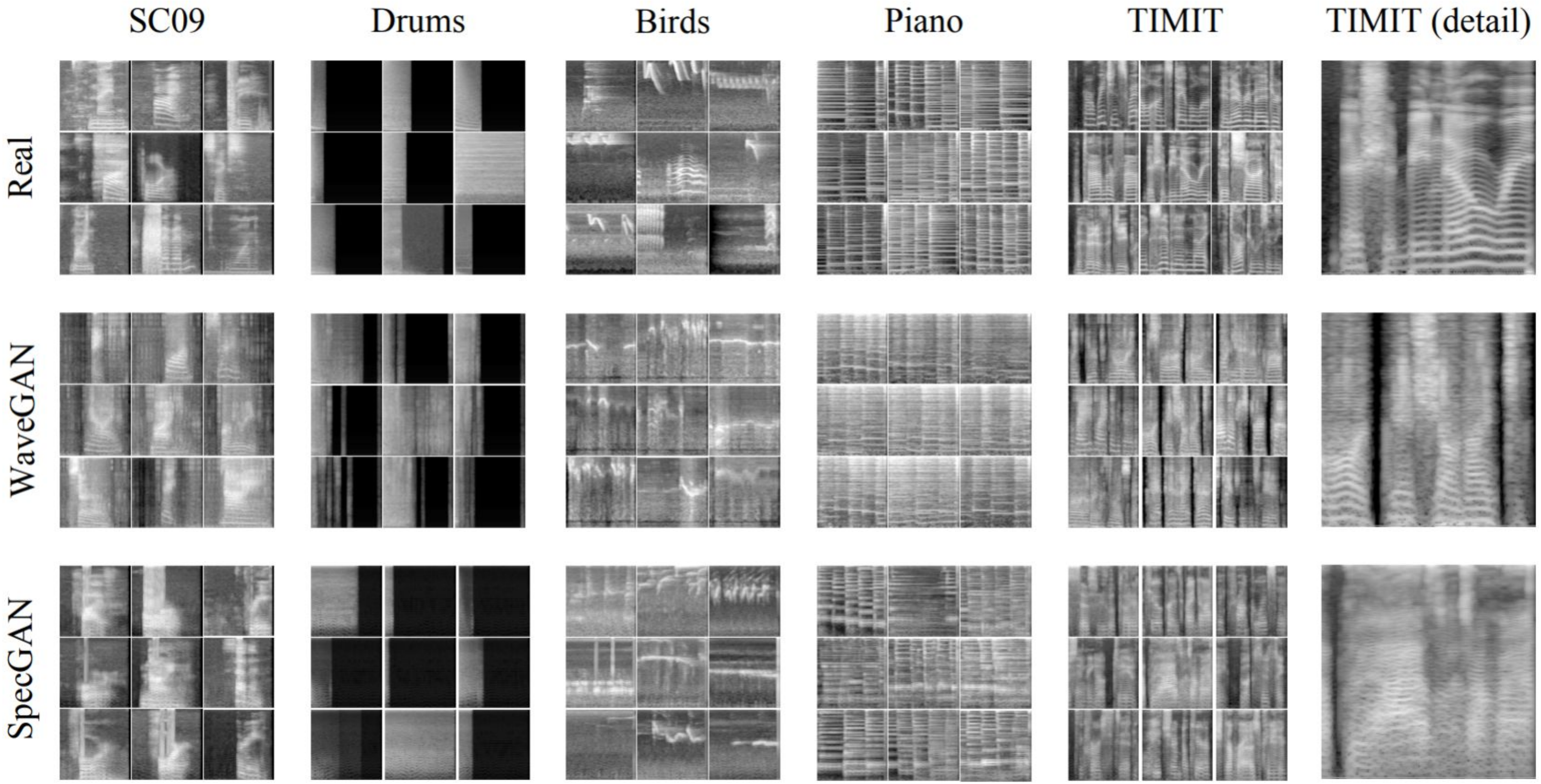
generating death metal
via livestream 24/7 to infinity

We make raw audio neural networks
that can imitate bands

Carr & Zukowski, 2018. “Generating Albums with SampleRNN to Imitate Metal, Rock, and Punk Bands”, arXiv.

Adversarial models

WaveGAN



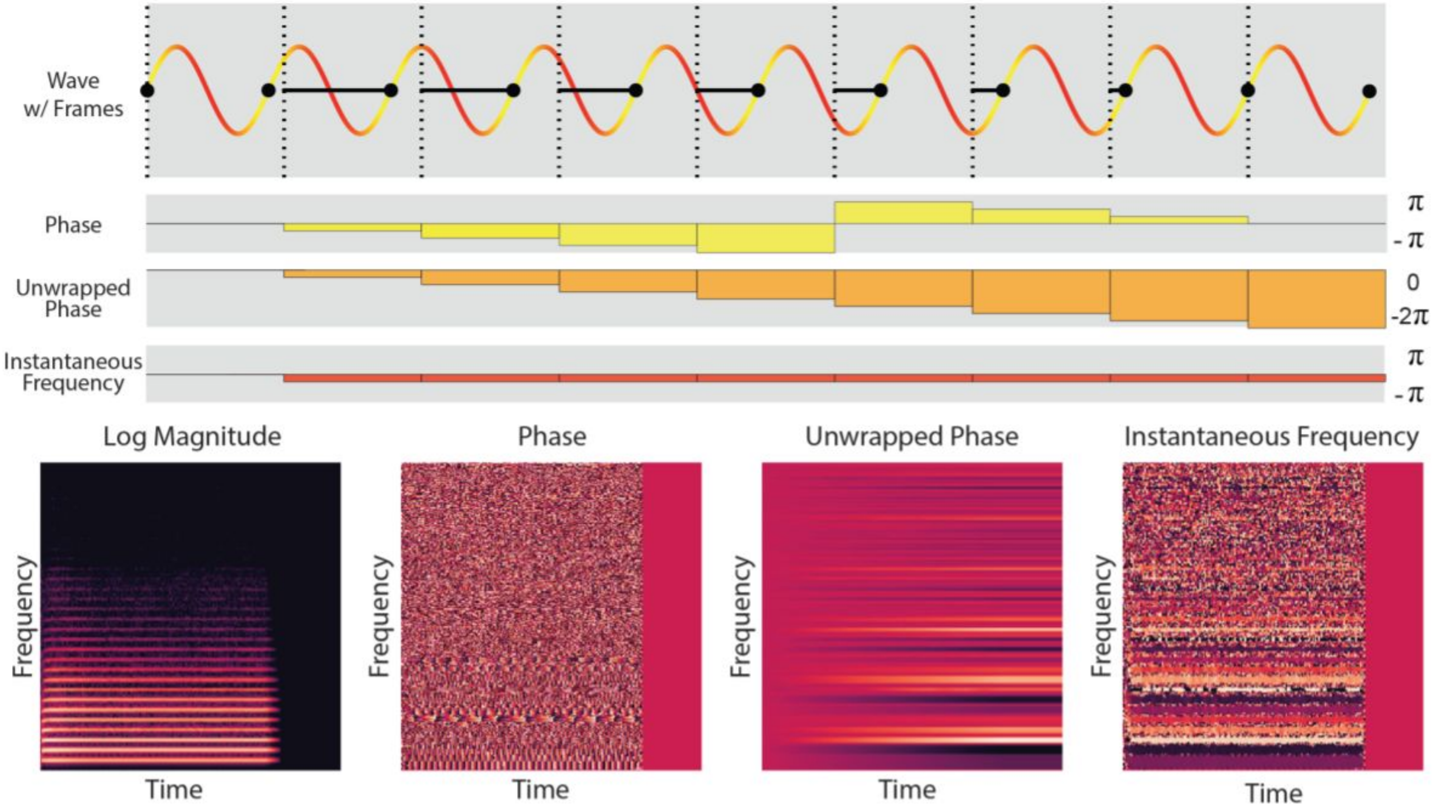
Donahue et al., 2019. “Adversarial Audio Synthesis”, ICLR.

WaveGAN

Non GAN-activated cat | WaveGAN activated cat | SpecGAN activated cat

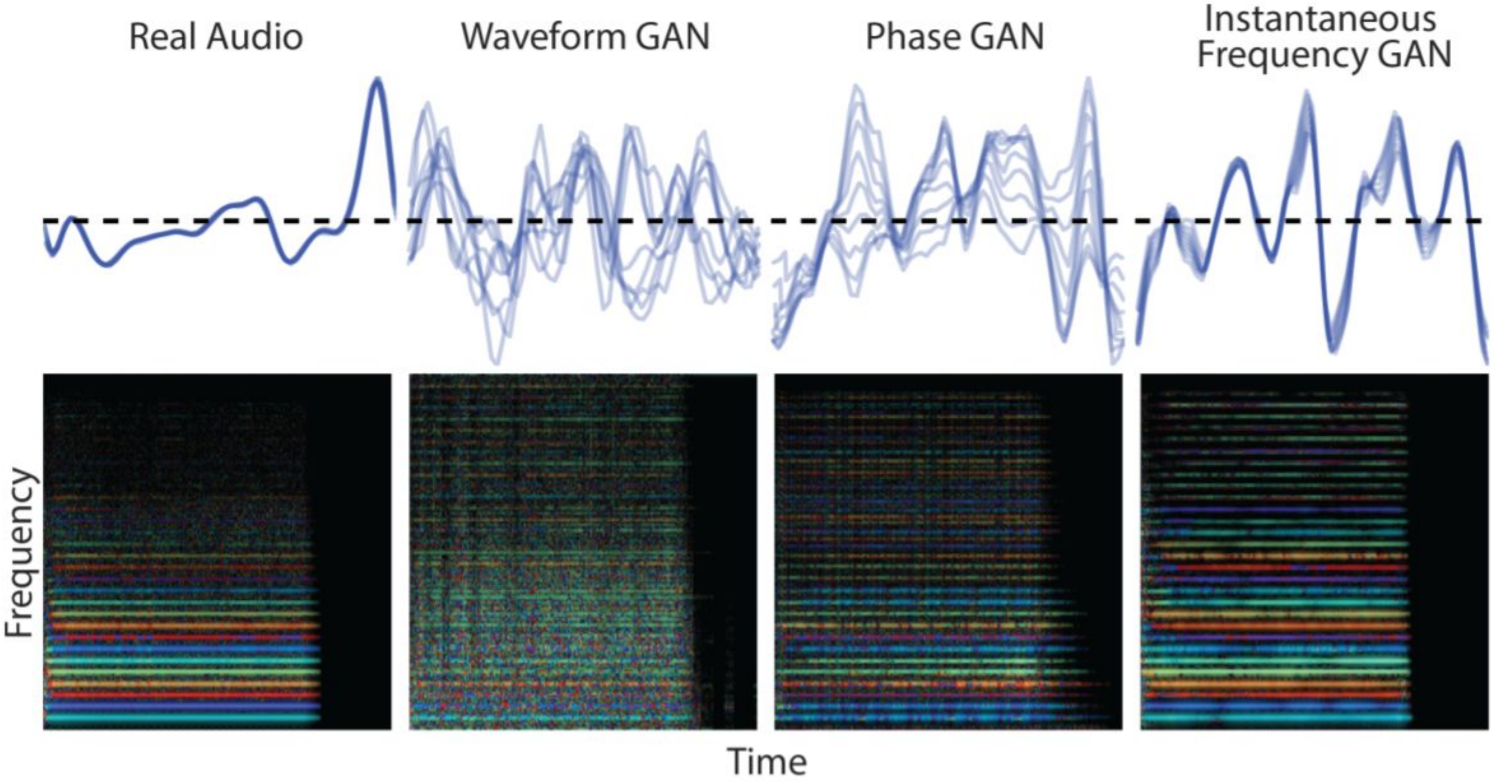
Donahue et al., 2019. “Adversarial Audio Synthesis”, ICLR.

GANSynth



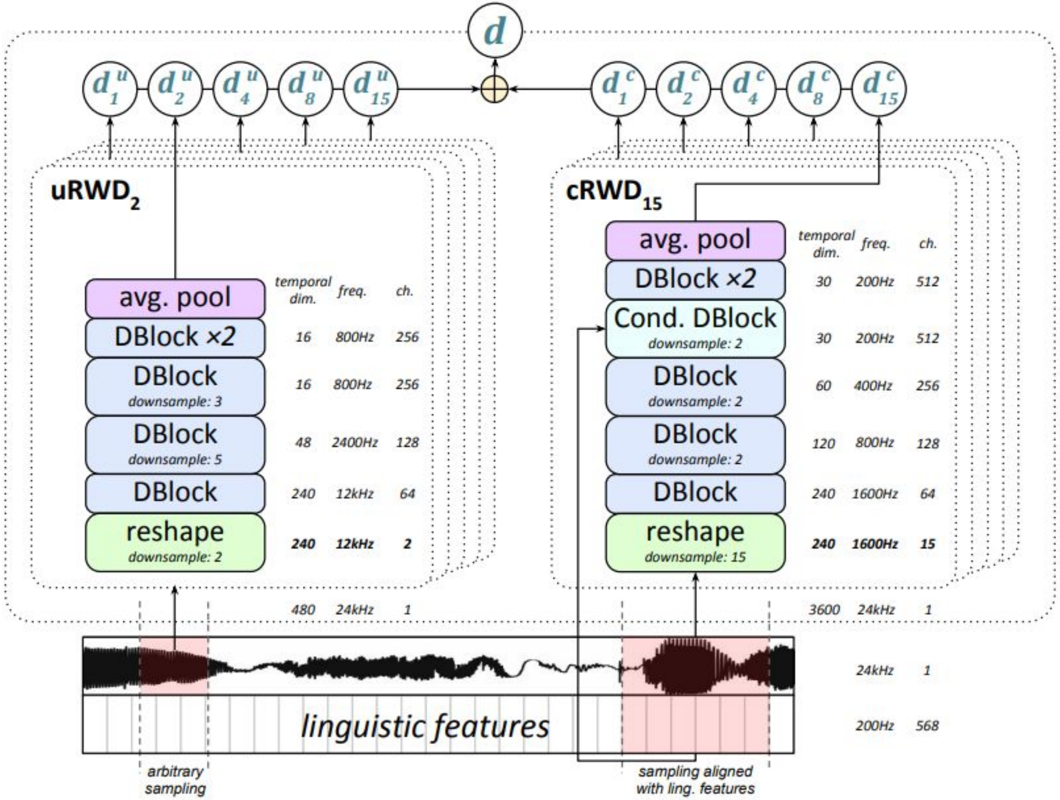
Engel et al., 2019. “GANSynth: Adversarial Neural Audio Synthesis”, ICLR.

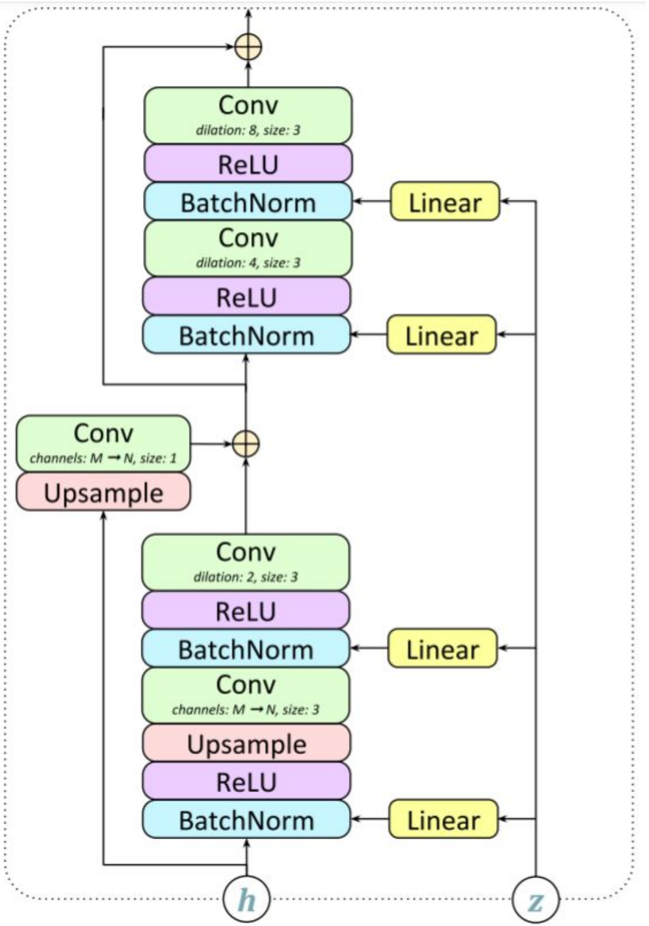
GANSynth



Engel et al., 2019. “GANSynth: Adversarial Neural Audio Synthesis”, ICLR.

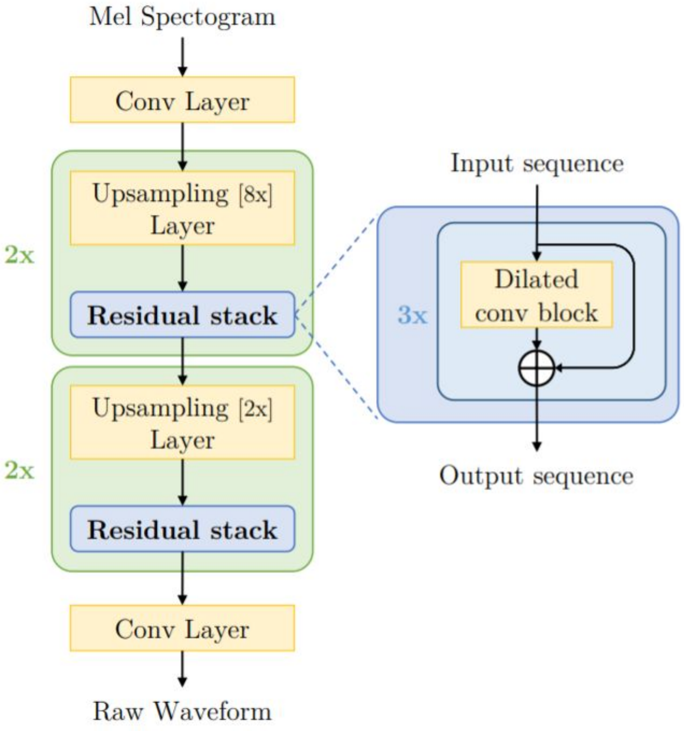
GAN-TTS



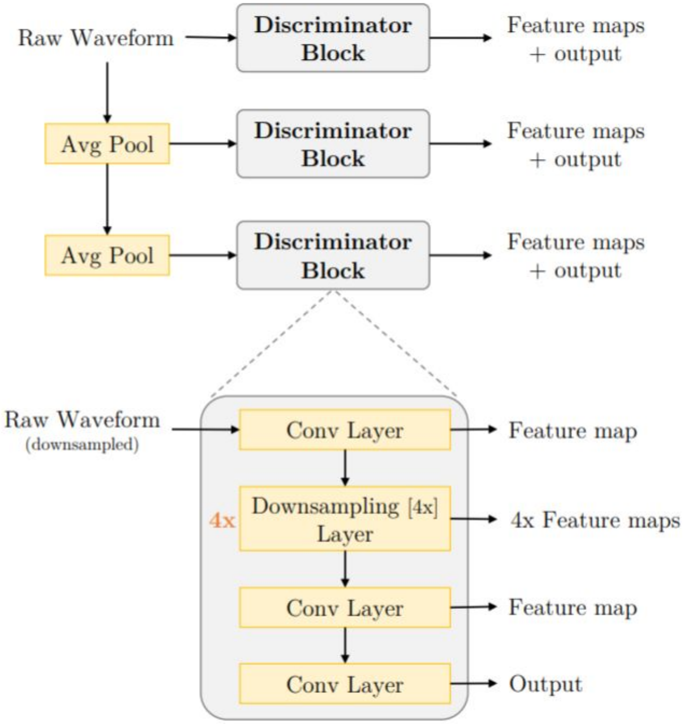


Binkowski et al., 2019. “High Fidelity Speech Synthesis with Adversarial Networks”, arXiv.

MelGAN



(a) Generator



(b) Discriminator

Kumar et al., 2019. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis", arXiv.

Why the emphasis on likelihood in music modelling?

Most popular generative modelling paradigm:

GANs

Why the emphasis on likelihood in music modelling?

Most popular generative modelling paradigm:

GANs

Most popular generative modelling paradigm for music:

likelihood-based (autoregressive)

Why the emphasis on likelihood in music modelling?

- We are still figuring out the right architectural priors for audio discriminators
    - For images, a stack of convolutions is all you need
    - What do we need for audio? Multiresolution? Dilation? Something phase shift invariant?

Ulyanov et al., 2018. “Deep Image Prior”, CVPR.
Pons et al., 2019. “Randomly weighted CNNs for (music) audio classification”, ICASSP.

Why the emphasis on likelihood in music modelling?

- We are still figuring out the right architectural priors for audio discriminators
  - For images, a stack of convolutions is all you need
  - What do we need for audio? Multiresolution? Dilation? Something phase shift invariant?

- The sparsely-conditioned setting is dominant
  - We care about “creativity” and capturing diversity
  - GANs are worse at this than likelihood-based models

Ulyanov et al., 2018. “Deep Image Prior”, CVPR.
Pons et al., 2019. “Randomly weighted CNNs for (music) audio classification”, ICASSP.

Alternatives to modelling raw audio directly

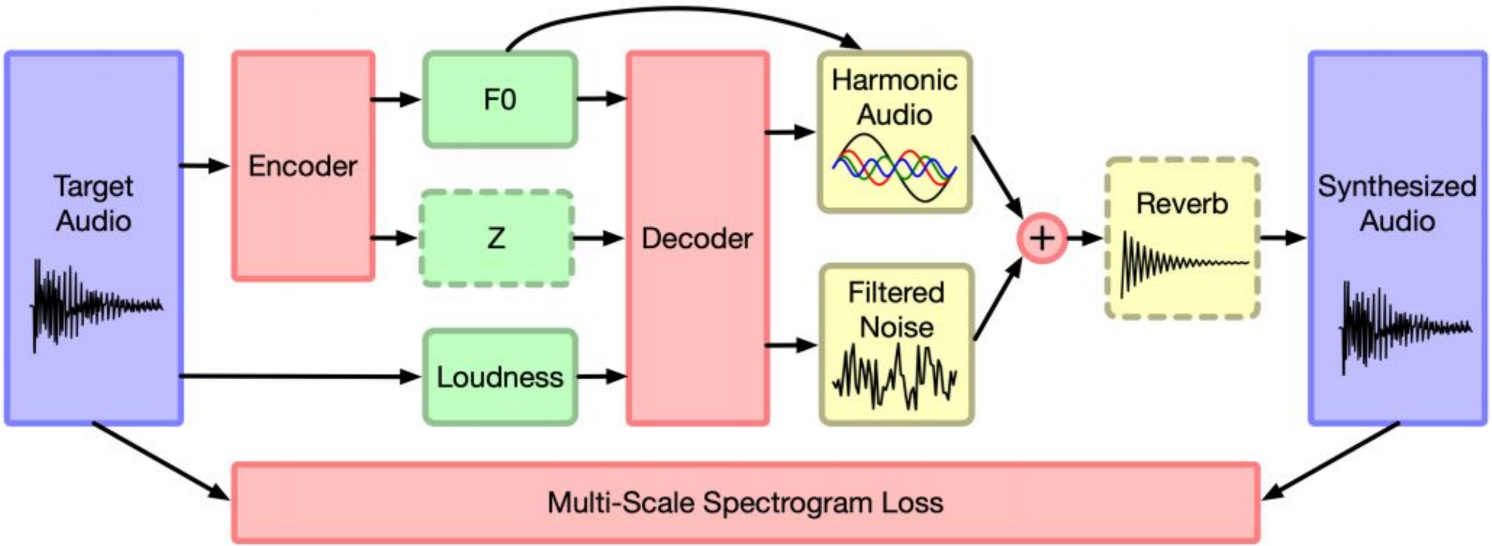
- Model complex-valued spectrograms and “deal” with phase
  (GANSynth)

- Model magnitude spectrograms
  Use a vocoder or Griffin-Lim to invert
  (Tacotron 1 & 2, MelGAN, MelNet, …)

Wang et al., 2017. “Tacotron: Towards end-to-end speech synthesis”, ISCA.
Shen et al., 2018. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, ICASSP.
Vasquez & Lewis, 2019. “MelNet: A Generative Model for Audio in the Frequency Domain”, arXiv.

Alternatives to modelling raw audio directly

- Differentiable Digital Signal Processing (anonymous authors, ICLR, in review)
  Use raw audio input, but put DSP components in the model
  https://openreview.net/forum?id=B1x1ma4tDr



Summary

- Generative modelling of raw audio is feasible, even in the sparsely conditioned setting

- Likelihood-based models dominate, but GANs are making in-roads in the densely conditioned setting

- Modelling large-scale structure from raw audio is an unsolved problem

Thank you

Sander Dieleman, Jordi Pons, Jongpil Lee