Arab Republic of Egypt
Ministry of Communications and Information Technology

Information Technology Institute

EPITA
SCHOOL OF ENGINEERING & COMPUTER SCIENCE

A.I. IN AUDIO & SIGNAL PROCESSING

Session 2: Deep learning for audio and speech processing

SESSION 2: DEEP LEARNING FOR AUDIO & SPEECH PROCESSING

Quick Summary

1. **Approaches for feature learning**
    a) Input représentations
    b) Filters shape
    c) Signal models
    d) Generative models
    e) How to visualize learnt représentations ?

2. **New learning paradigms**
    a) Classification
    b) Auto-encoders & variational auto-encoder
    c) Metric learning
    d) Semi-supervised learning

DEEP LEARNING FOR AUDIO AND SPEECH PROCESSING.

Approaches for feature learning

APPROACHES FOR FEATURE LEARNING

2D representation (time-frequency)

Representation

- spectrorgram (STFT magnitude)
- Mel-gram
- Constant-Q-transform

Basic idea

Considering time-frequency representation as a 2-D image as an input of a Conv2D network

Problem

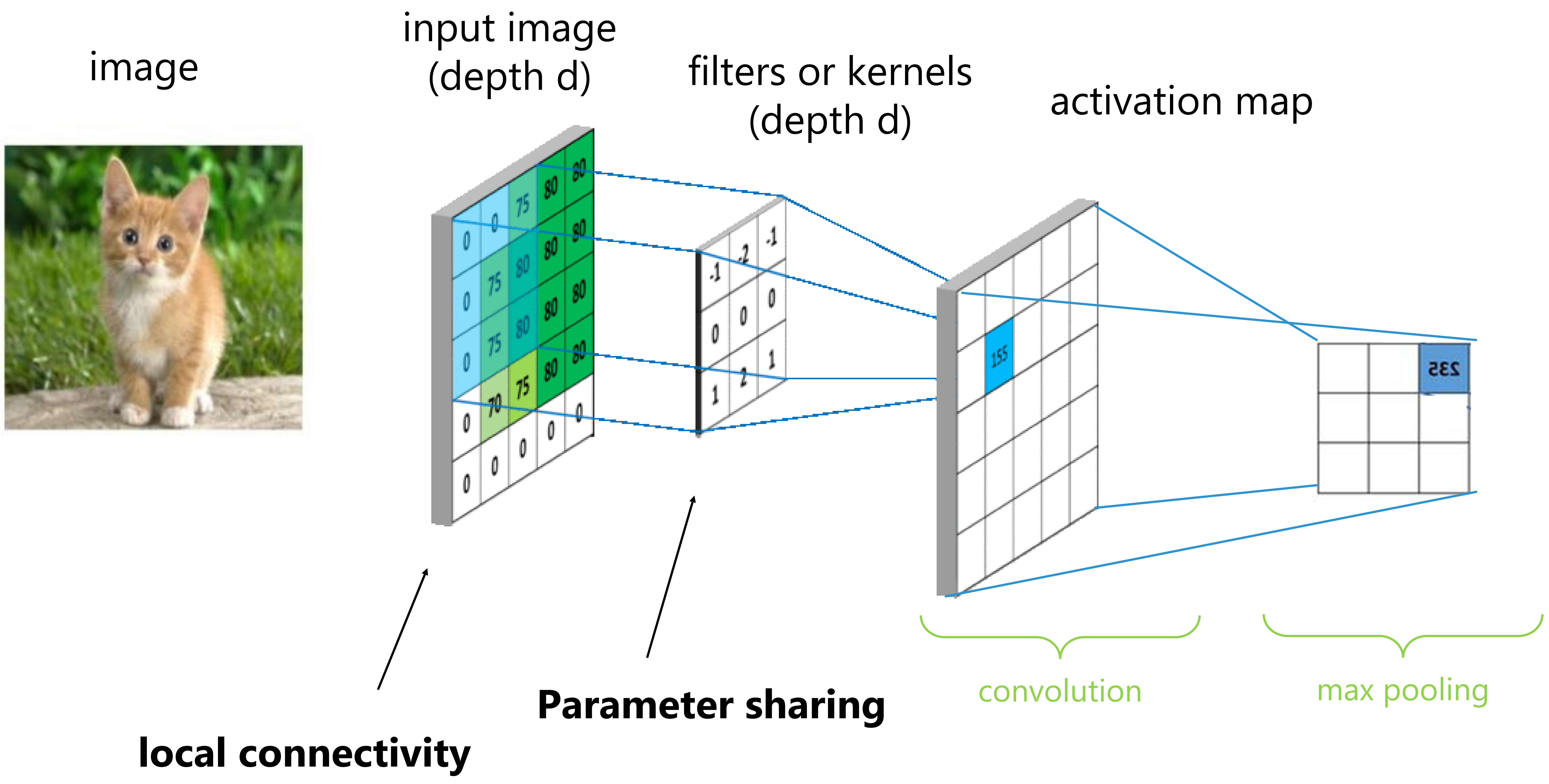
A time-frequency representation is not an image

Research direction

Choice of 1st Conv2D layer filters

APPROACHES FOR FEATURE LEARNING

Concept of Conventional Neural Network (CNN or ConvNet)



APPROACHES FOR FEATURE LEARNING

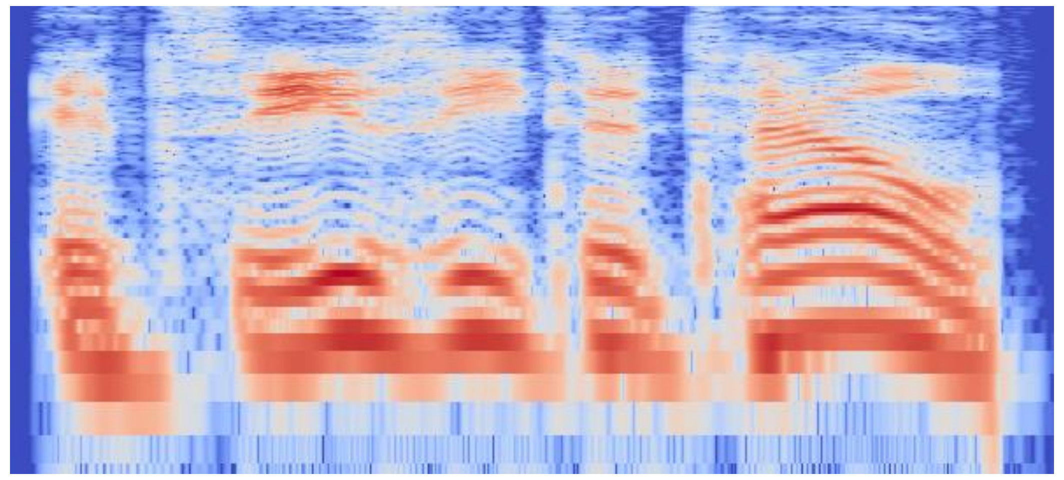
2D representation (time-frequency)

Natural images

- the 2 axis represent the same concept
- Whatever the position of an element is, it represents the same thing
  - → spatial invariance
  - → weight sharing in the 2 dimensions
- close pixels are often strongly correlated
- close and similar pixel usually belong to a same object

Harmonic sounds

- The 2 axis represent completely different things
- Properties of a sound event:
  - → same signification whenever it is played
  - → usually, different signification depending on its frequency
  - → no invariance in frequency (even in log-scale)
- Frequencies of a same source are not distributed locally on spectrogram (sparsity)

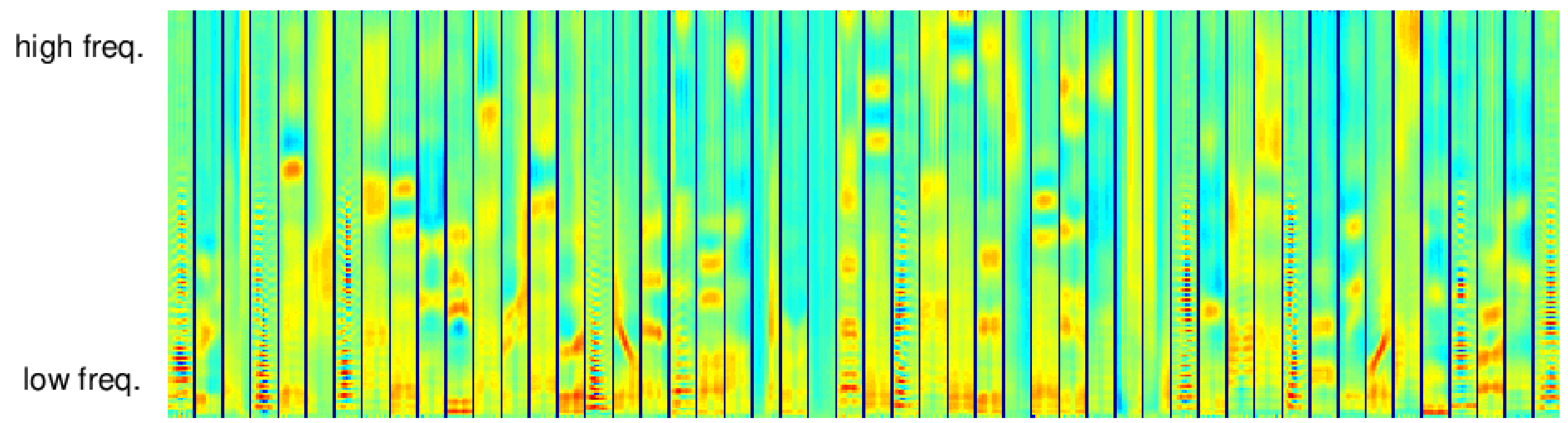


APPROACHES FOR FEATURE LEARNING

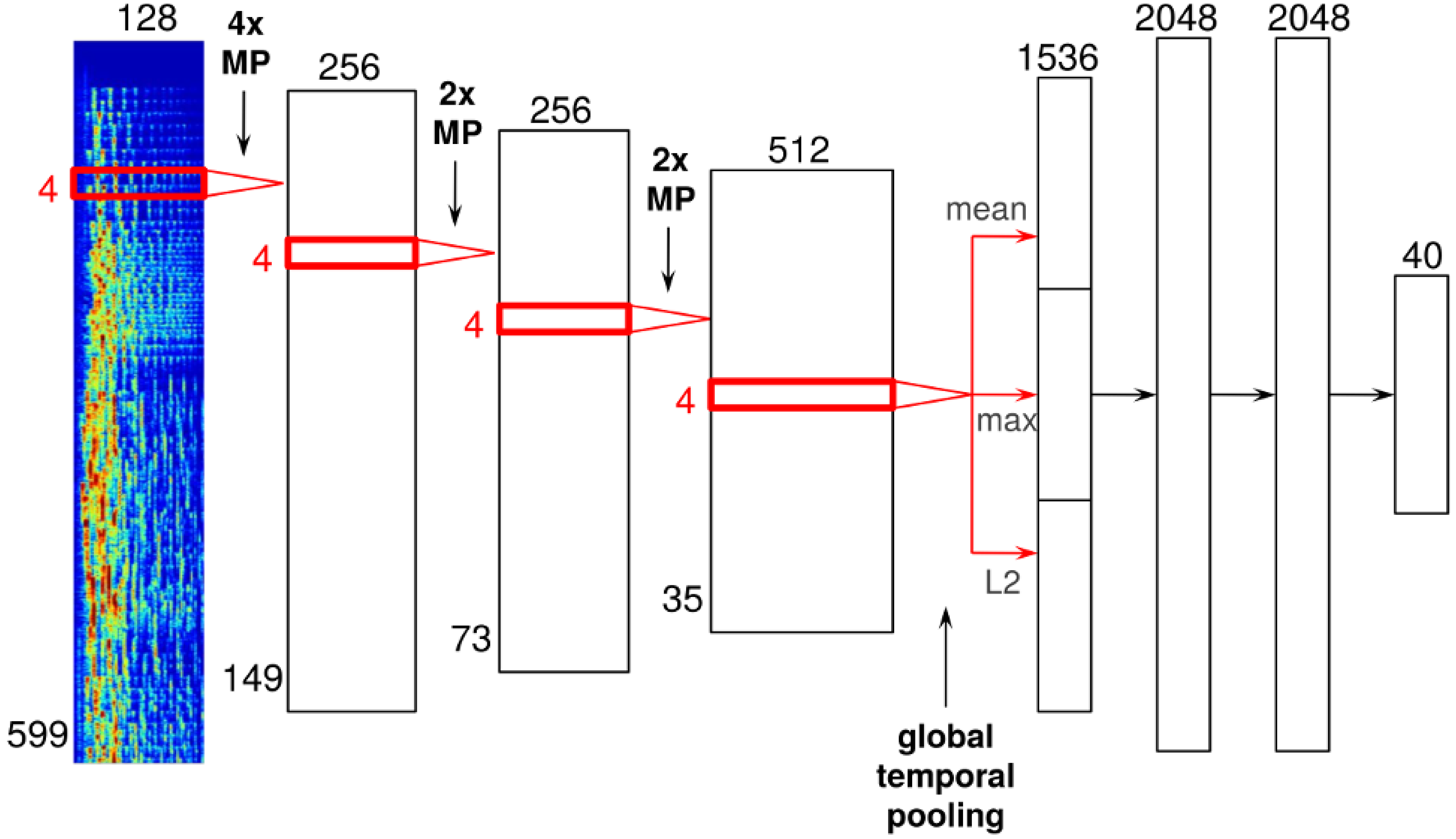
An adapted shape of 2D filters for 1st layer

1) Filter covering the entire frequency bandwidth

→ convolution only over time



Each column represents a « temporal receptive field » of a 1st layer basis in the spectrogram space



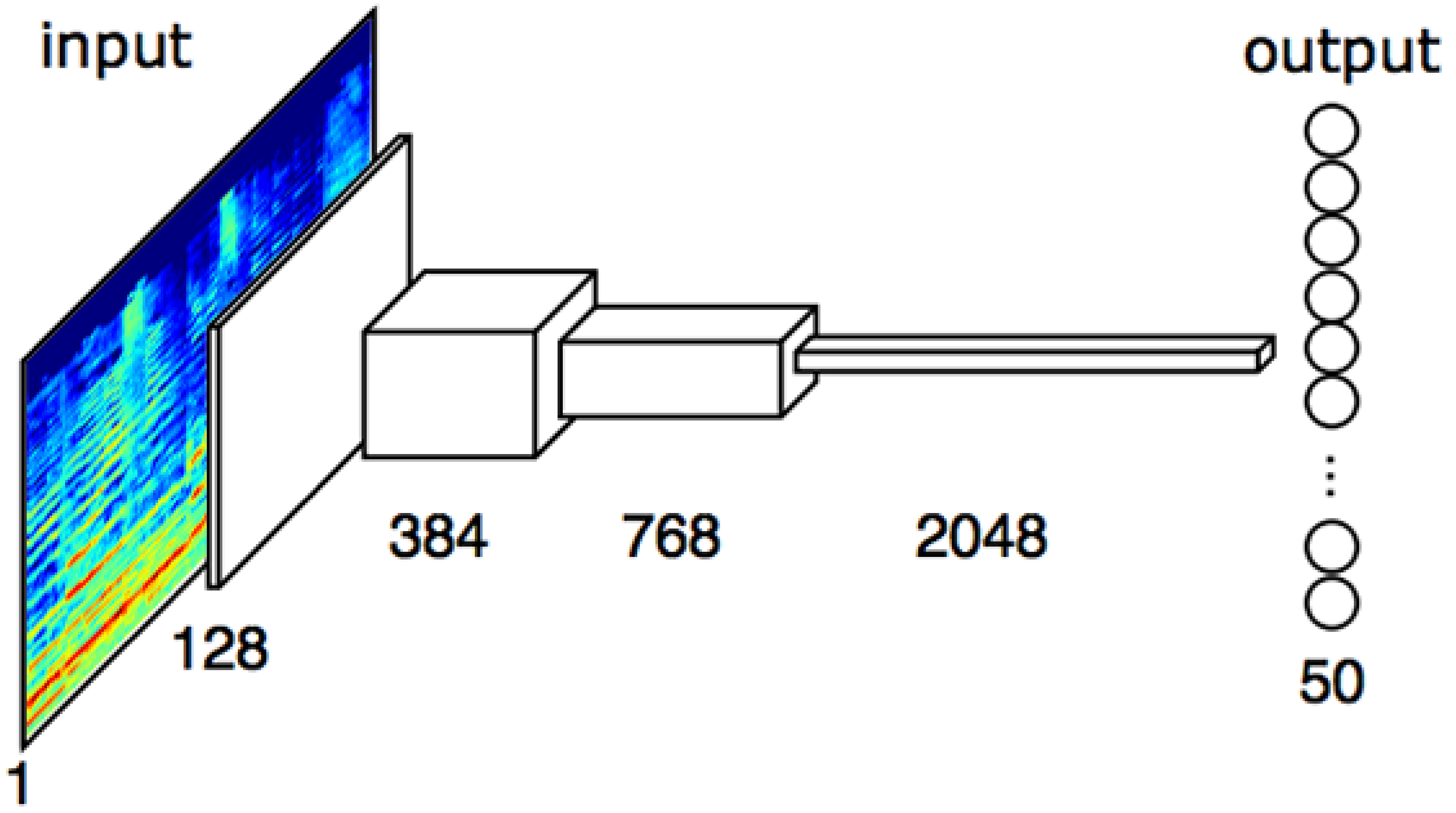
S. Dieleman. Recommending music on Spotify with deep learning

APPROACHES FOR FEATURE LEARNING

An adapted shape of 2D filters for 1st layer

2) 3x3 or 5x5 filters, used as in image processing

VGG-net on spectrum
→ for music automatic tagging (multi-label task)
→ using STFT, MFCC and Mel-gram representation as input



FCN-5	FCN-6	FCN-7
Mel-spectrogram *(input: 96×1366×1)*		
Conv 3×3×128		
MP (2, 4) *(output: 48×341×128)*		
Conv 3×3×256		
MP (2, 4) *(output: 24×85×256)*		
Conv 3×3×512		
MP (2, 4) *(output: 12×21×512)*		
Conv 3×3×1024		
MP (3, 5) *(output: 4×4×1024)*		
Conv 3×3×2048		
MP (4, 4) *(output: 1×1×2048)*		
·	Conv 1×1×1024	Conv 1×1×1024
	·	Conv 1×1×1024
Output 50×1 (sigmoid)		

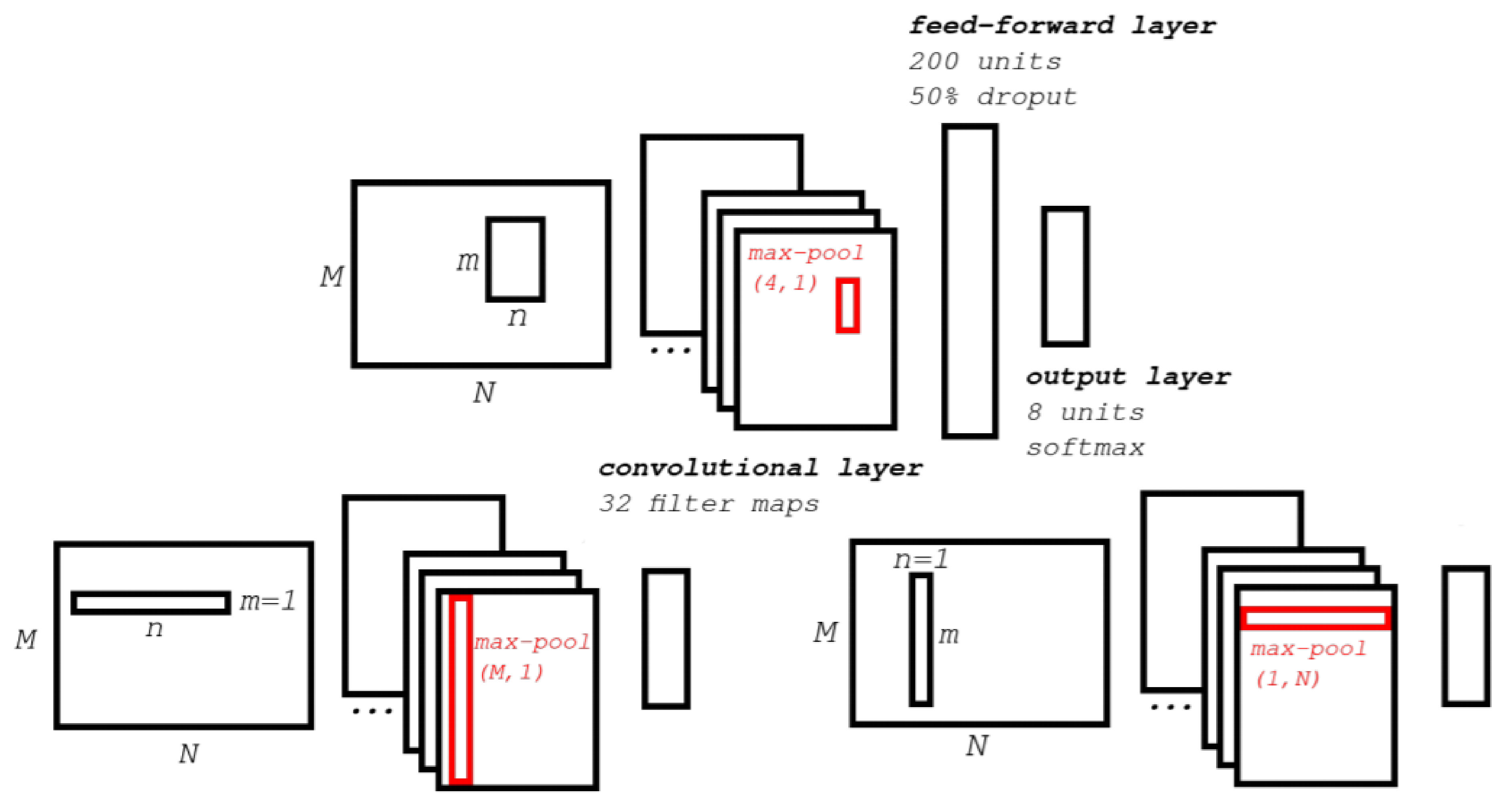
K. Choi, G. Fazekas and M. Sandler. Automatic tagging using deep convolutional neural networks. ISMIR 2016.
K. Simonyan and A. Zisserman. Very deep conventional networks for large-scake image recognition. 2015.

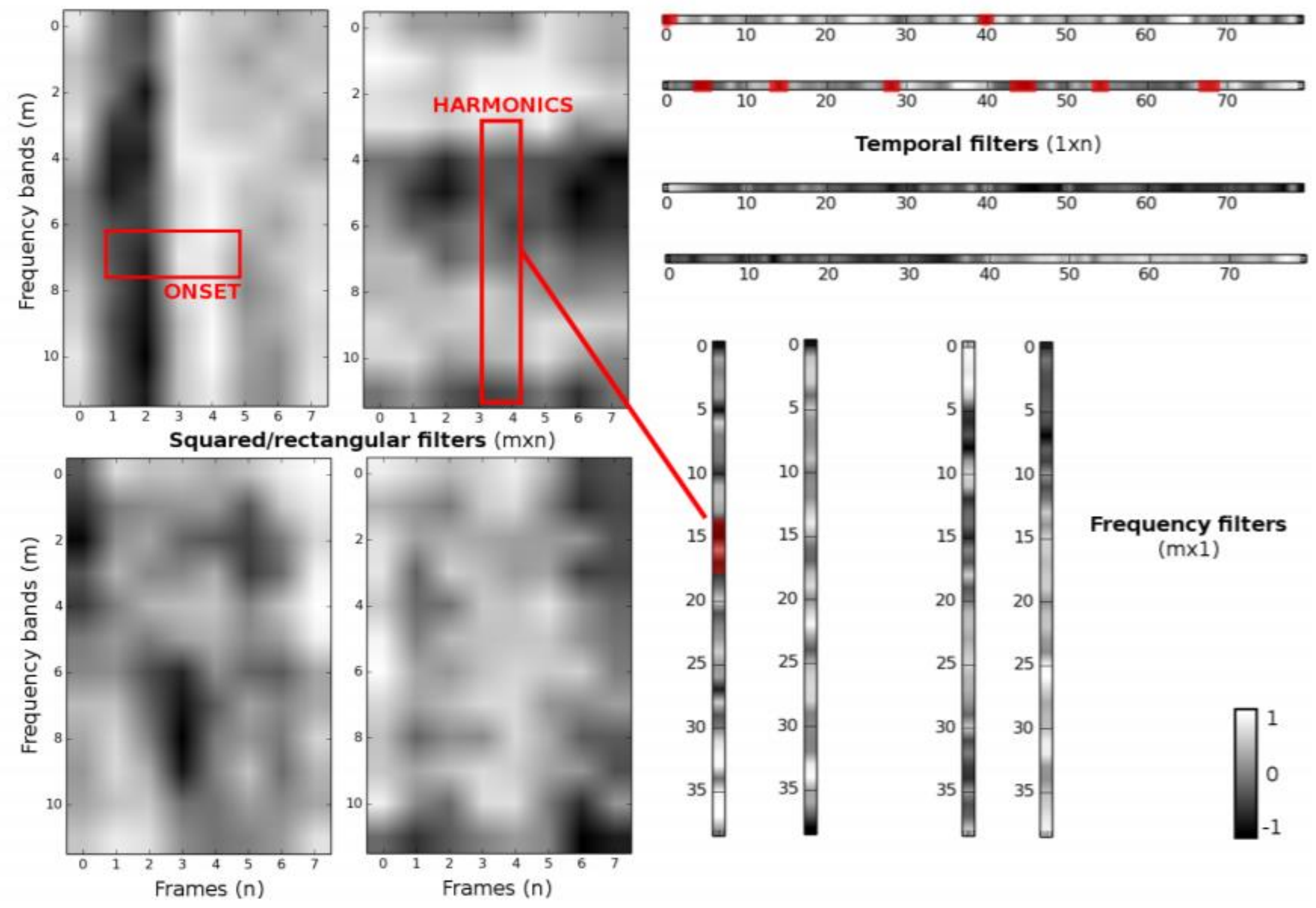
APPROACHES FOR FEATURE LEARNING

An adapted shape of 2D filters for 1st layer

3) Adapted filered forms to highlight specific properties

Musically-motivated CNN
→ the filters shape is suited to represent timbre (vertically) and to represent ryhtm (horizontally)





J. Pons , T. Lidy and X. Serra. Experimenting with Musically Motivated Convolutional Neural Networks. 2017.

APPROACHES FOR FEATURE LEARNING

<u>1D representation</u>

Representation

- Raw waveform (end-to-end learning)

Idea

- Learn the filter to be applied directly on the waveform to get the most appropriate representation for a given task

Problem

- How to model time invariance ?

Research still going on to explore this approach.

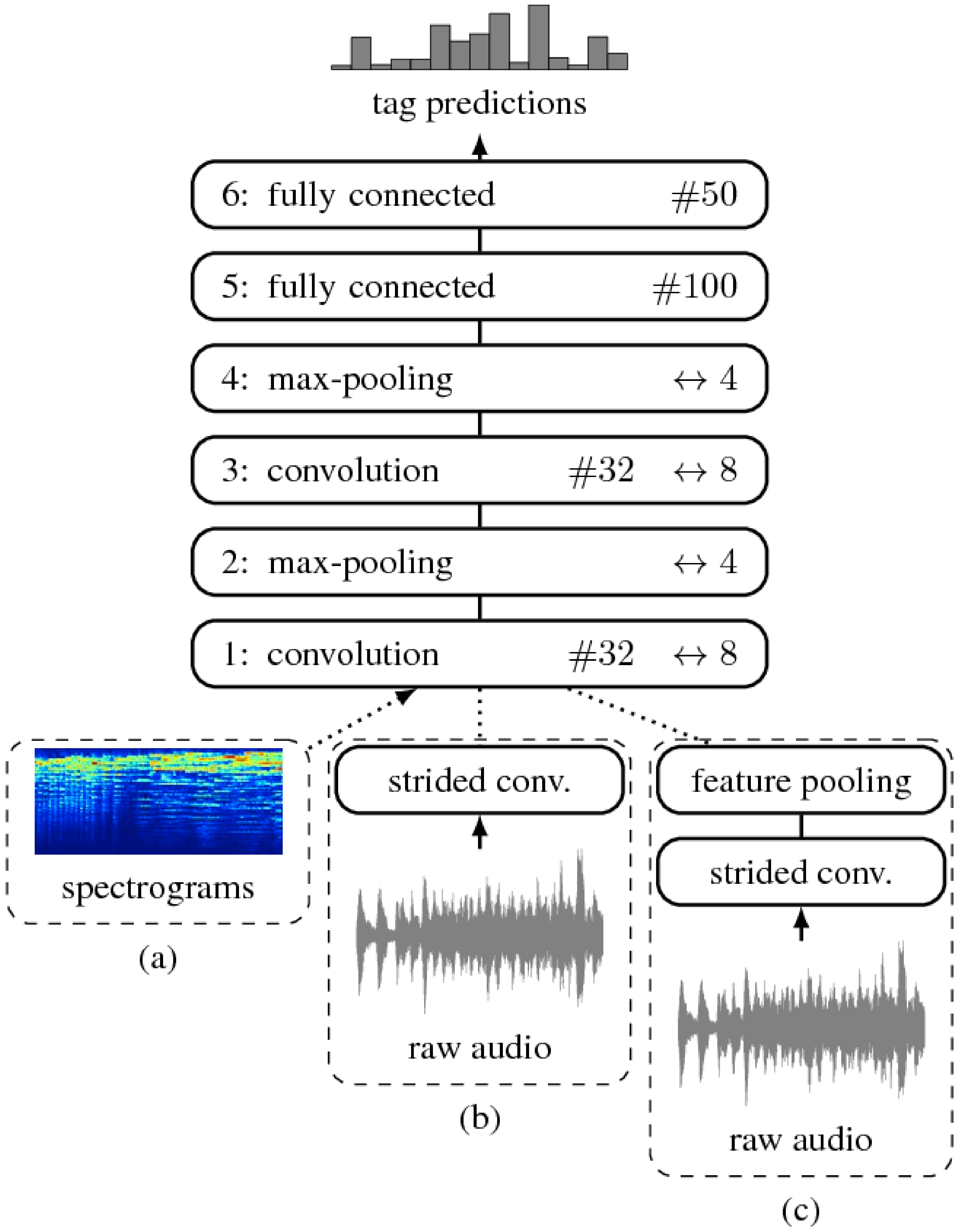
APPROACHES FOR FEATURE LEARNING

An adapted shape of 1D filters for 1st layer

1) Filter with unique size

Using waveform in input of a convolutional network (Conv1D, TDC)
→ filter size and stride: parameter of STFT

problem: is temporal invariance respected ?



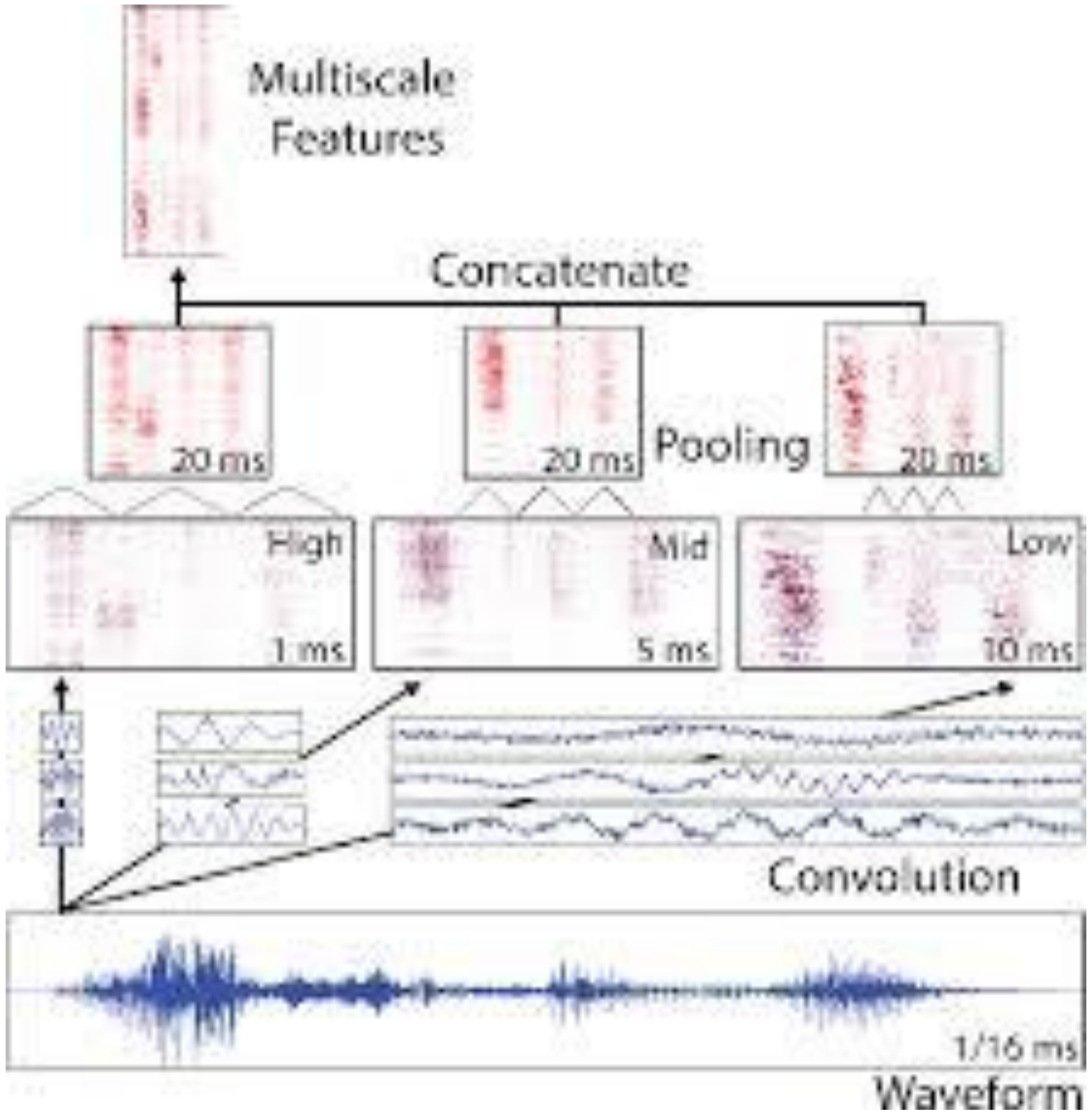
S. Dieleman, B. Schrauwen. End-to-end learning for music audio. 2014.

APPROACHES FOR FEATURE LEARNING

An adapted shape of 1D filters for 1st layer

2) Filters of different size

Multi-scale approach

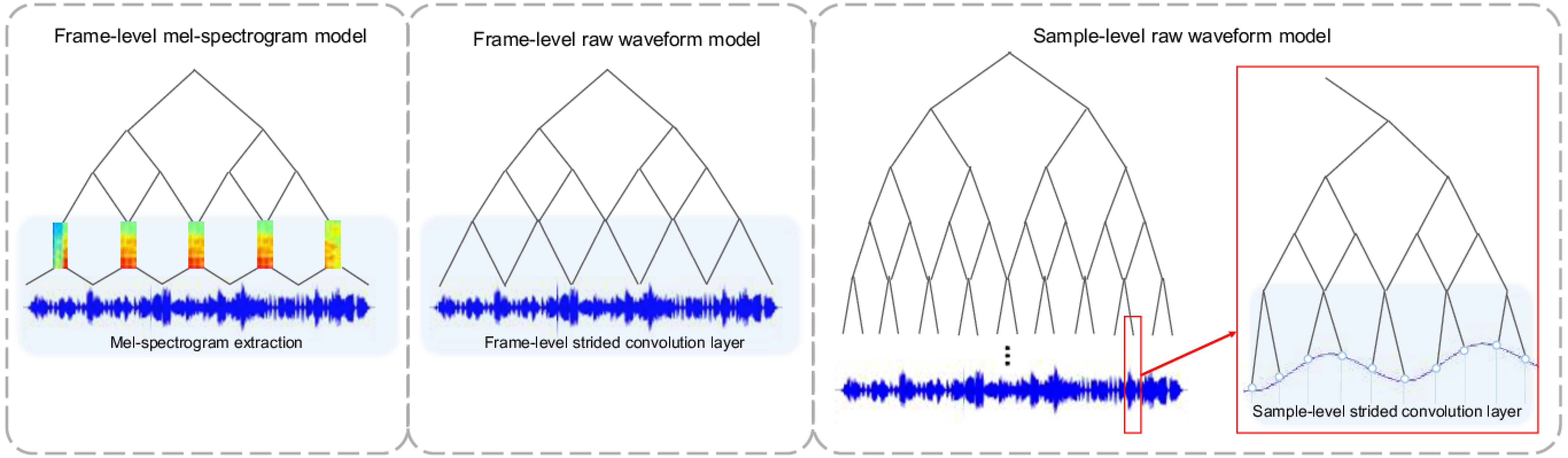


Z. Zhu, J. H. Engel, A. Hannun. Learning Multiscale Features Directly From Waveforms. 2016

APPROACHES FOR FEATURE LEARNING

An adapted shape of 1D filters for 1st layer

3)
Sample CNN: VGG-net on waveform
→ make it easier to ensure time invariance

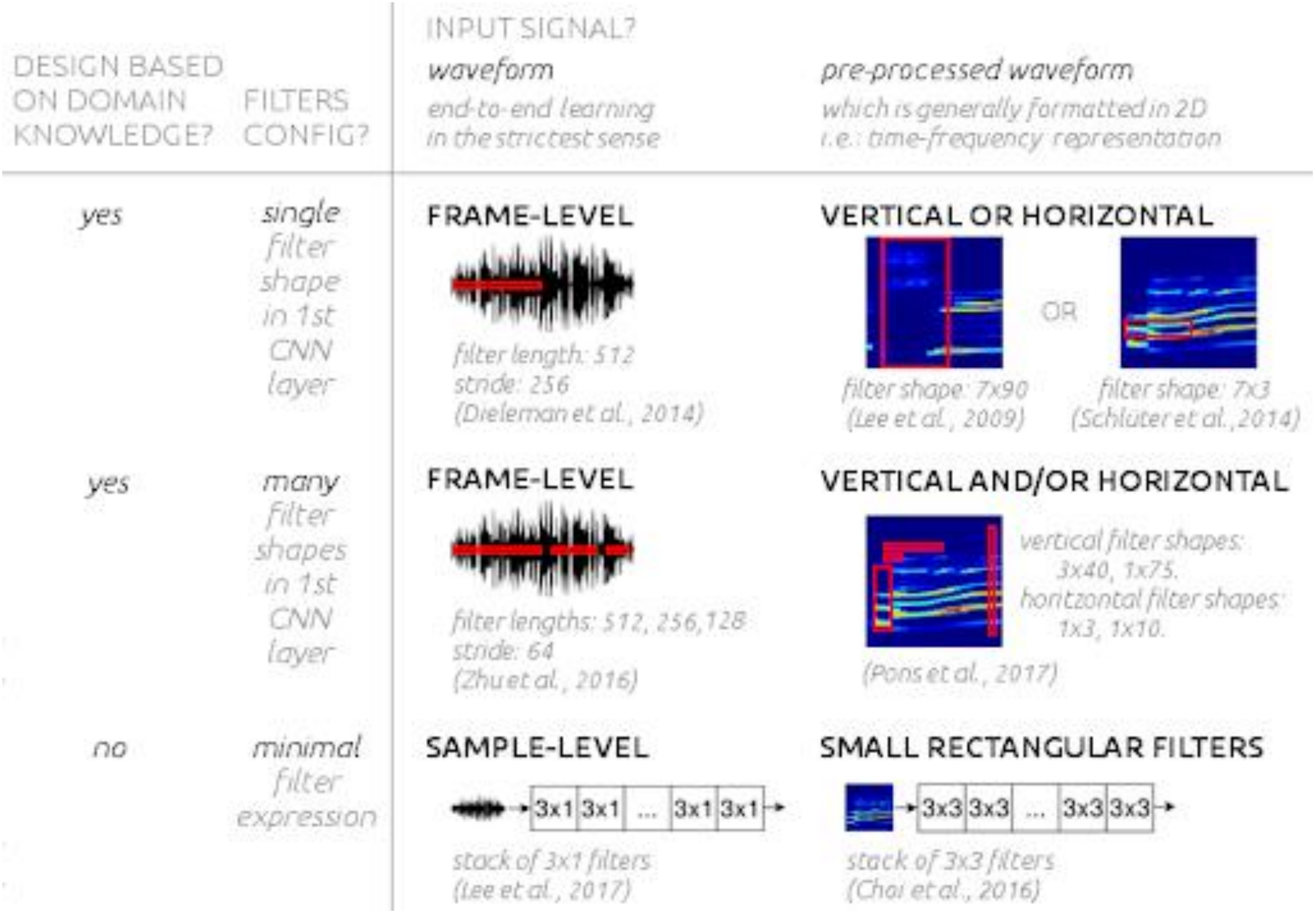


3^9 model, 19683 frames 59049 samples (2678 ms) as input			
layer	stride	output	# of params
conv 3-128	3	19683×128	512
conv 3-128 maxpool 3	1 3	19683×128 6561×128	49280
conv 3-128 maxpool 3	1 3	6561×128 2187×128	49280
conv 3-256 maxpool 3	1 3	2187×256 729×256	98560
conv 3-256 maxpool 3	1 3	729×256 243×256	196864
conv 3-256 maxpool 3	1 3	243×256 81×256	196864
conv 3-256 maxpool 3	1 3	81×256 27×256	196864
conv 3-256 maxpool 3	1 3	27×256 9×256	196864
conv 3-256 maxpool 3	1 3	9×256 3×256	196864
conv 3-512 maxpool 3	1 3	3×512 1×512	393728
conv 1-512 dropout 0.5	1 –	1×512 1×512	262656
sigmoid	–	50	25650
Total params			1.9×10^6

J. Lee, J. Park, K. Luke Kim, J. Nam. Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms 2017

APPROACHES FOR FEATURE LEARNING

An adapted shape of 1D filters for 1st layer

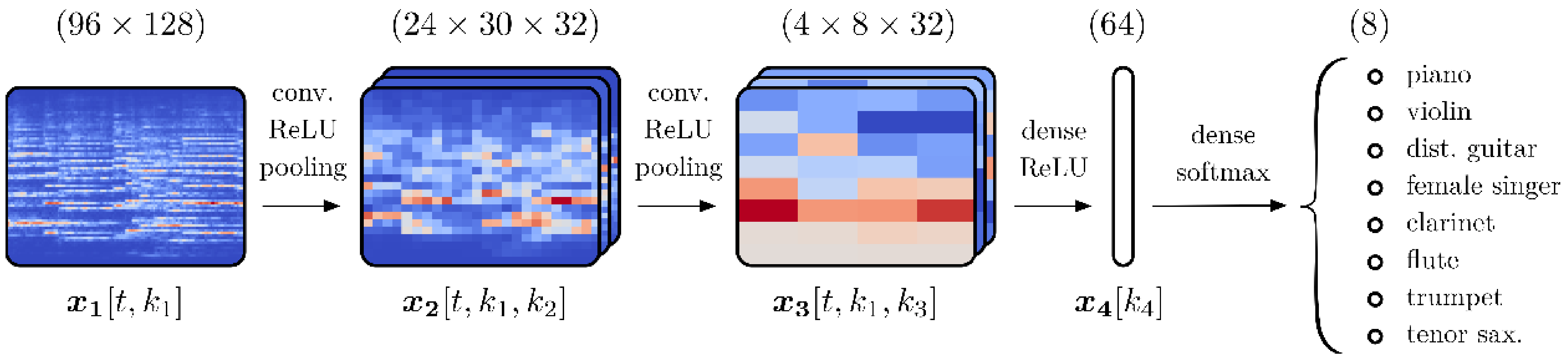


J. Pons 2019

APPROACHES FOR FEATURE LEARNING

Signal models

1) Source-filter + harmonic



Application : instruments recognition with ConvNet

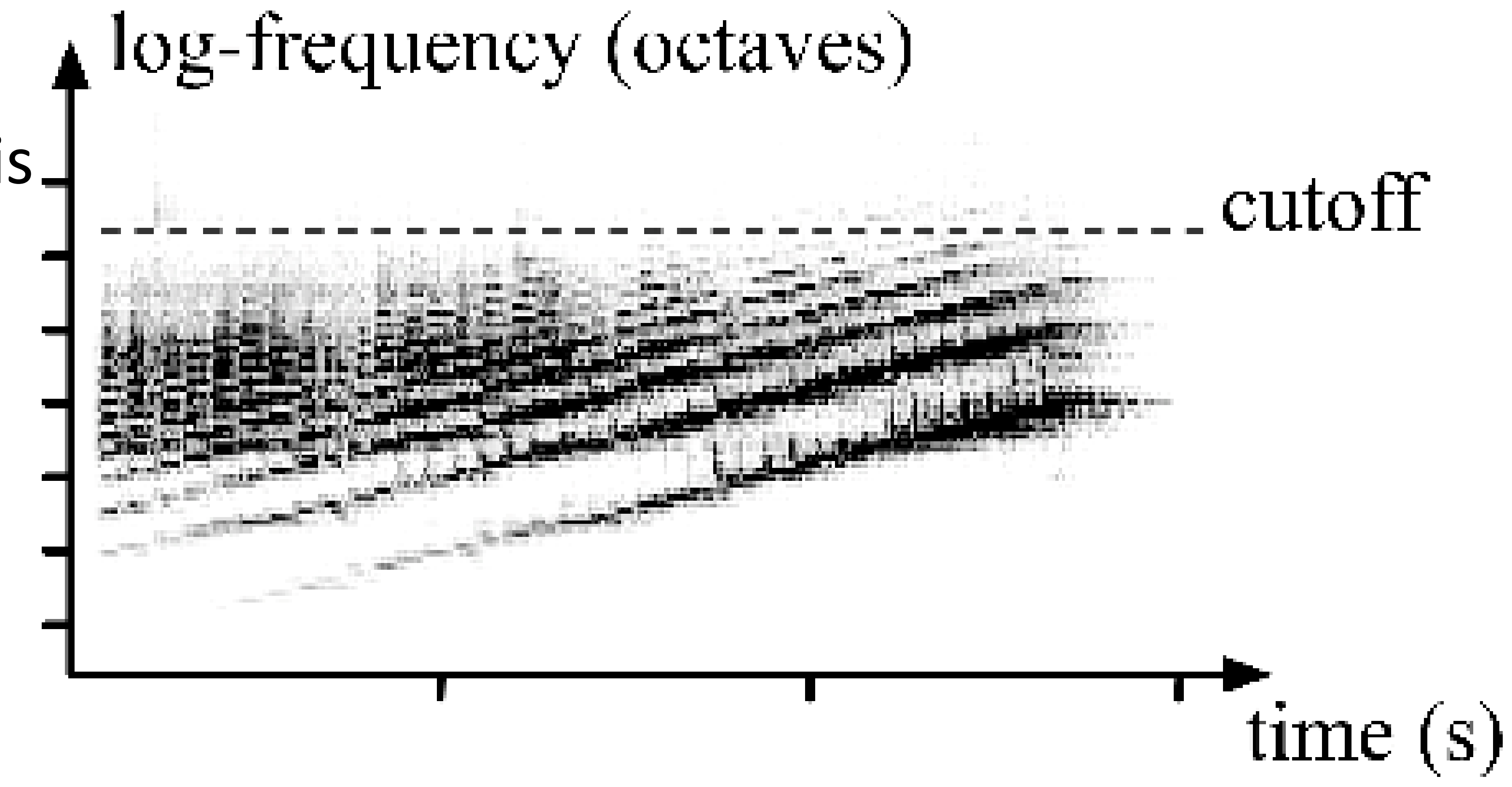
CQT: co-variant to transpositions

- chromatic scales are parallel diagonals
- low energy for frequencies over cut-off frequency

In high frequencies (f>cut-off):

- harmonics closed to each others, regularly distributed over frequency axis
- transposed sounds have similar spectra
- high correlation between CQT of different pitches
- the energy in a definite bandwidth is pitch-independant

  → 1D filter (convolution over time only)

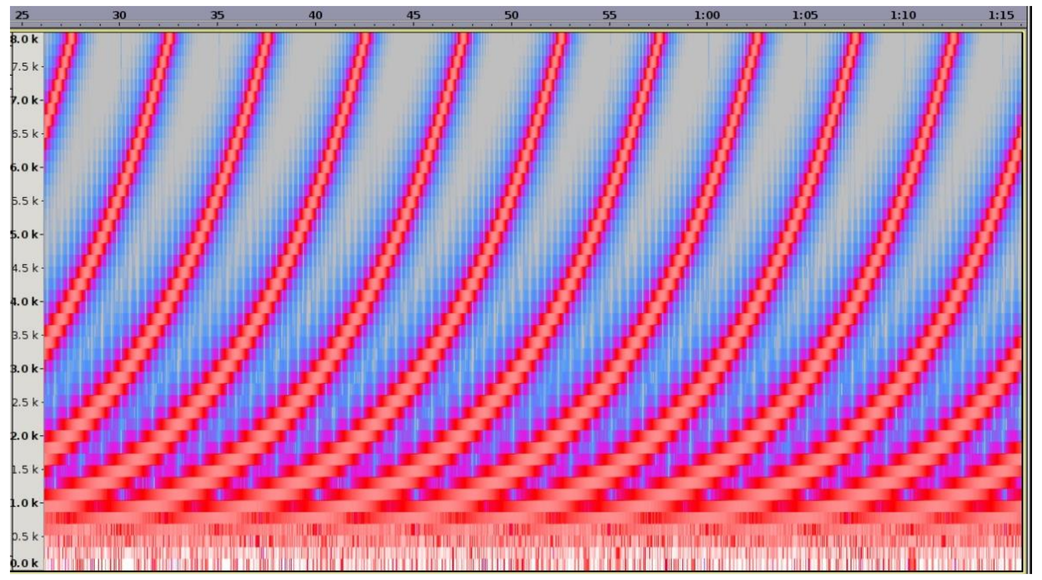


V. Lostanlen, C. E. Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. 2016.

APPROACHES FOR FEATURE LEARNING

Signal models

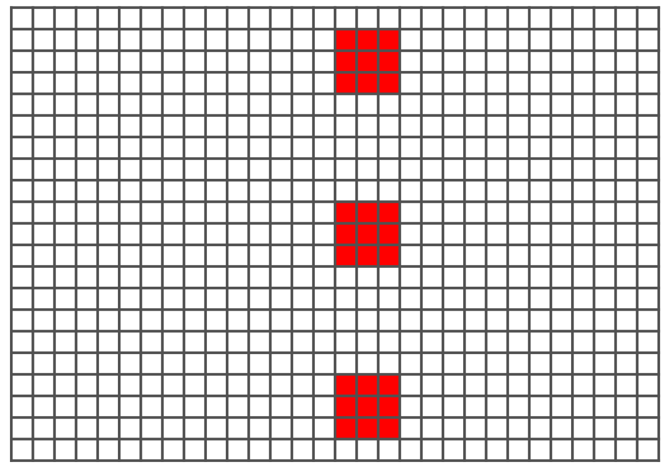
1) Source-filter + harmonic

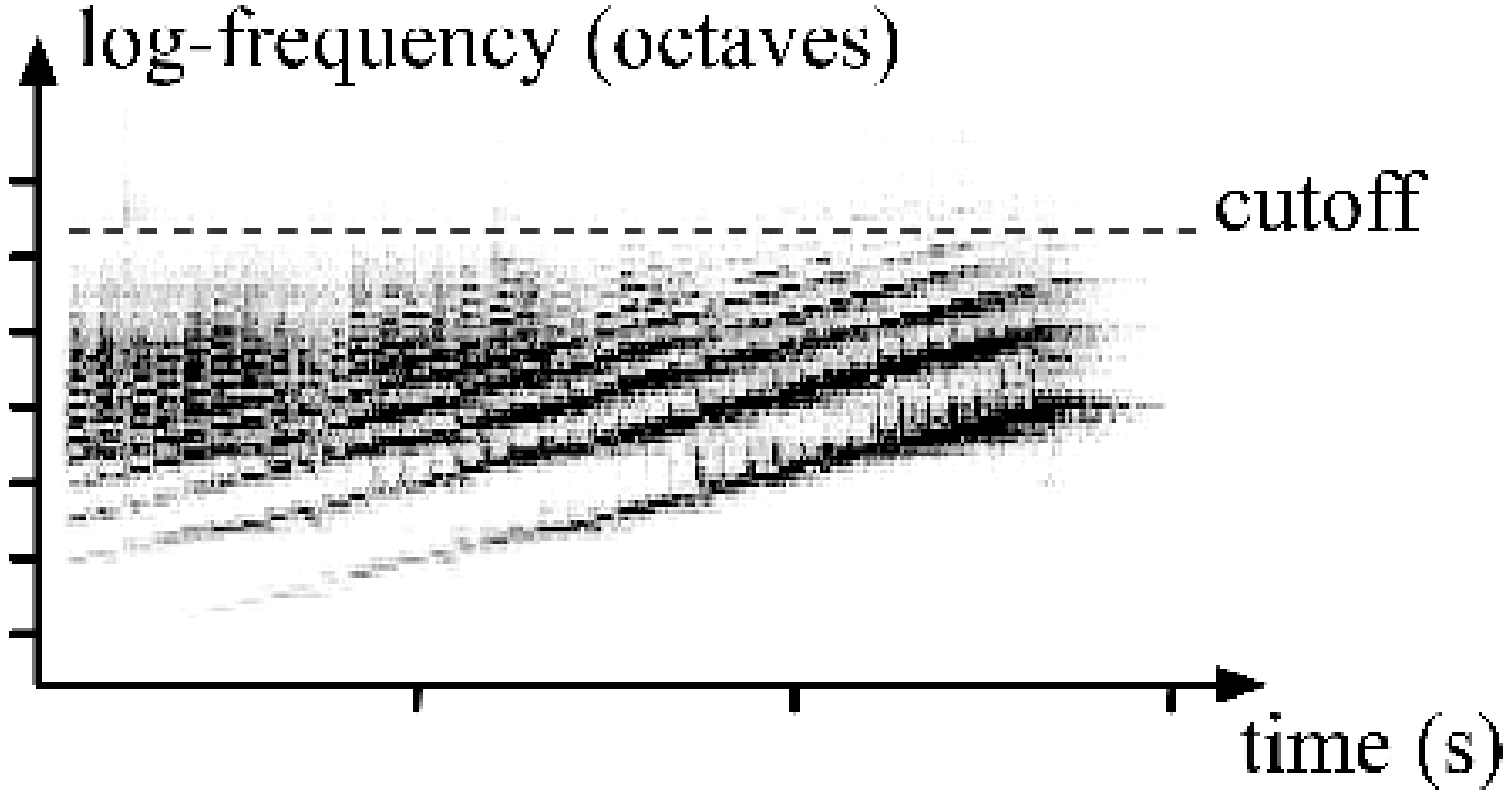


<u>In low frequencies (f<cut-off):</u>

- harmonic comb is sparse and co-variant with pitch
- harmonic structure is well described, measuring correlation between harmonics themselves
- log-frequency axis is rolled on Shepard's spiral
  → time-frequency filters have a 1-octave differences over frequencies

$$\boldsymbol{y_2}[t, k_1, k_2] = \boldsymbol{b_2}[k_2] + \sum_{\tau, \kappa_1, j_1} \boldsymbol{W_2}[\tau, \kappa_1, j_1, k_2] \times \boldsymbol{x_1}[t - \tau, k_1 - \kappa_1 - Qj_1]$$





V. Lostanlen, C. E. Cella. Deep convolutional networks on the pitch spiral for musical instrument recognition. 2016.

APPROACHES FOR FEATURE LEARNING

Signal models

2) Source-filter

Application: main melody estimation

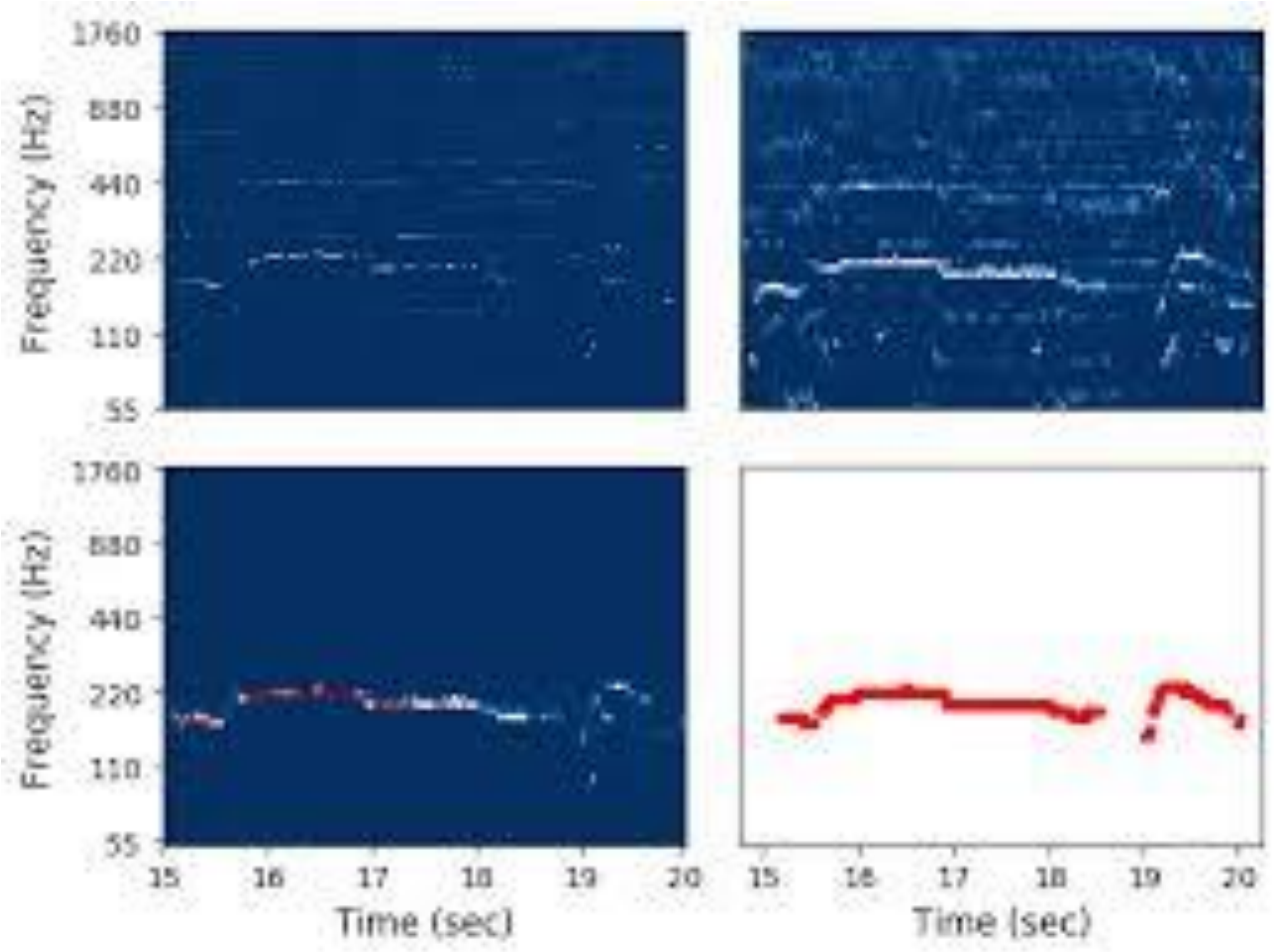
Idea:

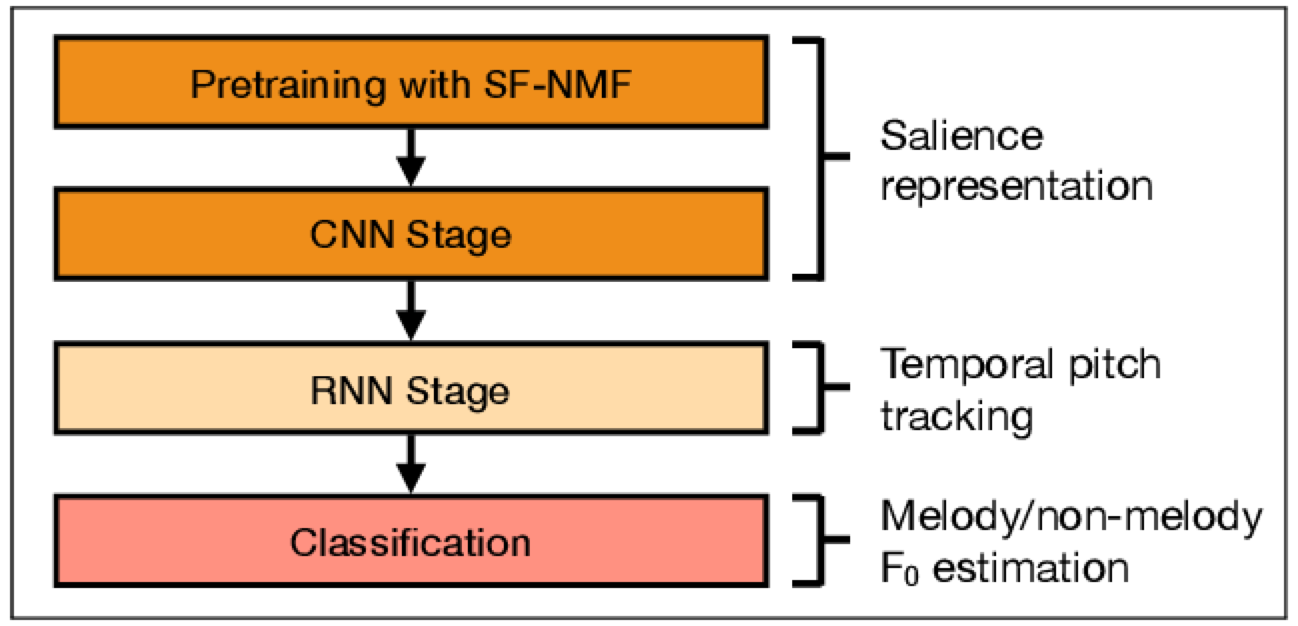
- Desomposition of signal according to NMF model [Durrieu, 2010]

$$\boldsymbol{V} \approx \widehat{\boldsymbol{V}} = \overset{\text{source}}{\boldsymbol{V}^{F_0}} \odot \overset{\text{filter}}{\boldsymbol{V}^{\Phi}} + \overset{\text{background}}{\boldsymbol{V}^{B}}$$

$$= \underset{\substack{\text{NMF}\\ \text{F0 basis}}}{\boldsymbol{W}^{F_0}} \, \underset{\substack{\text{source}\\ \text{F0 activations}}}{\boldsymbol{H}^{F_0}} \odot \underset{\text{NMF}}{\boldsymbol{W}^{\Phi}} \underset{\text{filter}}{\boldsymbol{H}^{\Phi}} + \underset{\text{NMF}}{\boldsymbol{W}^{B}} \underset{\text{background}}{\boldsymbol{H}^{B}}$$

- Estimated activations F_0 are used as inputs of a CNN or RNN





D. Basaran, S. Essid and G. Peeters. Main Melody Extraction with Source-Filter NMF and CRNN. ISMIR, 2018

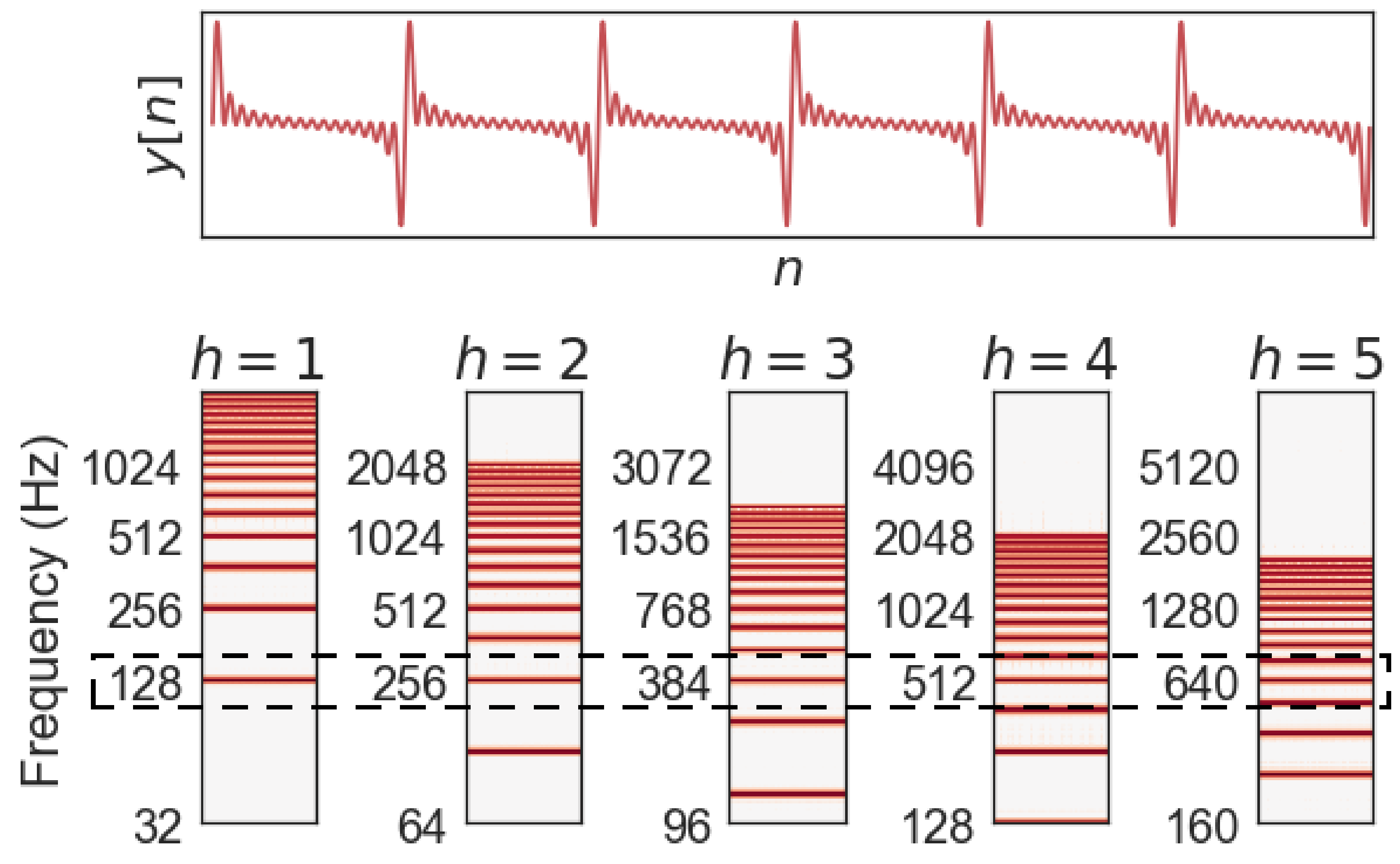
J. L. Durrieu, G. Richard, B. David and C. Fevotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. 2010.

APPROACHES FOR FEATURE LEARNING

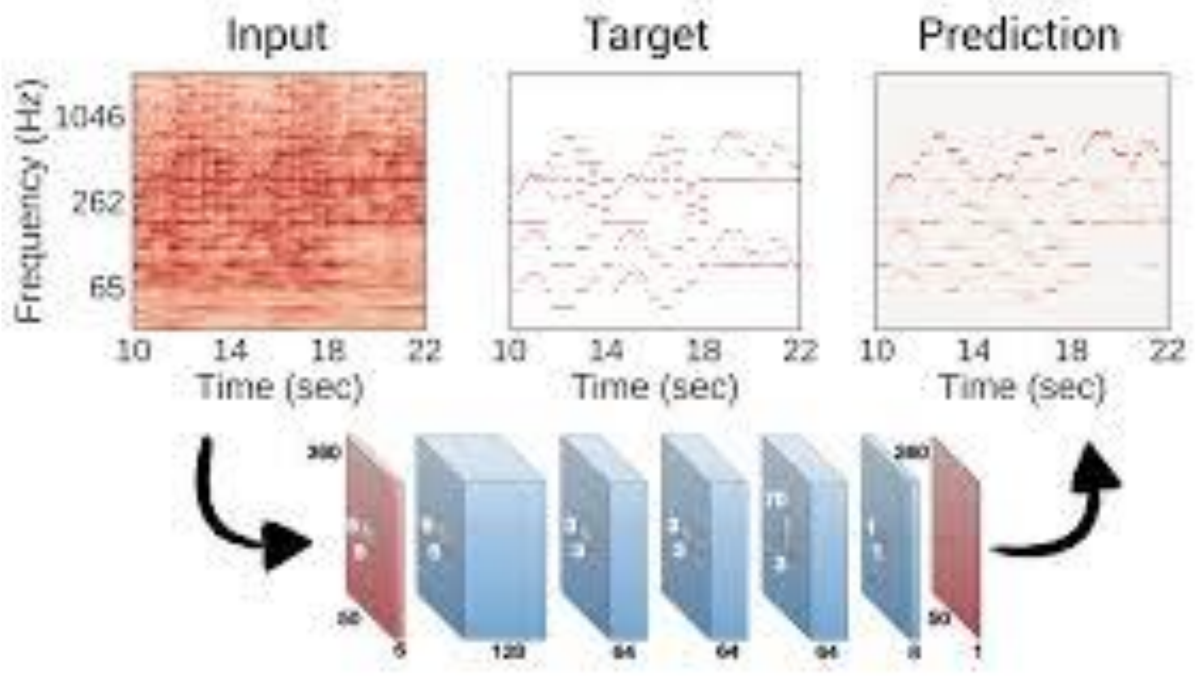
Signal models

3) Harmonic-Constant-Q-Transform (HCQT):

- Representation of audio signal using Constant-Q-Transform (CQT)
- Several CQT are computed for several minimal frequencies $hf_{\min}$
  → harmonics of hf_0 at same position in different CQTs
  → CQTs are stacked in input layer (RGB) depth



Application: multi-pitch estimation



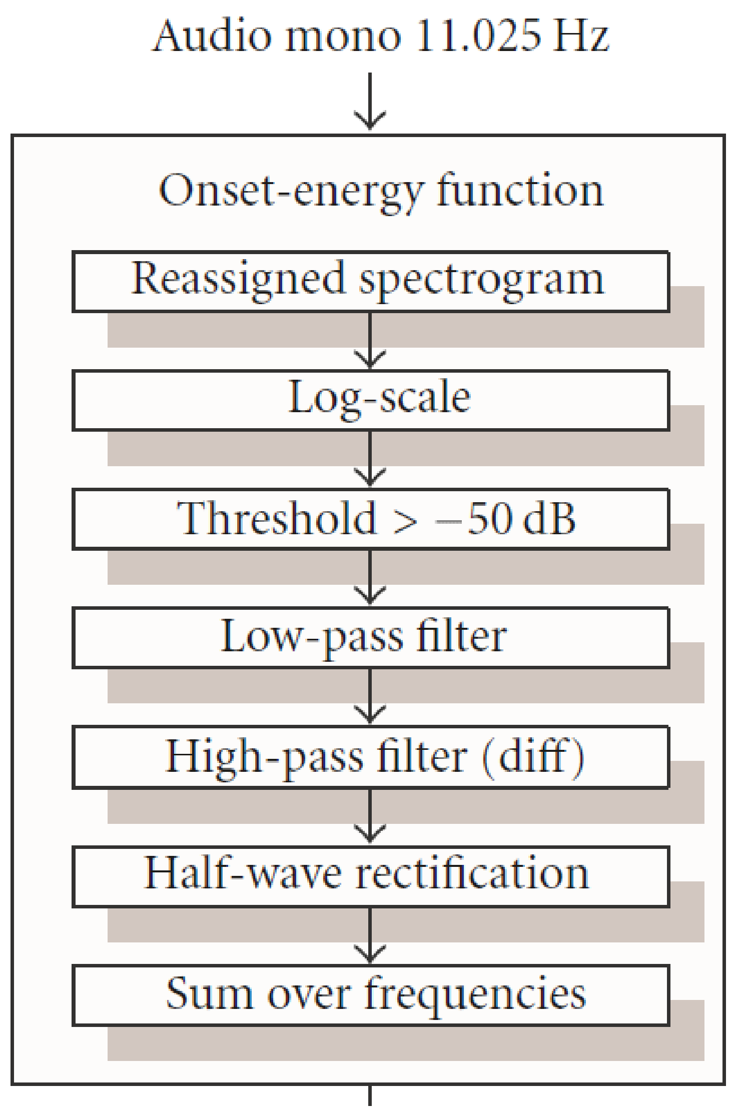
R. M. Bittner, B. McFee, J. Salamon, P. Li, J. P. Bello. Deep Salience Representations for F0 Estimation in Polyphonic Music. ISMIR 2017

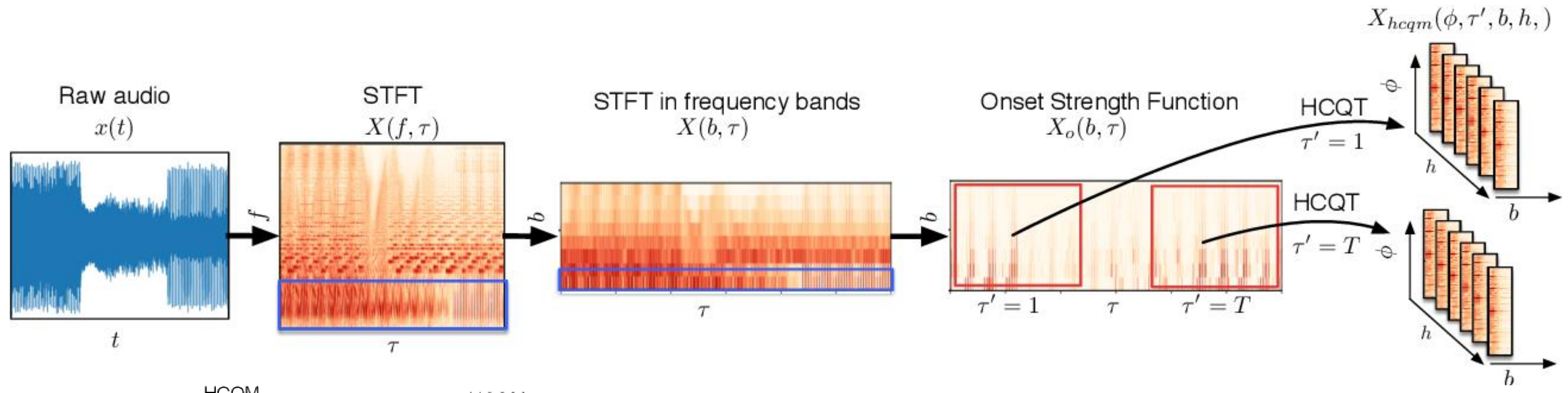
APPROACHES FOR FEATURE LEARNING

Signal models

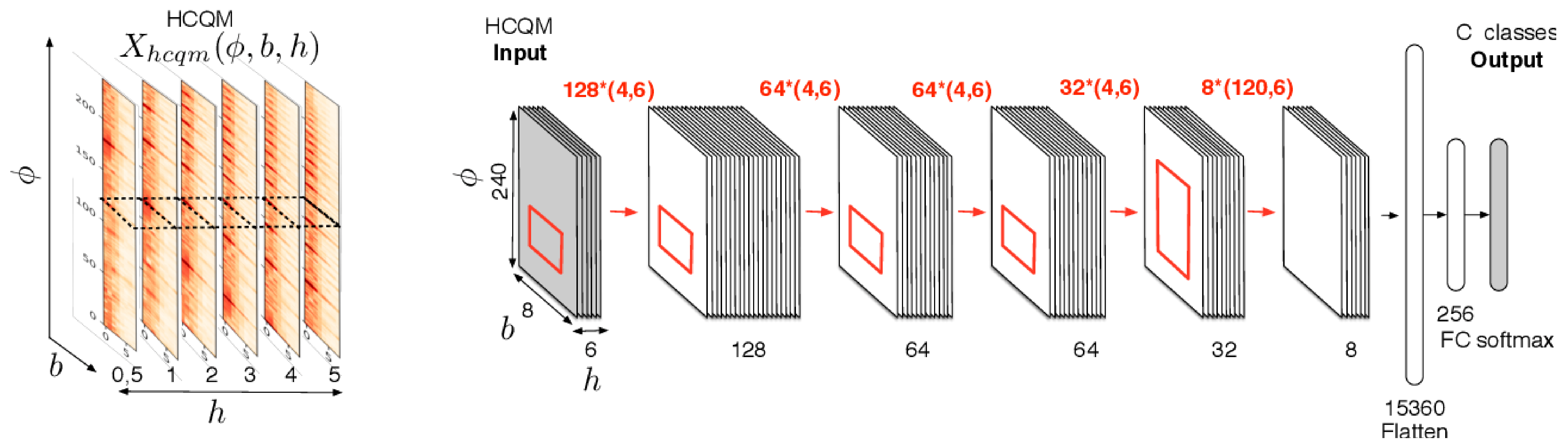
3) Harmonic-Constant-Q-Modulation (HCQM):

- In each percetive frequency bandwidth b, an onset function is computed $X_0(b,\tau)$
- The periodical content (tempo, metric, rythm) of each onset function is represented using an HCQT $X_{hcqm}(\phi,\tau',b,h)$ with ϕ modulation, h harmonic





<u>Application</u>:
- Tempo estimation
- rythm classification



H. Foroughmand, G. Peeters. Deep-Rhythm for Global Tempo Estimation in Music. 2019.

APPROACHES FOR FEATURE LEARNING

Generative models

A generative model unable to generate an audio signal x(m) using a z-representation
Usually:

- if z is a complex STFT
  → use DFT inverse

- If z is a spectrogram (magnitude of STFT)
  → use DFT inverse and try to reconstruct phase with Griffin and Lim algorithm
  (many artifacts)

Otherwise, do not use Fourier Transform anymore for that purpose

APPROACHES FOR FEATURE LEARNING

Generative models

1) Neural-Autoregressive Models: WaveNet

Generative model operating directly on audio samples

- « raw audio → challenging models
- Based on PixelCNN
- High resolution and long-term dependancies

Autoregressive model

next sample is almost reconstructed from linear convolution of past samples

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. 2016.

APPROACHES FOR FEATURE LEARNING

Generative models

1) Neural-Autoregressive Models: WaveNet
Causal convolution
requires many layers of large filters to increase receptive field

Dilated convolution (wholes)
increase the receptive field by orders of magnitude

Stacked dilated convolution
dilation doubled

Input/ouptut signal representation

Softmax layer

Conditional wavenet

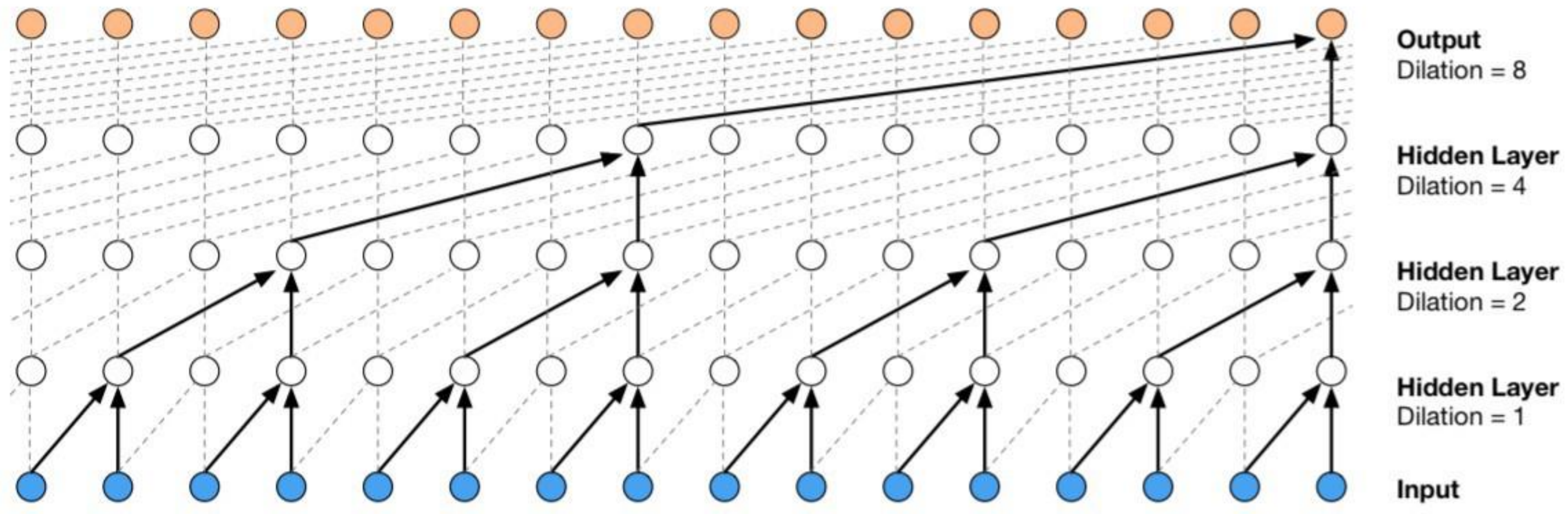
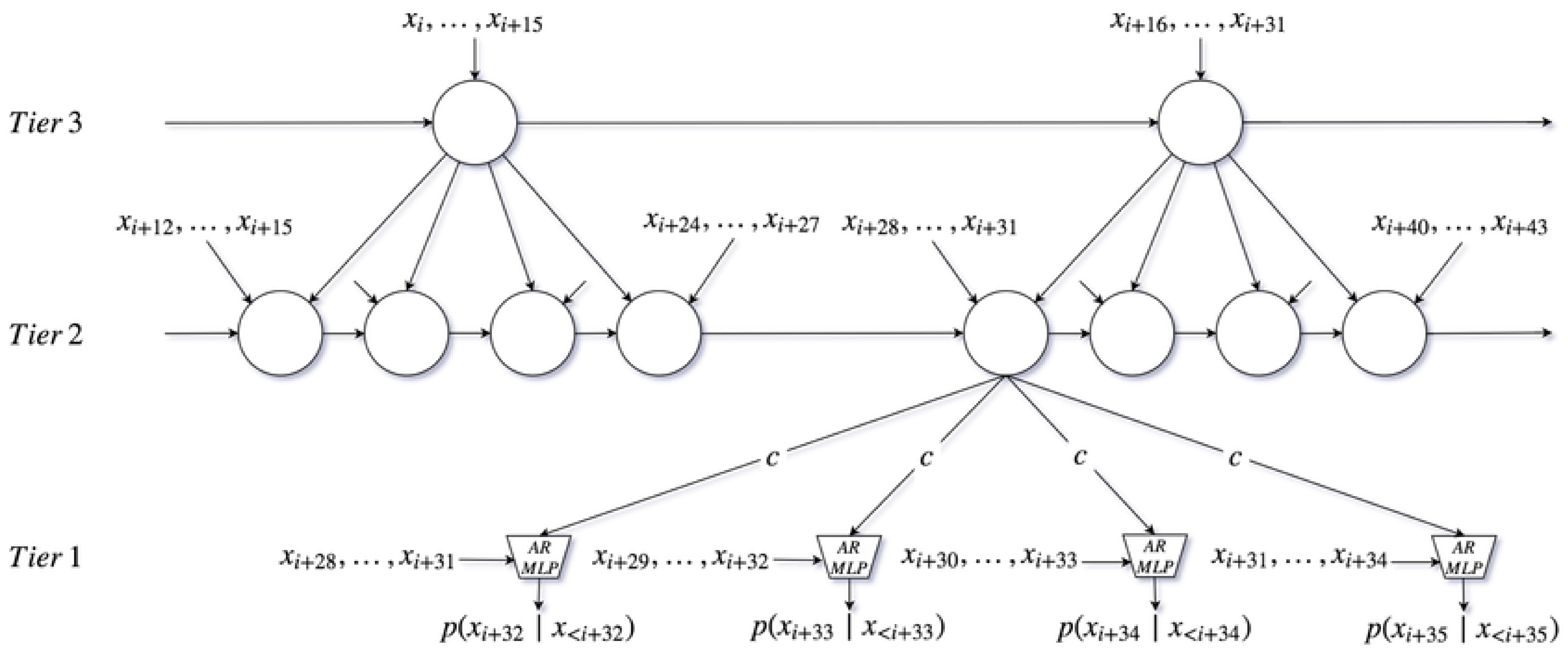


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

APPROACHES FOR FEATURE LEARNING

Generative models

2) Neural-Autoregressive Models: SampleRNN



S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, Y. Bengio. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. 2017.

DEEP LEARNING FOR AUDIO AND SPEECH PROCESSING.

New learning paradigms

NEW LEARNING PARADIGMS

Classification

- Binary classification

$$x^{(i)} \rightarrow \hat{y}^{(i)} = f_\theta(x^i)$$

$$\theta^* = \min_\theta \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

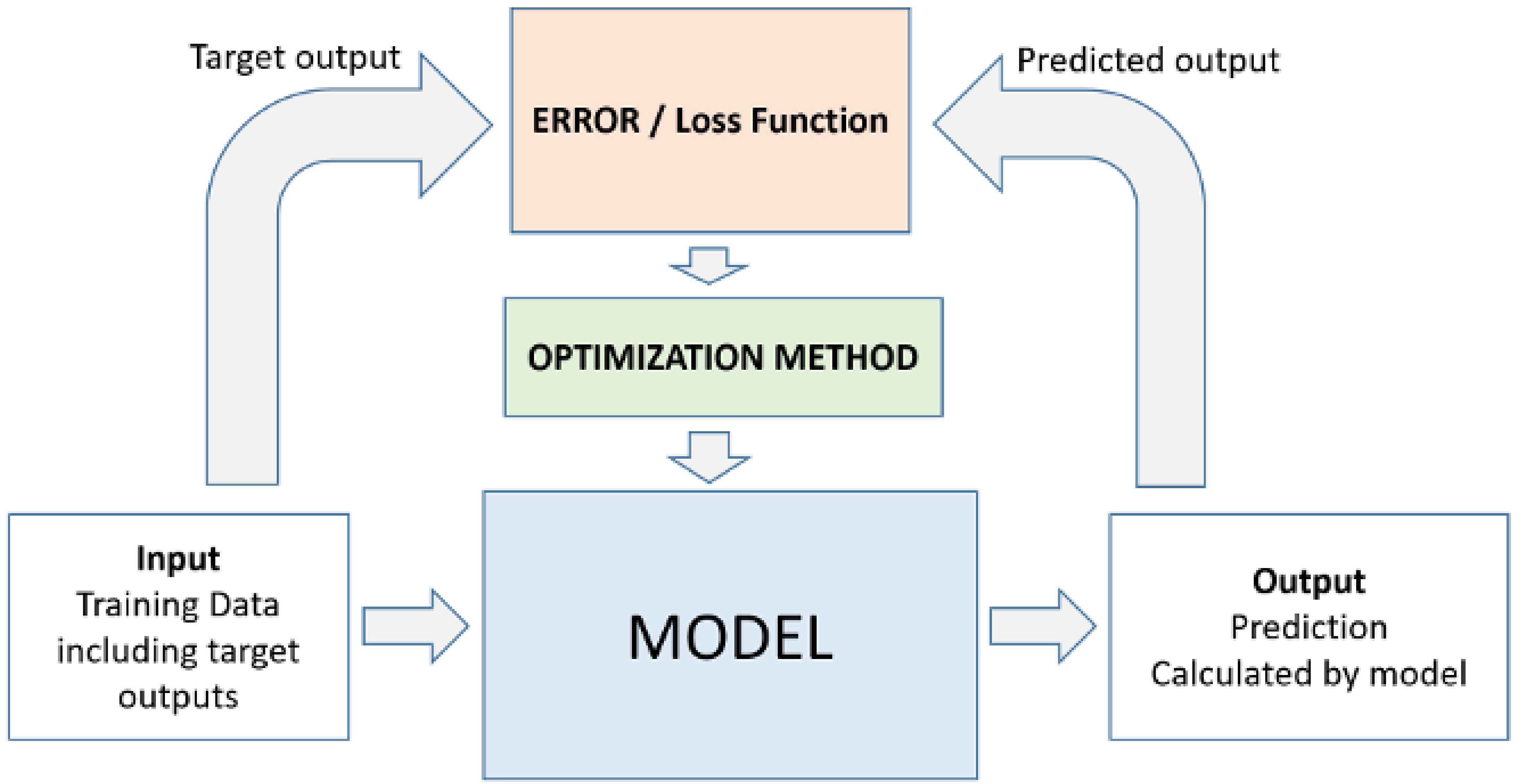
$$\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) = -\left(y^{(i)} \log \hat{y}^{(i)}\right) + \left(1 - y^{(i)} \log(1 - \hat{y}^{(i)})\right)$$

- Multi-class classification (single-label)

$$x^{(i)} \rightarrow \hat{y}^{(i)} = f_\theta(x^i)$$

$$\theta^* = \min_\theta \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

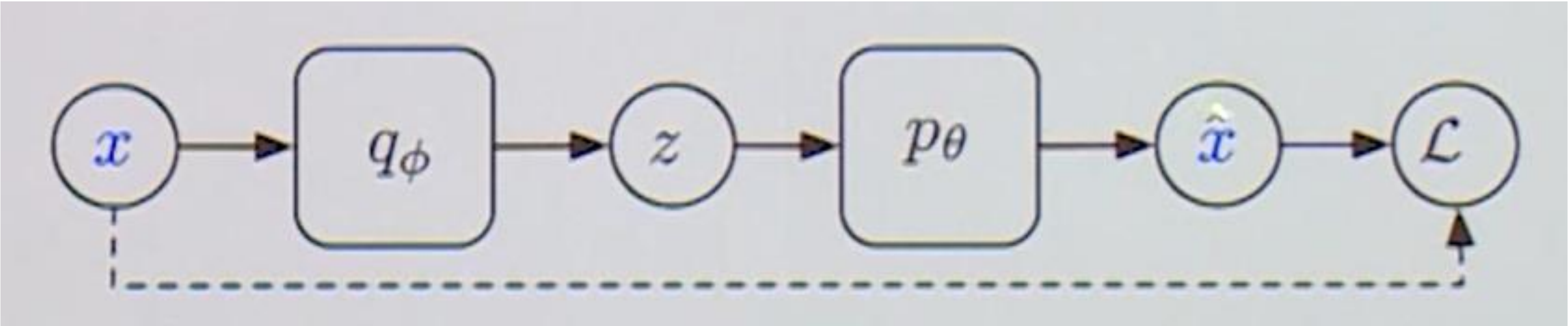
$$\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) = -\sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k{}^{(i)}$$



NEW LEARNING PARADIGMS

Encoder/Decoder (Auto-encoder/VAE)

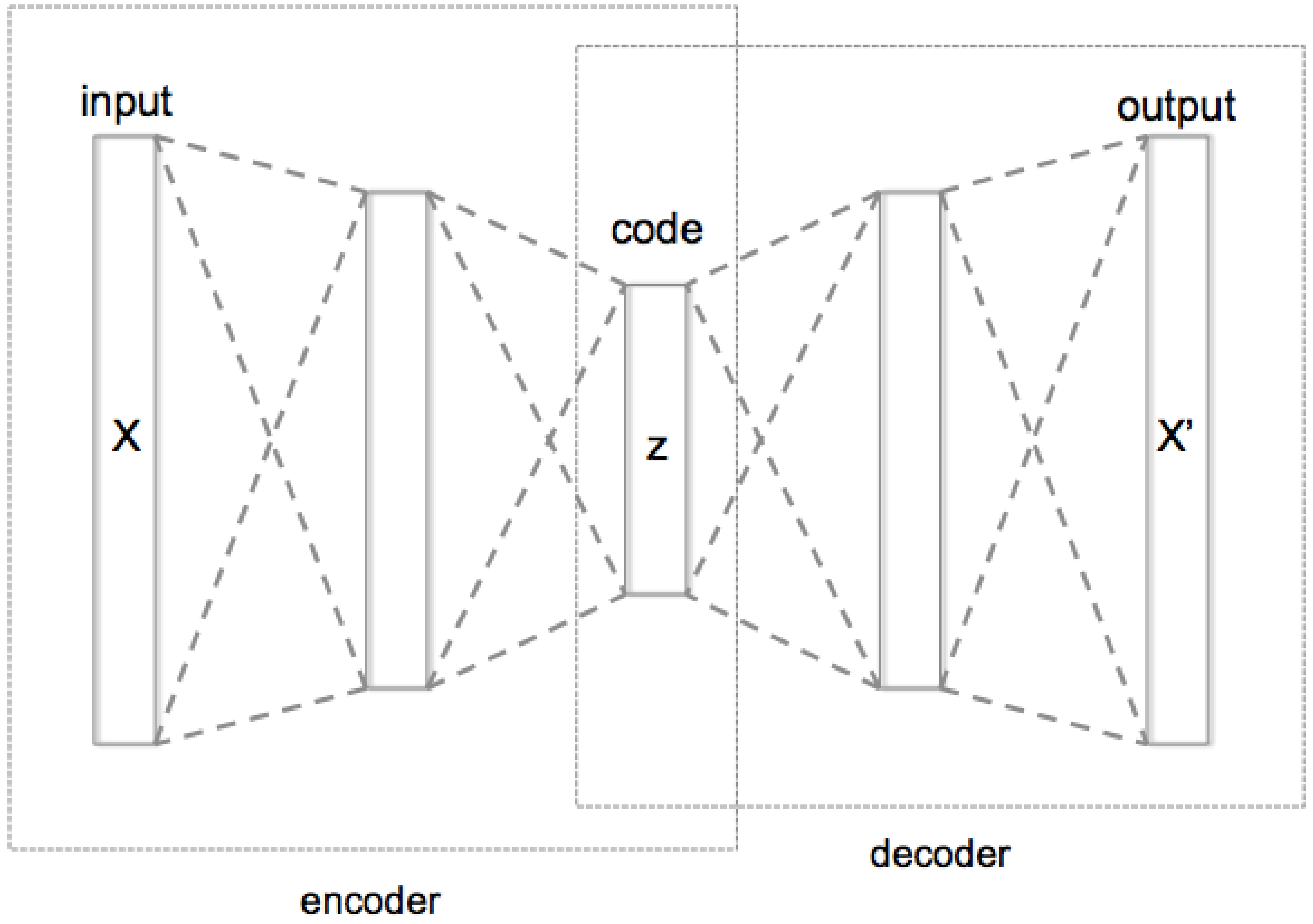
1) Auto-encoder



$$z = q_\phi(x)$$

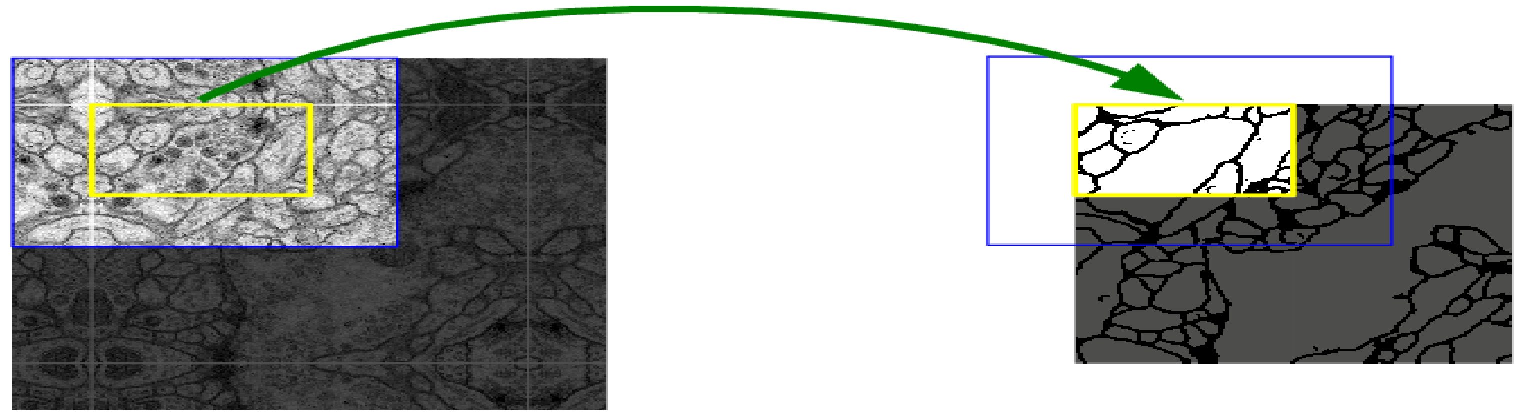
$$\hat{x} = p_\theta(x)$$

$$\mathcal{L} = \|x - \hat{x}\|_2{}^2$$



z refers to code, latent variable or latent representation (projection of x in a variety/manifold)

NEW LEARNING PARADIGMS



Encoder/Decoder (Auto-encoder/VAE)

2) Denoising auto-encoder
U-Net

- Contracting path to capture context and a symmetric expanding path that enables precise localization

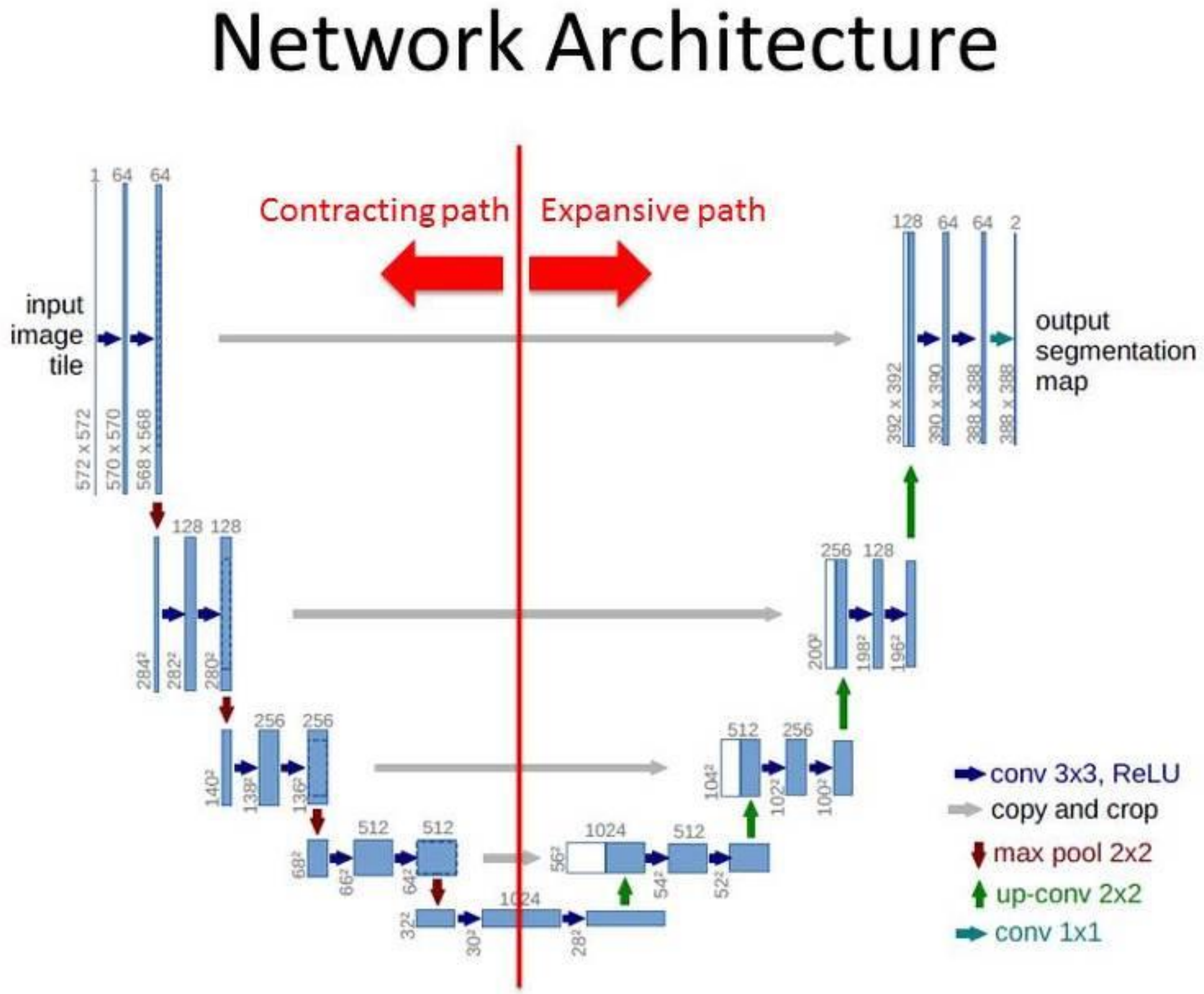
Separation of vocal and instrumental parts

- A time-frequency mask is learnt so that:

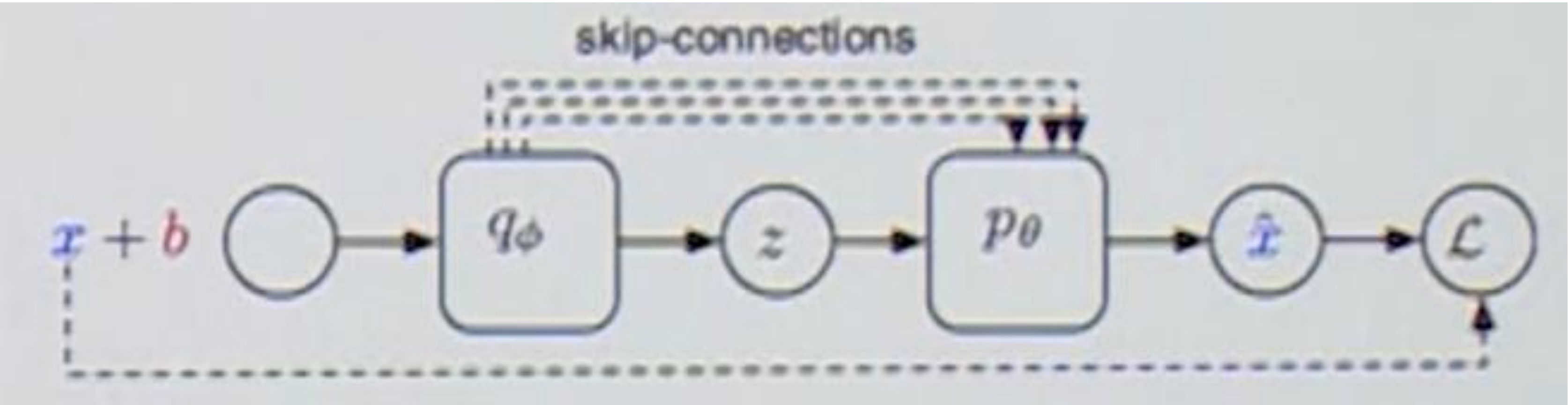
$$\boldsymbol{X} = \boldsymbol{X}_v + \boldsymbol{X}_i$$

$$\widehat{\boldsymbol{X}}_v = \mathbf{X} \otimes \boldsymbol{M}$$

$$\mathcal{L} = \left\| \boldsymbol{X}_v - \widehat{\boldsymbol{X}}_v \right\|_1$$



demo



A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, T. Weyde. Singing voice separation with deep U-Net convolutional networks. 2017.

O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

NEW LEARNING PARADIGMS

Encoder/Decoder (Auto-encoder/VAE)

3) Complex input / complex network

- Deep Complex U-Net for source separation (complex mask)

  Complex convolution

$$W = A + iB$$

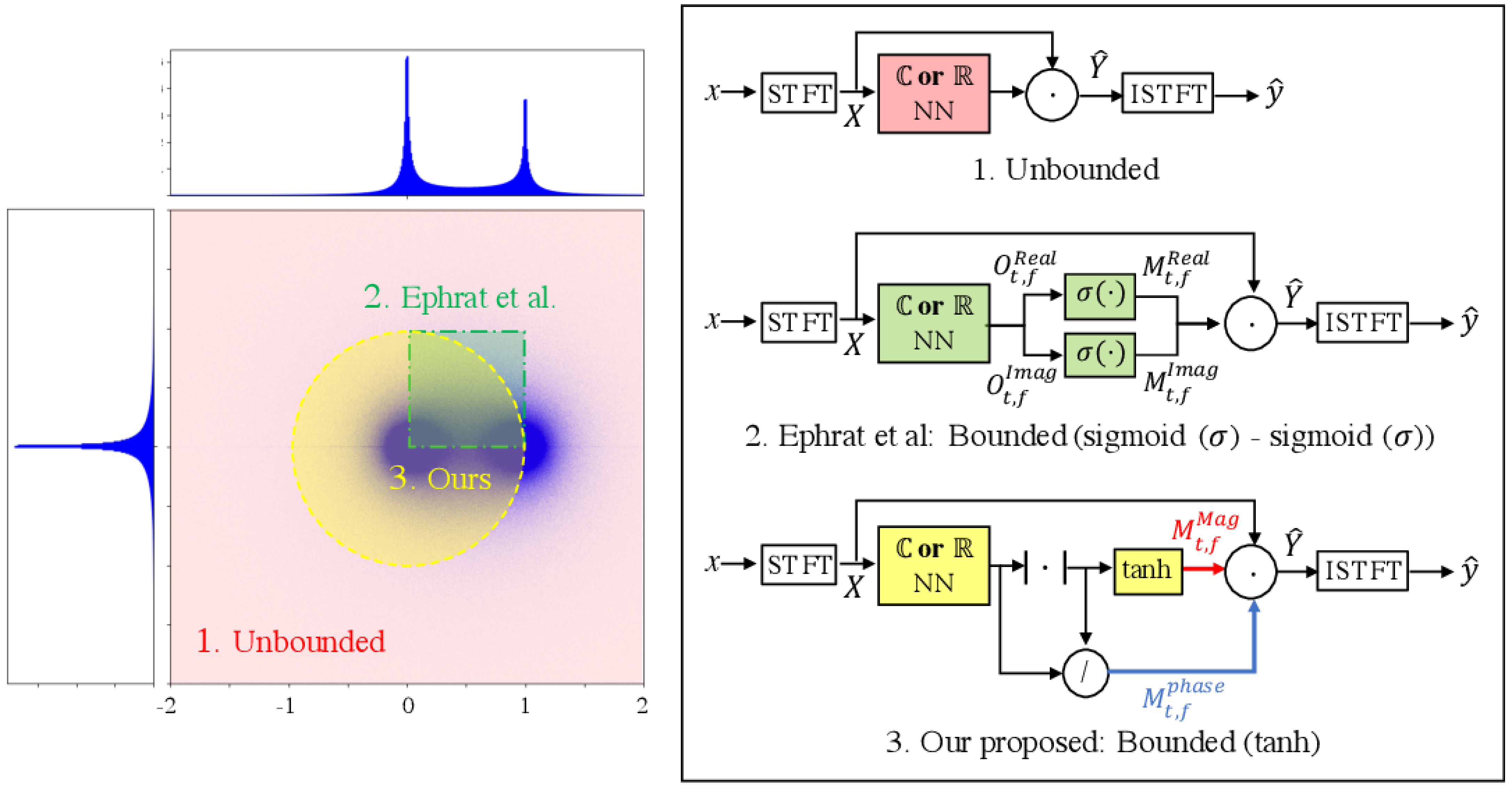
$$h = x + iy$$

$$W * h = (A * x - B * y) + i \cdot (B * x + a * y)$$

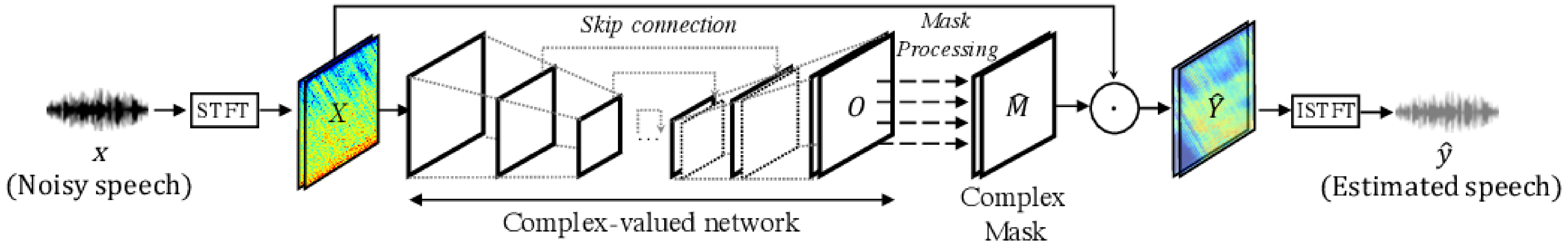
Complex masking

$$\hat{Y}_{t,f} = \widehat{M}_{t,f} \cdot X_{t,f}$$

$$= \left|\widehat{M}_{t,f}\right| \cdot \left|X_{t,f}\right| \cdot e^{i\phi_{\widehat{M}_{t,f}} + \phi_{X_{t,f}}}$$



(a) (b)



H. S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, K. Lee. Phase-aware Speech Enhancement with Deep Complex U-Net. 2019.

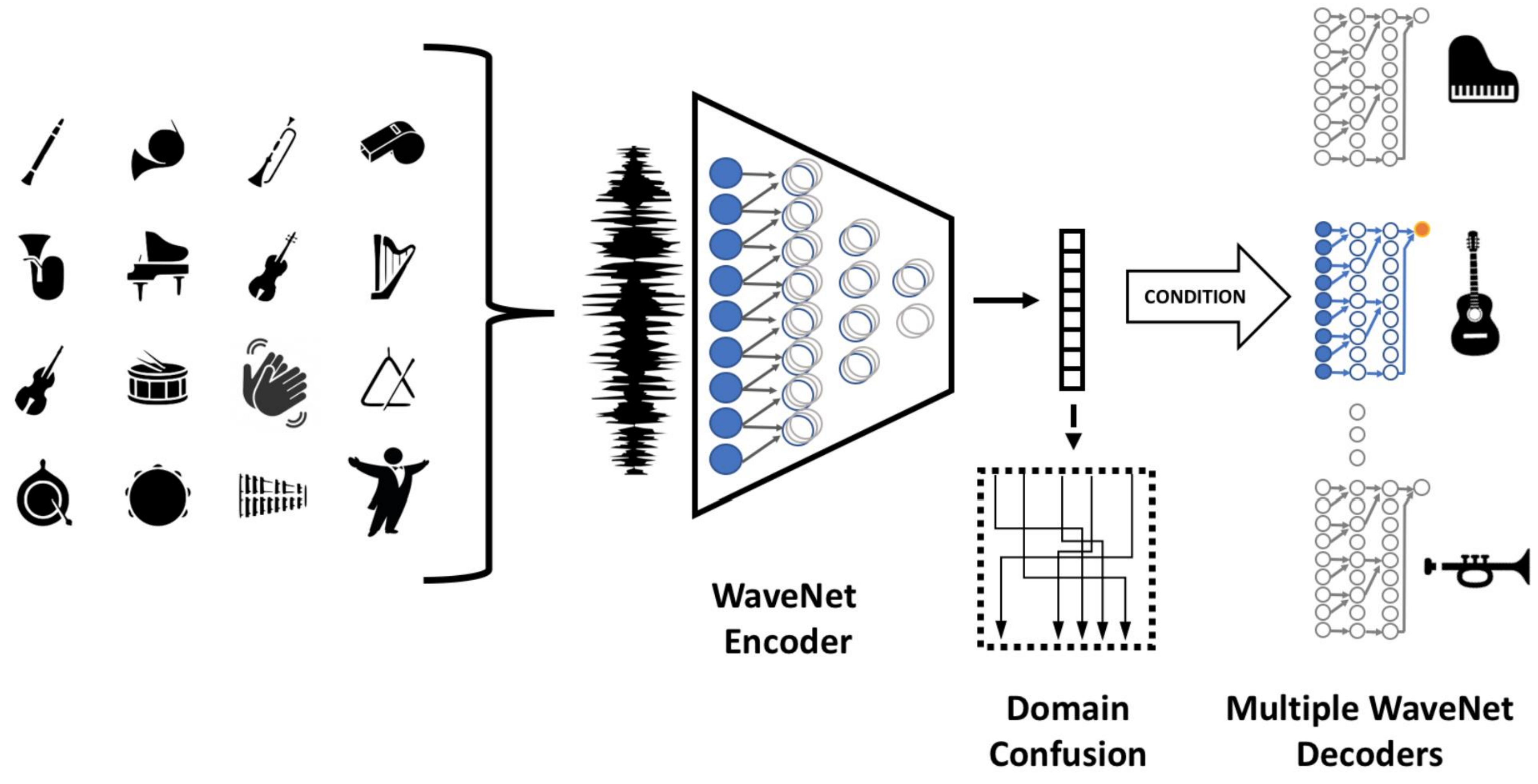
NEW LEARNING PARADIGMS

Encoder/Decoder (Auto-encoder/VAE)

4) Translating music across musical instruments, genres and styles

- $O(s,r)$:
  random augmentation of input s with seed r
- Encoder E:
  shared wavenet
- Disentangled latent space
  Domain classification C
- Decoder D^j (domain j):
  multiple wavenet, conditioned on the latent representation produced by E

- Adversial loss
  minimize reconstruction loss
  maximize domain classification
  → prevent the latent space to learn domain characteristic



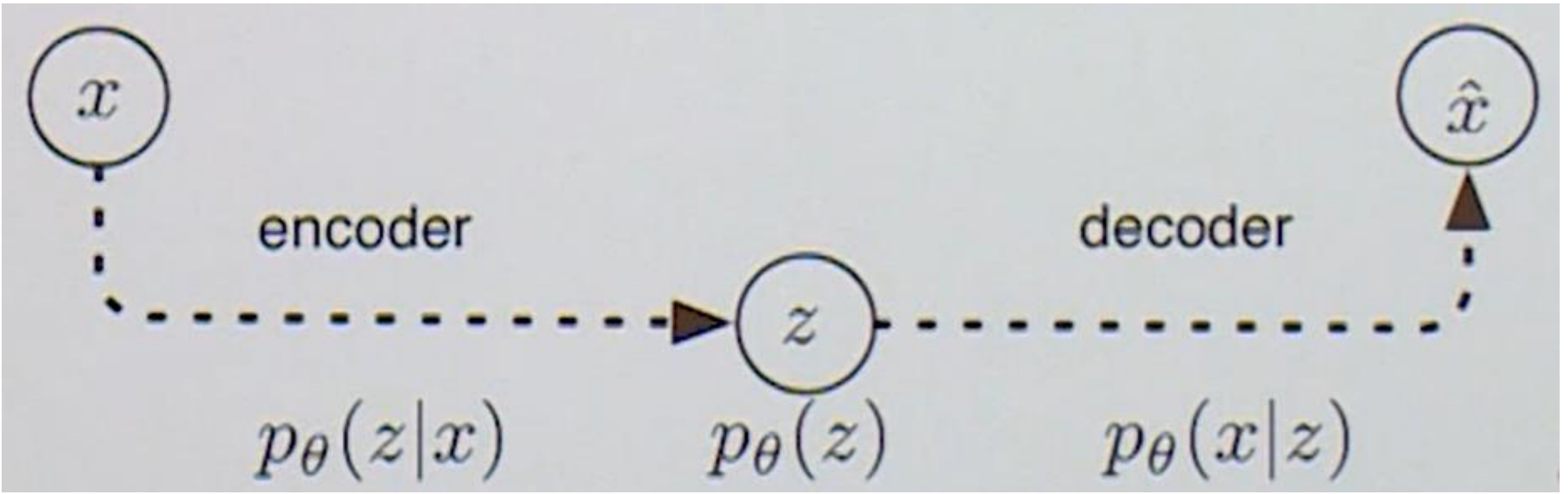
N. Mor, L. Wolf, A. Polyak, Y. Taigman. A Universal Music Translation Network. 2018.

NEW LEARNING PARADIGMS

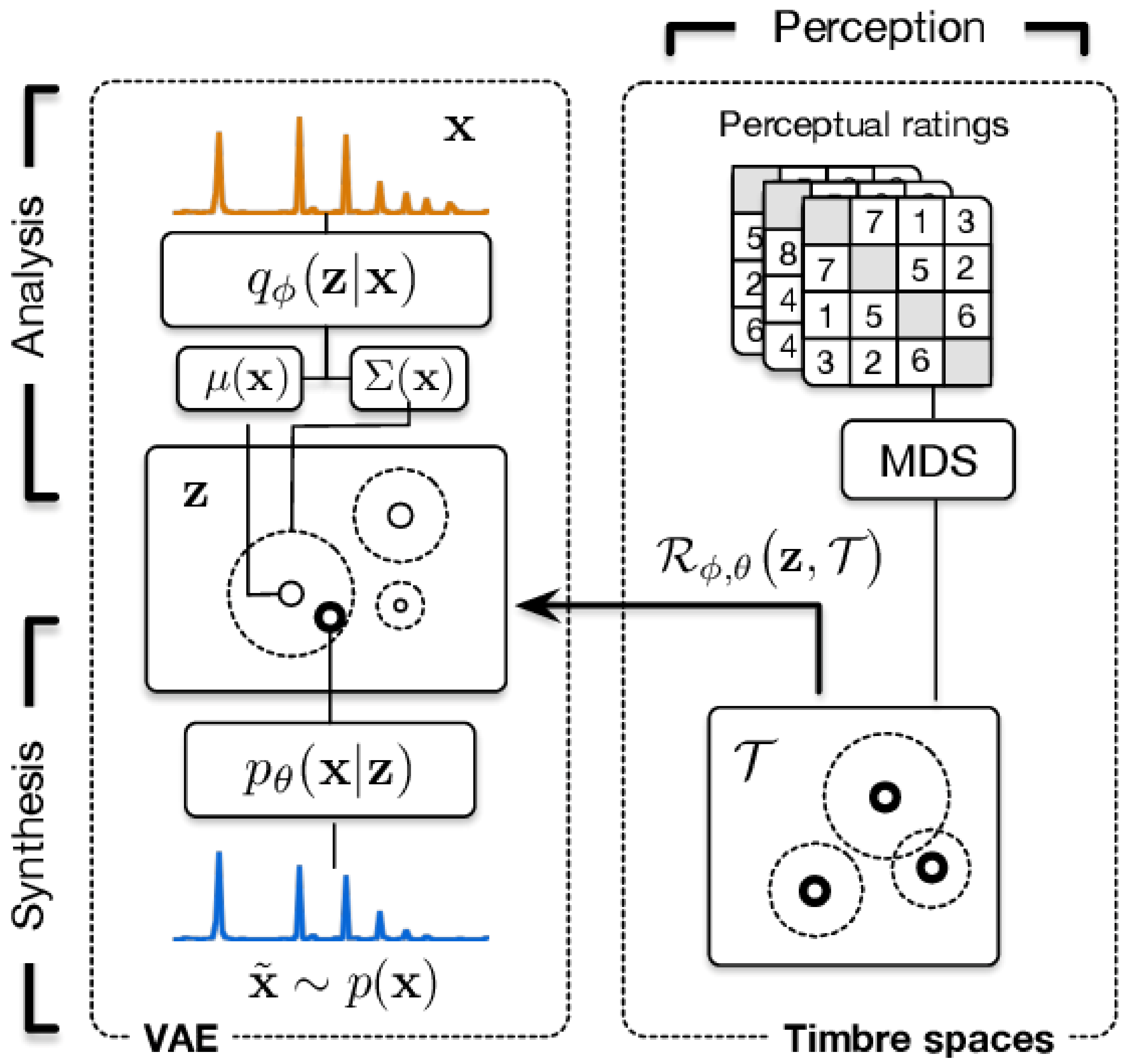
Encoder/Decoder (Auto-encoder/VAE)

5) Variational Auto-Encoder

- Latent variables are drawn from a prior
  $z_i \sim p(z)$
- data x have a likelihood that is conditioned on latent variables z:
  $x_i \sim p(x|z)$
- likelihood and prior: $p(x,z) = p(x|z)p(z) = p(xz)p(z)$



posterior prior



NEW LEARNING PARADIGMS

Metric learning

1) Triplet Loss

We train the network for a triplet of data anchor, positive, negative

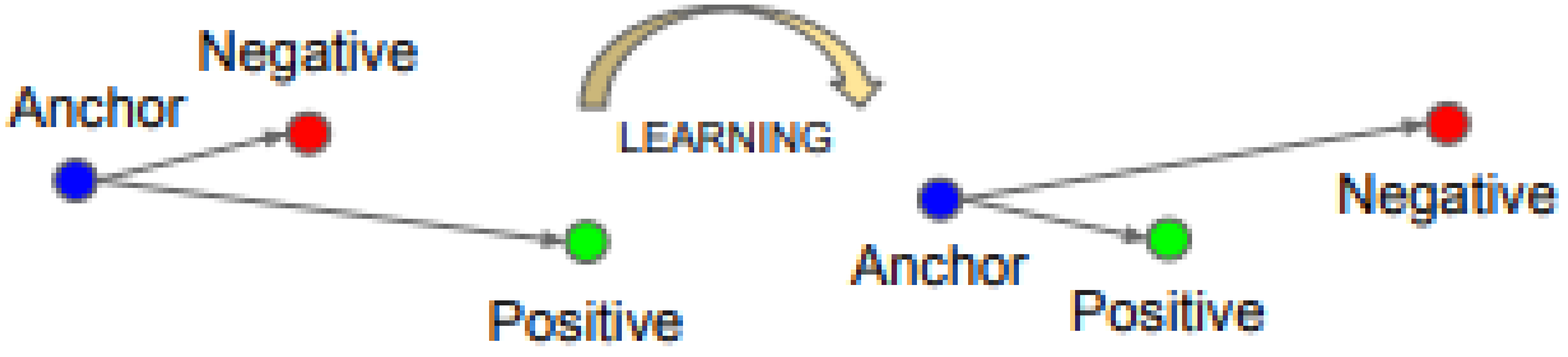


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

F. Schroff. FaceNet: A Unified Embedding for Face Recognition and Clustering

G. Doras, G. Peeters. Cover Detection using Dominant Melody Embeddings

DEEP LEARNING FOR AUDIO AND SPEECH PROCESSING.

Thank you for your attention.

Refenencences:

- Geoffroy Peeters, Telecom Paris Tech