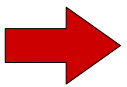
The Wavenet architecture
from text-to-speech to source separation

Jordi Pons

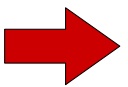
jordipons.me – @jordiponsdotme

Music Technology Group
Universitat Pompeu Fabra, Barcelona

WaveNet
(van den Oord et al., 2016)



WaveNet for Speech Denoising
in 2017

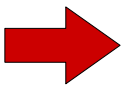


WaveNet for Source Separation
in 2018

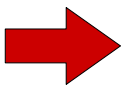
(show demonstration)

Is music source separation possible in the waveform domain?

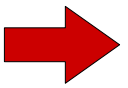
WaveNet
(van den Oord et al., 2016)



WaveNet for Speech Denoising
in 2017



WaveNet for Source Separation
in 2018



?

DeepConvSep for Source Separation
(Chandna et al., 2017)

&

Wave-U-Net for Source Separation
(Stoller et al., 2018)

Is music source separation possible in the waveform domain?

WaveNet
(van den Oord et al., 2016)

→ **WaveNet for Speech Denoising** in 2017

→ **WaveNet for Source Separation** in 2018

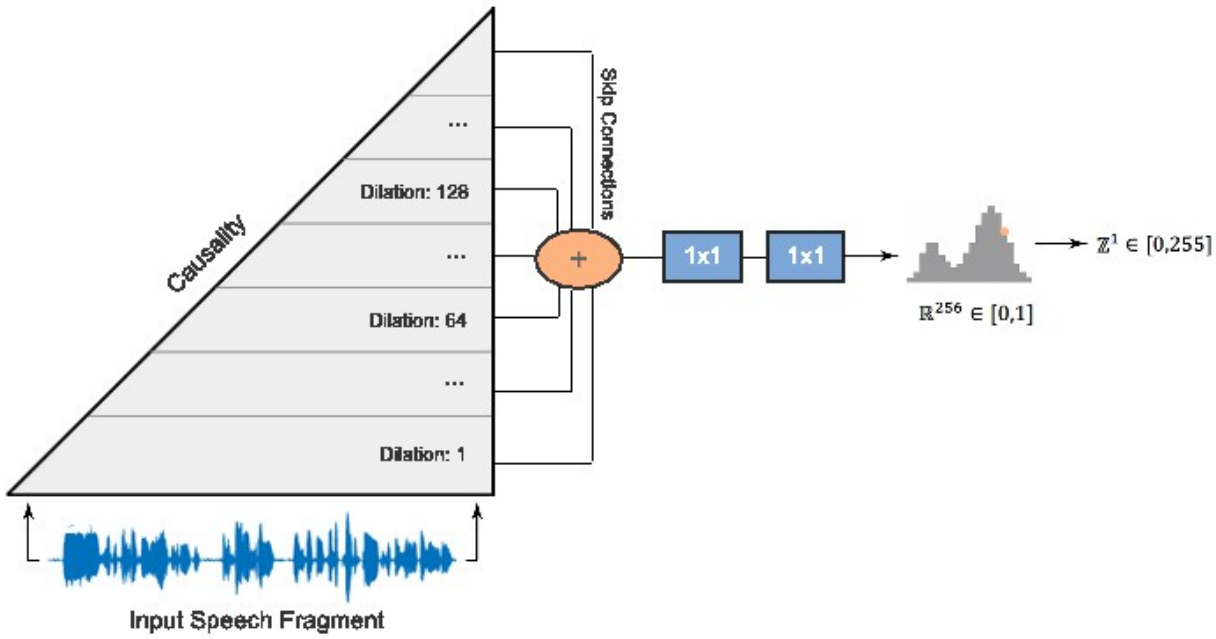
→ ?

DeepConvSep for Source Separation
(Chandna et al., 2017)

&

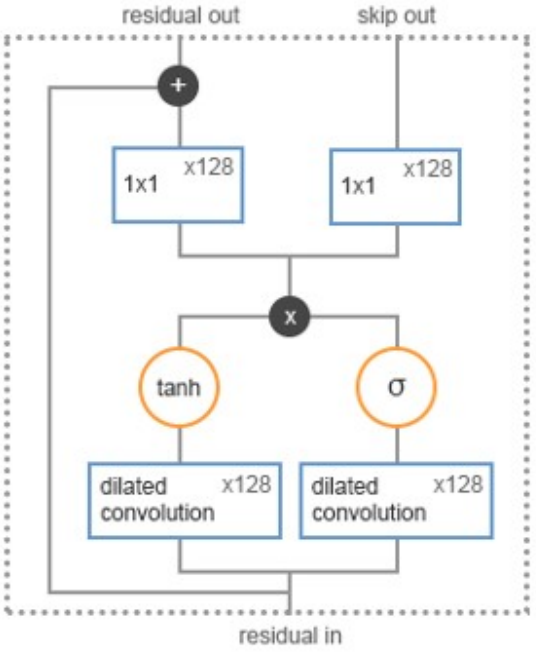
Wave-U-Net for Source Separation
(Stoller et al., 2018)

Introduction: WaveNet

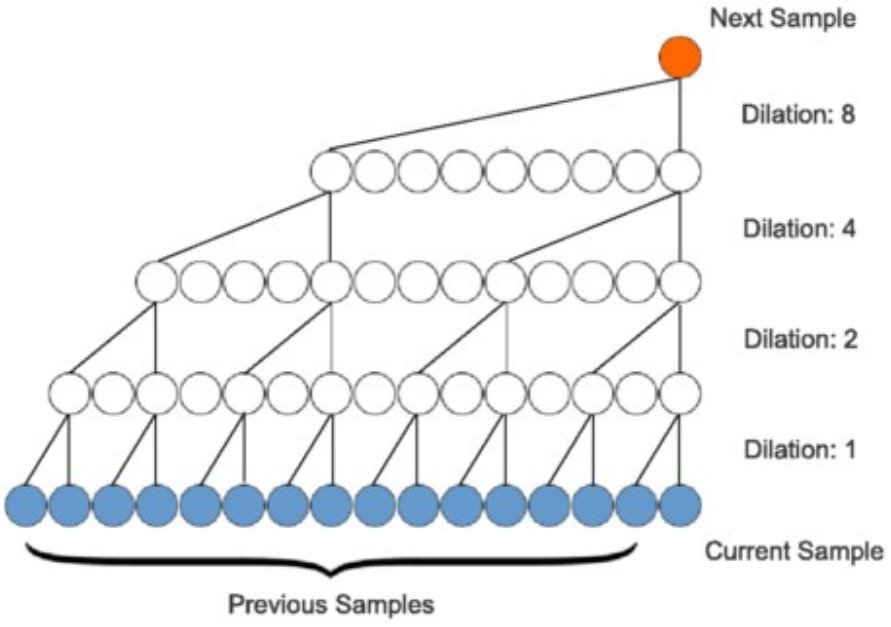


μ-law quantization: discrete softmax output distribution

Introduction: WaveNet

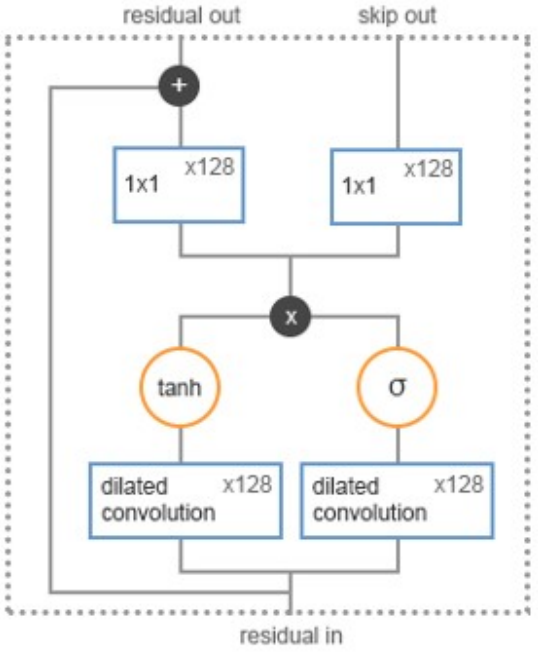


Residual layer



Causal, dilated convolutions

Introduction: Conditional WaveNet

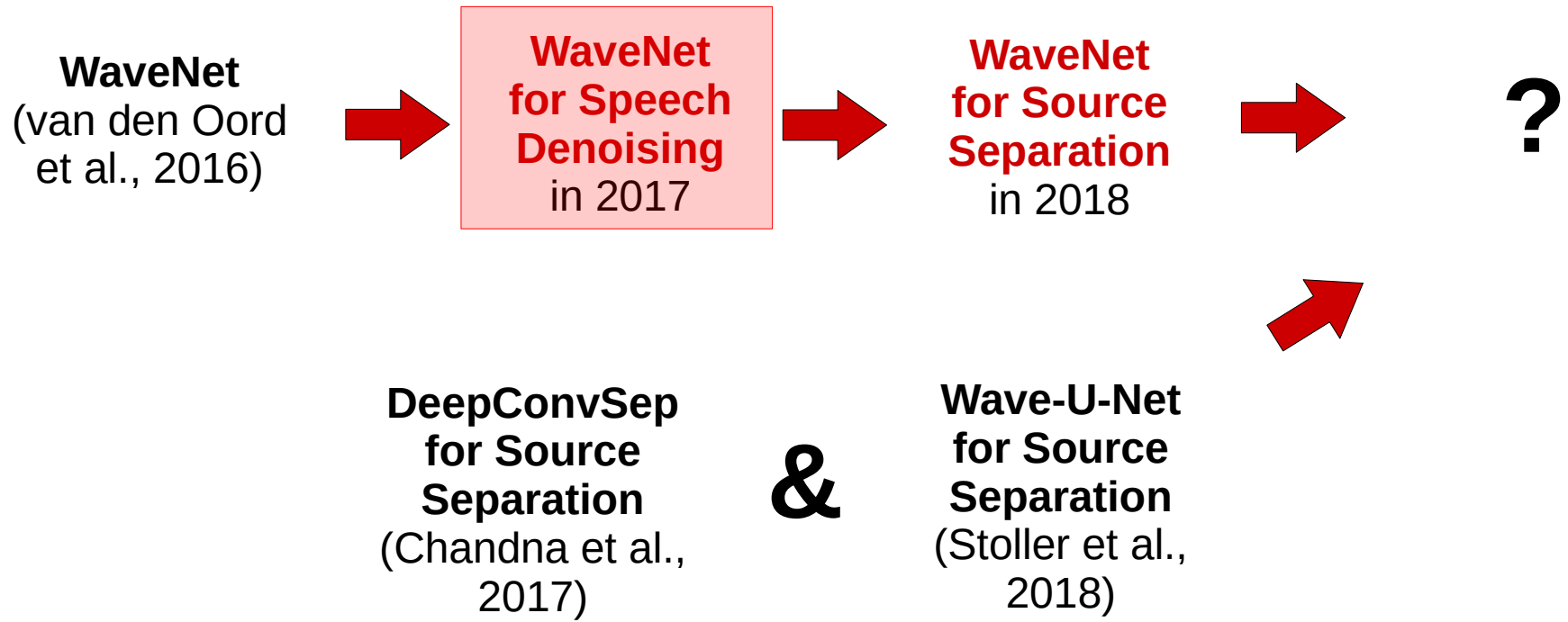


Residual layer

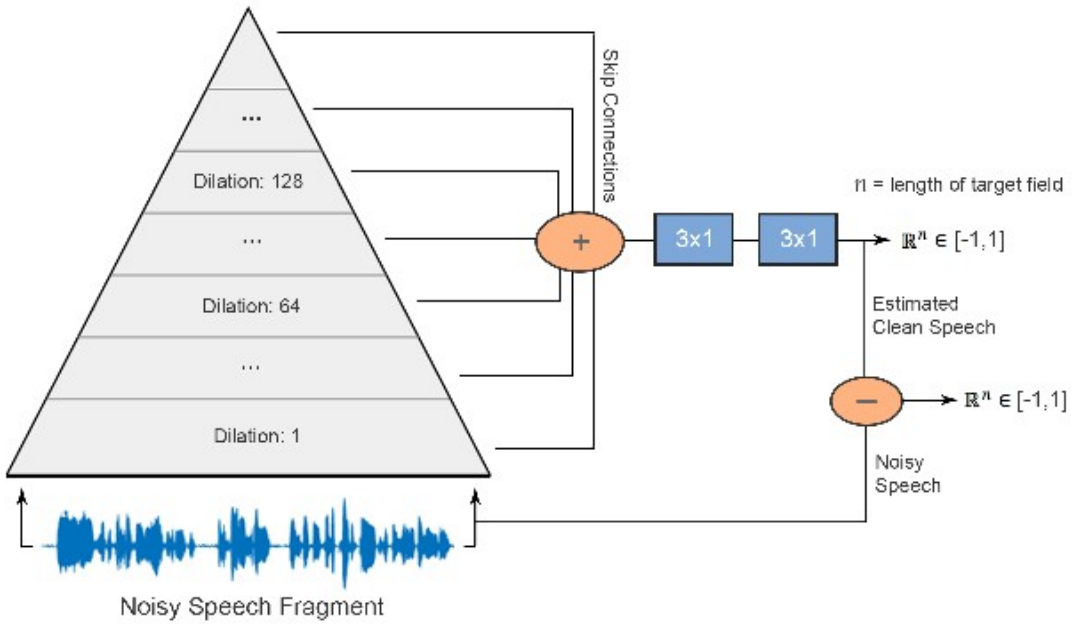
$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^{T}\mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^{T}\mathbf{h}\right)$$

Conditioning with a bias term
to perform text-to-speech

Is music source separation possible in the waveform domain?

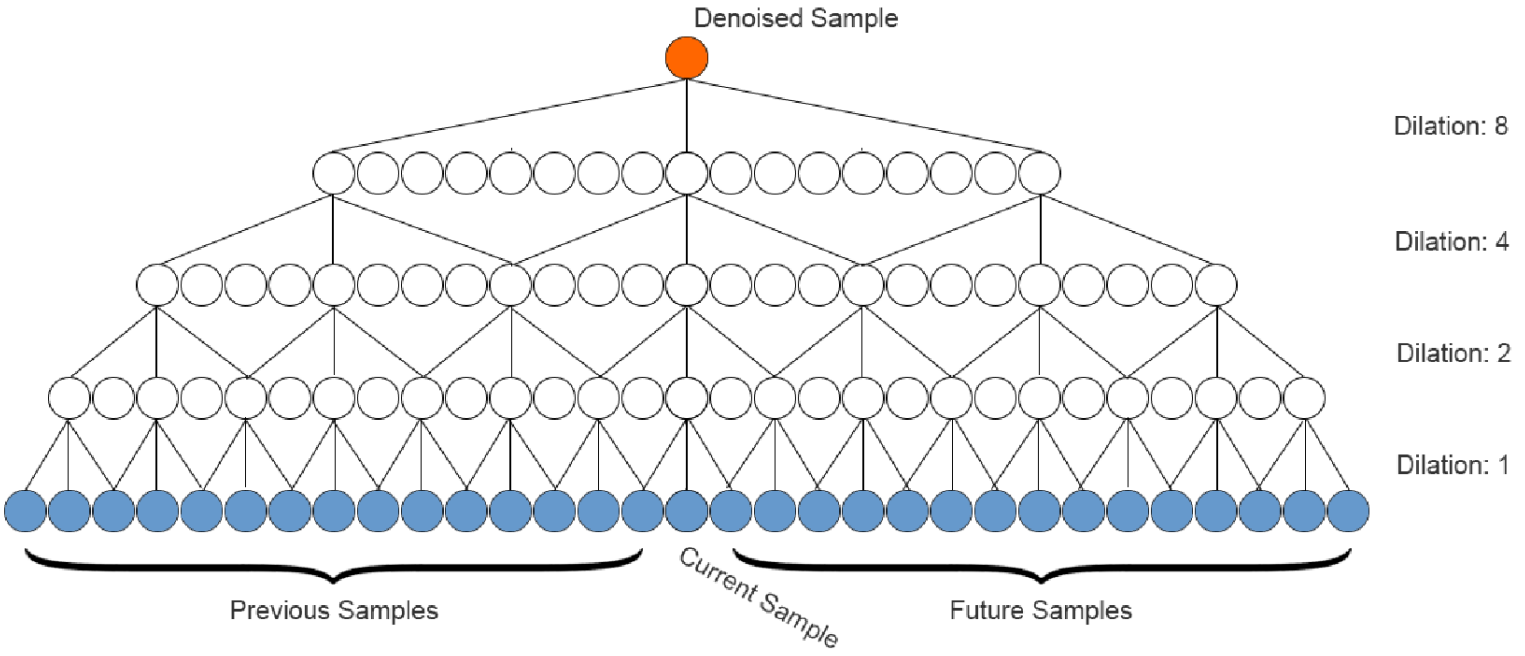


A WaveNet for Speech Denoising



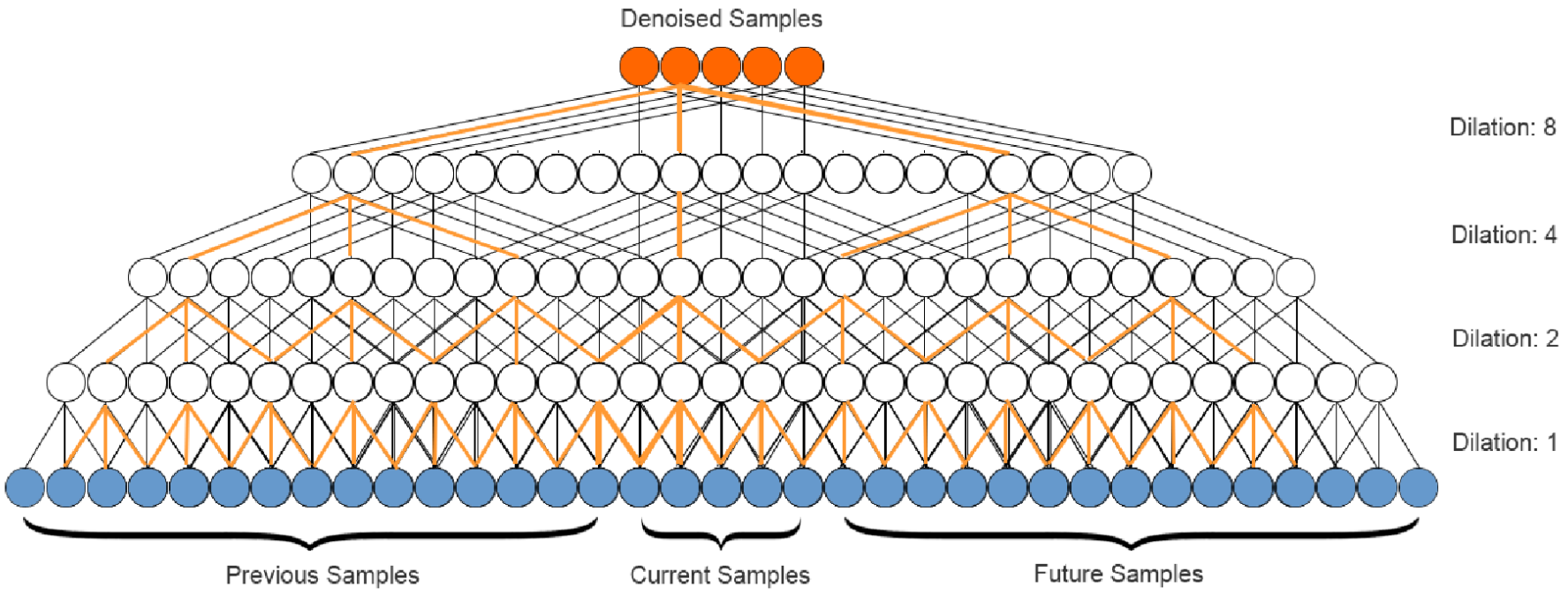
Non-causal 3x1 filters Supervised learning

Non-causal WaveNet: target field prediction



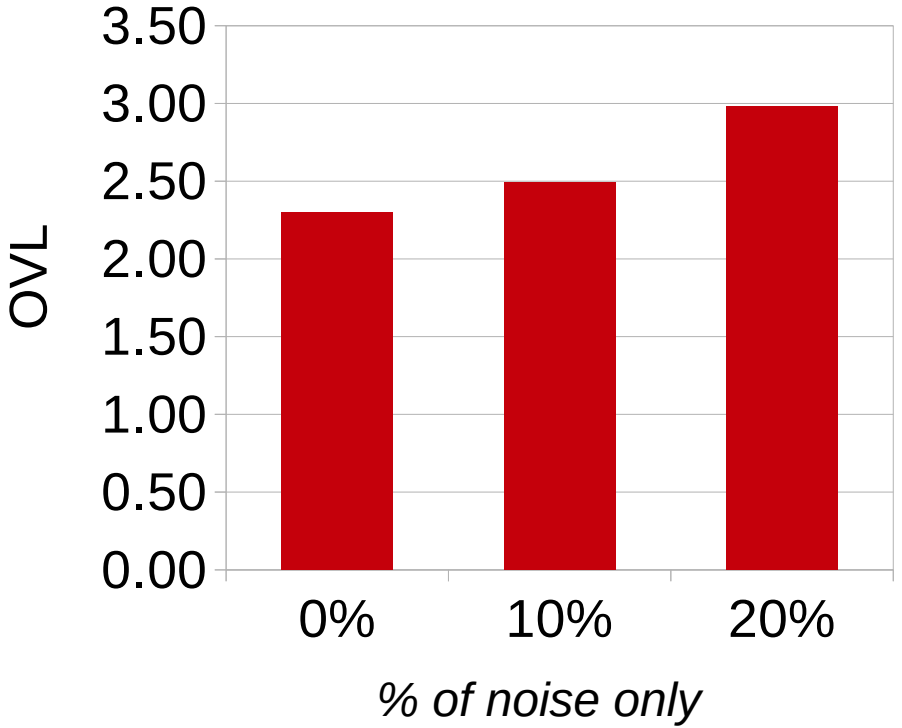
WaveNet is fully convolutional!

Non-causal WaveNet: target field prediction



Parallel inference on 1601 samples at once, results in a denoising time of ≈ 0.56 seconds per second of noisy audio on GPU.

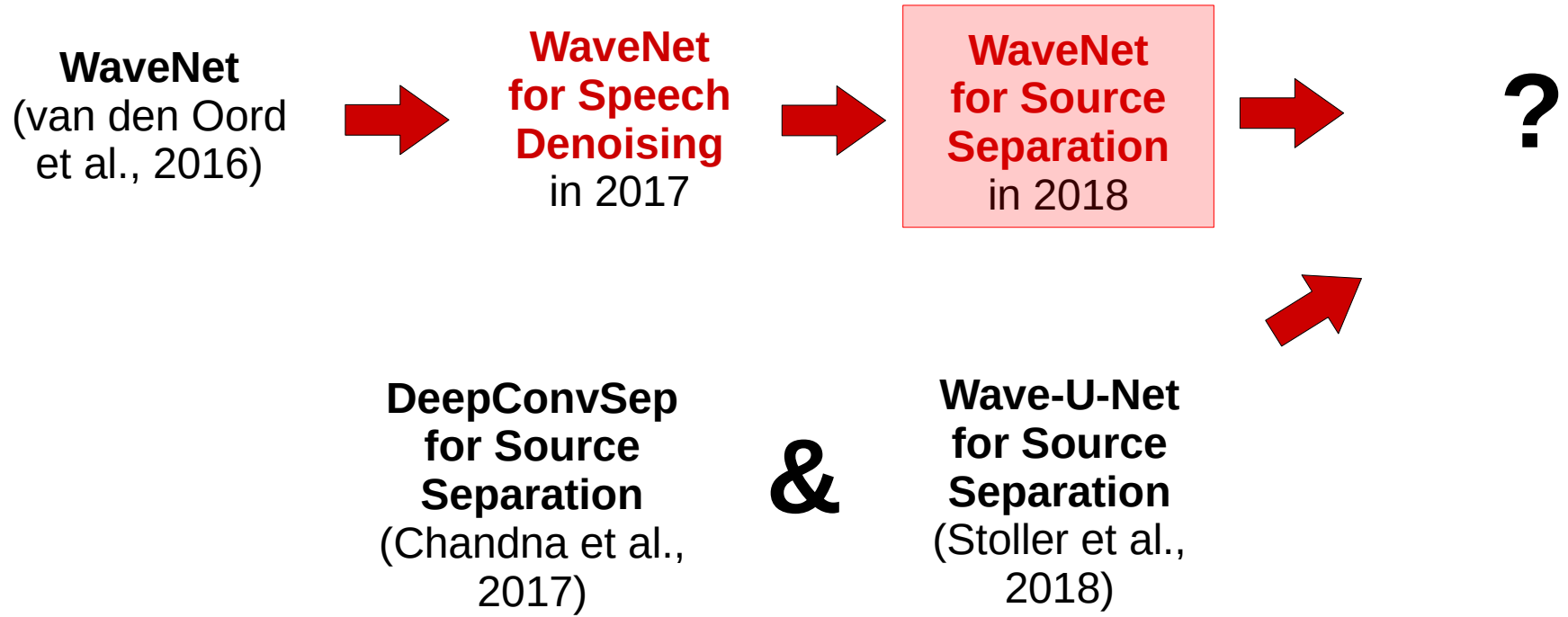
Key finding: control the model with data!



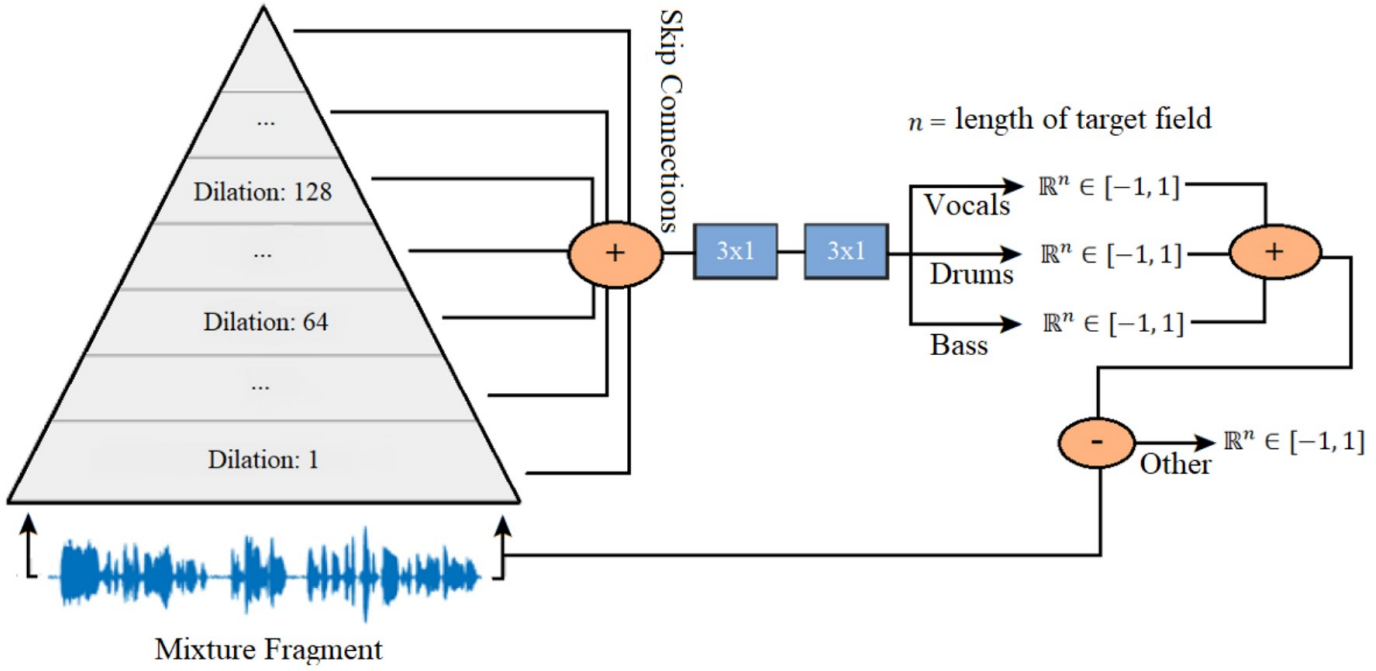
The model had difficulties
in producing silence

Data driven solution:
Noise only data augmentation

Is music source separation possible in the waveform domain?

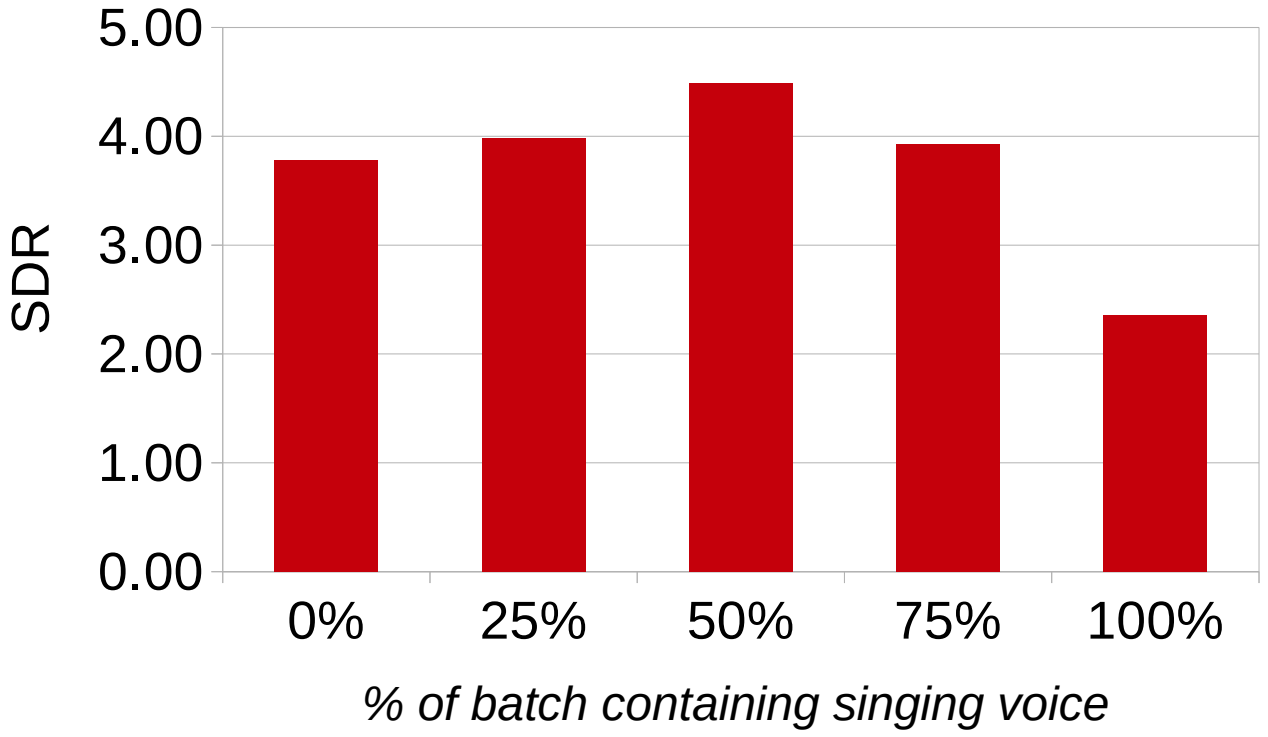


WaveNet for Music Source Separation



Capable to separate only vocals or multiple instruments

Let’s control the model with data!



The model had difficulties in producing continuous vocals

Data driven solution:
Control amount of singing voice in a batch

Is music source separation possible in the waveform domain?

WaveNet
(van den Oord et al., 2016)

WaveNet for Speech Denoising
in 2017

WaveNet for Source Separation
in 2018

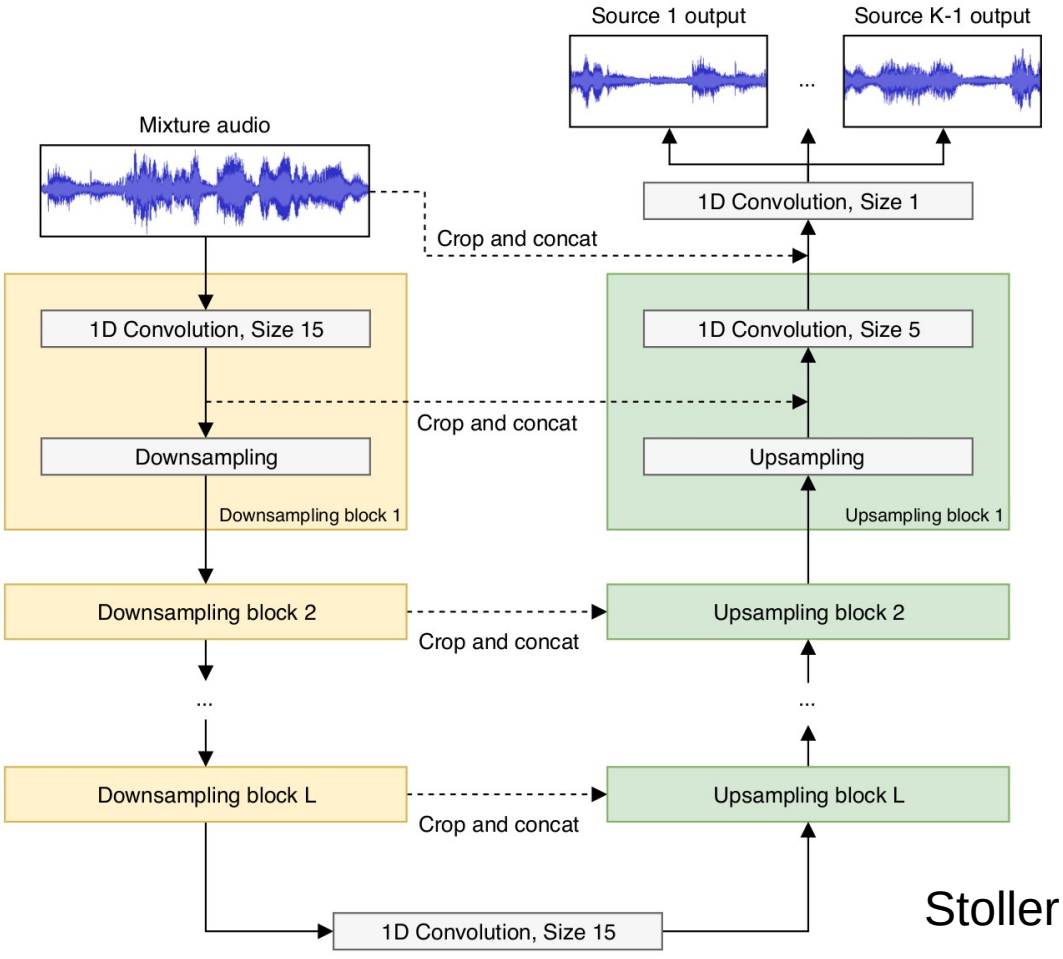
?

DeepConvSep for Source Separation
(Chandna et al., 2017)

&

Wave-U-Net for Source Separation
(Stoller et al., 2018)

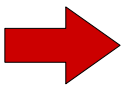
Wave-U-net for Music Source Separation



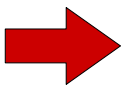
Stoller et al., 2018

Is music source separation possible in the waveform domain?

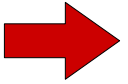
WaveNet
(van den Oord et al., 2016)



WaveNet for Speech Denoising
in 2017



WaveNet for Source Separation
in 2018



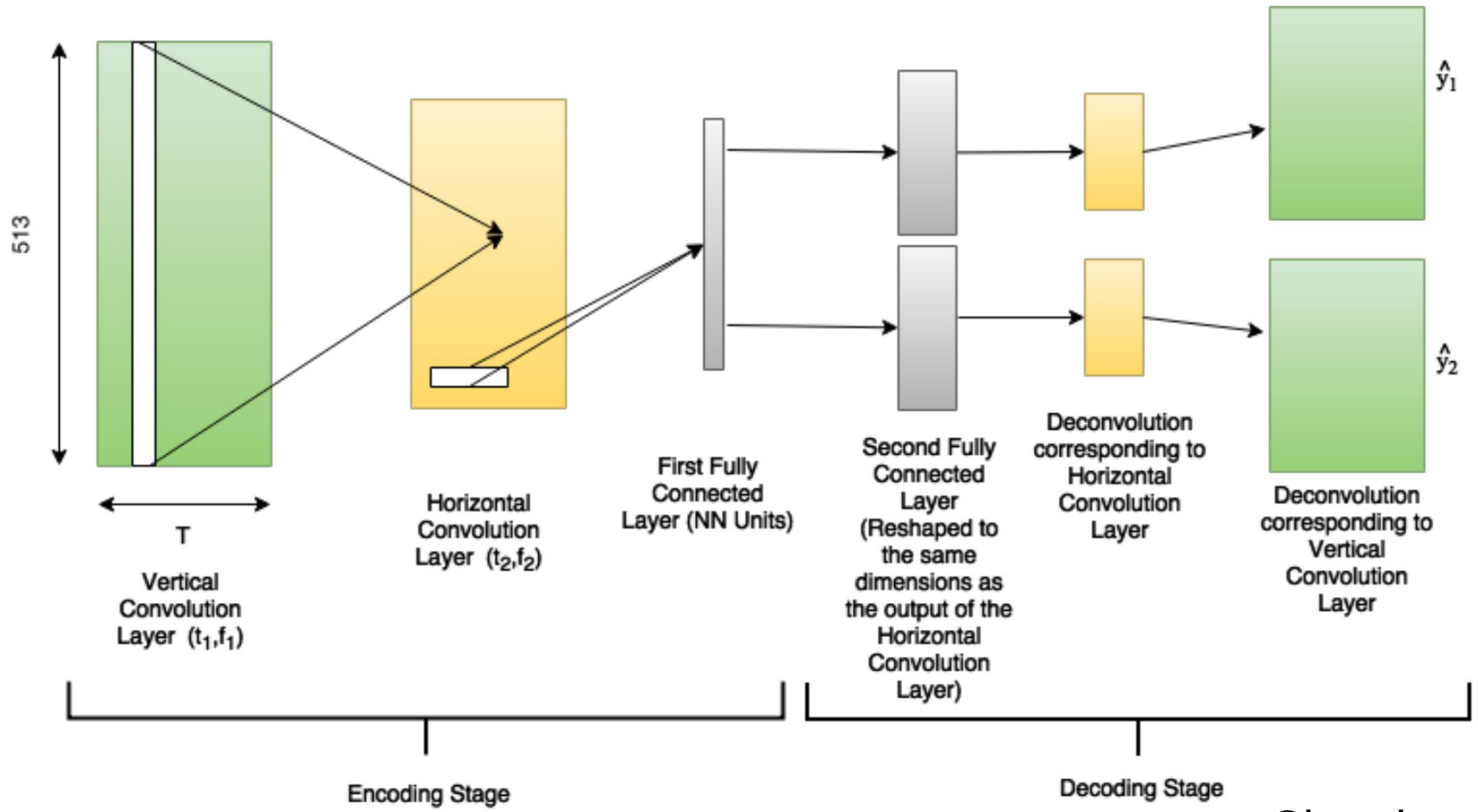
?

DeepConvSep for Source Separation
(Chandna et al., 2017)

&

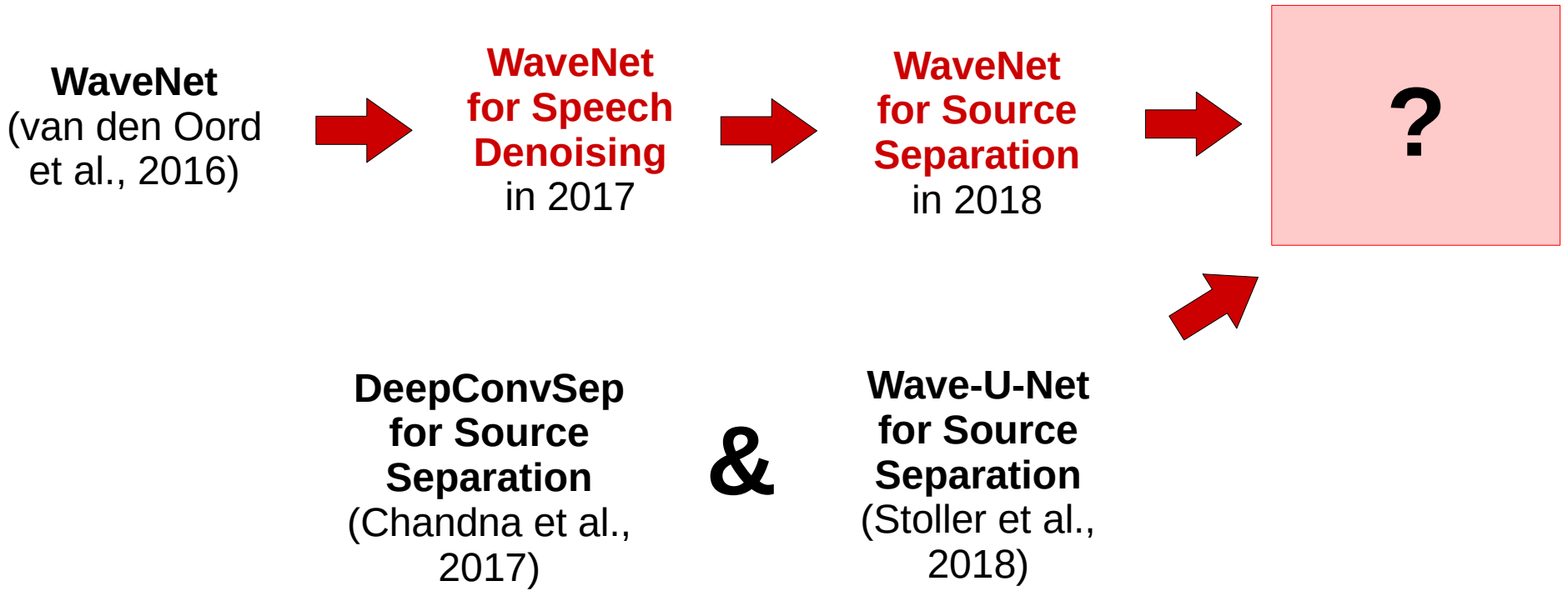
Wave-U-Net for Source Separation
(Stoller et al., 2018)

DeepConvSep: a spectrogram-based model



Chandna et al., 2017

Is music source separation possible in the waveform domain?



The Wavenet architecture
from text-to-speech to source separation

Jordi Pons

jordipons.me – @jordiponsdotme

May 2019