大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Tutorial on end-to-end text-to-speech synthesis

Part 1 – Neural waveform modeling

Xin WANG

National Institute of Informatics, Japan

2019-01-27

contact: wangxin@nii.ac.jp
we welcome critical comments, suggestions, and discussion

SELF-INTRODUCTION

- 国立情報学研究所
- 山岸研究室
- 特任研究員

ワン　シン
王　鑫

- PhD (2015-2018)　総研大・国立情報学研究所
- M.A (2012-2015)　中国科学技術大学

- http://tonywangx.github.io

CONTENTS

- Introduction: text-to-speech synthesis
- Neural waveform models
- Summary & software

INTRODUCTION

Text-to-speech synthesis (TTS) [1,2]

Text → Text-to-speech (TTS) → Speech waveform

NII's TTS systems

Text →  → Waveform

Text → NII TTS system → Waveform

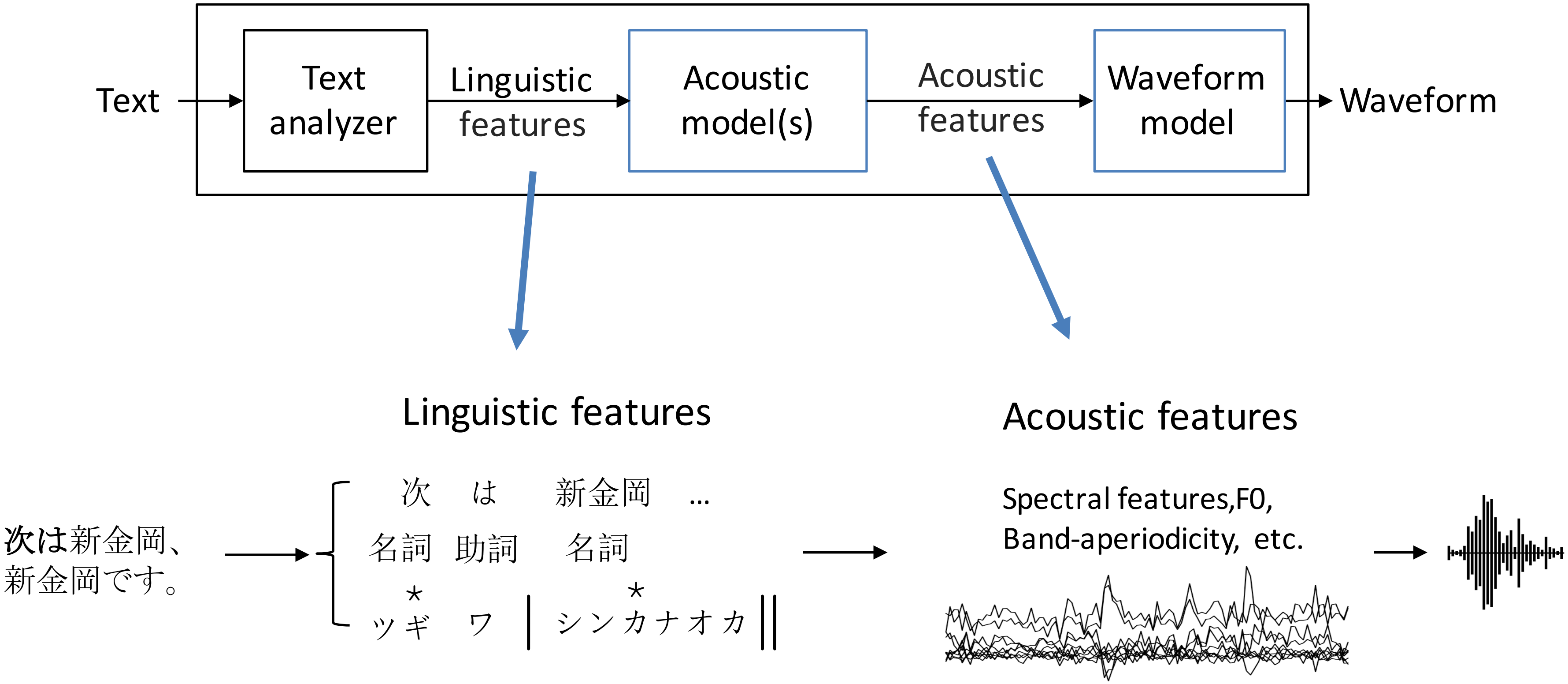
2013 2016 2018 Year

[1] P. Taylor. Text-to-Speech Synthesis. Cambridge University Press, 2009.
[2] T. Dutoit. An Introduction to Text-to-speech Synthesis. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

INTRODUCTION

Text-to-speech synthesis (TTS)

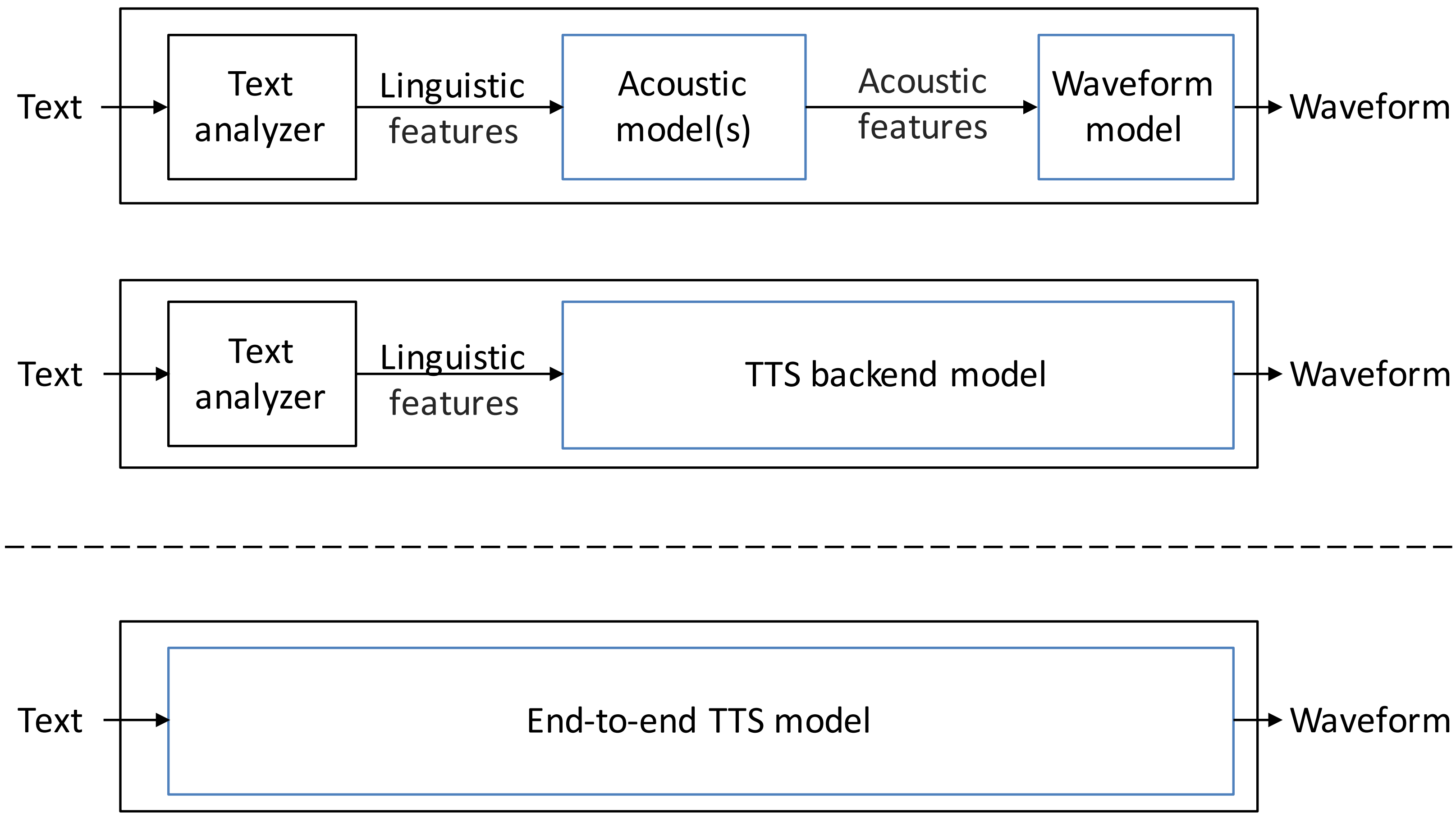
- Architectures



INTRODUCTION

Text-to-speech synthesis (TTS)

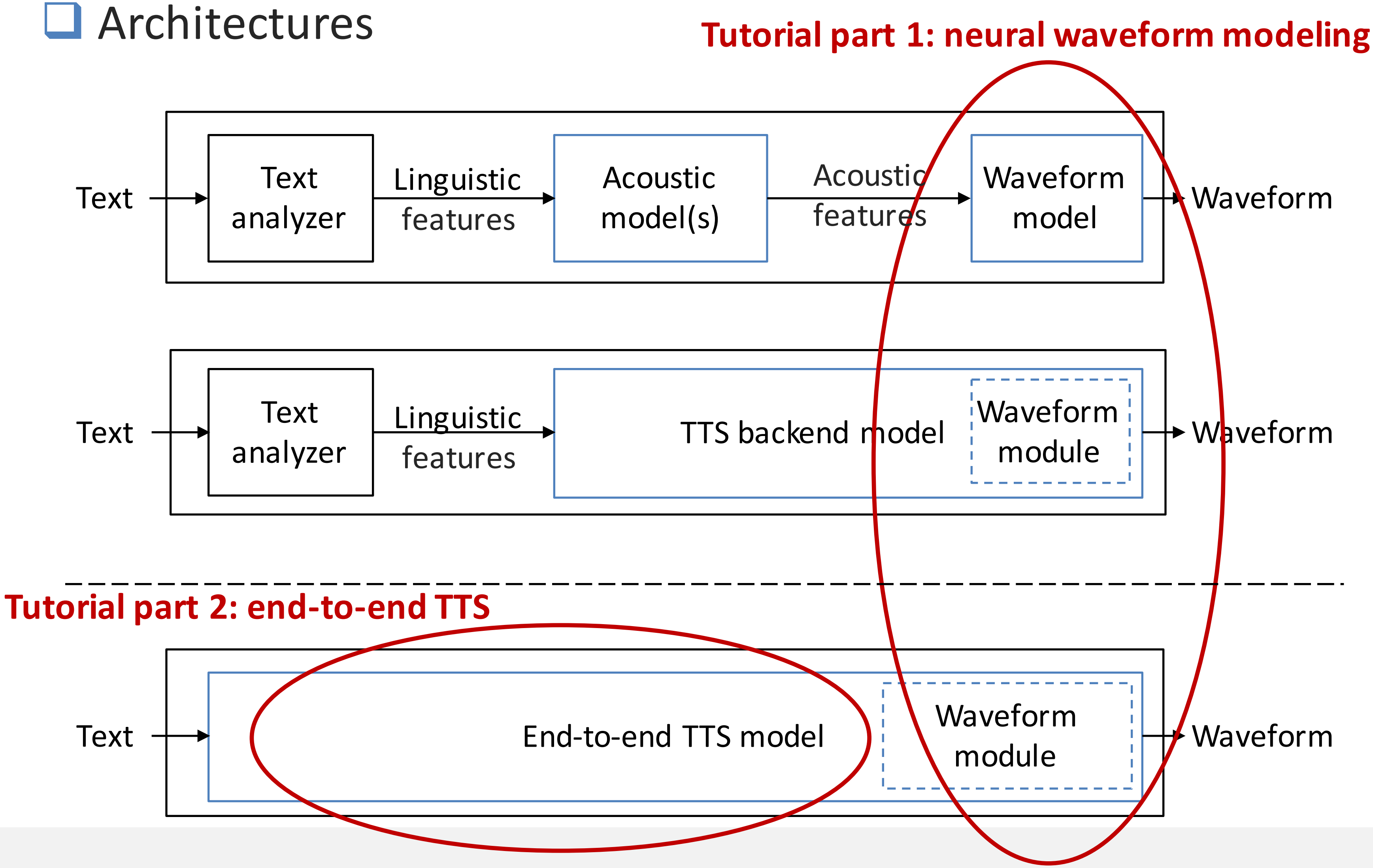
- Architectures



INTRODUCTION

Text-to-speech synthesis (TTS)

- Architectures

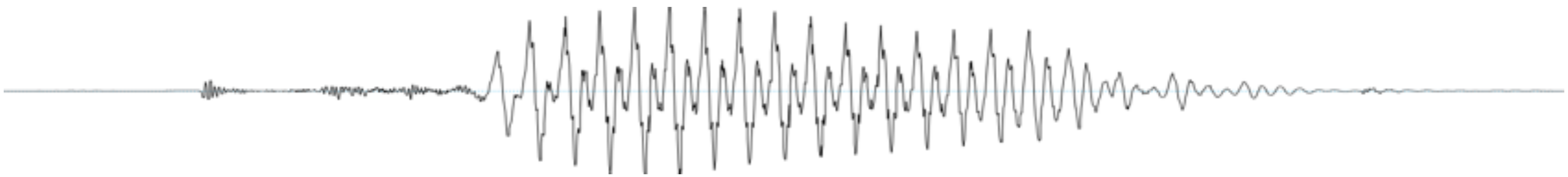


CONTENTS

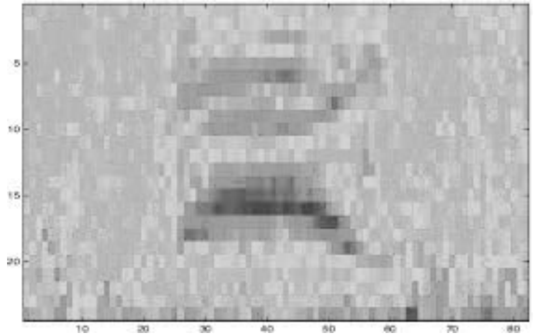
- Introduction: text-to-speech synthesis
- Neural waveform modeling
  - Overview
  - Autoregressive models
  - Normalizing flow
  - STFT-based training criterion
- Summary

OVERVIEW

Task definition

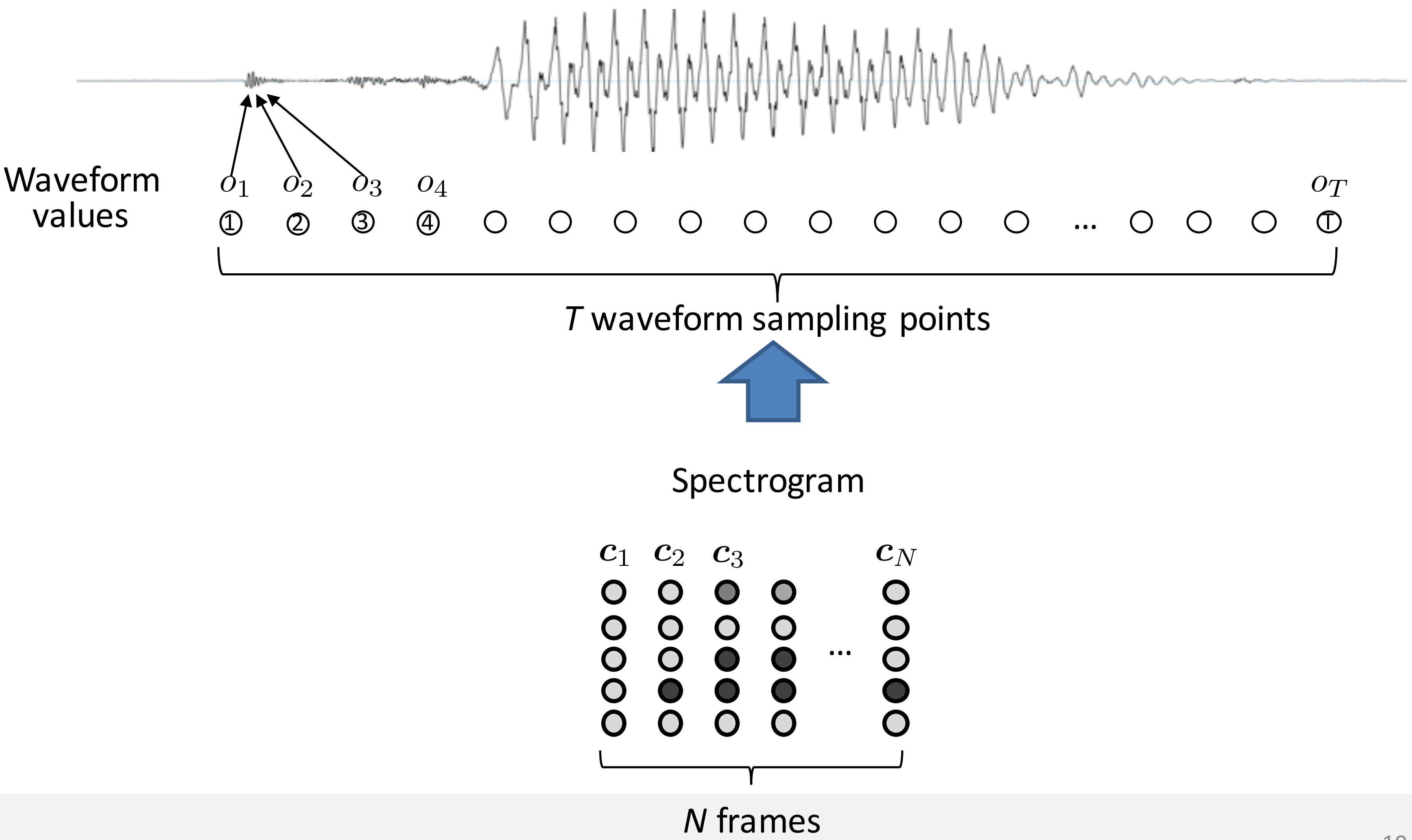


Spectrogram



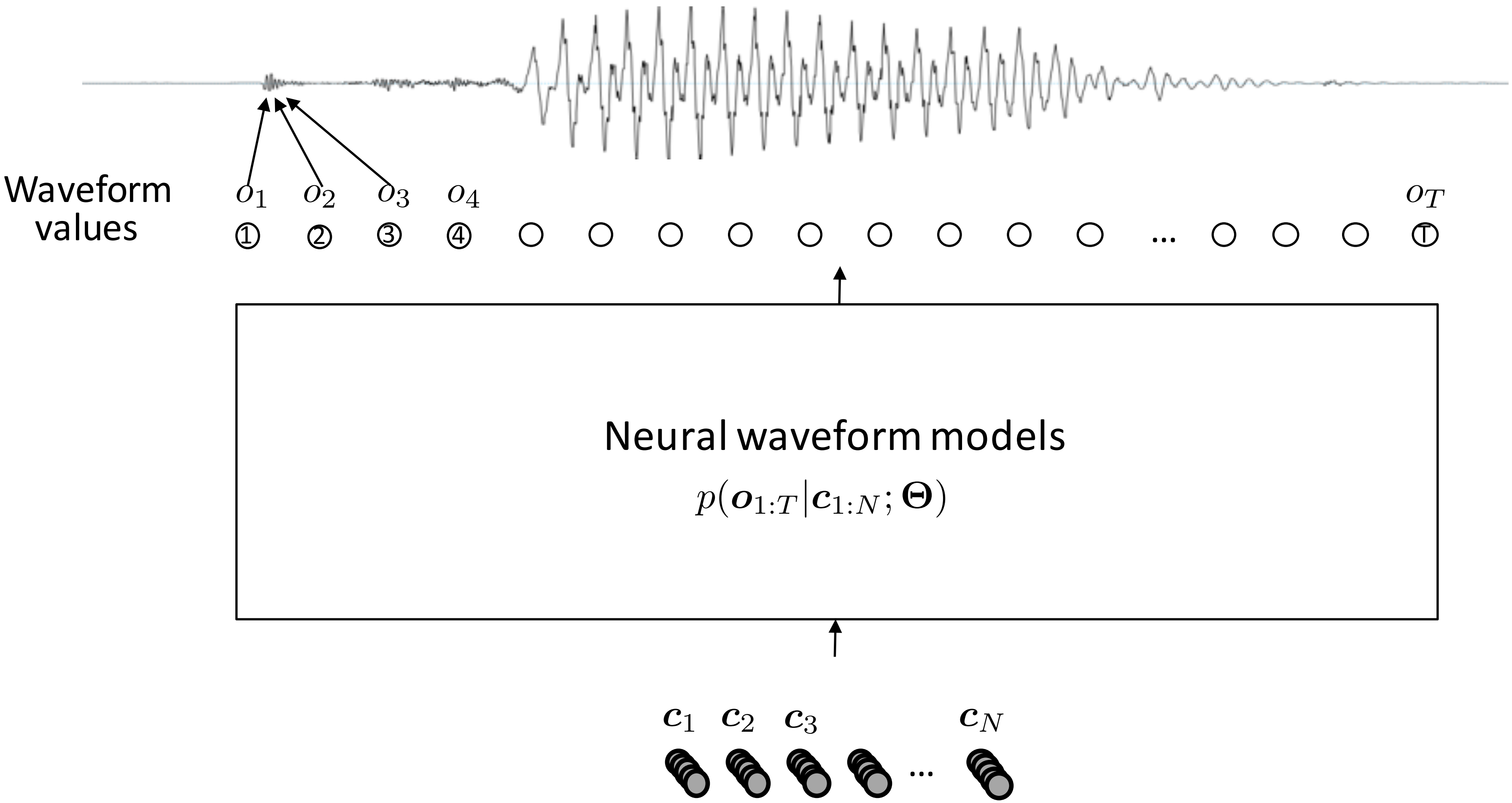
OVERVIEW

Task definition



OVERVIEW

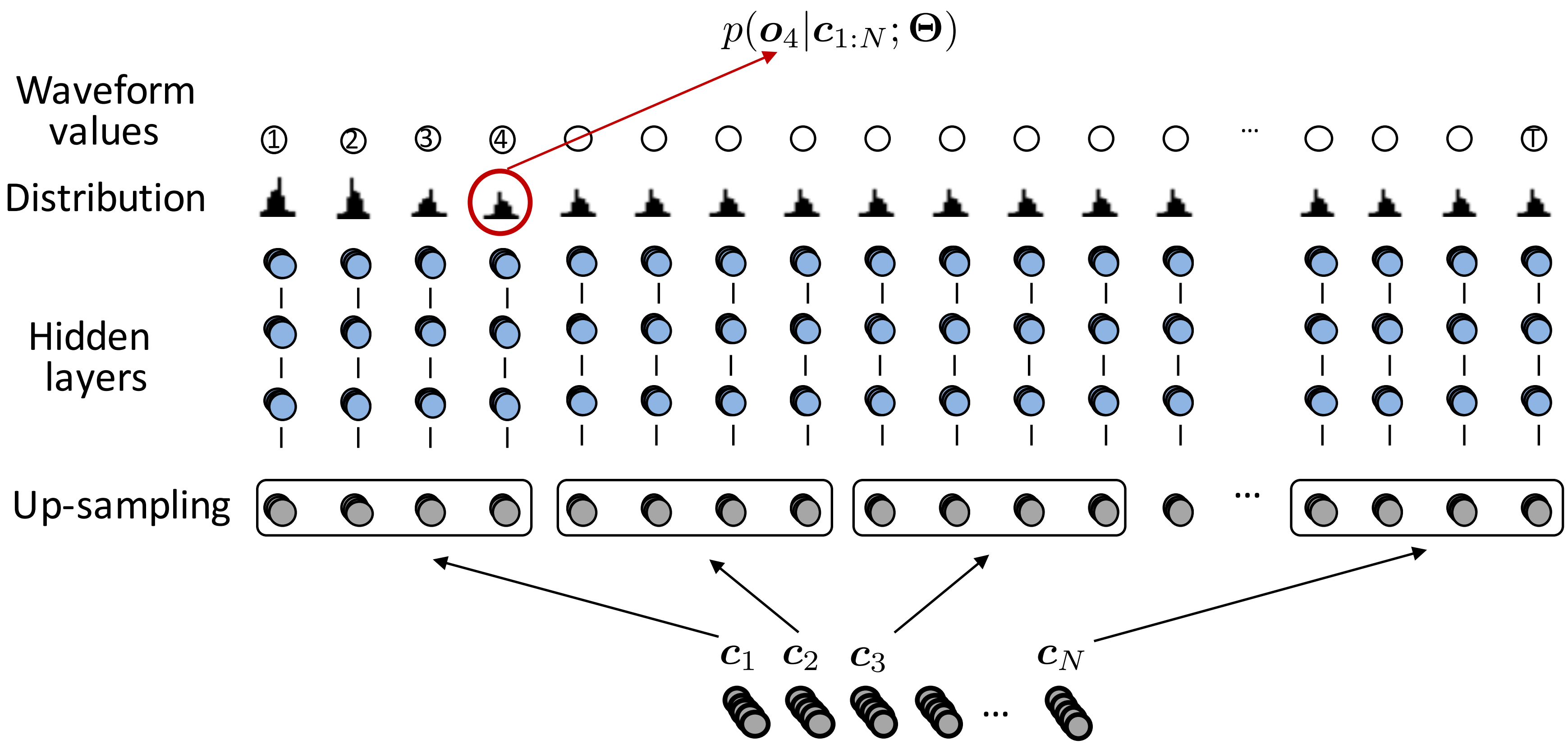
Task definition



OVERVIEW

Naïve model

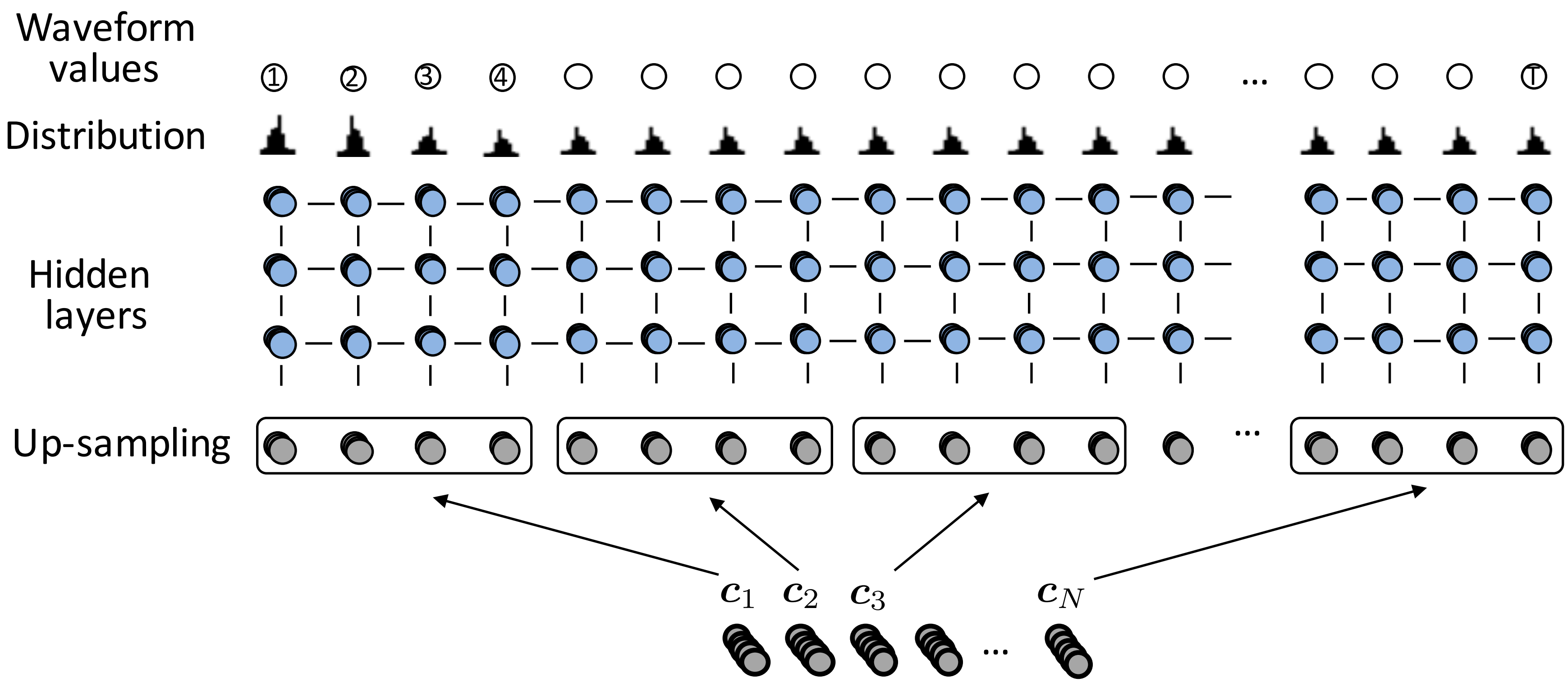
- Simple network + maximum-likelihood
  - Feedforward network



Overview

Naïve model

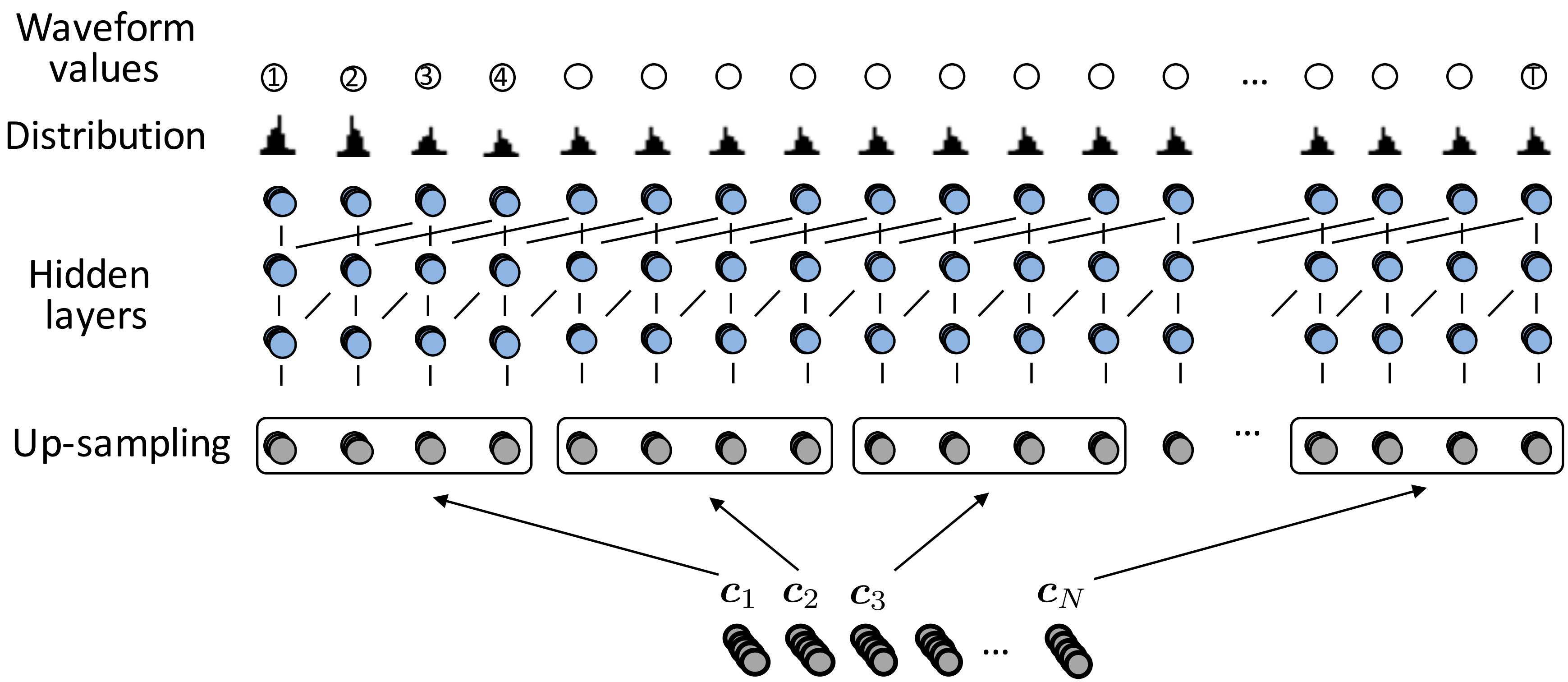
- Simple network + maximum-likelihood
  - RNN



Overview

Naïve model

- Simple network + maximum-likelihood
  - Dilated CNN [1,2]



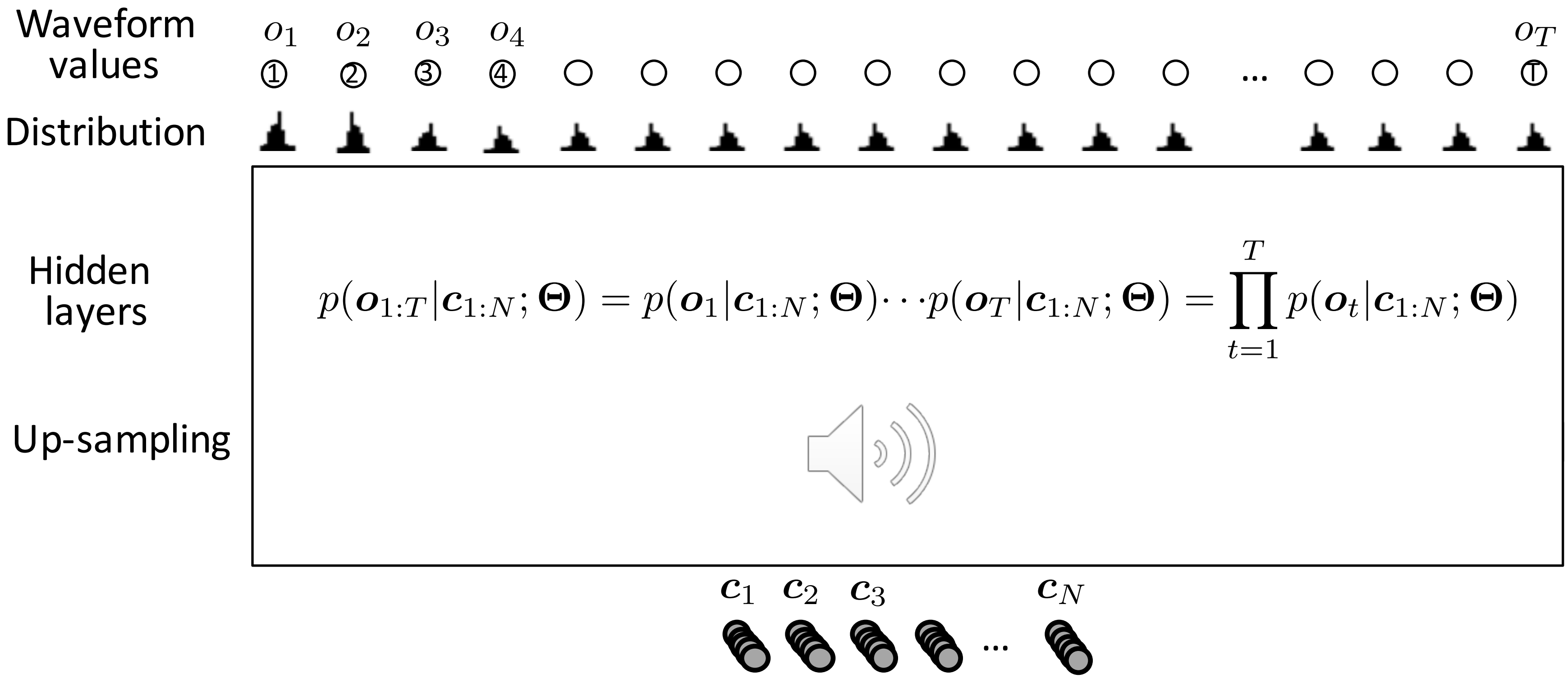
[1] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.
[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3):328–339, 1989.

OVERVIEW

Naïve model

mismatch

Weak temporal model <---------------> strongly correlated data $\boldsymbol{o}_{1:T}$

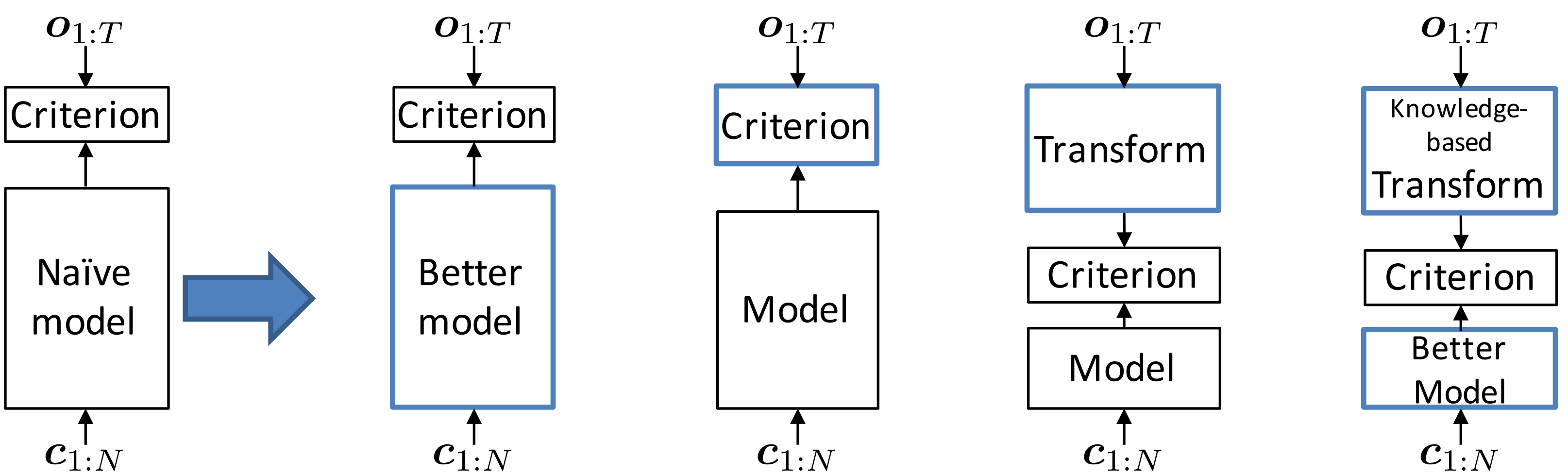


OVERVIEW

Towards better models

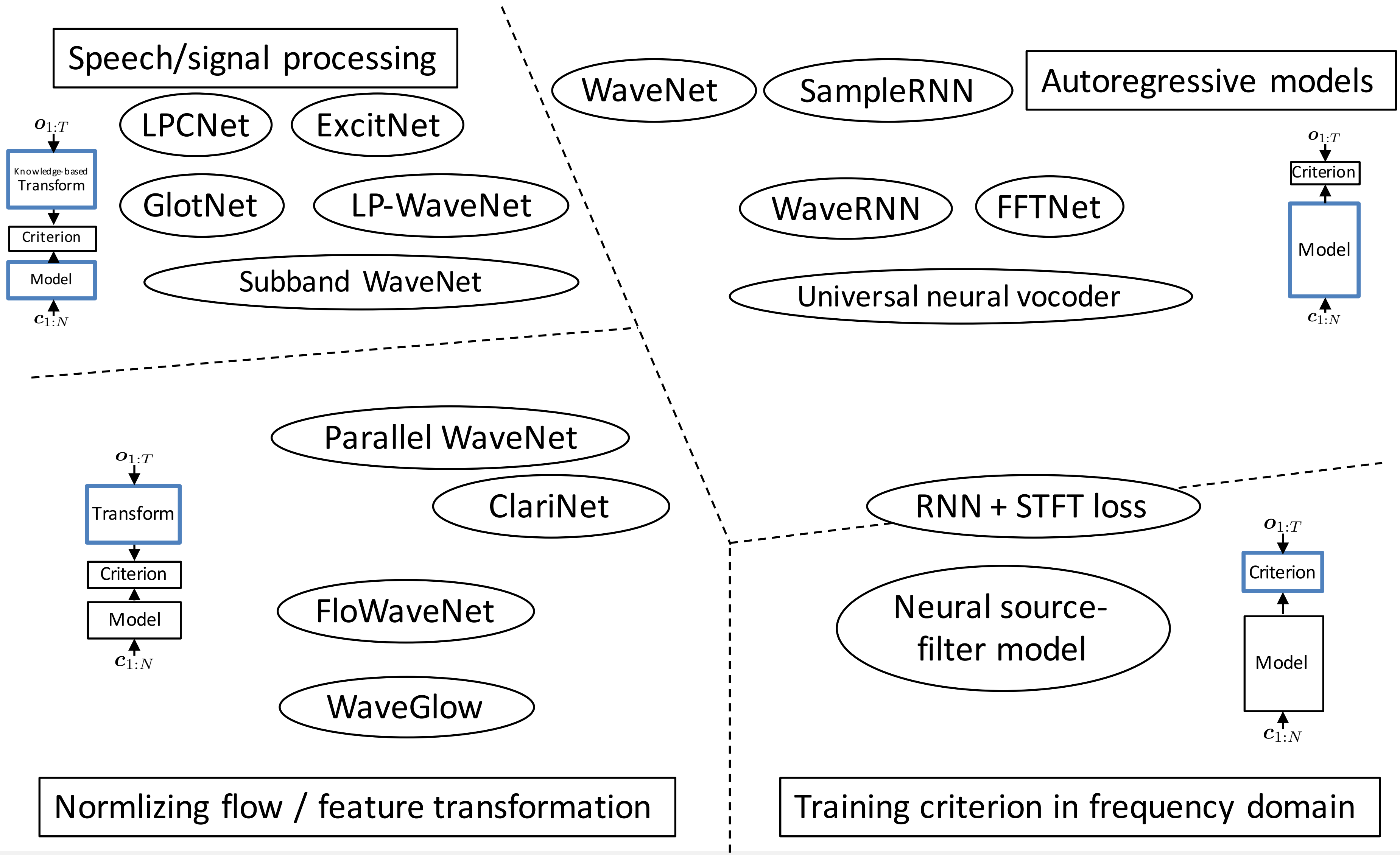
mismatch

Weak temporal model <---------------> strongly correlated data



Better model Better criterion Easier target Combination

OVERVIEW



☛ Please check reference in the appendix

CONTENTS

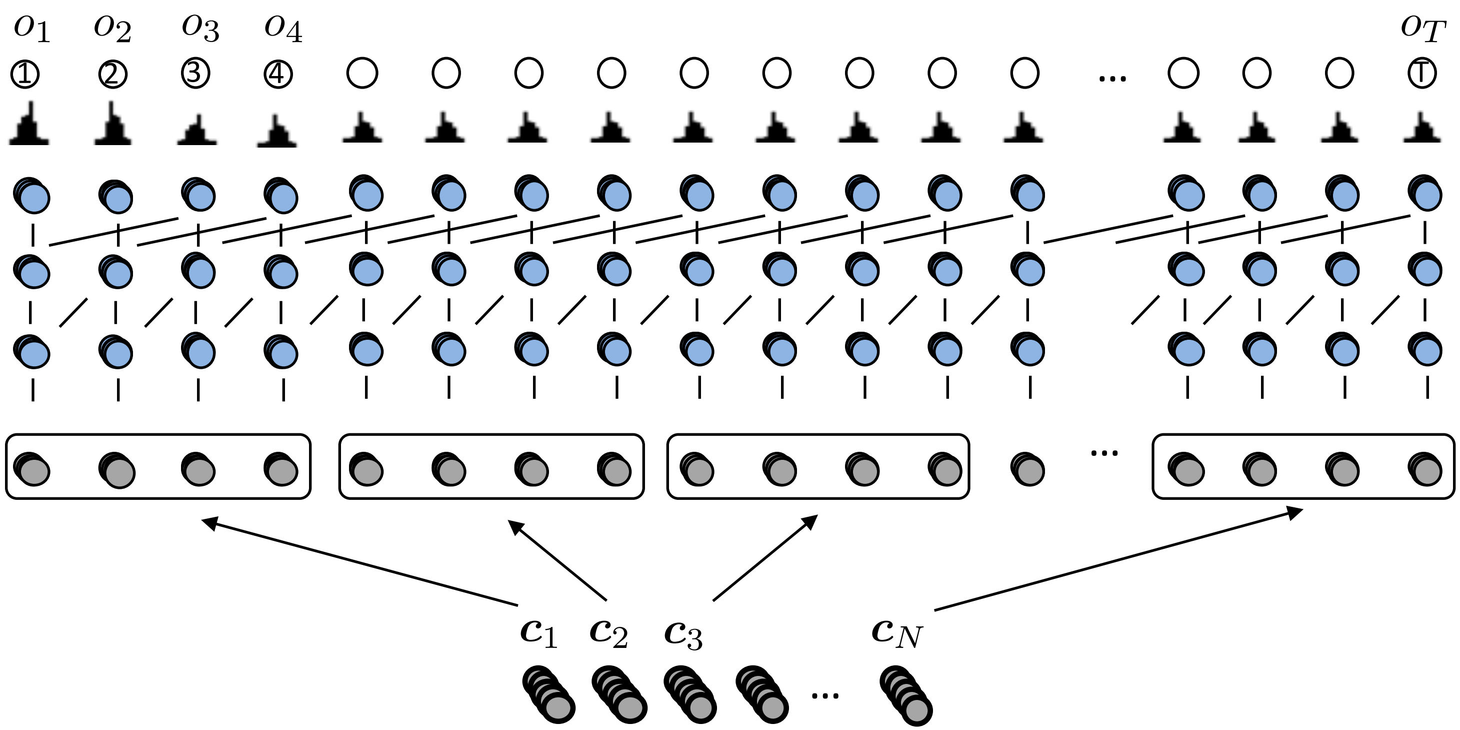
- Introduction: text-to-speech synthesis
- Neural waveform models
  - Overview
  - Autoregressive models
  - Normalizing flow
  - STFT-based training criterion
- Summary & software

PART I: AUTOREGRESSIVE MODELS

Core idea

- Naïve model

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{c}_{1:N};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\boldsymbol{c}_{1:N};\boldsymbol{\Theta})$$

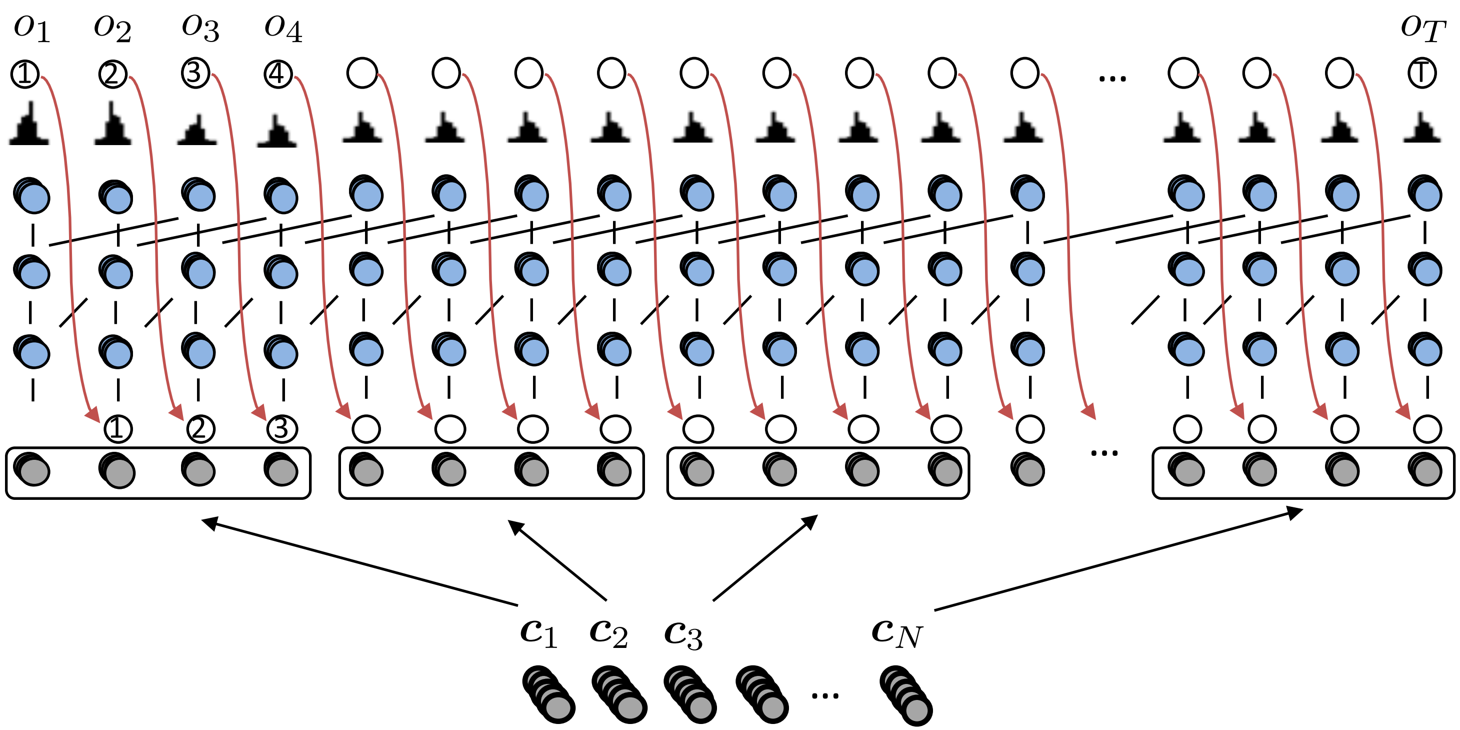


PART I: AUTOREGRESSIVE MODELS

Core idea

- Autoregressive (AR) model

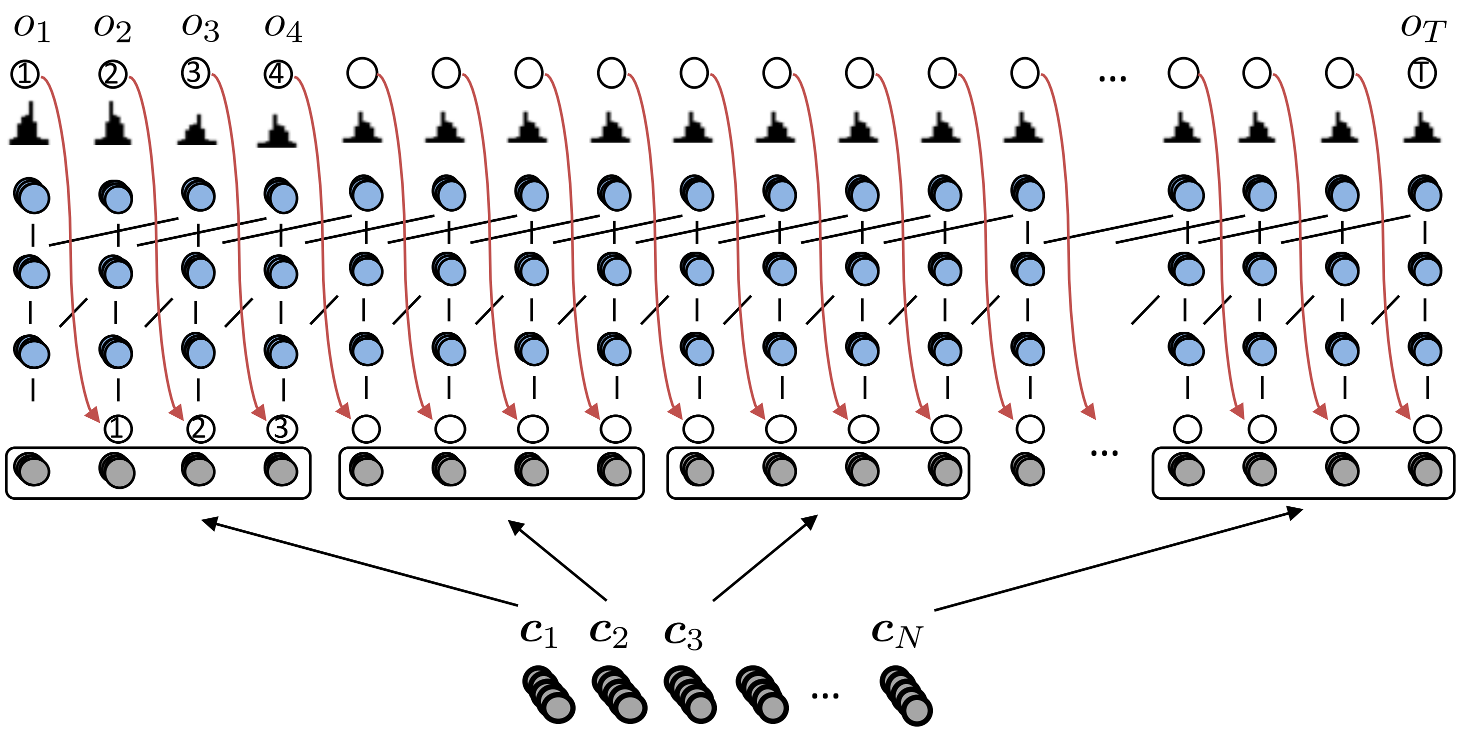
$$p(\boldsymbol{o}_{1:T}|\boldsymbol{c}_{1:N};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\underline{\boldsymbol{o}_{1:t-1},}\,\boldsymbol{c}_{1:N};\boldsymbol{\Theta})$$



PART I: AUTOREGRESSIVE MODELS

Core idea

- Autoregressive (AR) model
  - Training: use natural waveform for feedback (teacher forcing [1])

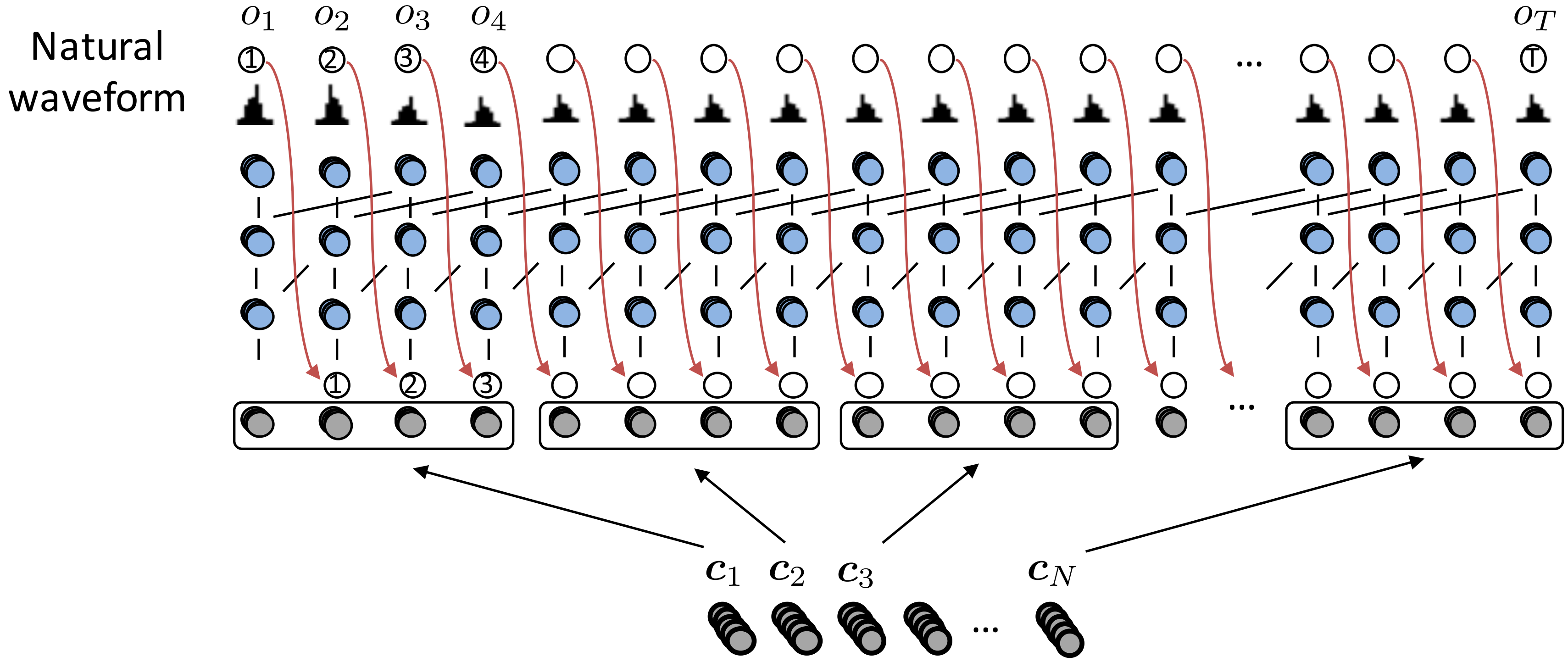


[1] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280, 1989.

PART I: AUTOREGRESSIVE MODELS

Core idea

- Autoregressive (AR) model
  - Training: use natural waveform for feedback (teacher forcing [1])



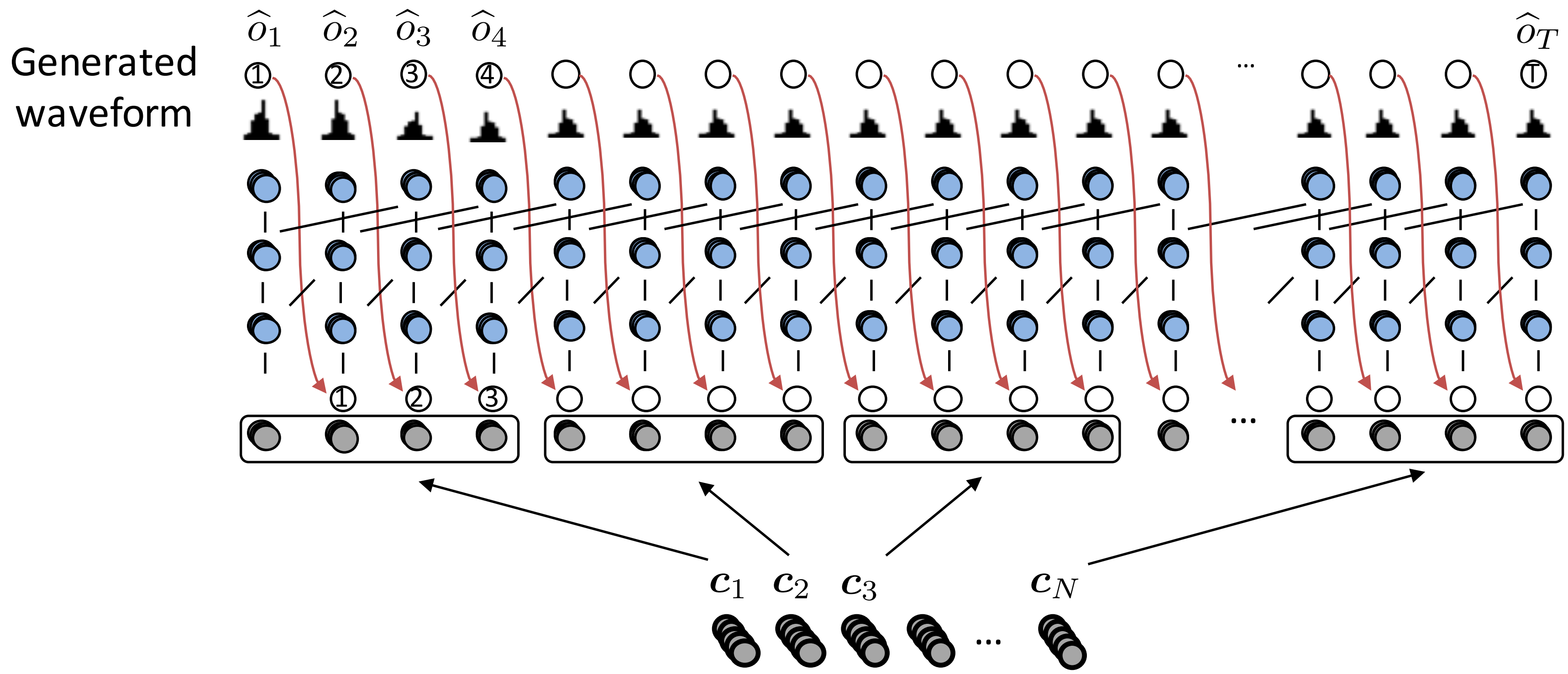
[1] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280, 1989.

PART I: AUTOREGRESSIVE MODELS

Core idea

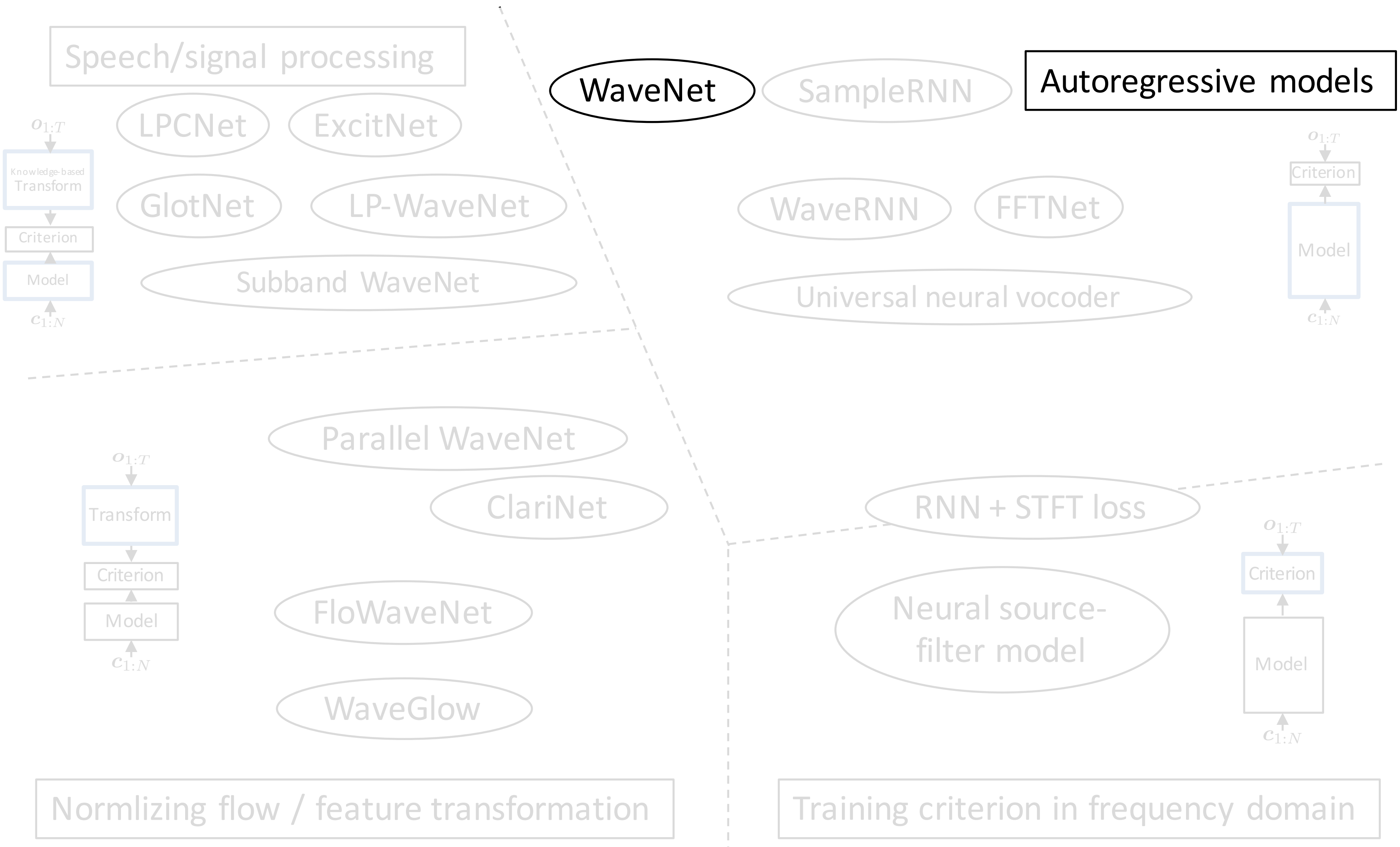
❑ Autoregressive (AR) model

- Training: use natural waveform for feedback (teacher forcing [1])
- Generation: $p(\widehat{\boldsymbol{o}}_{1:T}|\boldsymbol{c}_{1:N};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\widehat{\boldsymbol{o}}_t|\widehat{\boldsymbol{o}}_{t-P:t-1}, \boldsymbol{c}_{1:N};\boldsymbol{\Theta})$



[1] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280, 1989.

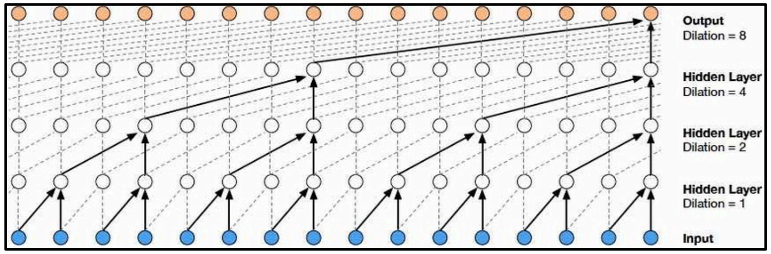
PART I: AUTOREGRESSIVE MODELS

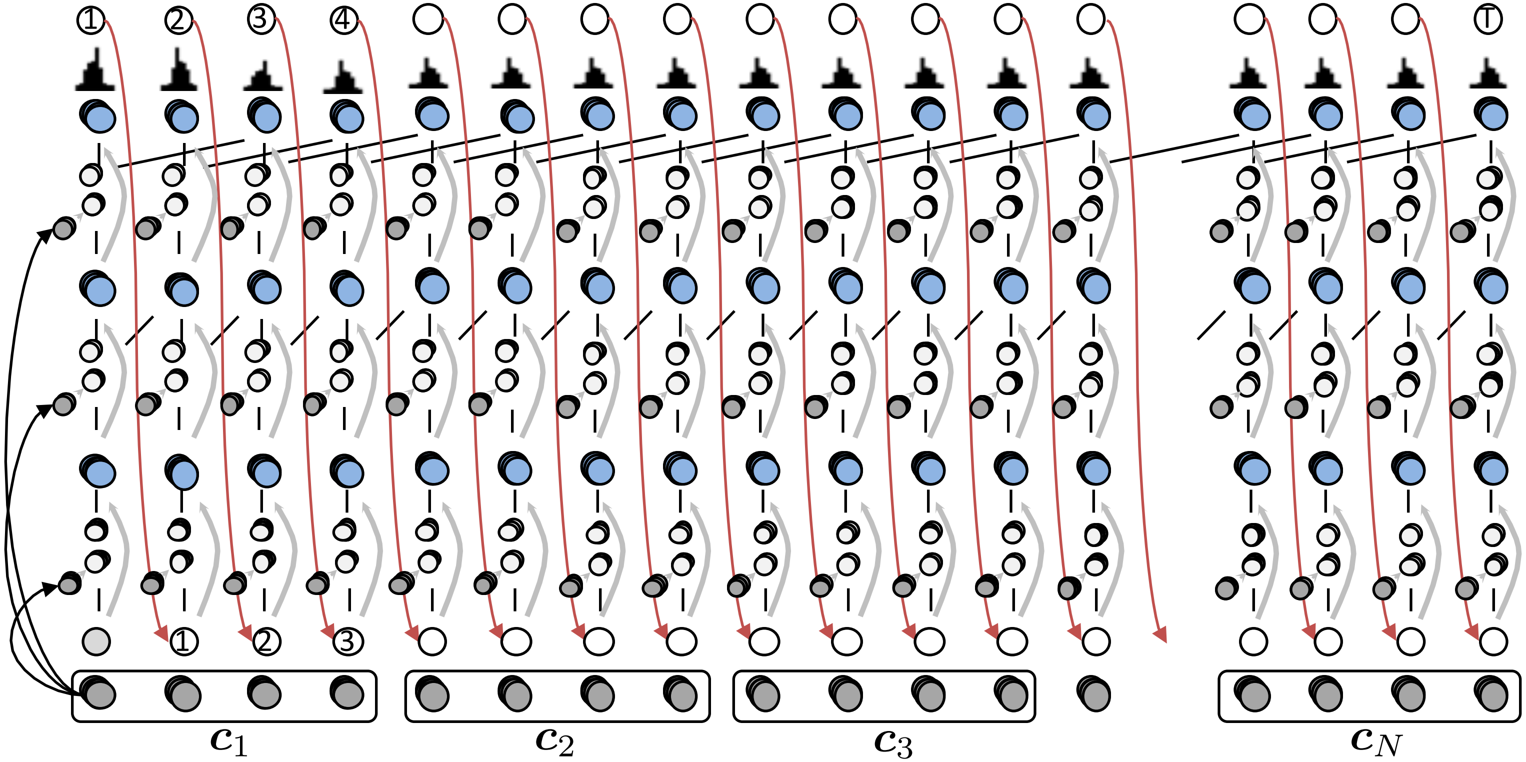


PART I: AUTOREGRESSIVE MODELS

WaveNet

❑ Network structure

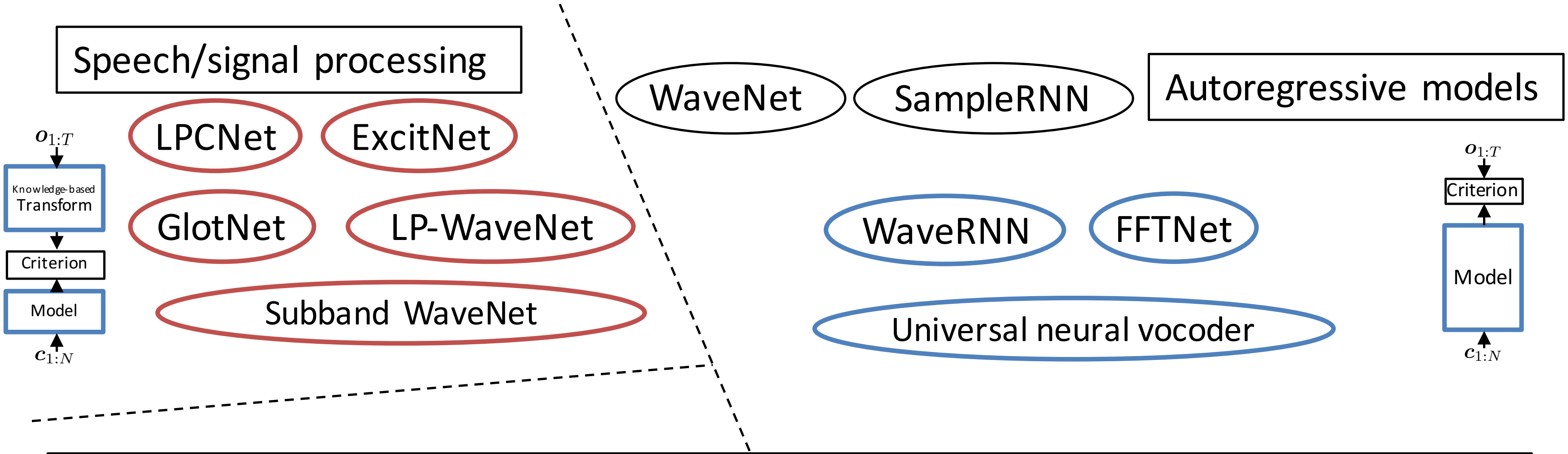




- Details: http://id.nii.ac.jp/1001/00185683/
http://tonywangx.github.io/pdfs/wavenet.pdf

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

PART I: AUTOREGRESSIVE MODELS



- Issues with WaveNet & Sample RNN
  - ! Generation is very very … very slow
- Acceleration?
  - **Engineering-based**: parallelize/simplify computation
  - **Knowledge-based**: speech / signal processing theory

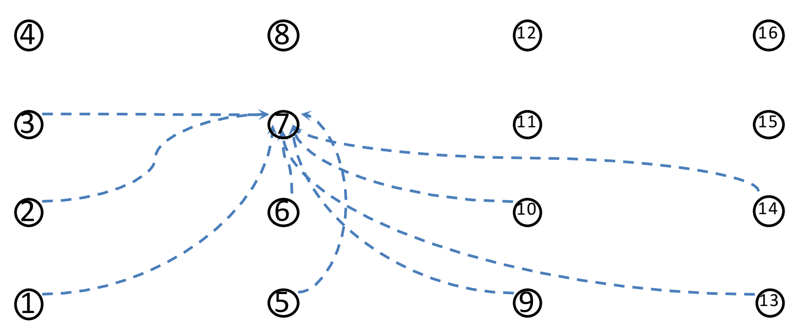
☛ Please check SampleRNN in the appendix

PART I: AUTOREGRESSIVE MODELS

WaveRNN

- Linear time AR dependency in WaveNet

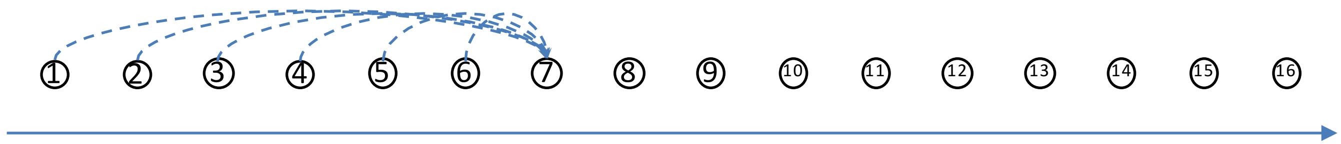
- Subscale AR dependency + batch sampling



PART I: AUTOREGRESSIVE MODELS

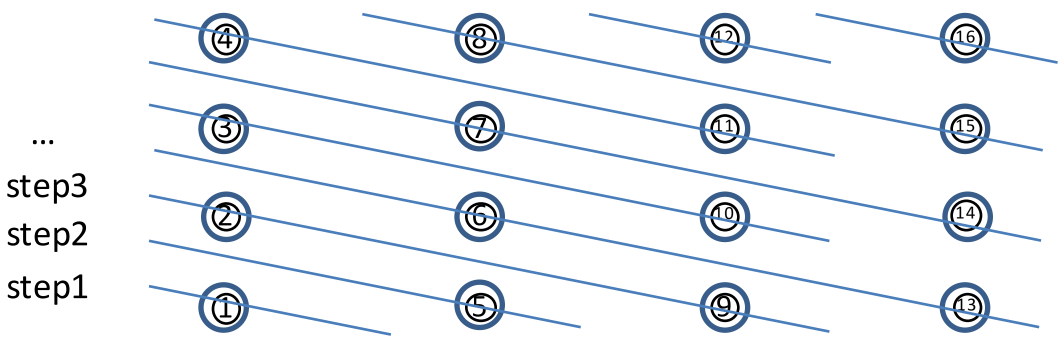
WaveRNN

- Linear time AR dependency in WaveNet



16 steps to generate

- Subscale AR dependency + batch sampling

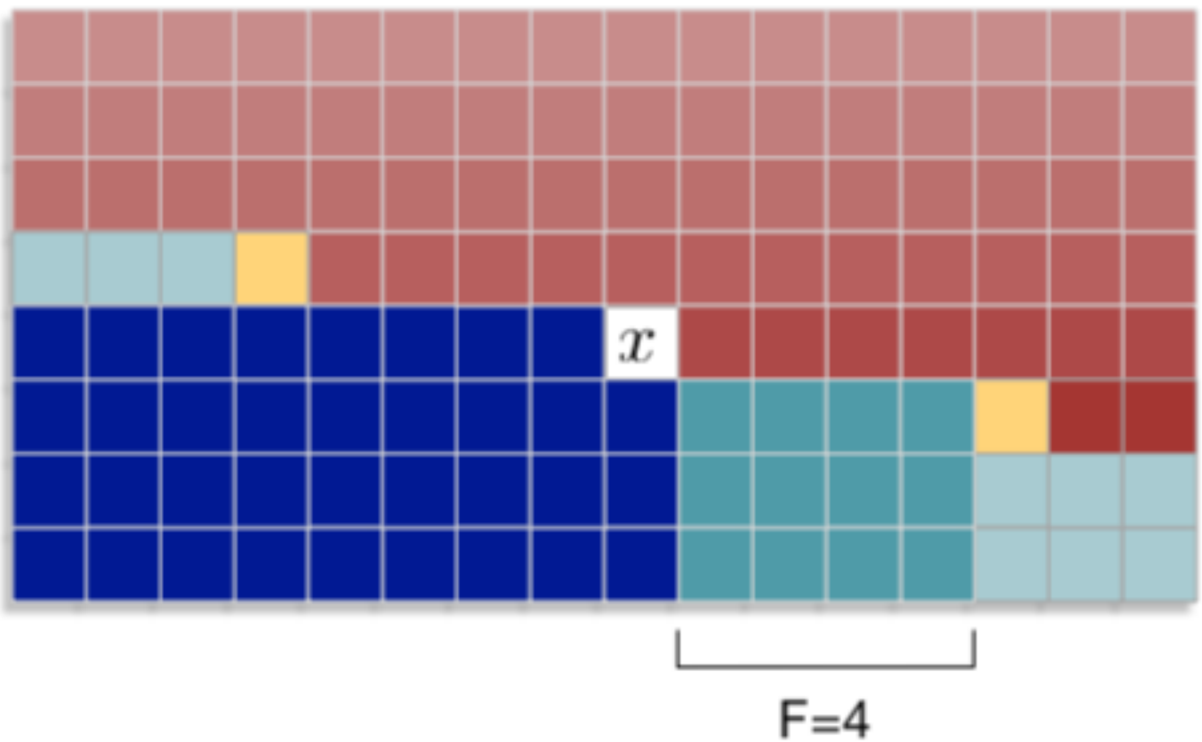


10 steps to generate

- Generate multiple samples at the same time

WaveRNN

- Linear time AR dependency in WaveNet



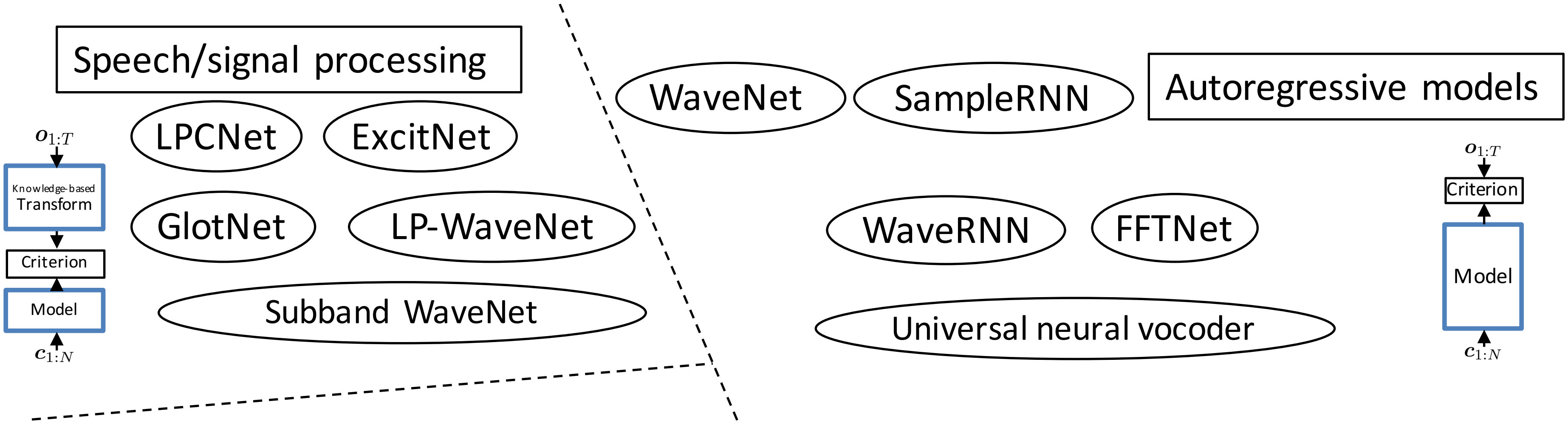
generate

- Subs

generate

- Generate multiple samples at the same time

PART I: AUTOREGRESSIVE MODELS

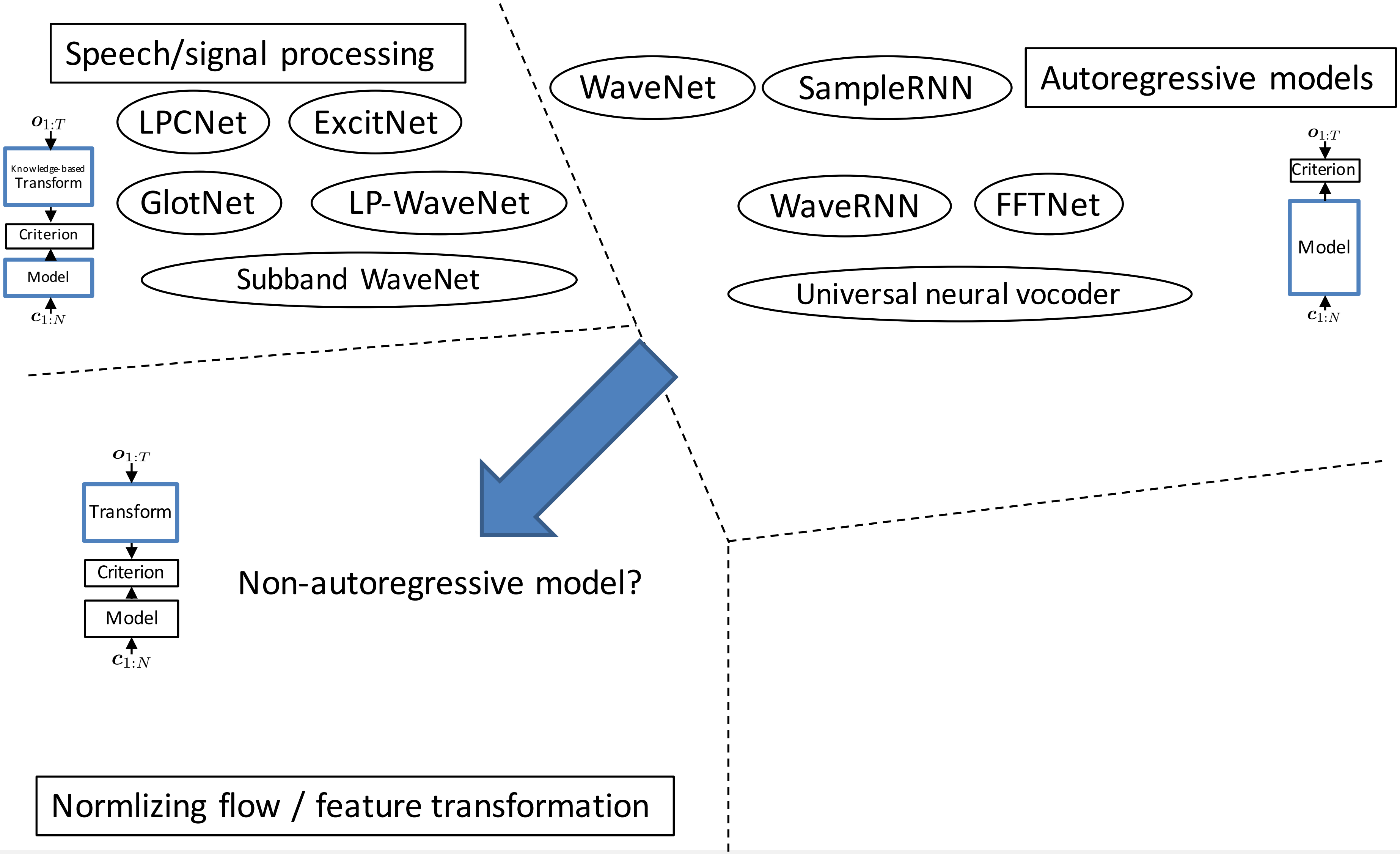


- AR models

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{c}_{1:N};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\boldsymbol{o}_{1:t-1},\boldsymbol{c}_{1:N};\boldsymbol{\Theta})$$

! Generation is still slow

PART I: AUTOREGRESSIVE MODELS

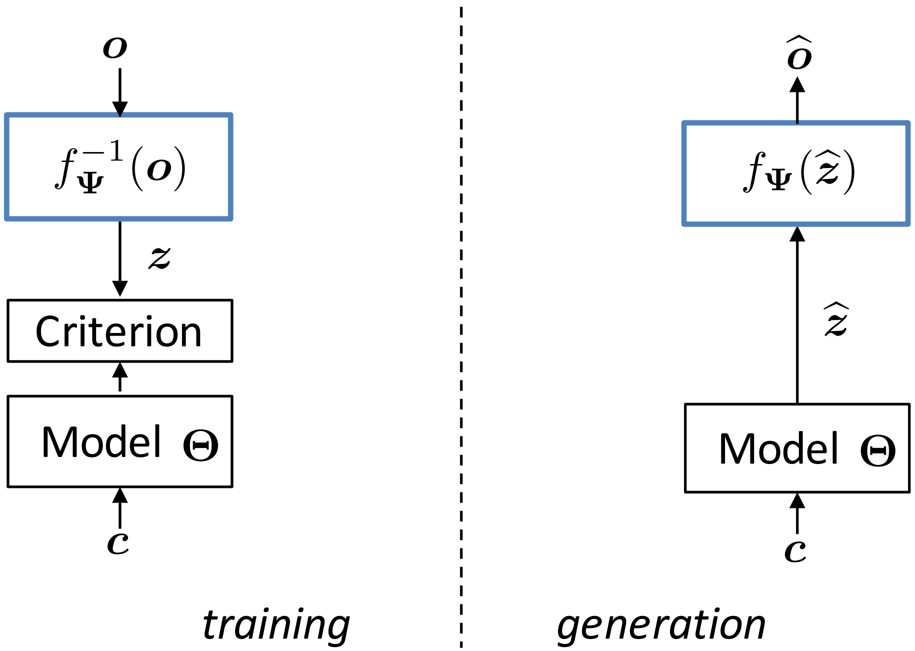


CONTENTS

PART II: NORMALIZING FLOW-BASED MODELS

General idea

$$p_o(\boldsymbol{o}|\boldsymbol{c};\boldsymbol{\Theta},\boldsymbol{\Psi}) = p_z(\boldsymbol{z} = f_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{o})|\boldsymbol{c};\boldsymbol{\Theta})\left|\det\frac{\partial f_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{o})}{\partial \boldsymbol{o}}\right|$$

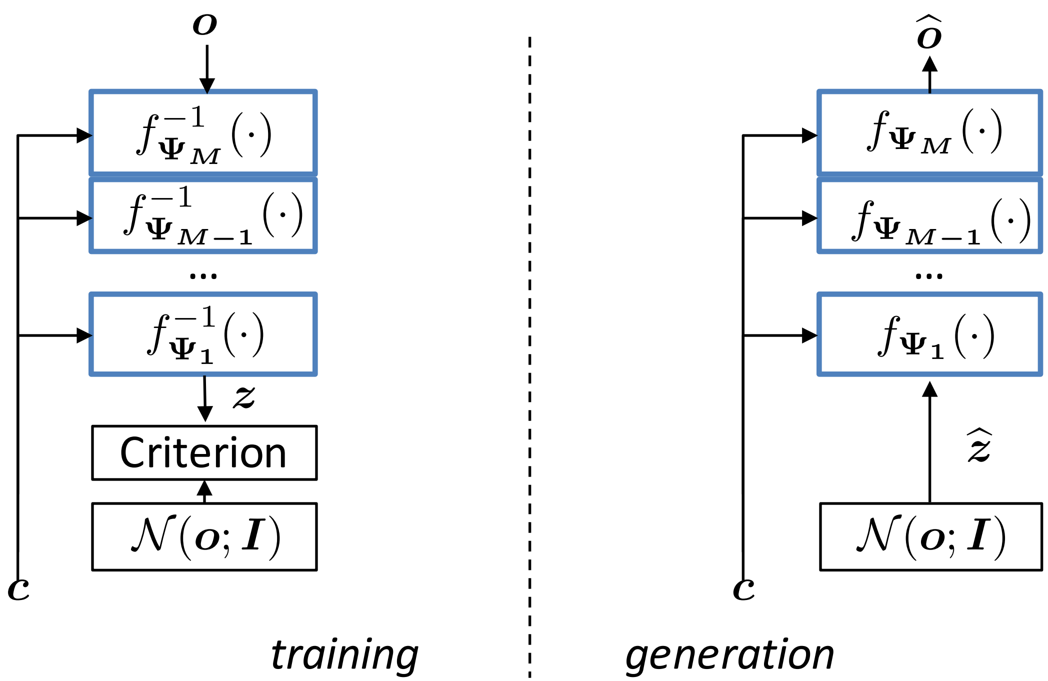


- *o* with strong temporal correlation ---> *z* with weak temporal correlation
- Principle of changing random variable

PART II: NORMALIZING FLOW-BASED MODELS

General idea

$$p_o(\boldsymbol{o}|\boldsymbol{c};\boldsymbol{\Psi}_1,\cdots,\boldsymbol{\Psi}_M) = p_z(\boldsymbol{z} = f^{-1}_{\boldsymbol{\Psi_1}} \circ \cdots \circ f^{-1}_{\boldsymbol{\Psi}_{\boldsymbol{M-1}}} \circ f^{-1}_{\boldsymbol{\Psi}_M}(\boldsymbol{o})|\boldsymbol{c}) \prod_{m=1}^{M} \left| \det \boldsymbol{J}(f^{-1}_{\boldsymbol{\Psi_m}}) \right|$$

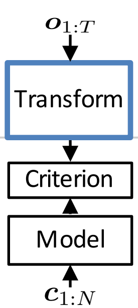


- *o* ---> Gaussian noise sequence *z*
- Multiple transformations: normalizing flow

Part II: Normalizing flow-based models

$$p_o(\boldsymbol{o}|\boldsymbol{c};\boldsymbol{\Psi}_1,\cdots,\boldsymbol{\Psi}_M) = p_z(\boldsymbol{z} = f^{-1}_{\boldsymbol{\Psi}_1} \circ \cdots \circ f^{-1}_{\boldsymbol{\Psi}_{M-1}} \circ f^{-1}_{\boldsymbol{\Psi}_M}(\boldsymbol{o})|\boldsymbol{c}) \prod_{m=1}^{M} \left| \det \boldsymbol{J}(f^{-1}_{\boldsymbol{\Psi}_m}) \right|$$

- Different time complexity $f_{\boldsymbol{\Psi}_m}(\cdot)$, different training strategies



Parallel WaveNet
ClariNet

$f_{\boldsymbol{\Psi}_m}(\cdot)$ in linear time domain

FloWaveNet
WaveGlow

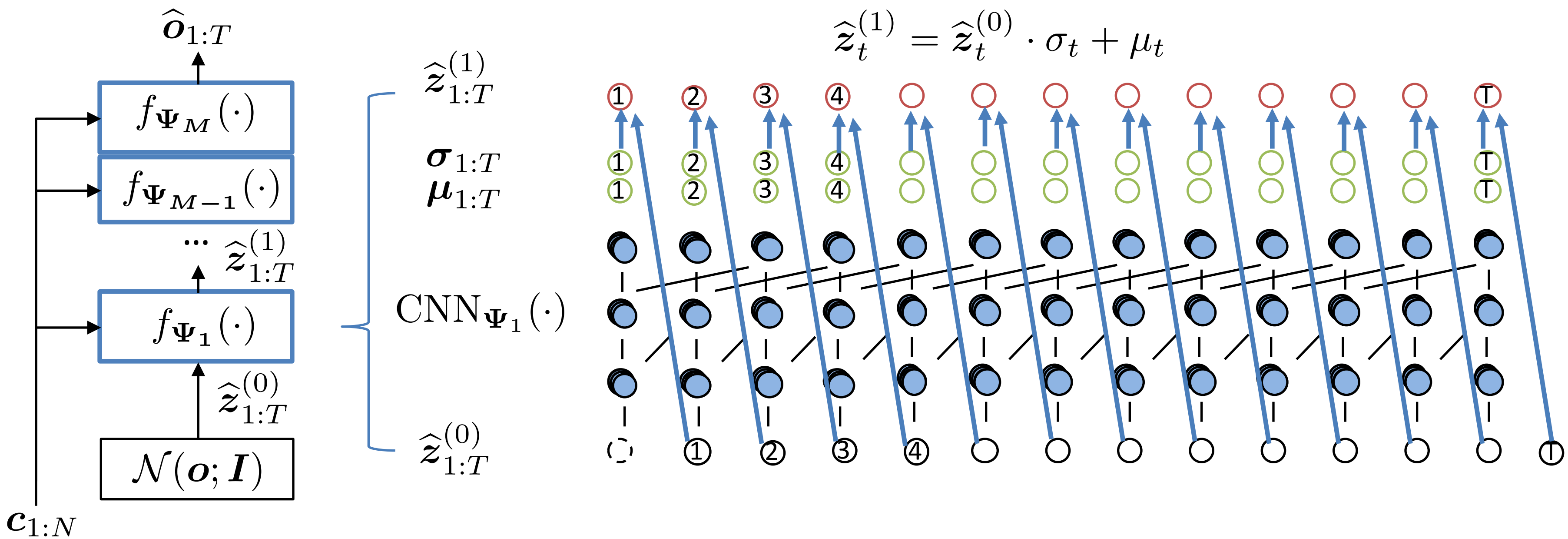
$f_{\boldsymbol{\Psi}_m}(\cdot)$ in subscale time domain

A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, et al. Parallel WaveNet: Fast high-fidelity speech synthesis. arXiv preprint arXiv:1711.10433, 2017.
W. Ping, K. Peng, and J. Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281, 2018.
S. Kim, S.-g. Lee, J. Song, and S. Yoon. Flowavenet: A generative flow for raw audio. arXiv preprint arXiv:1811.02155, 2018.
R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. arXiv preprint arXiv:1811.00002, 2018.

PART II: NORMALIZING FLOW-BASED MODELS

ClariNet & parallel WaveNet

- Generation process



✓ Parallel computation

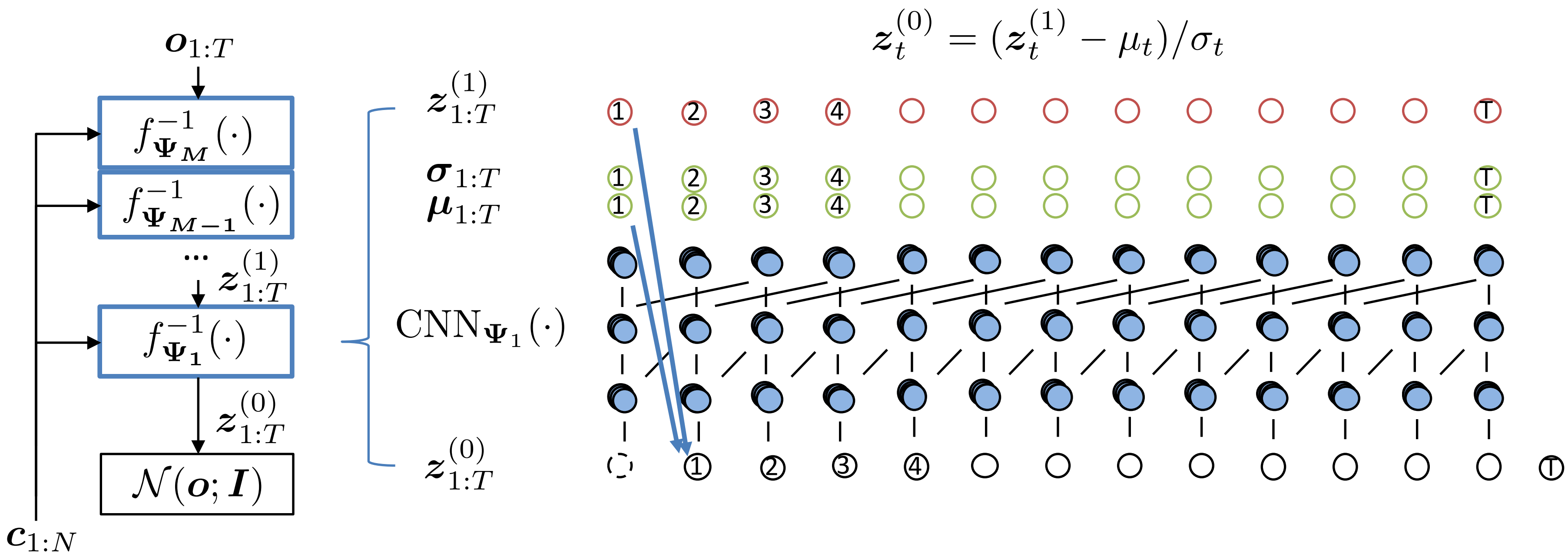
✓ Fast generation

* Initial condition may be implemented differently

PART II: NORMALIZING FLOW-BASED MODELS

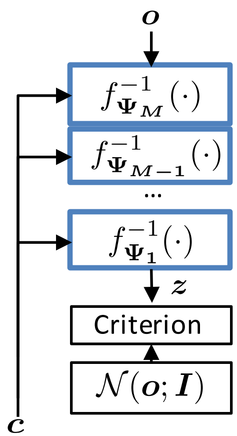
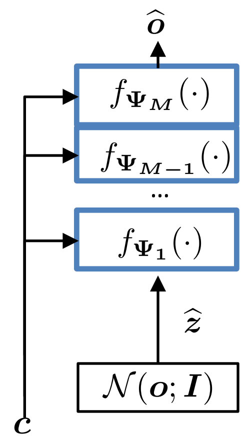
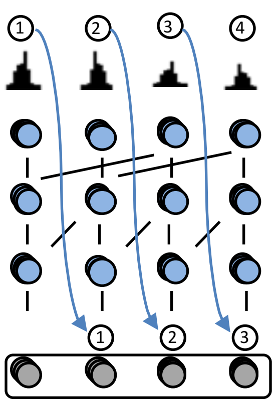
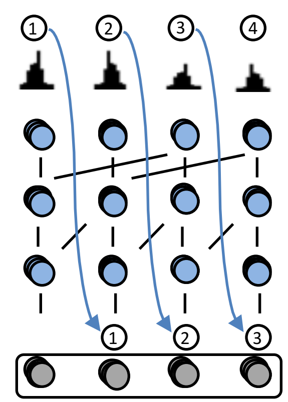
ClariNet & parallel WaveNet

- Naïve training process



! Training time ~ O(T)

PART II: NORMALIZING FLOW-BASED MODELS

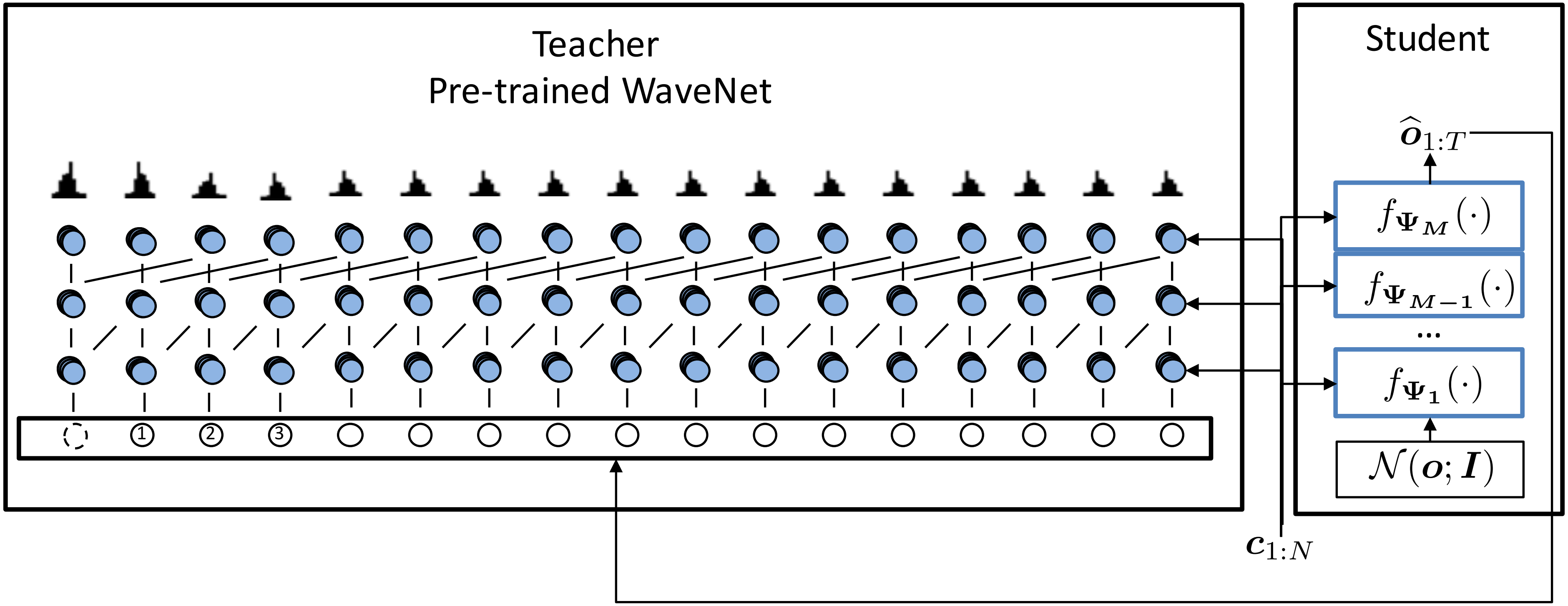
	Training		Generation	
(Inverse-AR [1]) Normalizing flow		! ~ O(T)		✓ ~ O(1)
Original WaveNet		✓ ~ O(1)		! ~ O(T)

[1] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In Proc. NIPS, pages 4743–4751, 2016.

PART II: NORMALIZING FLOW-BASED MODELS

ClariNet & parallel WaveNet

- Fast training : knowledge distilling

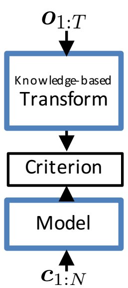


- Teacher gives $p(\widehat{o}_t|\widehat{\boldsymbol{o}}_{1:t-1}, \boldsymbol{c}_{1:T}, \text{teacher})$
- Student gives $p(\widehat{o}_t|\boldsymbol{z}_{1:T}, \boldsymbol{c}_{1:T}, \text{student})$

Student learns from teacher

PART II: NORMALIZING FLOW-BASED MODELS

Speech/signal processing



LPCNet

ExcitNet

GlotNet

LP-WaveNet

Subband WaveNet

WaveNet

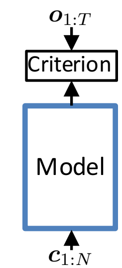
SampleRNN

Autoregressive models

WaveRNN

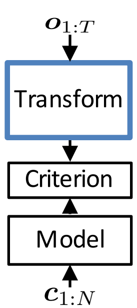
FFTNet

Universal neural vocoder



Parallel WaveNet

ClariNet



Without teacher WaveNet ?

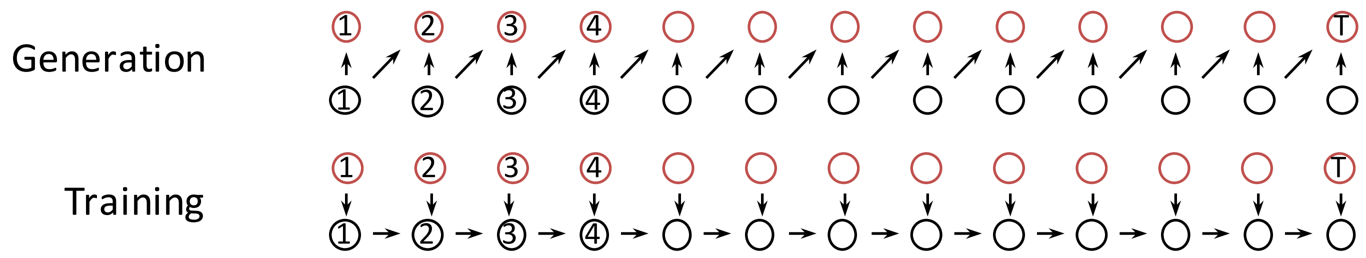
- Parallel WaveNet & ClariNet
  - ✓ No AR dependency in generation
  - ! Need pre-trained WaveNet

Normlizing flow / feature transformation

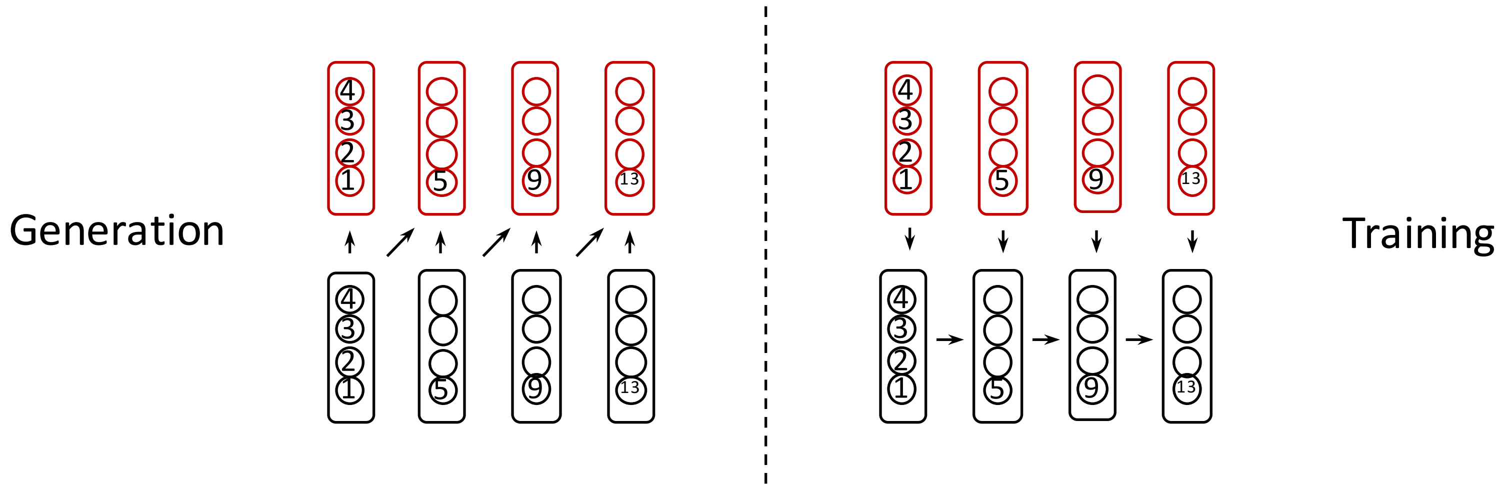
PART II: NORMALIZING FLOW-BASED MODELS

WaveGlow

- Why $f^{-1}_{\Psi_1}(\cdot)$ ~ O(T) ? Dependency in linear time domain



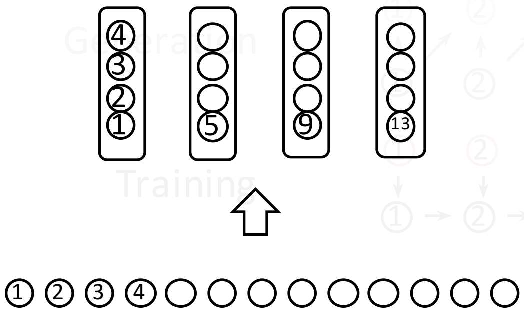
- Reduce T? Dependency in subscale time domain



R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. arXiv preprint arXiv:1811.00002, 2018.

PART II: NORMALIZING FLOW-BASED MODELS

WaveGlow



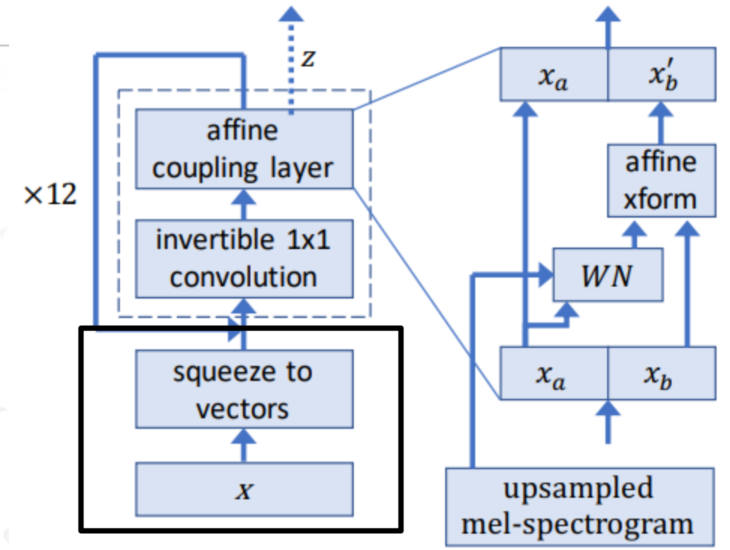
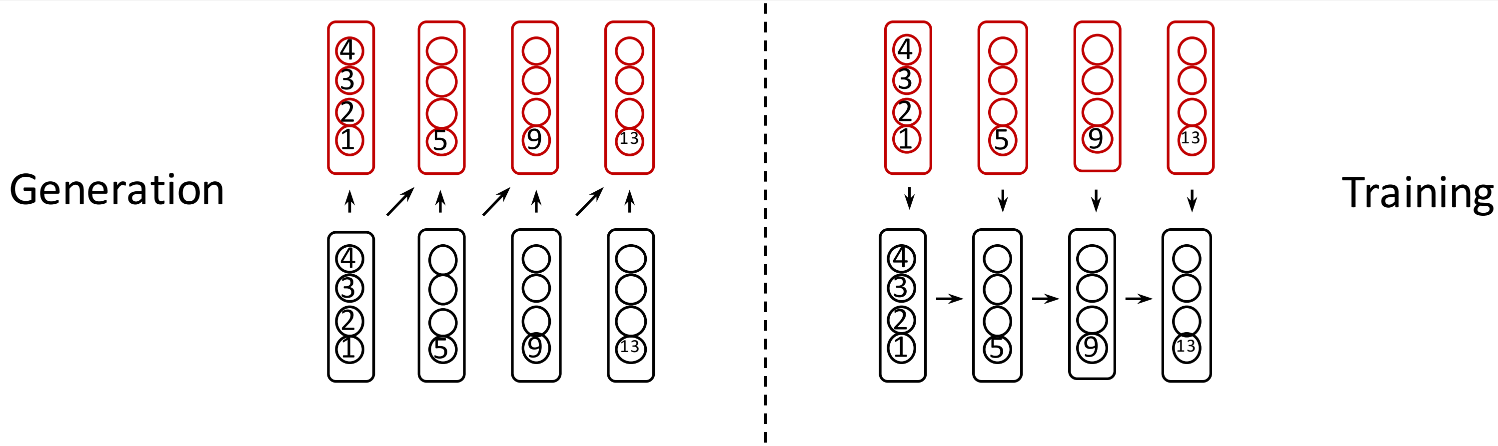
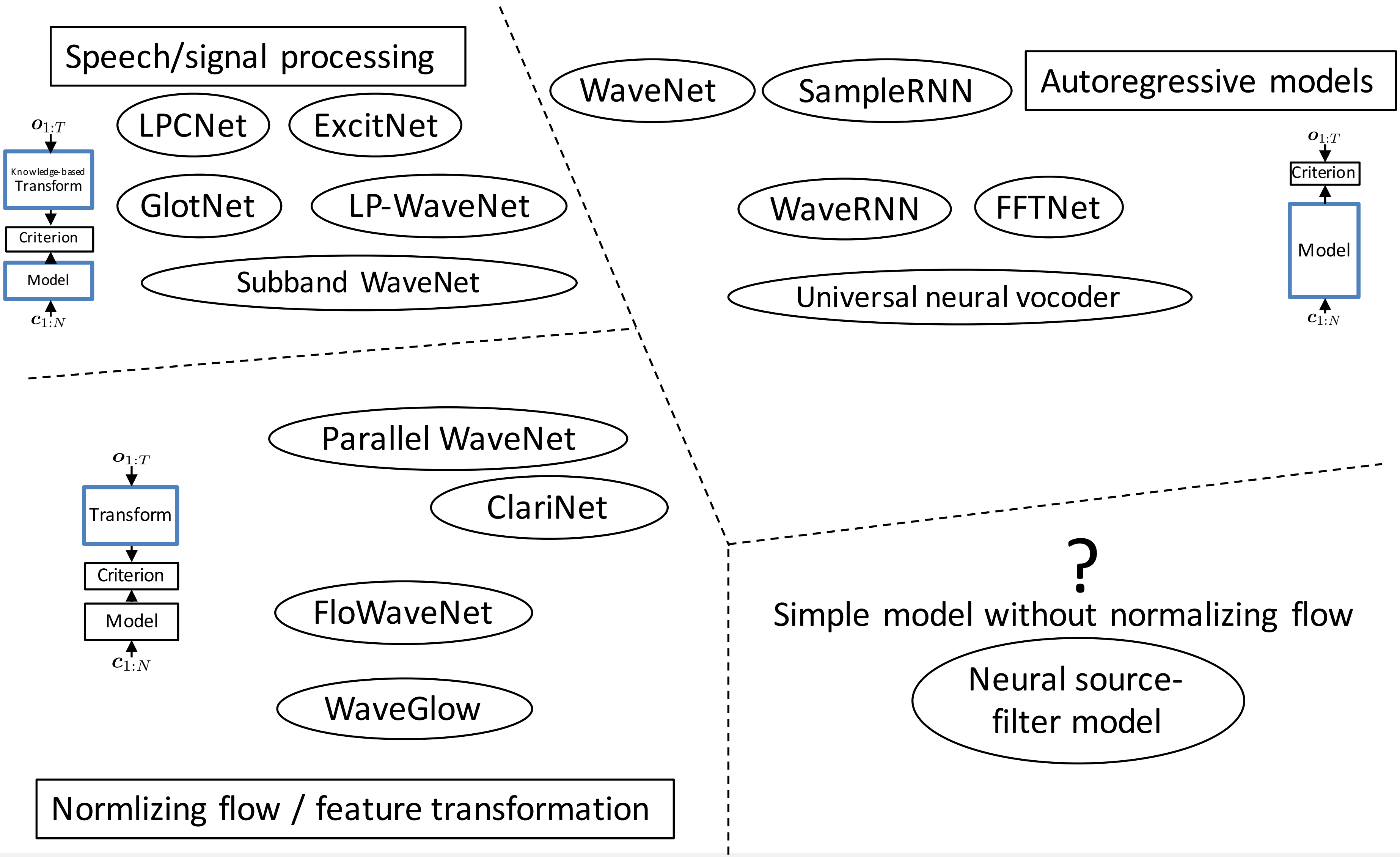


Fig. 1: WaveGlow network



R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. arXiv preprint arXiv:1811.00002, 2018.

PART II: NORMALIZING FLOW-BASED MODELS

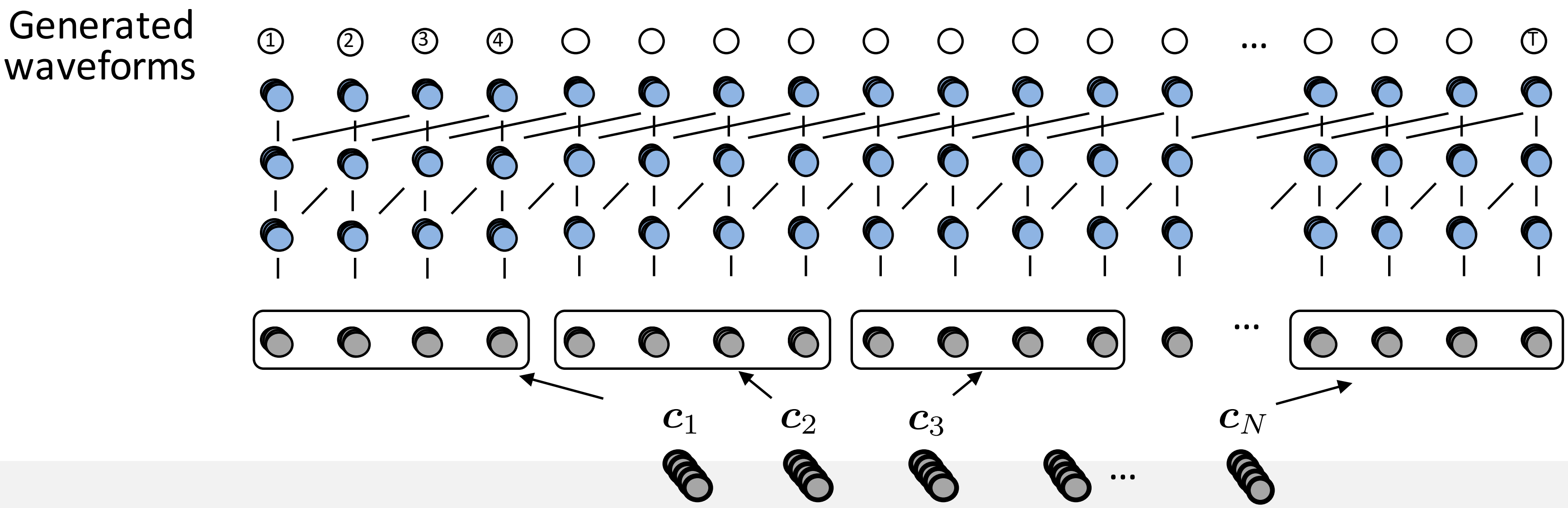


CONTENTS

Part III: STFT-based training criterion

Neural source-filter model

- Naïve model and STFT-based criterion

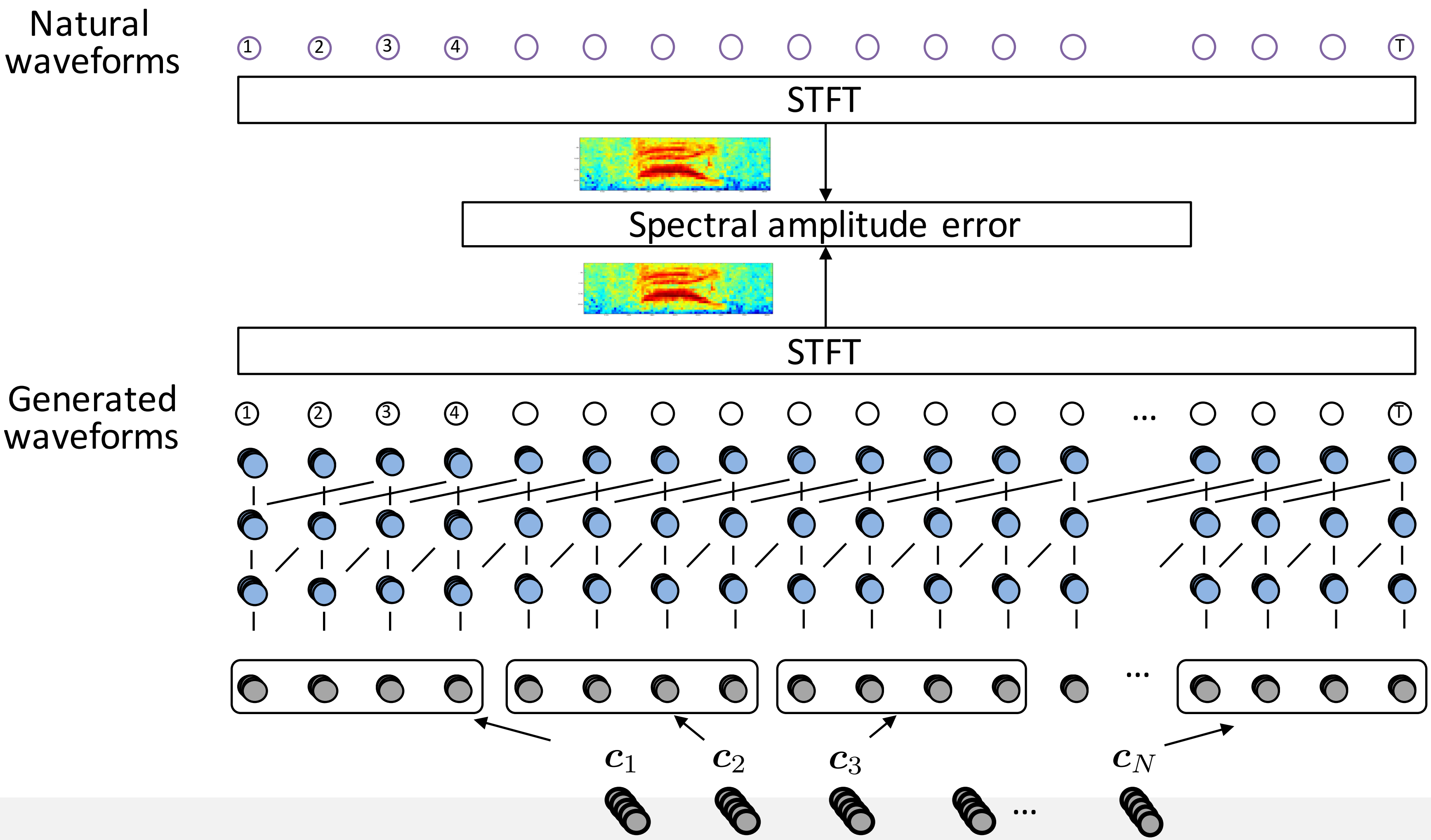


X. Wang, S. Takaki, and J. Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. arXiv preprint arXiv:1810.11946, 2018.

PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Naïve model and STFT-based criterion

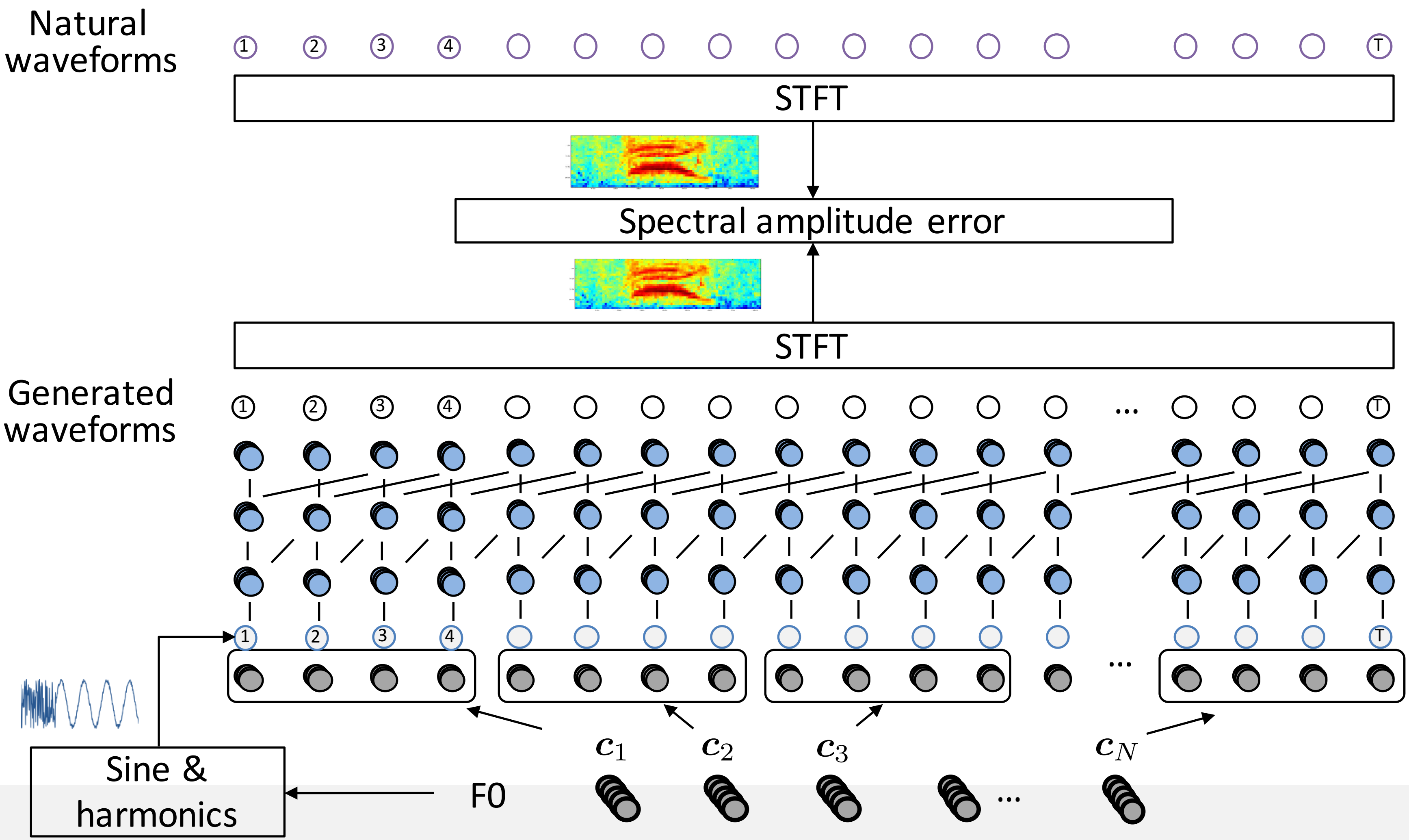


X. Wang, S. Takaki, and J. Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. arXiv preprint arXiv:1810.11946, 2018.

PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

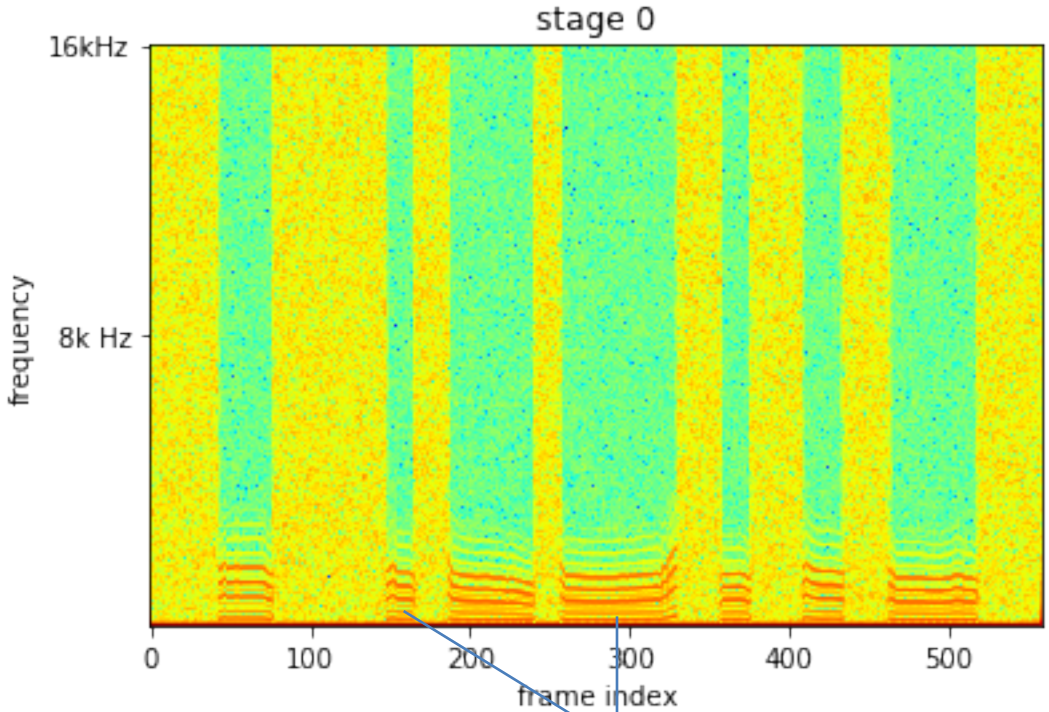
- Naïve model and STFT-based criterion

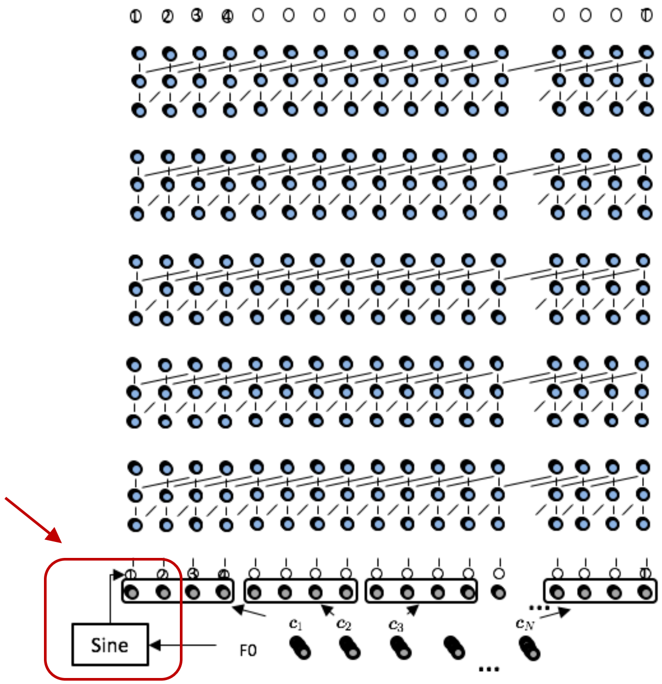


PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Examples



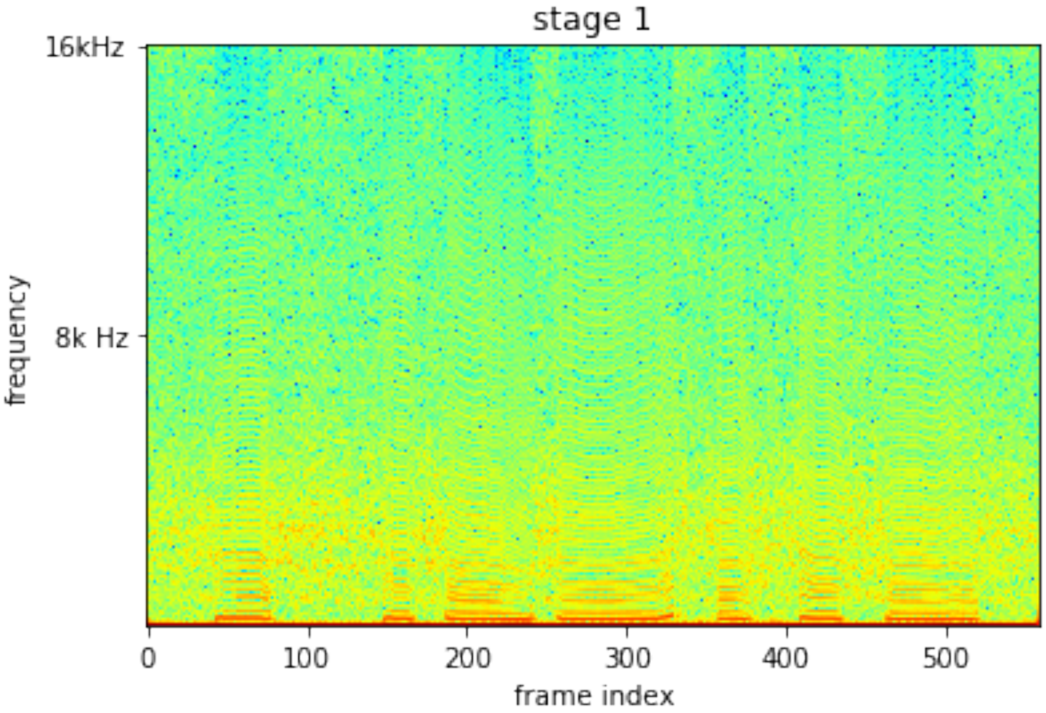


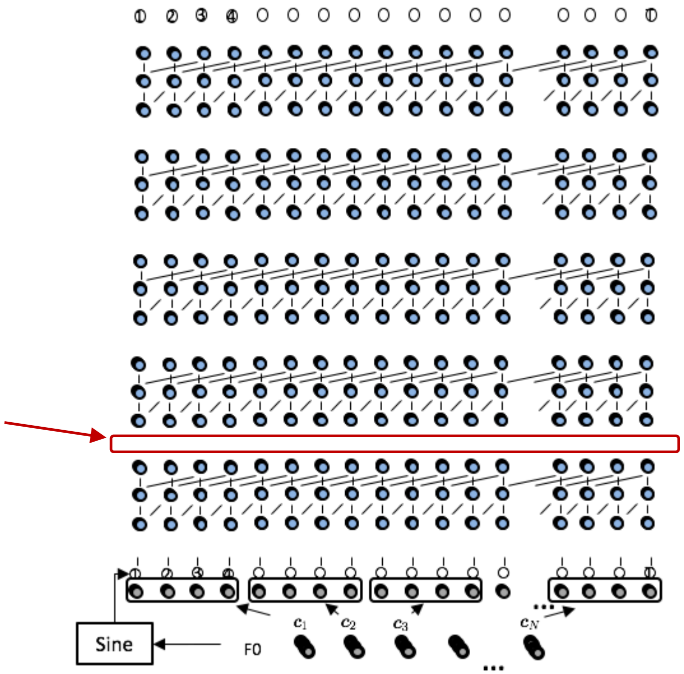
Only harmonics
No formants

PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Examples

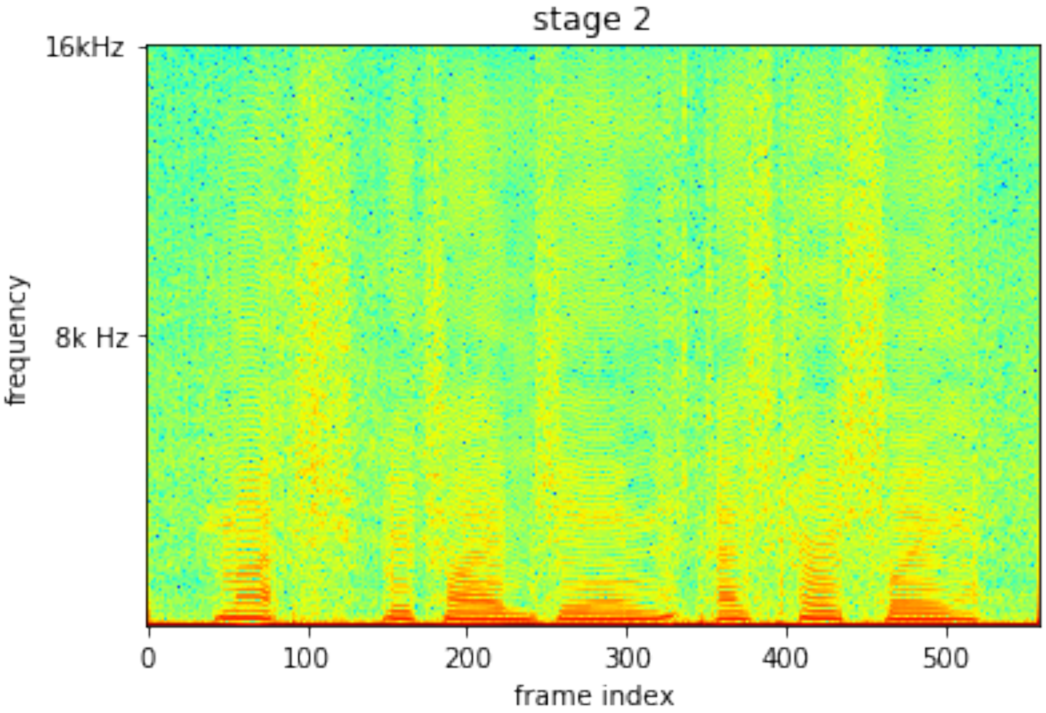


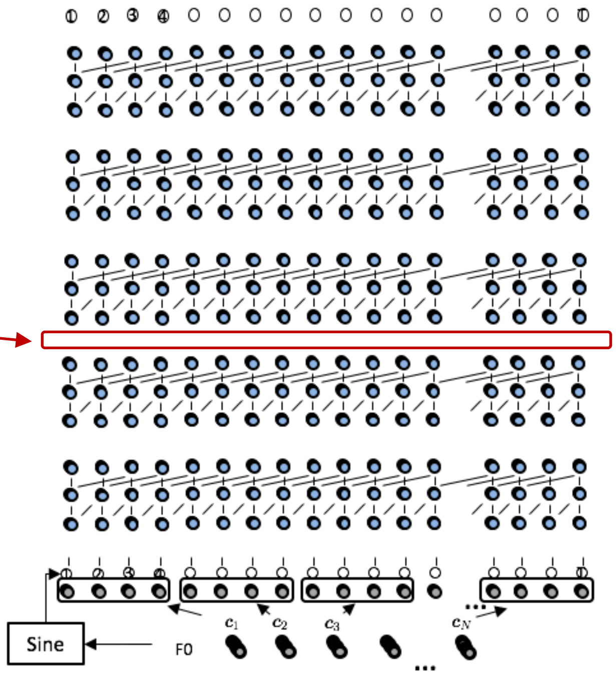


PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Examples

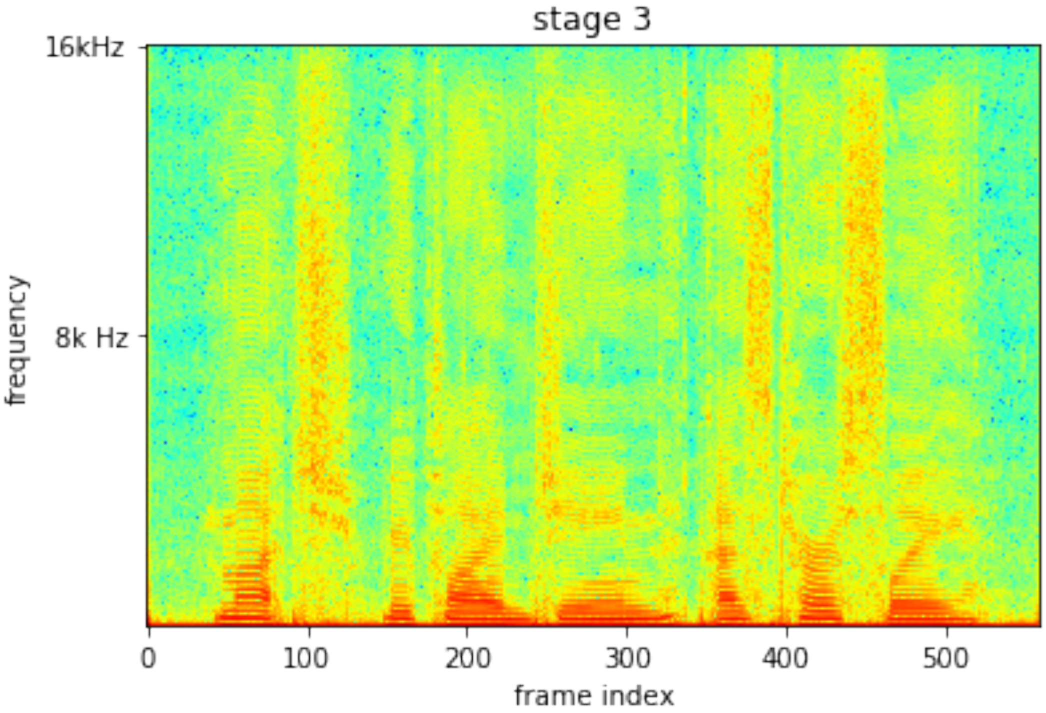


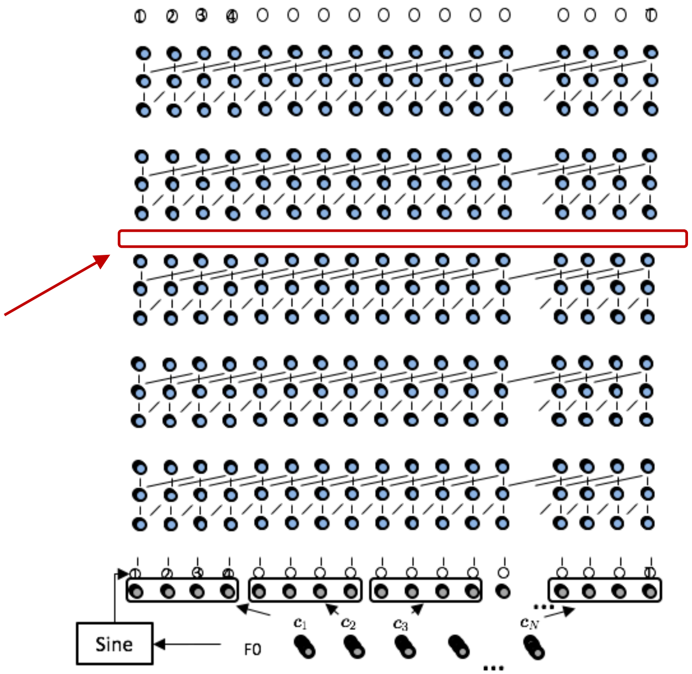


PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Examples

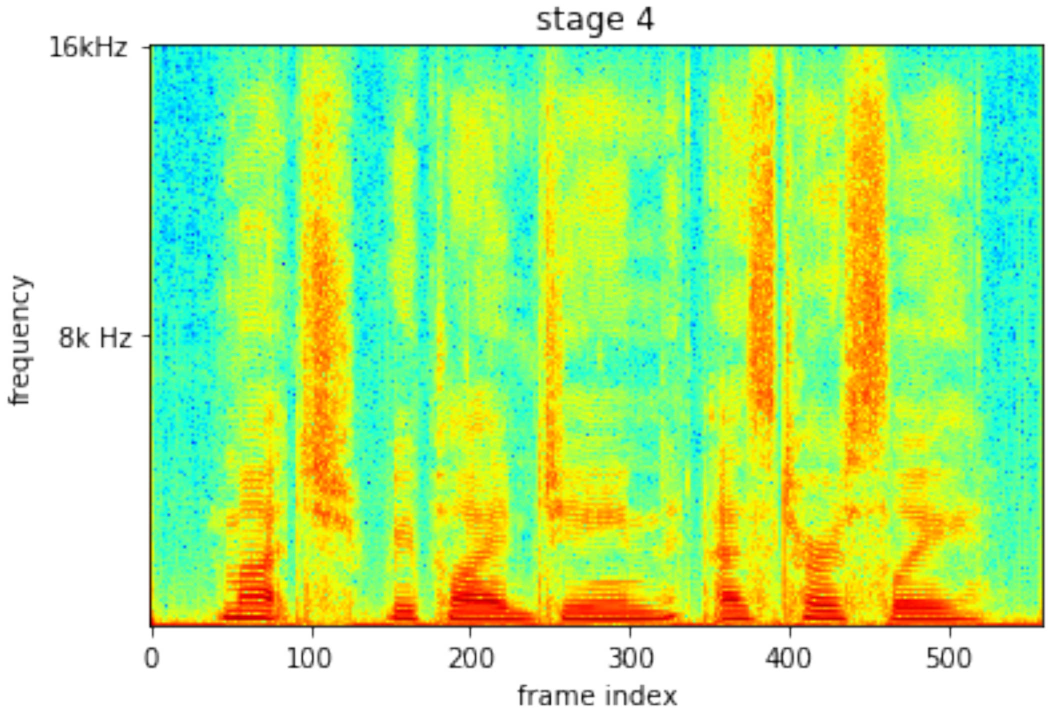


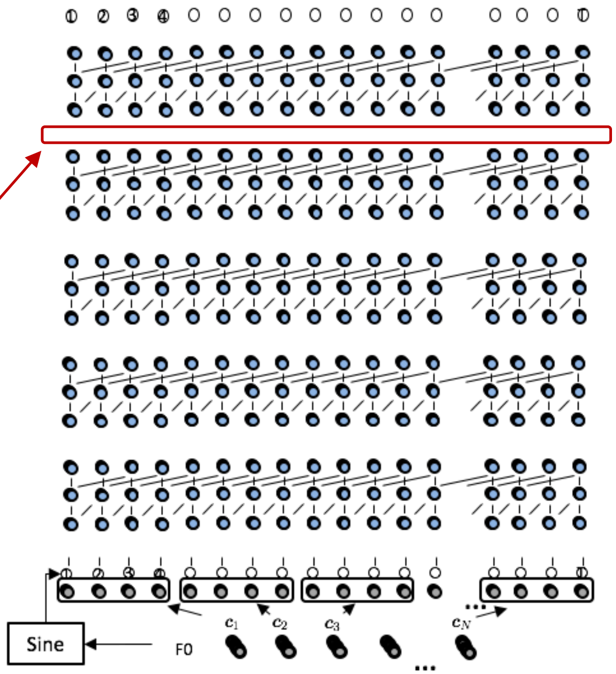


PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

- Examples

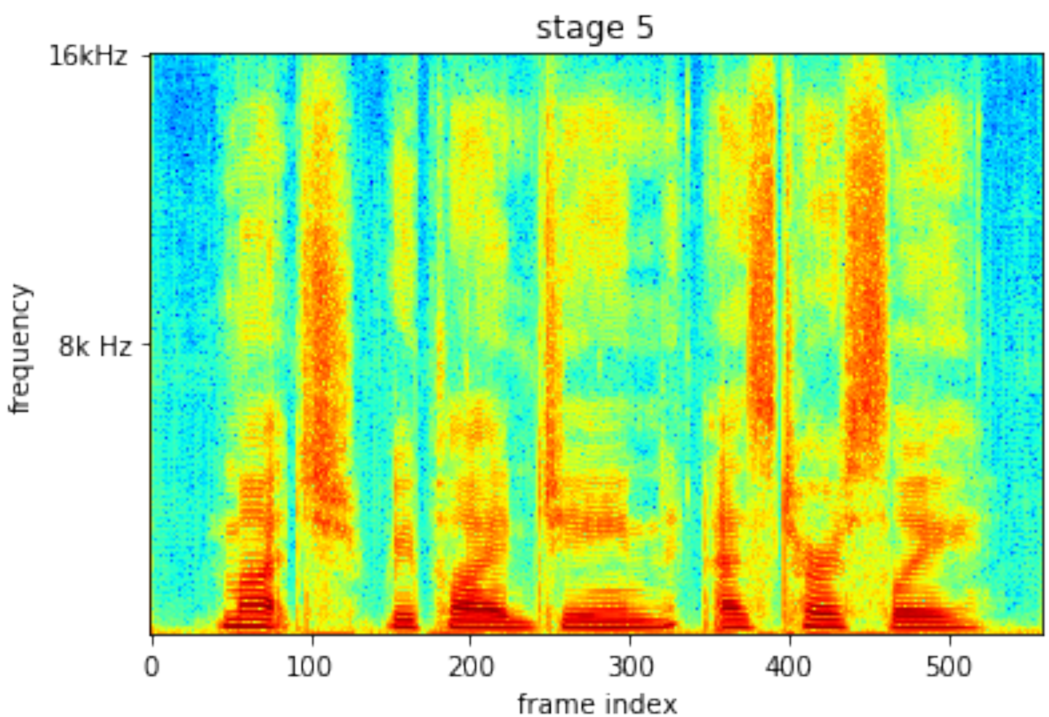


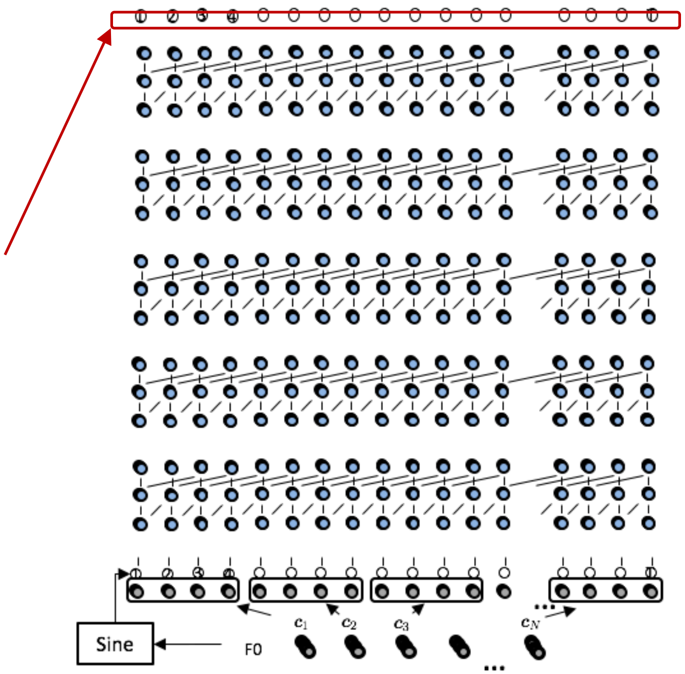


Part III: STFT-based training criterion

Neural source-filter model

- Examples



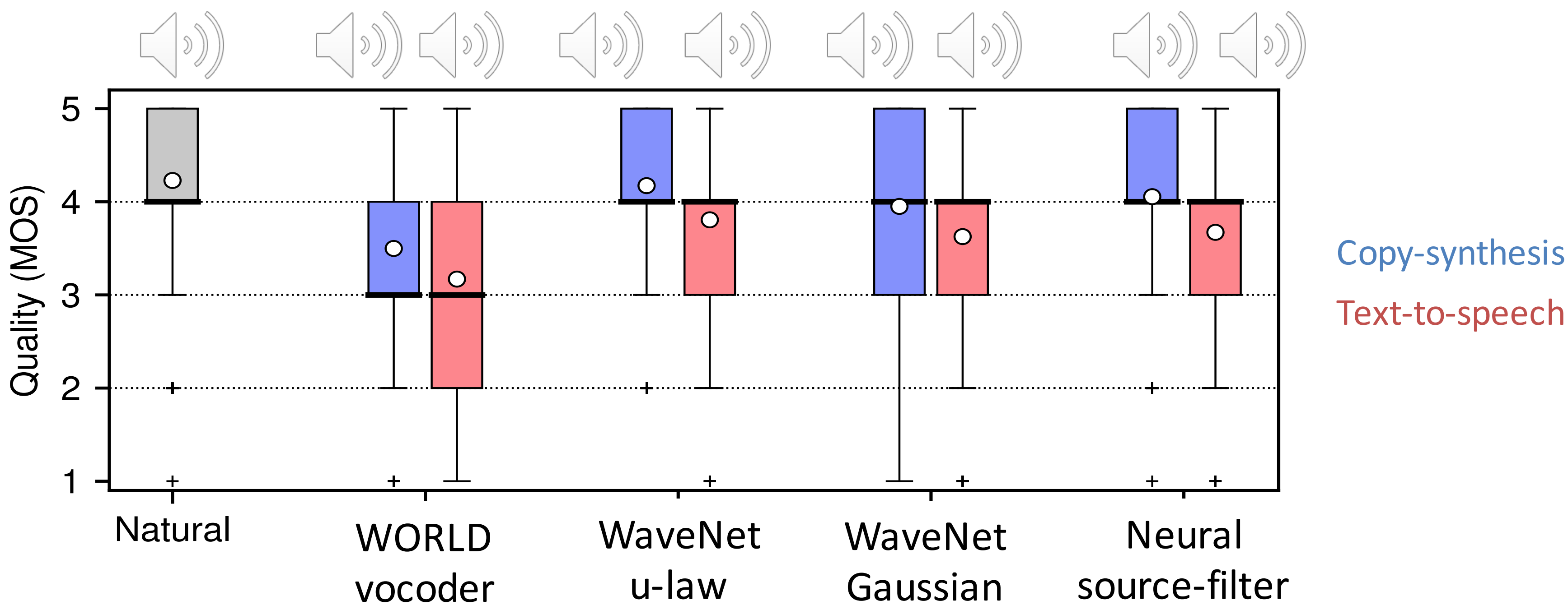


PART III: STFT-BASED TRAINING CRITERION

Neural source-filter model

❑ Results

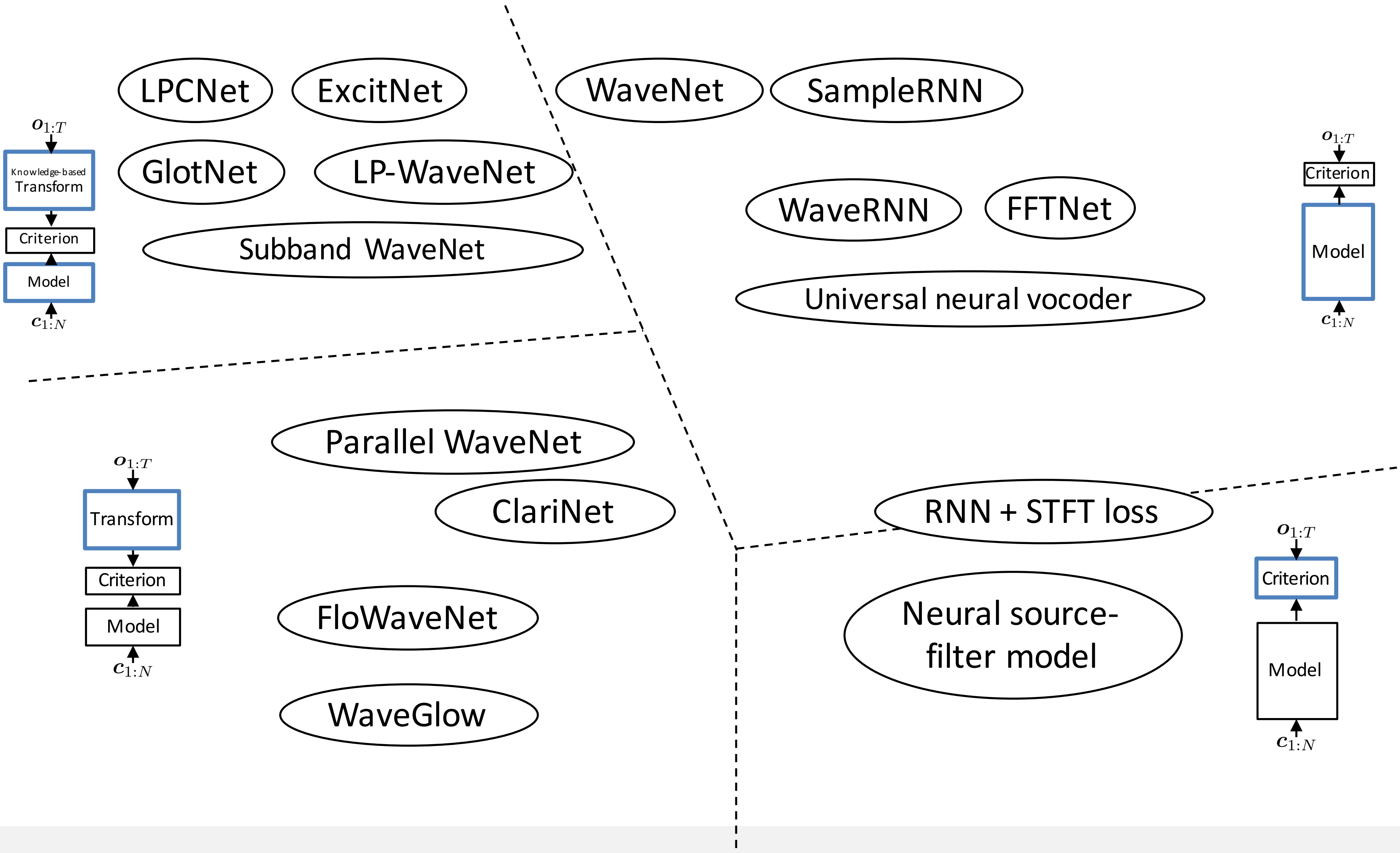
- Faster than WaveNet (at least **100** times)
- Smaller than WaveNet
- Speech quality is similar to WaveNet



https://nii-yamagishilab.github.io/samples-nsf/index.html

CONTENTS

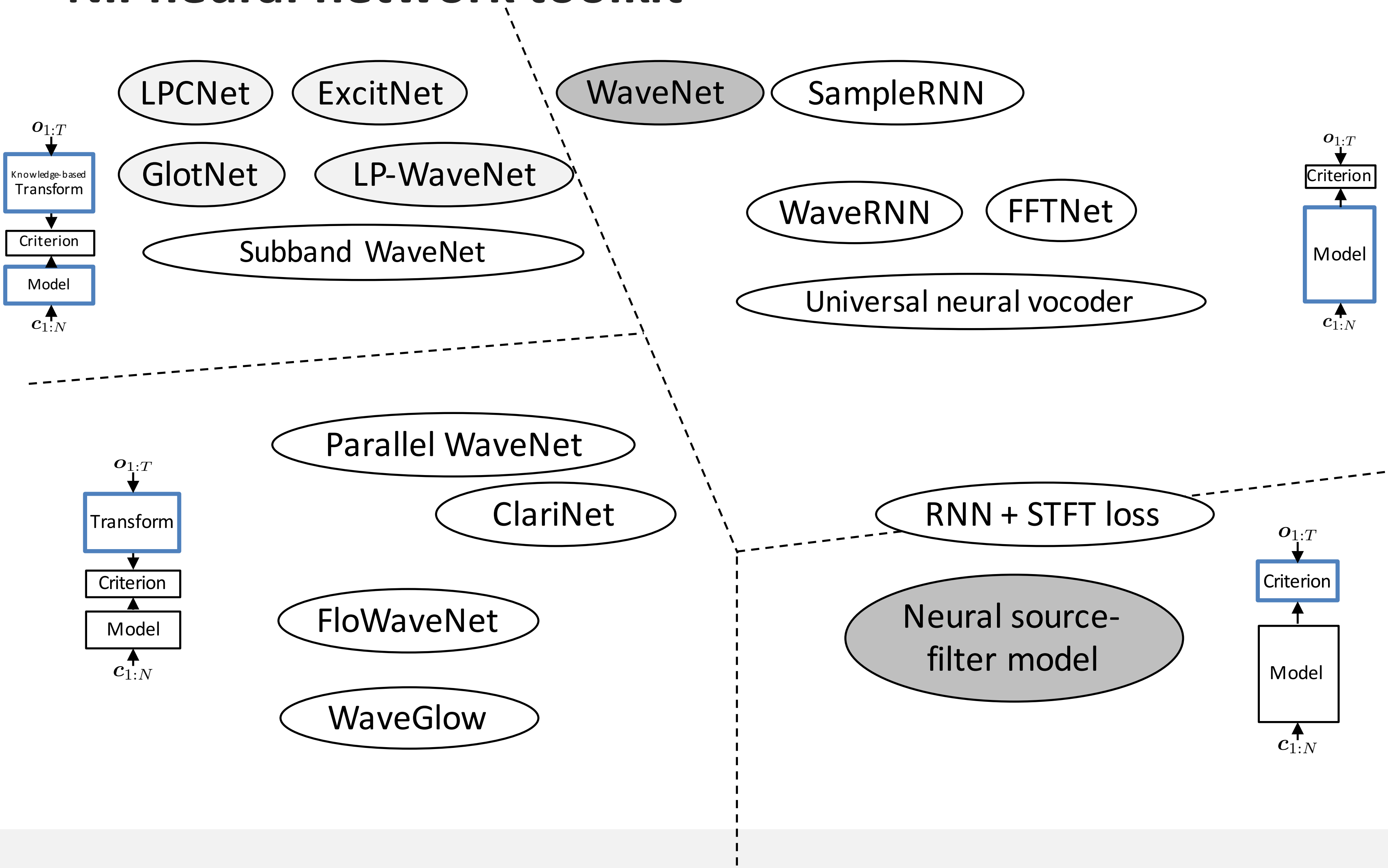
SUMMARY



$o_{1:T}$
Knowledge-based Transform
Criterion
Model
$c_{1:N}$
LPCNet
ExcitNet
GlotNet
LP-WaveNet
Subband WaveNet
WaveNet
SampleRNN
WaveRNN
FFTNet
Universal neural vocoder
$o_{1:T}$
Criterion
Model
$c_{1:N}$
$o_{1:T}$
Transform
Criterion
Model
$c_{1:N}$
Parallel WaveNet
ClariNet
FloWaveNet
WaveGlow
RNN + STFT loss
Neural source-filter model
$o_{1:T}$
Criterion
Model
$c_{1:N}$

SOFTWARE

NII neural network toolkit



SOFTWARE

NII neural network toolkit

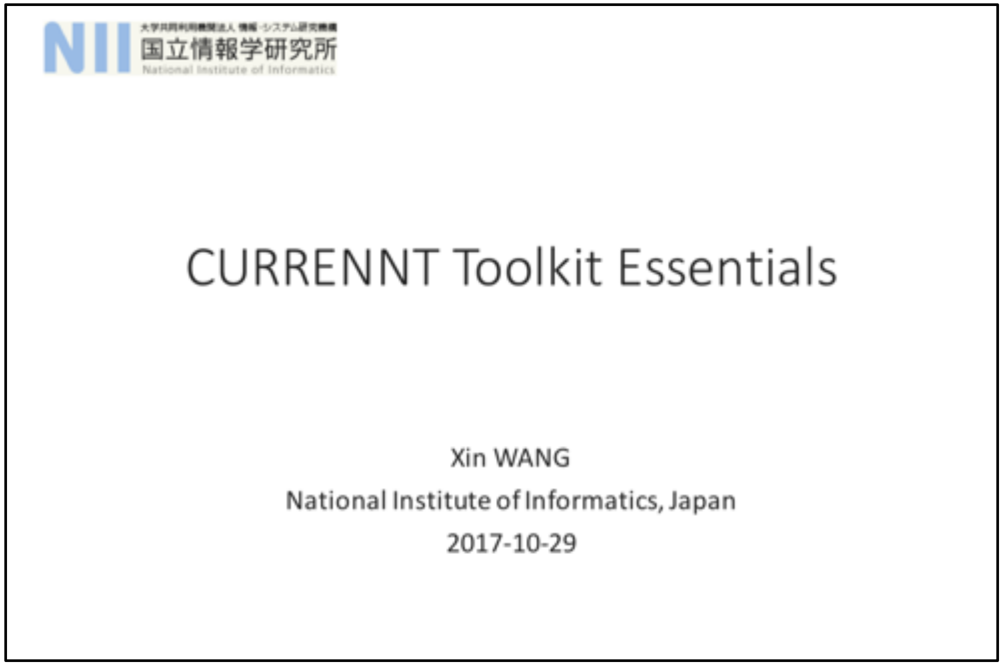
- Neural waveform models
  1. Toolkit cores: https://github.com/nii-yamagishilab/project-CURRENNT-public.git
  2. Toolkit scripts: https://github.com/nii-yamagishilab/project-CURRENNT-scripts

SOFTWARE

NII neural network toolkit

❑ Useful slides

- http://tonywangx.github.io/pdfs/CURRENNT_TUTORIAL.pdf



- Other related slides http://tonywangx.github.io/slides.html

REFERENCE

WaveNet: A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

SampleRNN: S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837, 2016.

WaveRNN: N. Kalchbrenner, E. Elsen, K. Simonyan, et.al. Efficient neural audio synthesis. In J. Dy and A. Krause, editors, Proc. ICML, volume 80 of Proceedings of Machine Learning Research, pages 2410–2419, 10–15 Jul 2018.

FFTNet: Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu. FFTNet: A real-time speaker-dependent neural vocoder. In Proc. ICASSP, pages 2251–2255. IEEE, 2018.

Universal vocoder: J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, and R. Barra-Chicote. Robust universal neural vocoding. arXiv preprint arXiv:1811.06292, 2018.

Subband WaveNet: T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai. An investigation of subband wavenet vocoder covering entire audible frequency range with limited acoustic features. In Proc. ICASSP, pages 5654–5658. 2018.

Parallel WaveNet: A. van den Oord, Y. Li, I. Babuschkin, et. al.. Parallel WaveNet: Fast high-fidelity speech synthesis. In Proc. ICML, pages 3918–3926, 2018.

ClariNet: W. Ping, K. Peng, and J. Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281, 2018.

FlowWaveNet: S. Kim, S.-g. Lee, J. Song, and S. Yoon. Flowavenet: A generative flow for raw audio. arXiv preprint arXiv:1811.02155, 2018.

WaveGlow: R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. arXiv preprint arXiv:1811.00002, 2018.

RNN+STFT: S. Takaki, T. Nakashika, X. Wang, and J. Yamagishi. STFT spectral loss for training a neural speech waveform model. In Proc. ICASSP (submitted), 2018.

NSF: X. Wang, S. Takaki, and J. Yamagishi. Neural source-filter-based waveform model for statistical para- metric speech synthesis. arXiv preprint arXiv:1810.11946, 2018.

LP-WavNet: M.-J. Hwang, F. Soong, F. Xie, X. Wang, and H.-G. Kang. Lp-wavenet: Linear prediction-based wavenet speech synthesis. arXiv preprint arXiv:1811.11913, 2018.

GlotNet: L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku. Speaker-independent raw waveform model for glottal excitation. arXiv preprint arXiv:1804.09593, 2018.

ExcitNet: E. Song, K. Byun, and H.-G. Kang. Excitnet vocoder: A neural excitation model for parametric speech synthesis systems. arXiv preprint arXiv:1811.04769, 2018.

LPCNet: J.-M. Valin and J. Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. arXiv preprint arXiv:1810.11846, 2018.

End of Part 1

Codes, scripts, slides

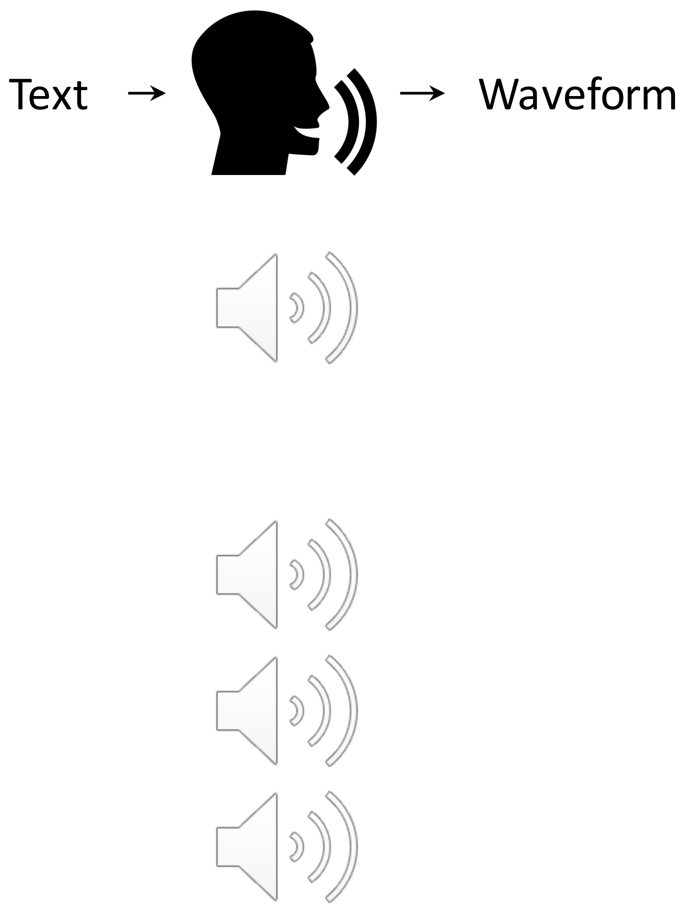
http://nii-yamagishilab.github.io

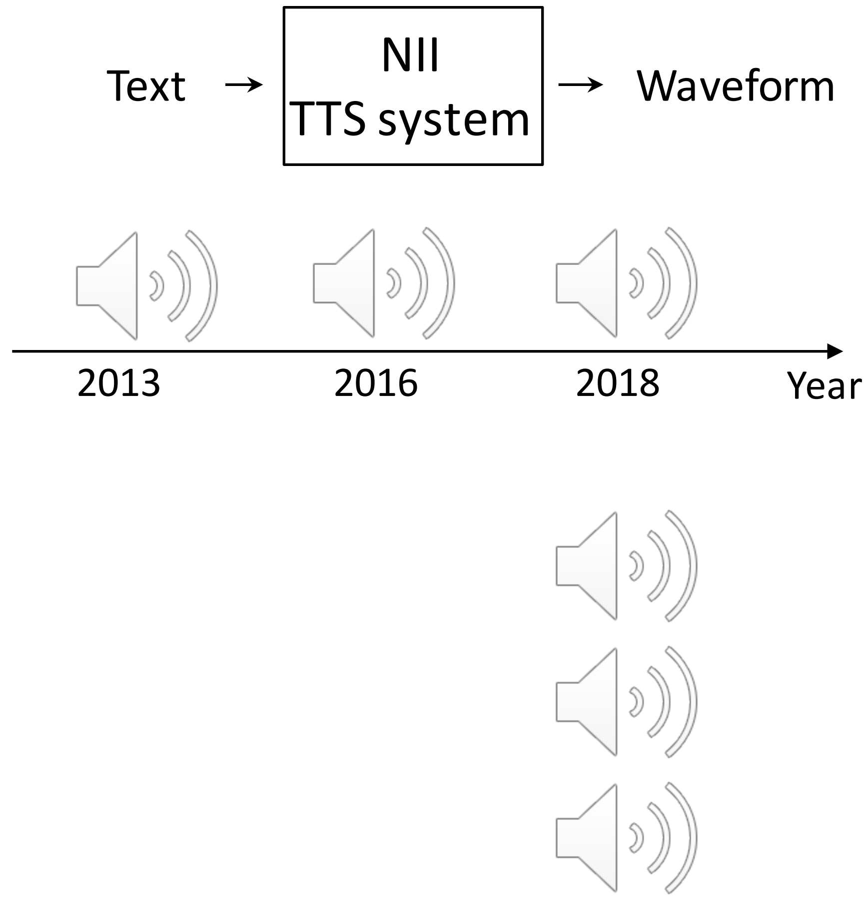
This work was partially supported by JST CREST Grant Number JPMJCR18A6, Japan and by MEXT KAKENHI Grant Numbers (16H06302, 16K16096, 17H04687, 18H04120, 18H04112, 18KT0051), Japan.

INTRODUCTION

Text-to-speech (TTS)

- Speech samples from NII's TTS system





PART I: AUTOREGRESSIVE MODELS

SampleRNN

- Idea
  - Hierarchical / multi-resolution dependency

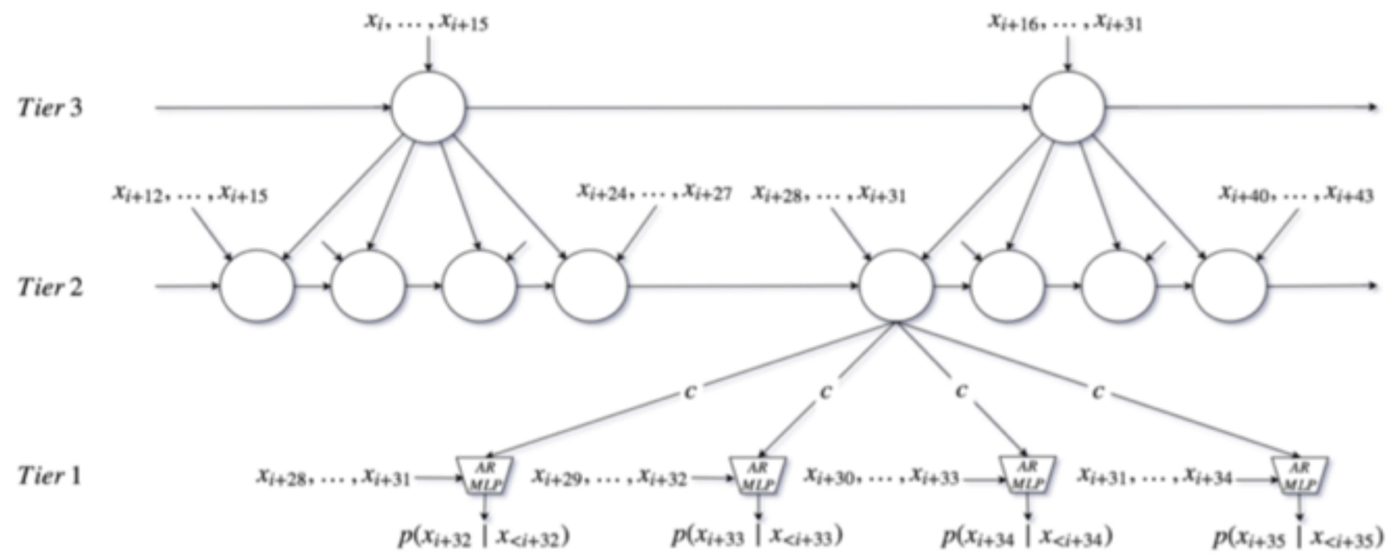


Figure 1: Snapshot of the unrolled model at timestep i with $K = 3$ tiers. As a simplification only one RNN and up-sampling ratio $r = 4$ is used for all tiers.

S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. Samplernn: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837, 2016.

PART I: AUTOREGRESSIVE MODELS

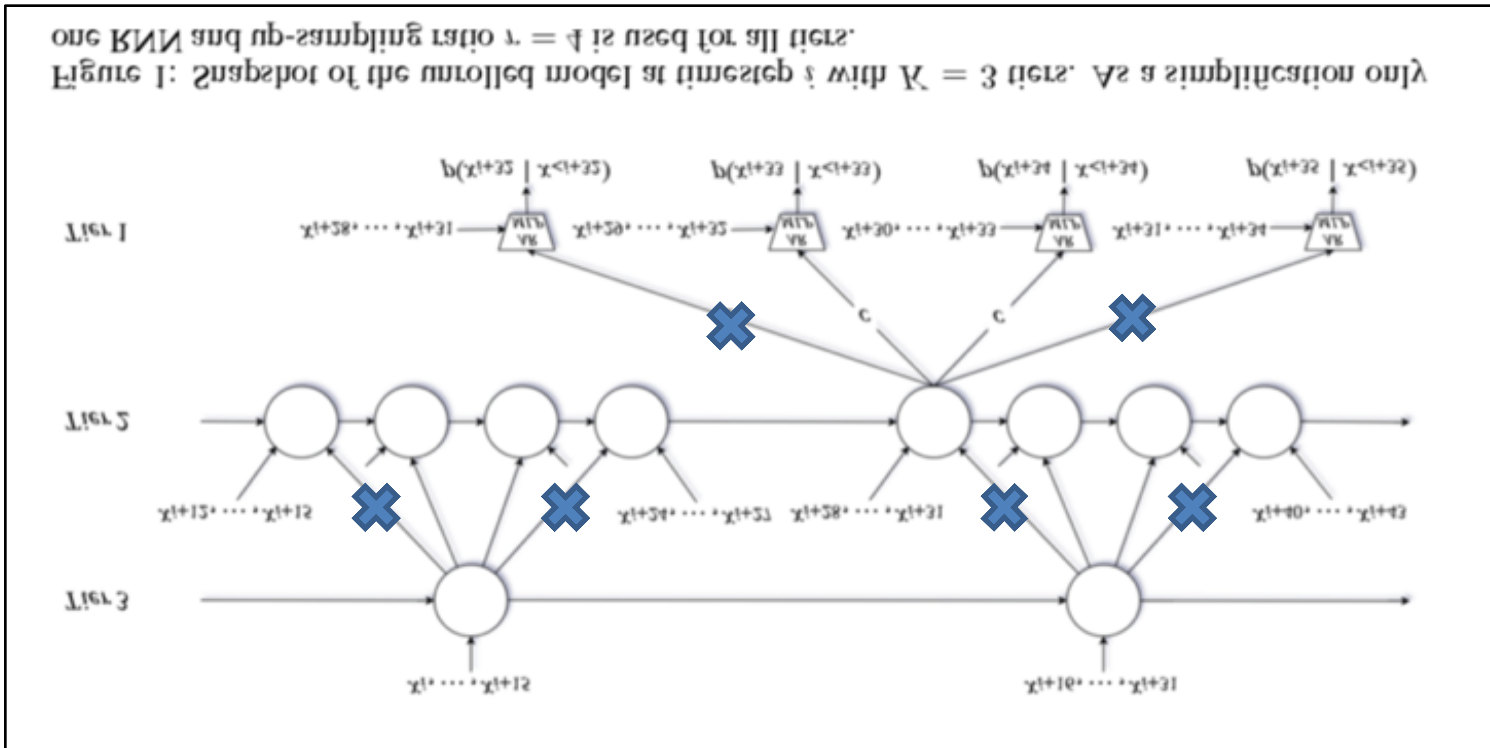
SampleRNN

- Example network structure

Tie 1

Tie 2

Tie 3

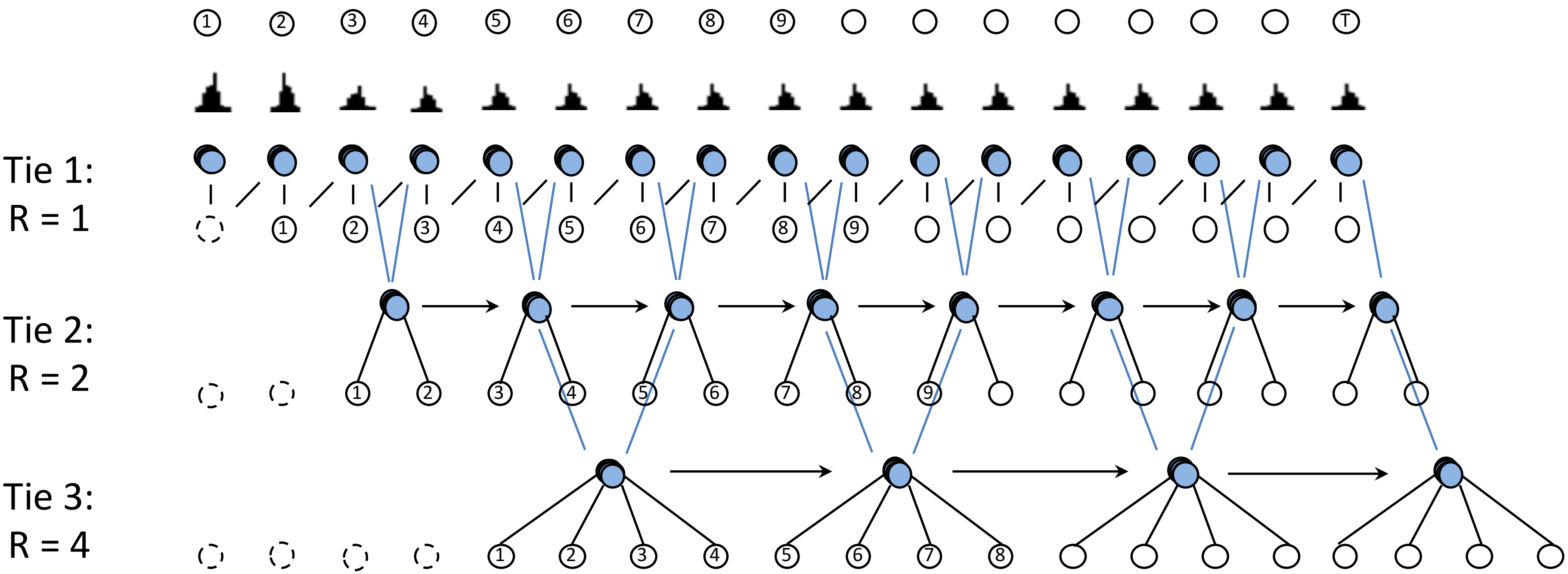


- R: time resolution increased by * 2

PART I: AUTOREGRESSIVE MODELS

SampleRNN

- Example network structure
  - Training

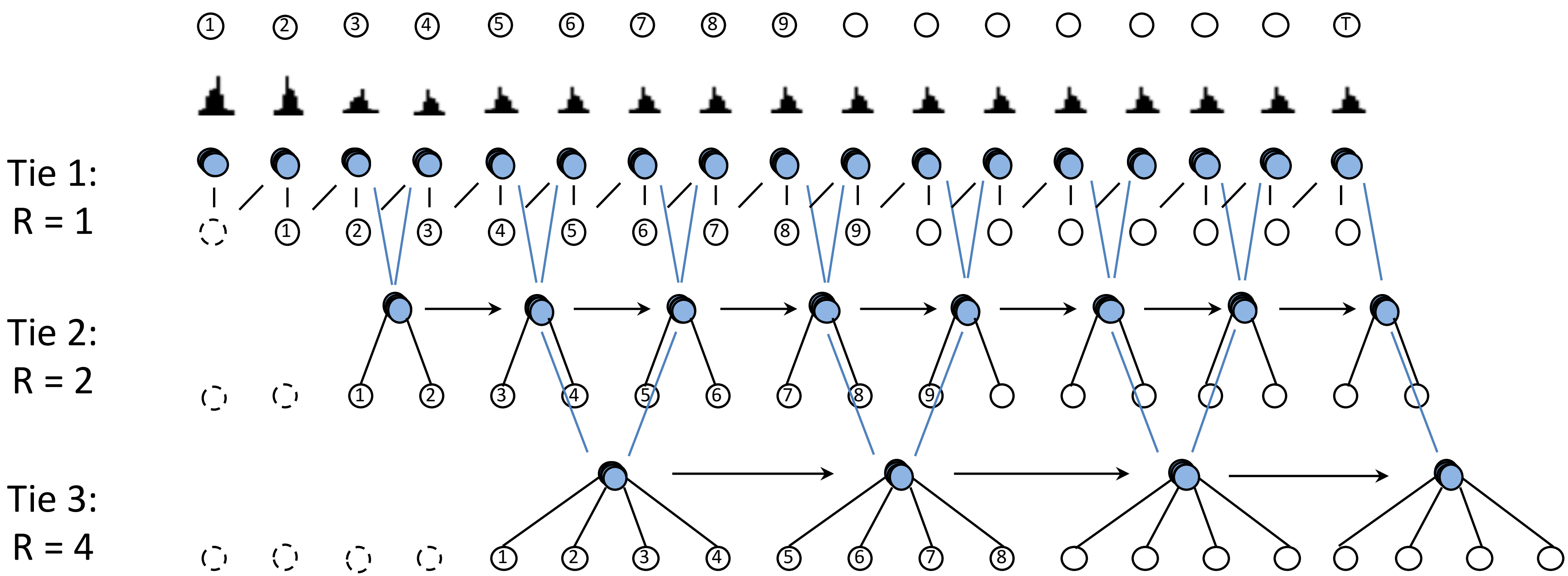


  - R: time resolution increased by * 2

PART I: AUTOREGRESSIVE MODELS

SampleRNN

- Example network structure
  - Generation process

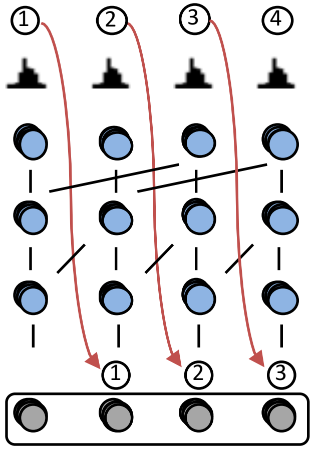


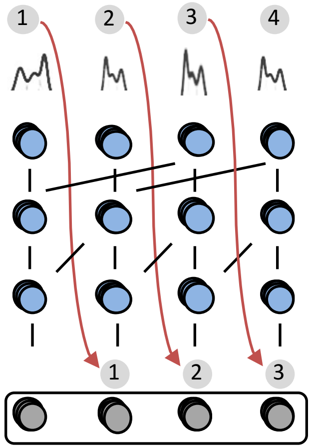
PART I: AUTOREGRESSIVE MODELS

WaveNet

❑ Variants

- *u*-Law discrete waveform ---> continuous-valued waveform
  - Mixture of logistic distribution [1]
  - GMM / Single-Gaussian [2]





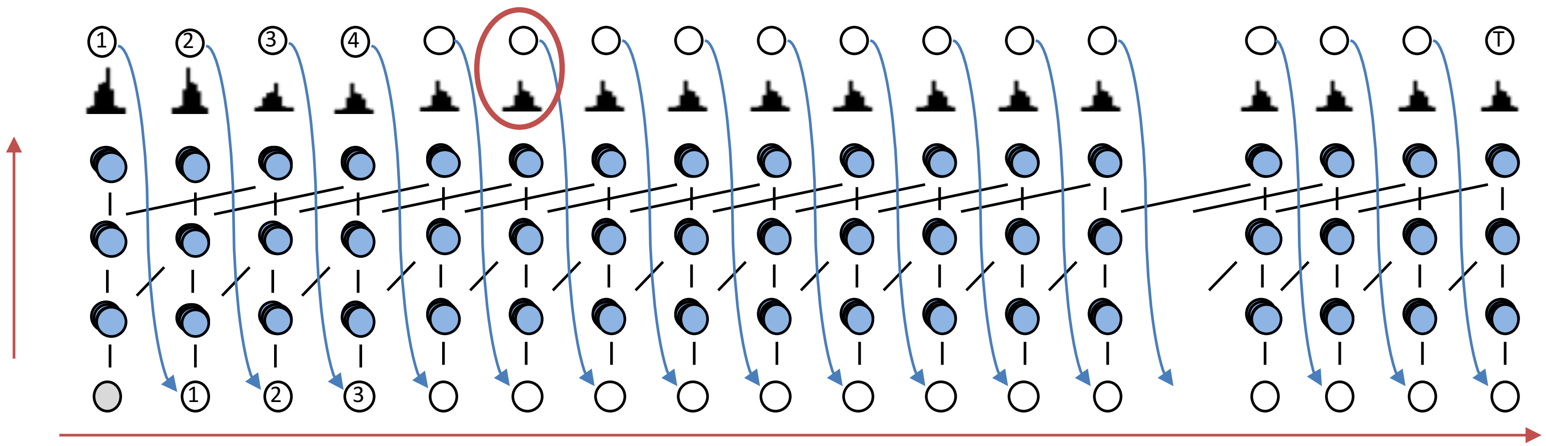
- Quantization noise shaping [3], related noise shaping method [4]

[1] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. arXiv preprint arXiv:1701.05517, 2017.
[2] W. Ping, K. Peng, and J. Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281, 2018.
[3] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(7):1173–1180, 2018.
[4] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai. An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation. In Proc. ICASSP, pages 5664–5668. IEEE, 2018.

PART I: AUTOREGRESSIVE MODELS

WaveRNN

❑ WaveNet is inefficient

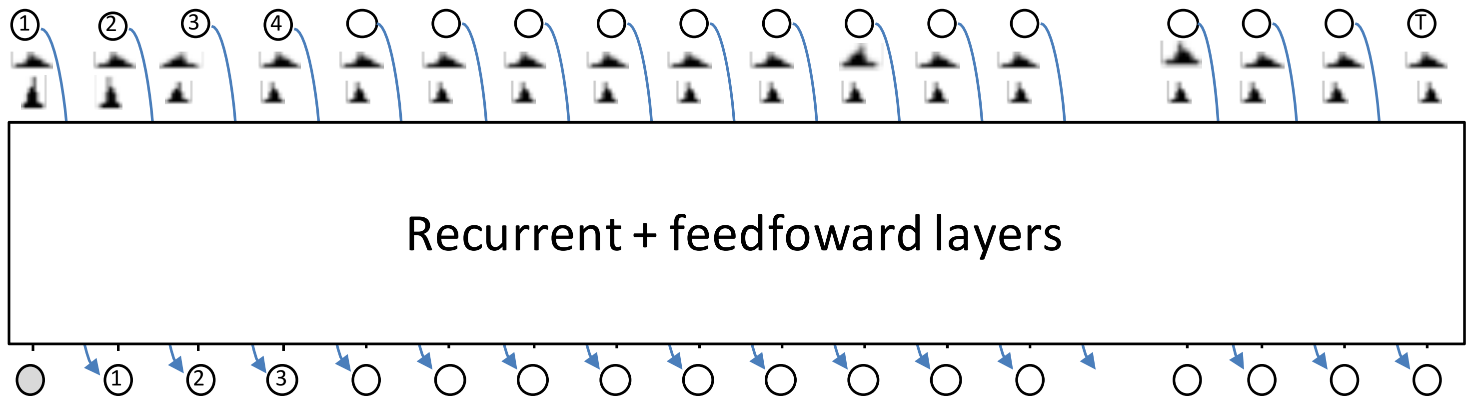


- Computation cost
  1. Impractical for 16 bit PCM (softmax of size 65536)
  2. Very deep network (50 dilated CNN …)
  3. …
- Time latency
  1. Generation time ~ O(waveform_length)

PART I: AUTOREGRESSIVE MODELS

WaveRNN

- WaveRNN strategies



- Computation cost
  1. ~~Impractical for 16 bit PCM (softmax of size 65536)~~ Two-level softmax
  2. ~~Very deep network (50 dilated CNN ...)~~ RNN + Feedforward
  3. ...
- Time latency
  1. ~~Generation time ~ O(waveform_length)~~ Subscale dependency + batch

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. In J. Dy and A. Krause, editors, Proc. ICML, volume 80 of Proceedings of Machine Learning Research, pages 2410–2419, Stock- holmsm̈assan, Stockholm Sweden, 10–15 Jul 2018. PMLR.