ICASSP 2018
Calgary, Canada

Comparing Recent Waveform Generation and Acoustic Modeling Methods for Neural-network-based Speech Synthesis

Xin WANG, Jaime Lorenzo-Trueba, Shinji TAKAKI, Lauri Juvela, Junichi YAMAGISHI

National Institute of Informatics, Japan & Aalto University, Finland

2018-04-17

OVERVIEW

Motivation

- Better modules for the statistical parametric speech synthesis (SPSS) framework?

Method

- Plug and test new acoustic models and waveform generators

Results

- Best combination
  - WaveNet-based vocoder
  - Autoregressive (AR) acoustic models
- Quality: as good as vocoded speech (at 16kHz)

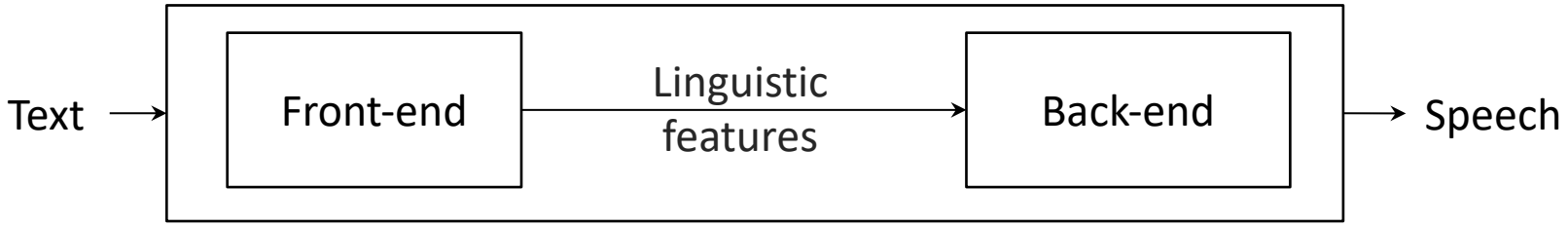
CONTENTS

- Introduction
- Models and modules
- Experiments
- Summary

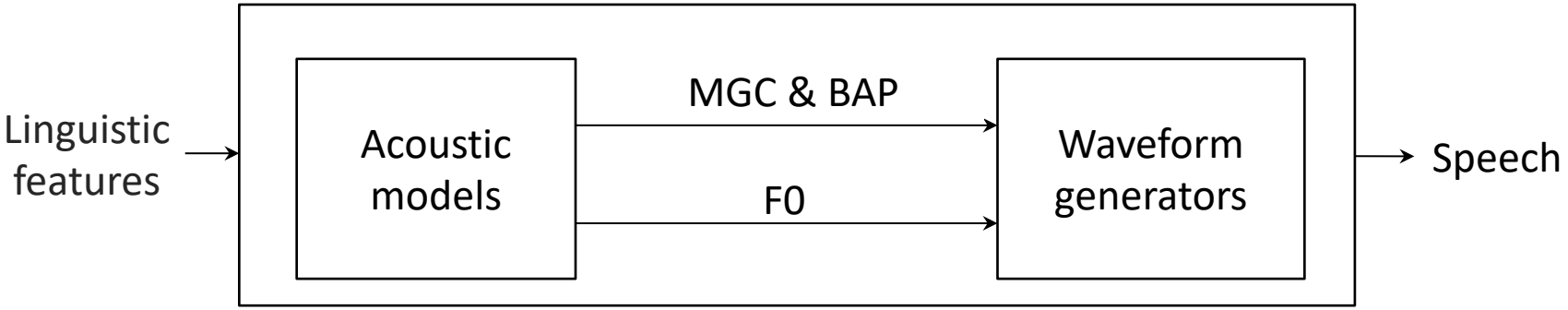
INTRODUCTION

Background

- Conventional TTS pipeline [1]



- SPSS back-end [2,3]



- MGC: Mel-generalized cepstral coefficients [4]
- BAP: band-aperiodicity

[1] T. Dutoit. An Introduction to Text-to-speech Synthesis. Kluwer Academic Publishers, Norwell, MA, USA, 1997.
[2] Tokuda, K., et al., (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*, *101*(5), 1234–1252.
[3] Zen, H., et al. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*, 1039–1064.
[4] Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis a unified approach. In Proc. ICSLP, pages 1043–1046.

INTRODUCTION

Topic of this work

- Better modules for SPSS back-end?

Recurrent neural networks (RNNs)

Autoregressive (AR) models

General adversarial network (GAN)

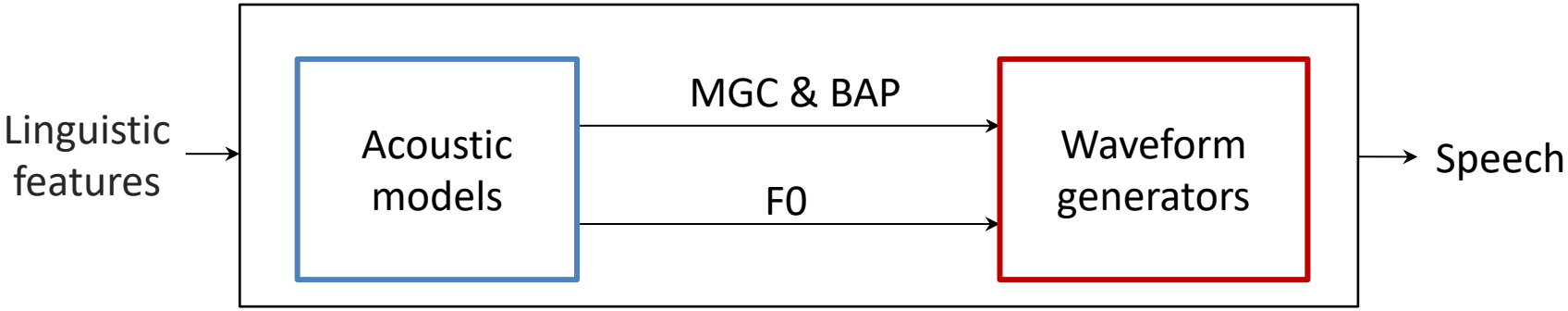
…

WORLD vocoder

+ Phrase recovery

WaveNet-based vocoder

…

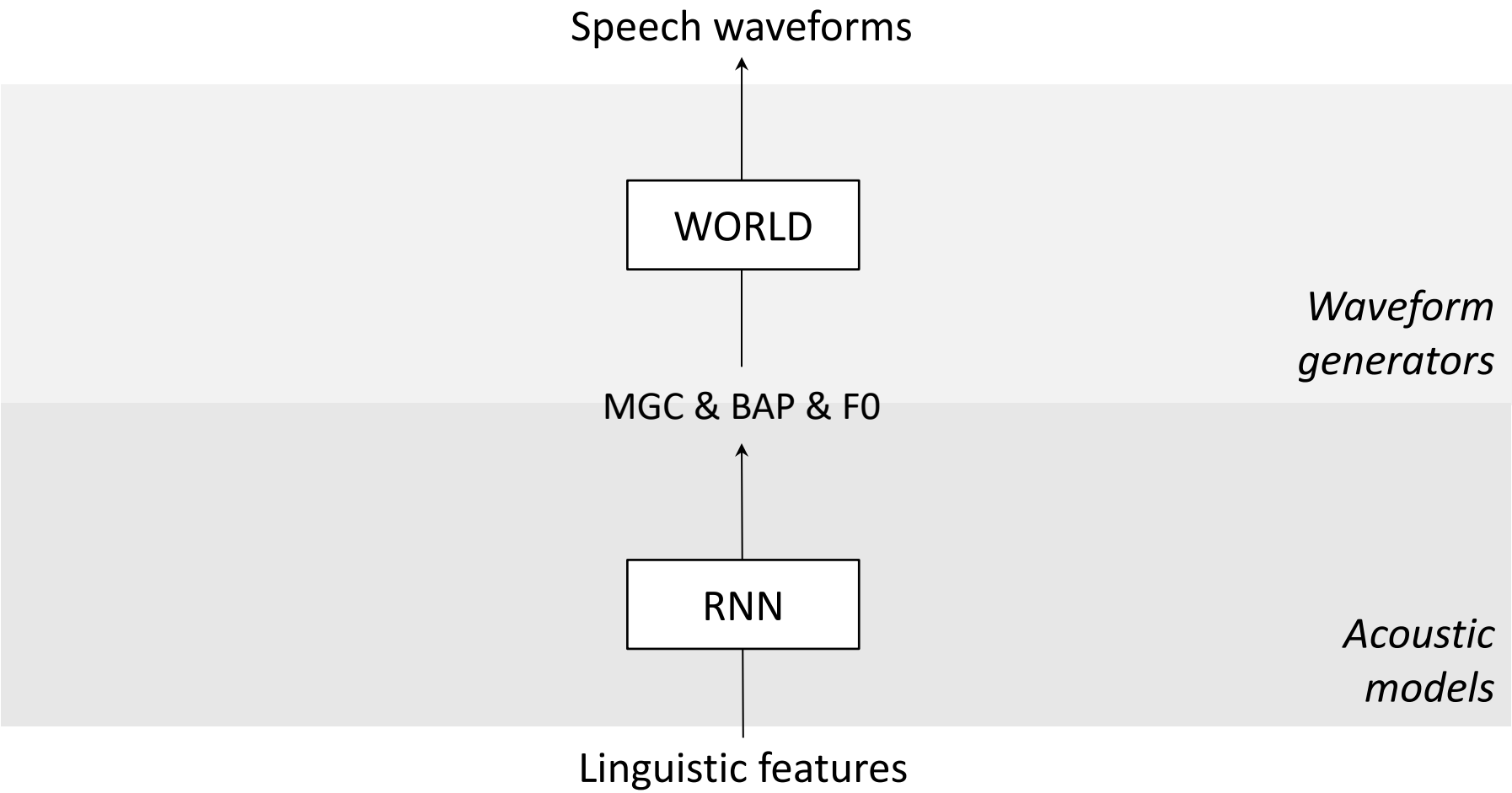


- MGC: Mel-generalized cepstral coefficients
- BAP: band-aperiodicity

CONTENTS

- Introduction
- Models and modules
- Experiments
- Summary

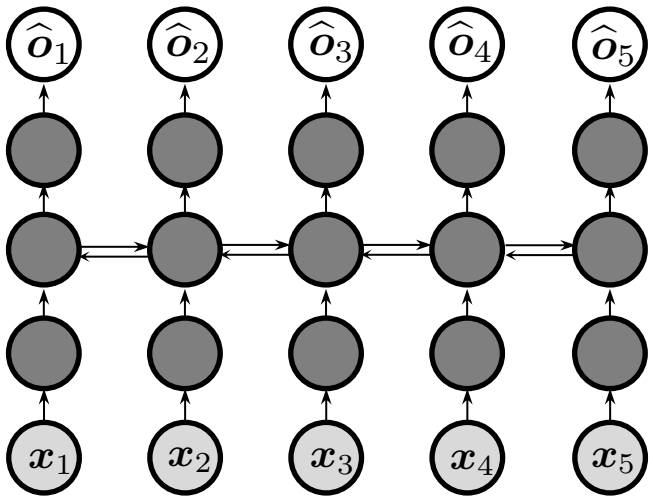
Models & Methods



Models & Methods

Acoustic models

- Baseline RNN

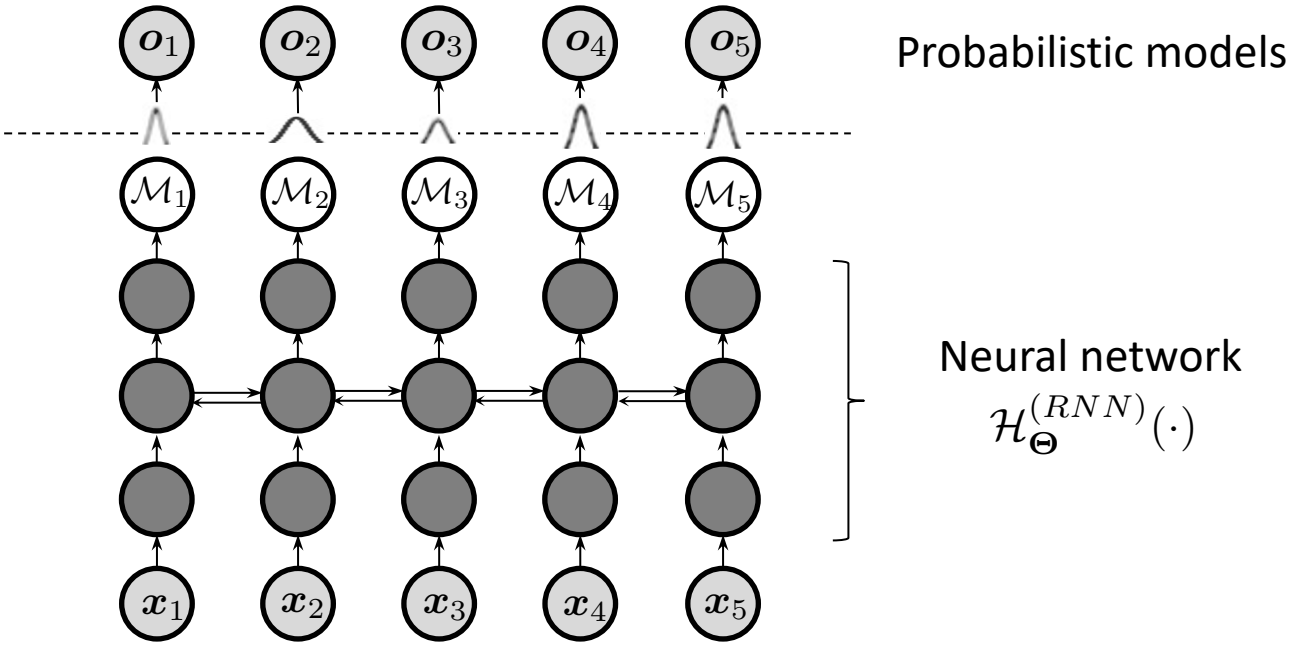


- Sequence of linguistic features $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots\}$
- Sequence of generated acoustic features $\{\widehat{\boldsymbol{o}}_1, \cdots, \widehat{\boldsymbol{o}}_t, \cdots\}$

Models & Methods

Acoustic models

- Baseline RNN



$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\boldsymbol{x}_{1:T};\boldsymbol{\Theta}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t;\boldsymbol{\mu}_t,\boldsymbol{I})$$

$$\boldsymbol{\mathcal{M}}_t = \{\boldsymbol{\mu}_t\}, \quad \text{where} \quad \boldsymbol{\mu}_t = \mathcal{H}_{\boldsymbol{\Theta}}^{(RNN)}(\boldsymbol{x}_{1:T}, t)$$

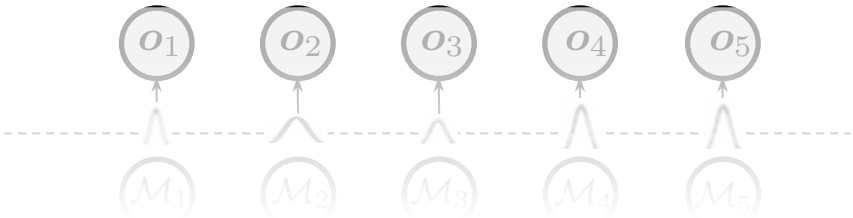
$$\widehat{\boldsymbol{o}}_t = \boldsymbol{\mu}_t$$

[5] C. M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995.

Models & Methods

Acoustic models

- Baseline RNN



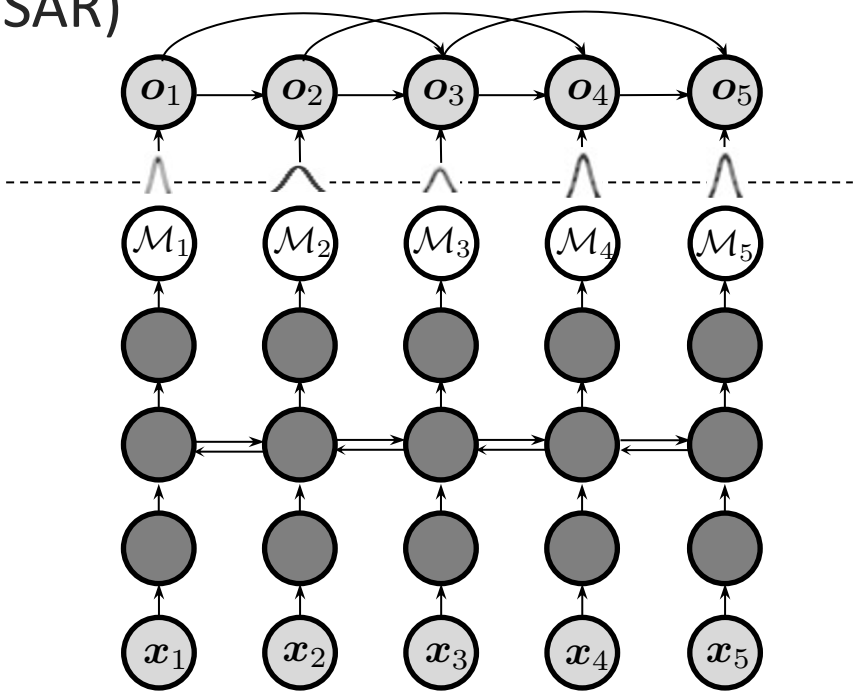
- Limitations
  1. Conditional independence ⇨ AR models
  2. Maximum-likelihood training ⇨ GAN

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\boldsymbol{x}_{1:T}; \boldsymbol{\Theta}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t, \boldsymbol{I})$$

Models & Methods

Acoustic models

- Shallow AR (SAR)



K = 2

$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T};\boldsymbol{\Theta},\boldsymbol{\Psi}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\underline{\boldsymbol{o}_{t-K:t-1}},\boldsymbol{x}_{1:T};\boldsymbol{\Theta}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t;\boldsymbol{\mu}_t + \underline{f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{t-K:t-1})},\boldsymbol{I})$$

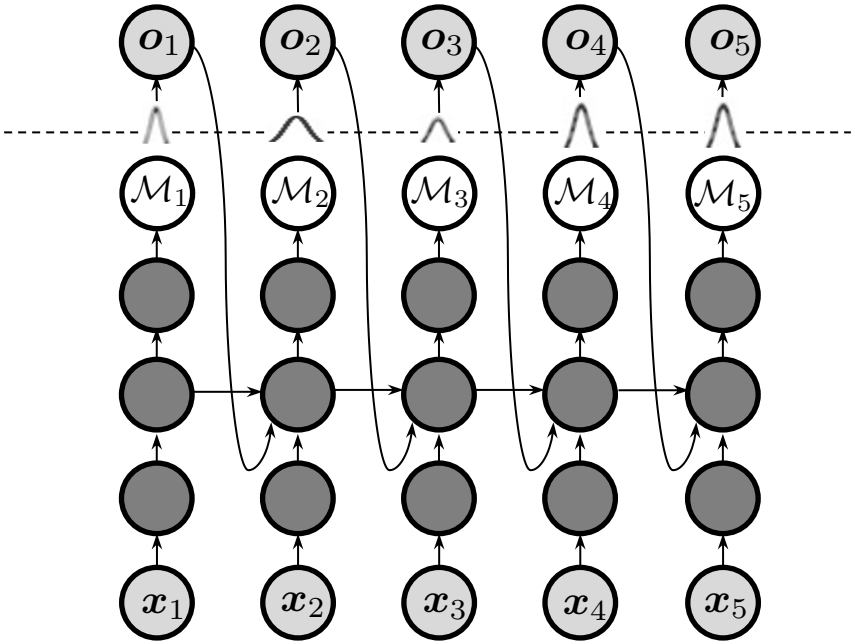
- Alternative interpretation: trainable filter + RNN

[6] X. Wang, S. Takaki, and J. Yamagishi. An autoregressive recurrent mixture density network for parametric speech synthesis. In Proc. ICASSP, pages 4895–4899, 2017.

Models & Methods

Acoustic models

- Deep AR (DAR)



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

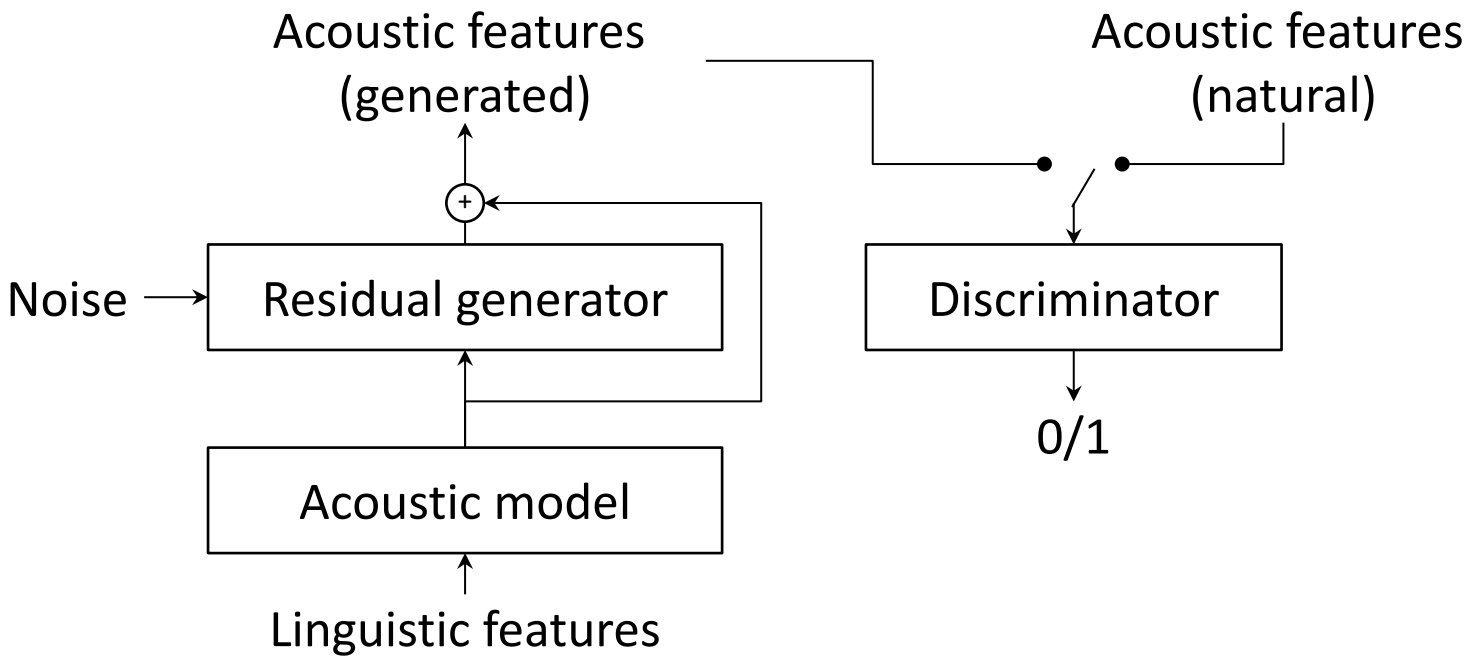
- Only for (quantized) F0 modeling

[7] X. Wang, S. Takaki, and J. Yamagishi. Autoregressive neural F0 model for statistical parametric speech synthesis. IEEE Transactions on Audio, Speech and Language Processing. (Accepted)

Models & Methods

Acoustic models

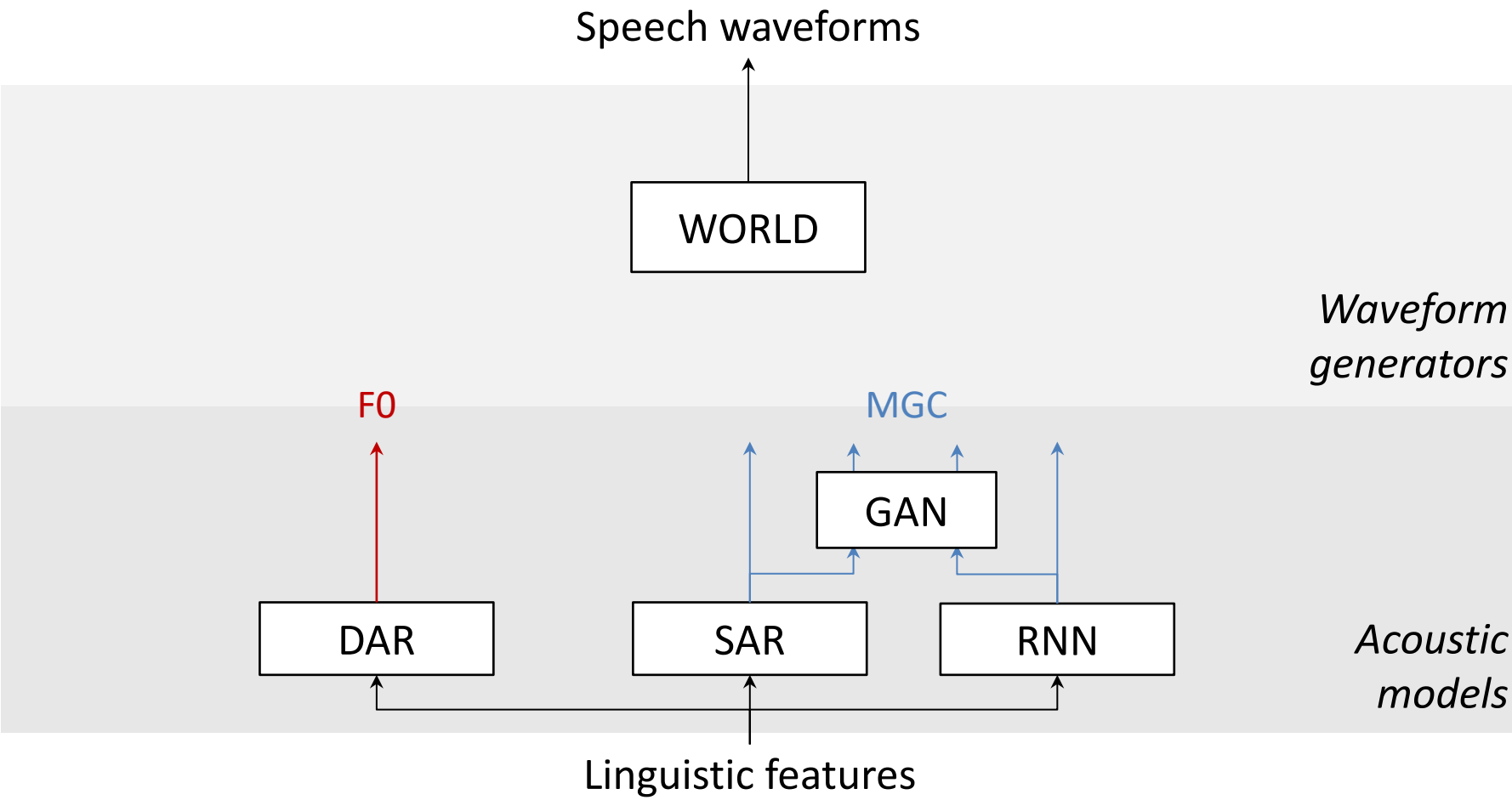
- GAN
  - GAN-based post-filter [8]



[8] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In Proc. ICASSP, pages 4910–4914, 2017.

Models & Methods

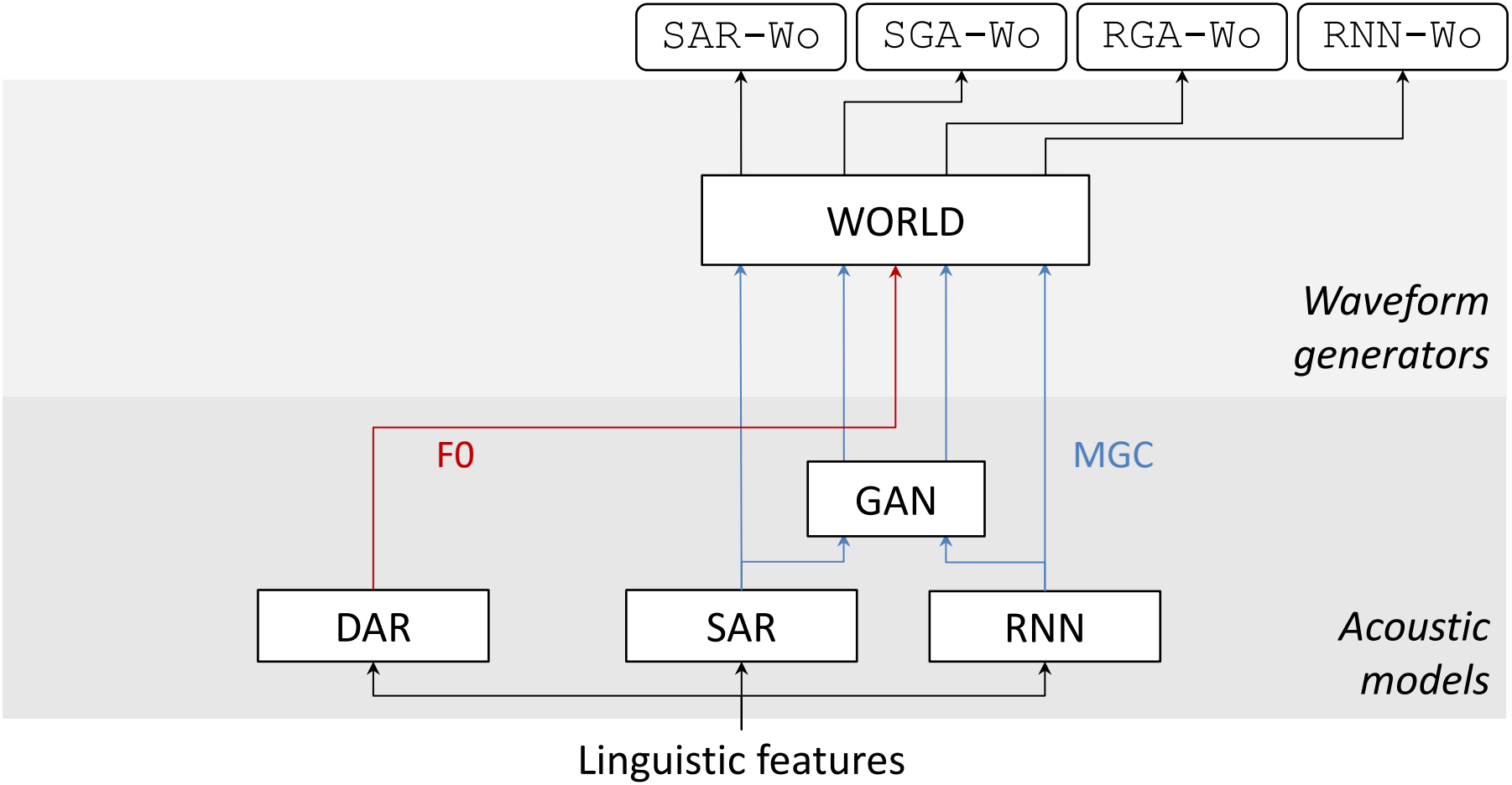
Acoustic models



- BAP is not shown

Models & Methods

Acoustic models

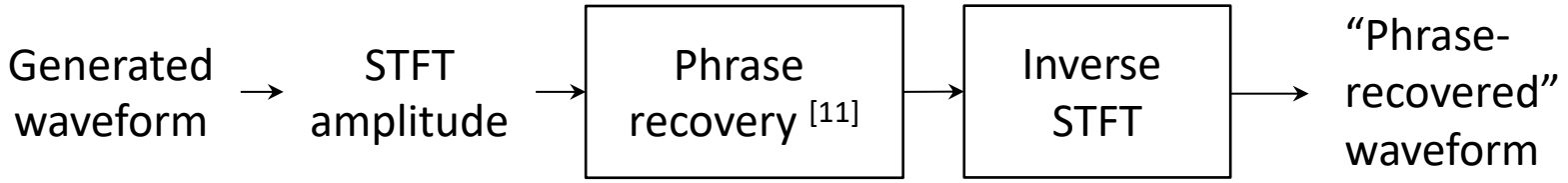


- BAP is not shown

Models & Methods

Waveform generators

- Deterministic approaches
  - WORLD [9]
    - Binary voicing decision
    - Minimum phase
  - A log domain pulse model (PML) [10]
    - Source-filter model, additive in log-domain
    - Binary noisy mask
  - WORLD + phrase recovery

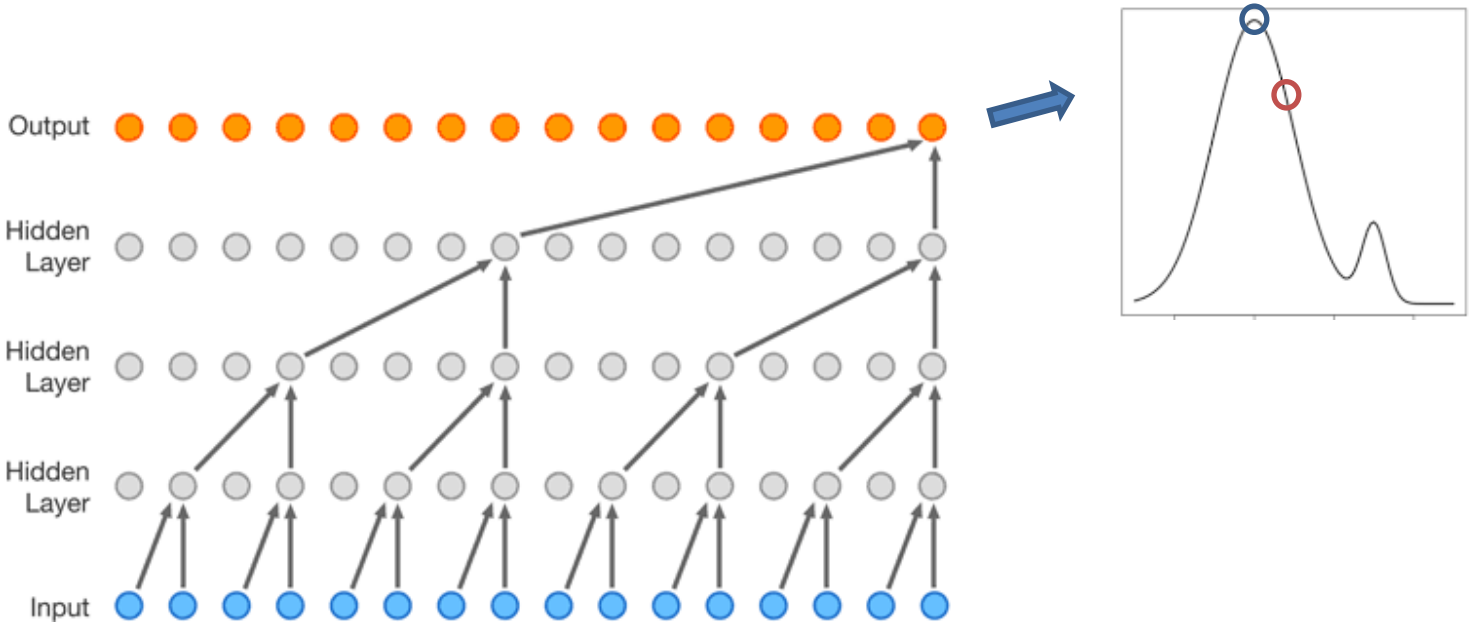


[9] M. Morise, et al. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. on Information and Systems, 99(7):1877–1884, 2016.
[10] G. Degottex, et al. A log domain pulse model for parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017.
[11] D. Griffin and J. Lim. Signal estimation from modified short-time Fourier transform. IEEE Trans. ASSP, 32(2):236–243, 1984.

Models & Methods

Waveform generators

- WaveNet-based vocoder [12, 13]



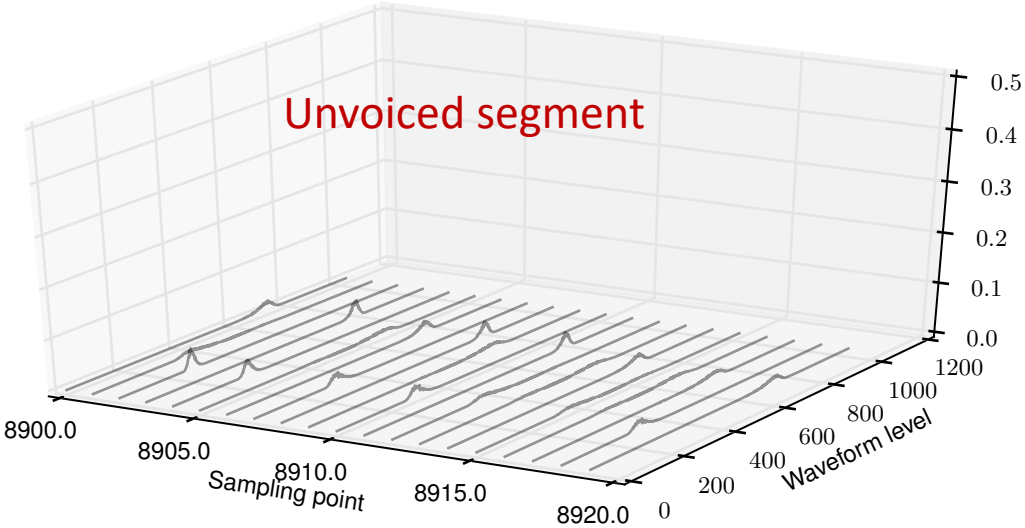
- How to generate (search) a waveform:
  1. Exploration: sampling
  2. Exploitation: picking one-best

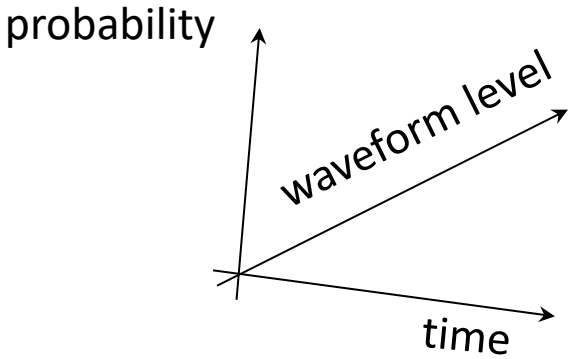
[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.
[13] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda. Speaker-dependent WaveNet vocoder. In Proc. Interspeech, pages 1118–1122, 2017.

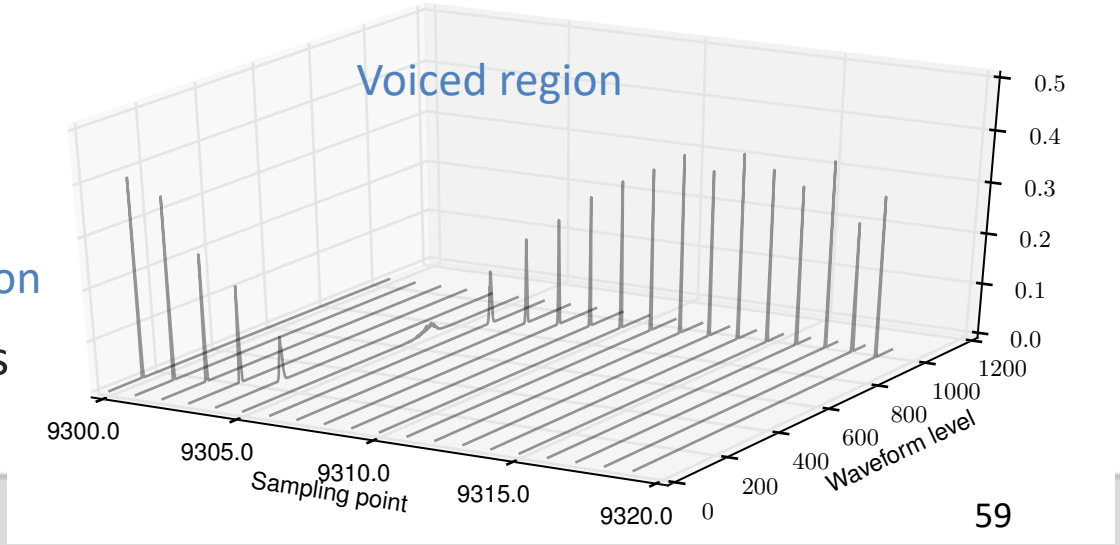
Models & Methods

Waveform generators

- WaveNet-based vocoder



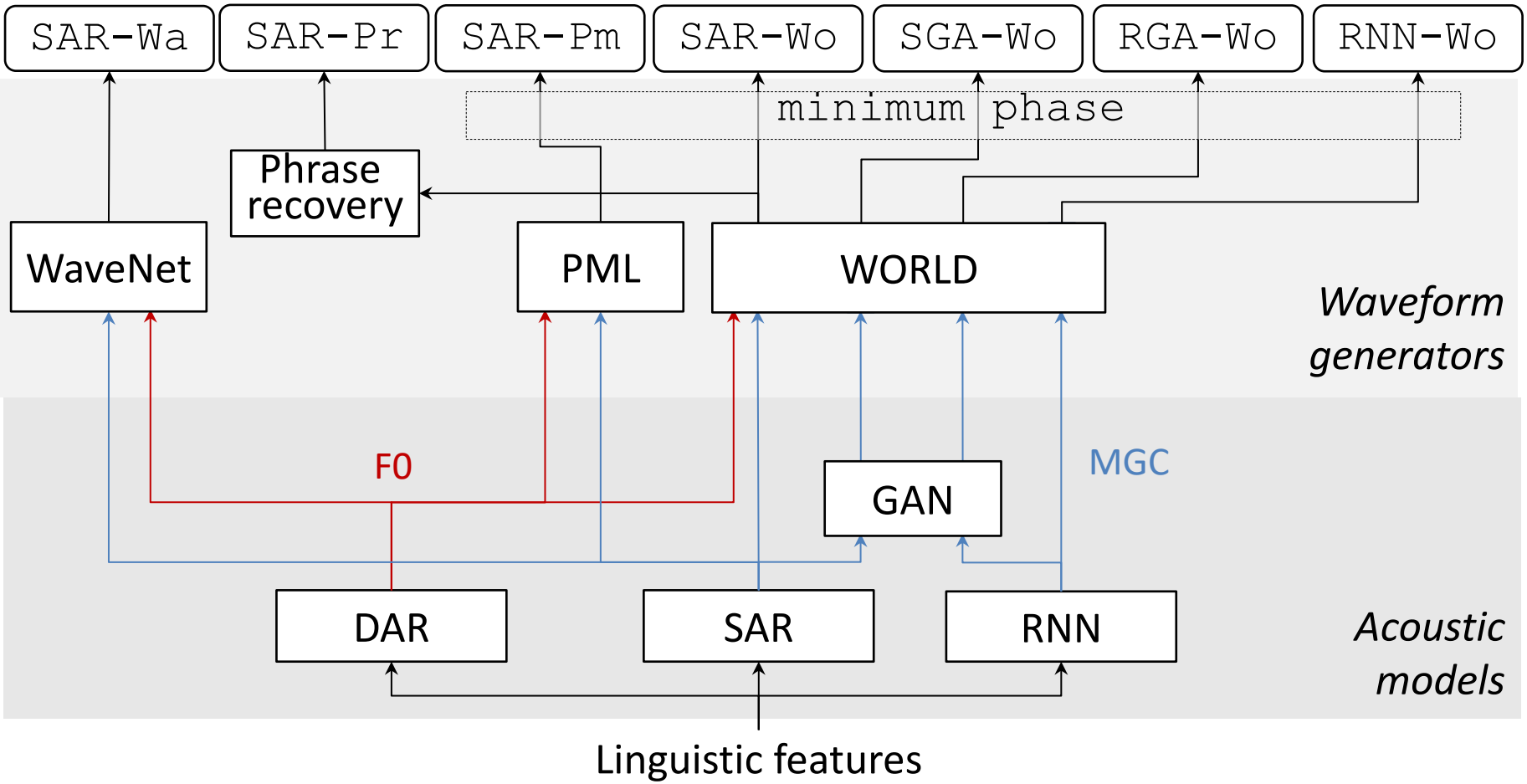




- Sampling in unvoiced region
- Picking one-best in voiced region
- Less distortion of harmonics
  ☛ appendix & paper

Models & Methods

Acoustic models



CONTENTS

- Introduction
- Models and modules
- Experiments
- Summary

EXPERIMENTS

Configuration

- Data

Corpus	Size	Note
ATR Ximera F009 voice [14]	~30,000 utterances 48 hours	Sampling rate: **48kHz** Japanese, neutral style, reading

- Recording period: over 1 year

	Feature	Dimension
Linguistic	Phone identity, prosodic tags ...	~ 390
Acoustic	MGC	60
	BAP	25
	F0	1

- Front-end: Open JTalk [15]

[14] Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K. (2004). Ximera: A new TTS from ATR based on corpus-based technologies. In Proc. SSW5, pages 179–184.
[15] HTS Working Group. The Japanese TTS System 'Open JTalk', 2015.

EXPERIMENTS

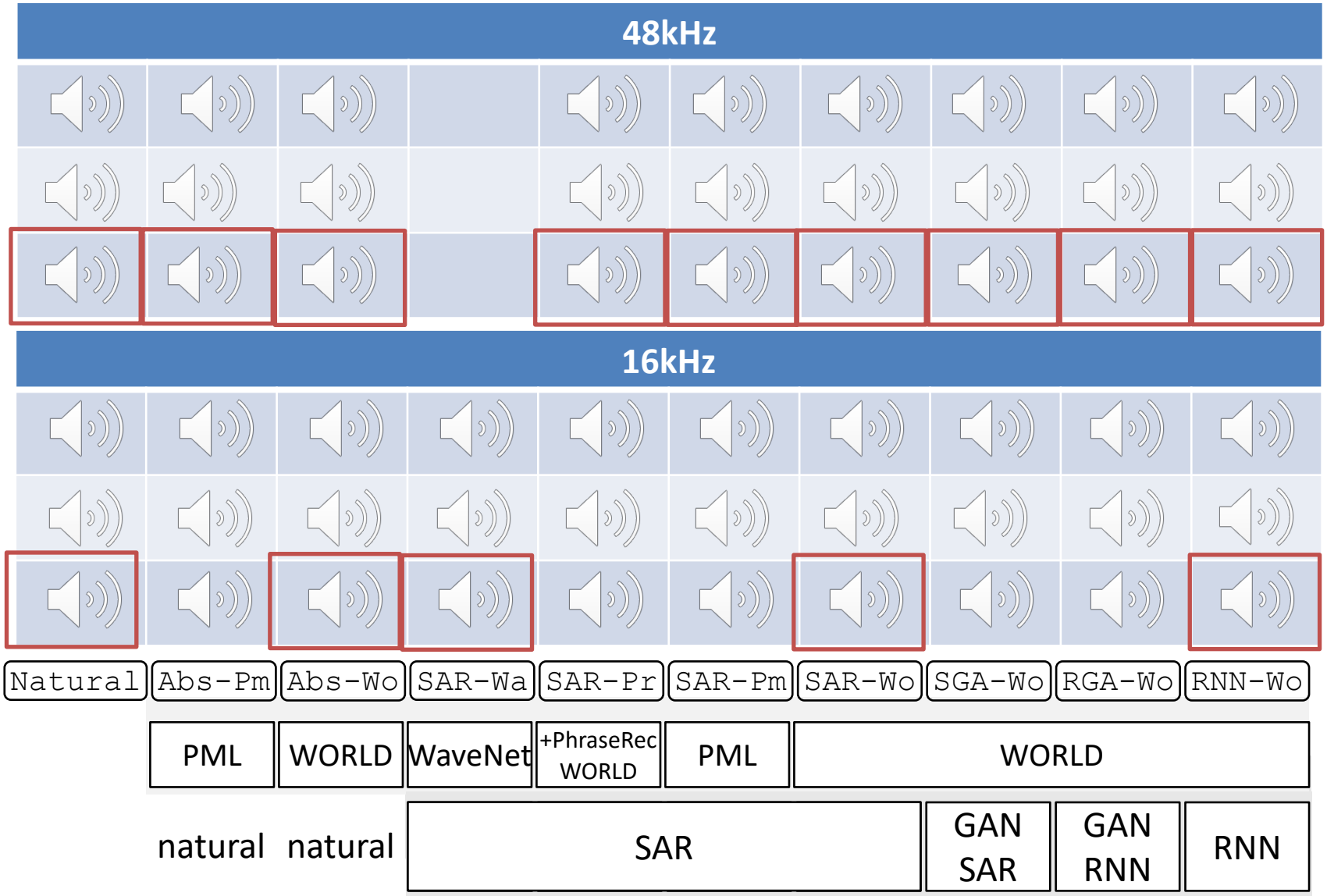
Configuration

- Listening test
  - Quality: MOS (1-5)
  - Similarity: rate 1-5, natural reference 48kHz
  - Participants: **235** native Japanese listeners, **1500** sets of results

- Systems
  - Common network configuration (cf. the paper)
  - Without Δ, Δ^2, nor formant enhancement
  - Sampling rate: 48kHz & 16kHz, except **SAR-Wa** at 16kHz (10 bits, μ-law)

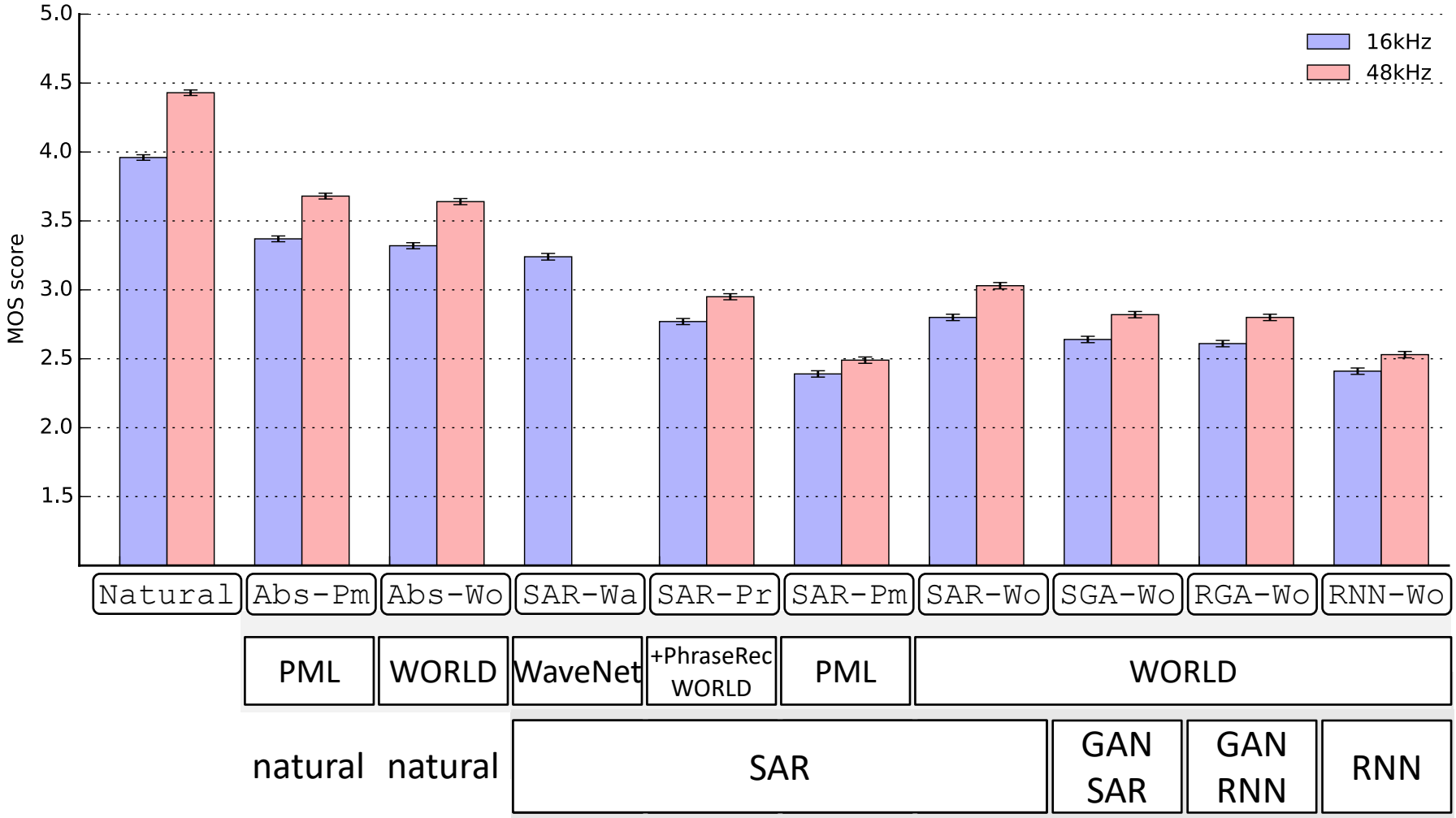
Natural	Abs-Pm	Abs-Wo	SAR-Wa	SAR-Pr	SAR-Pm	SAR-Wo	SGA-Wo	RGA-Wo	RNN-Wo
	PML	WORLD	WaveNet	+PhraseRec WORLD	PML	WORLD			
	natural	natural	SAR				GAN SAR	GAN RNN	RNN

EXPERIMENTS



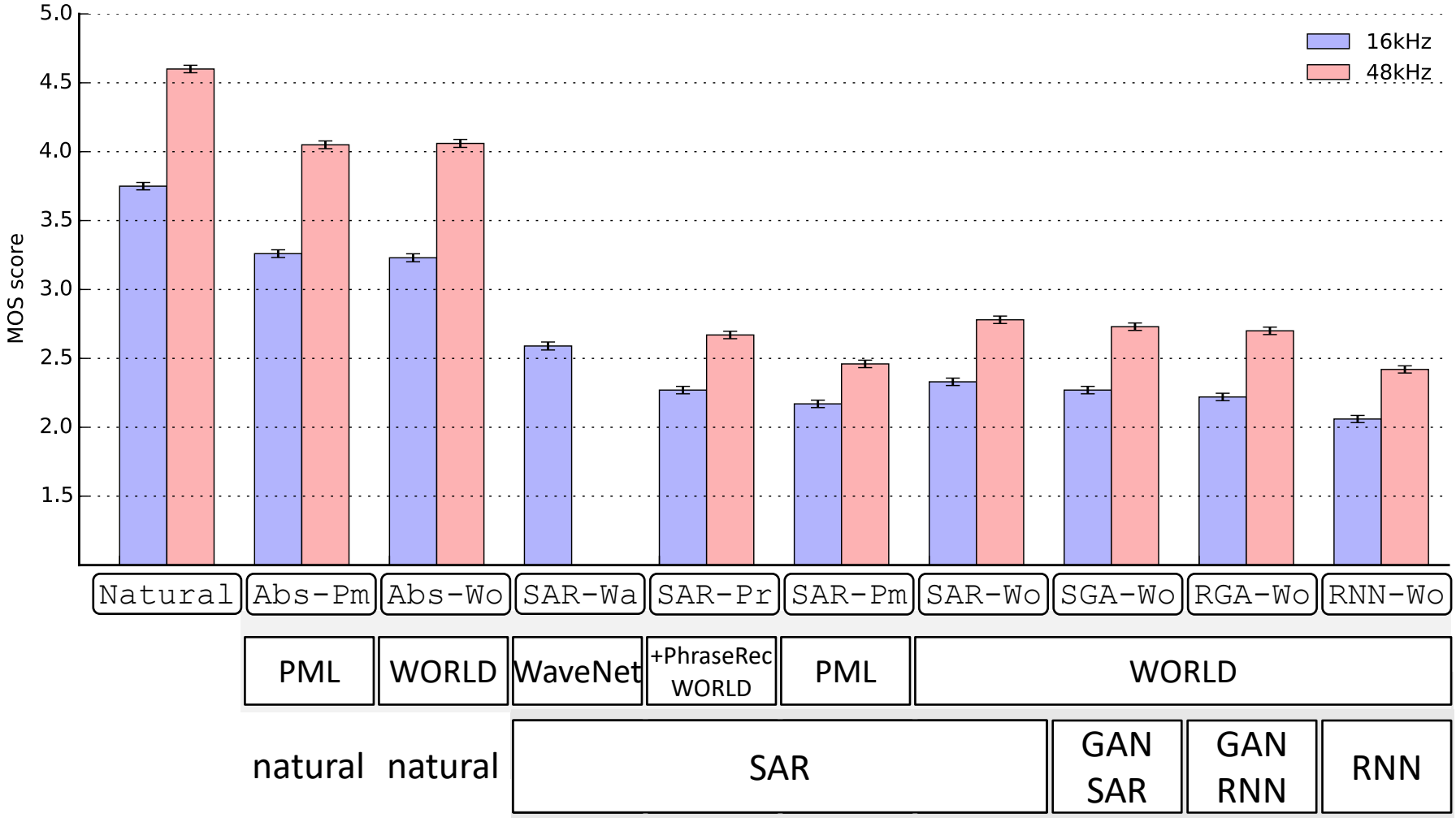
EXPERIMENTS

Quality scores



EXPERIMENTS

Similarity scores



CONTENTS

- Introduction
- Models and methods
- Experiments
- Summary

SUMMARY

Plug and test

- SAR:
  - Avoid conditional independence assumption
  - Alleviate the over-smoothing effect ☛ appendix & paper
- WaveNet
  - Generation method: one-best generation + random sampling
  - Less distortion of harmonics ☛ appendix & paper
- WaveNet vocoder + SAR & DAR
  - Better than other combinations
  - Worse than natural speech
  - Close to vocoded speech

FURTHER IMPROVEMENT?

Recent work ☛ appendix

- ❑ SAR
  - Special case of volume-preserving normalizing flow [16]
  - Extended SAR = time-variant filter + RNN
- ❑ WaveNet-vocoders
  - Training based on generated conditional features
- ❑ Annotated linguistic features
  - Reduce the gap between natural & synthetic speech

Future work?

- ❑ Reduce the variability of recordings
- ❑ Use complex-valued neural models

[16] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In Proc. ICML, pages 1530–1538, 2015.

MESSAGE

Code, recipes, slides

- Acoustic models & WaveNet (CUDA/C++)

  https://github.com/TonyWangX/CURRENNT_MODIFIED
  https://github.com/TonyWangX/CURRENNT_Recipes

- Simple explanation on WaveNet and acoustic models

  http://tonywangx.github.io/slides.html

tonywangx.github.io
wangxin@nii.ac.jp

Thank you for your attention

Q & A

tonywangx.github.io
wangxin@nii.ac.jp

APPENDIX - WAVENET

Conditional network

WaveNet-Backend

Learning curve

- No cherry picking
- All samples based on generated acoustic features or automatically inferred linguistic features

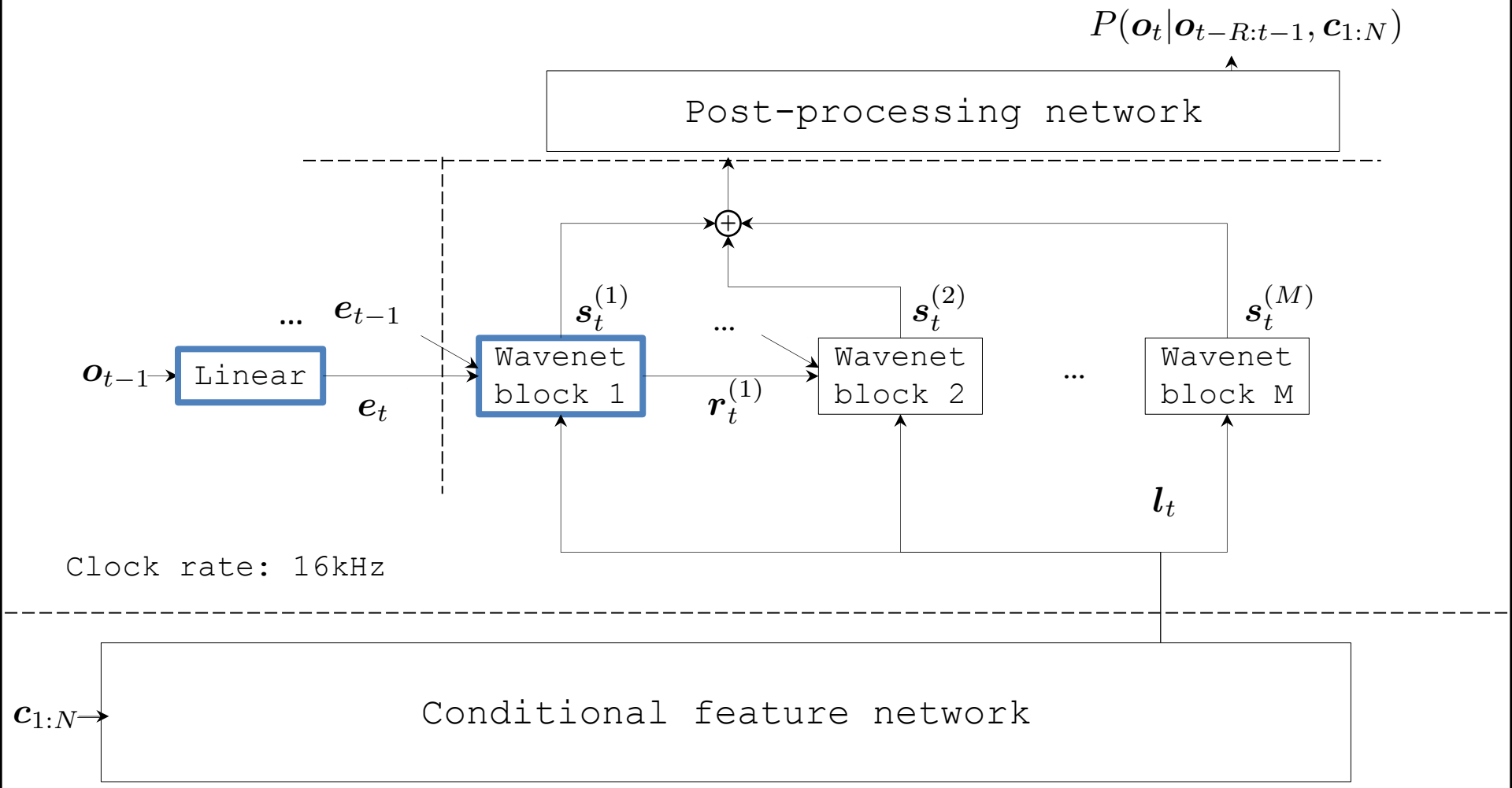
Generation method

Generation with other acoustic models

Training based on generated features

APPENDIX - WAVENET

Structure



APPENDIX - WAVENET

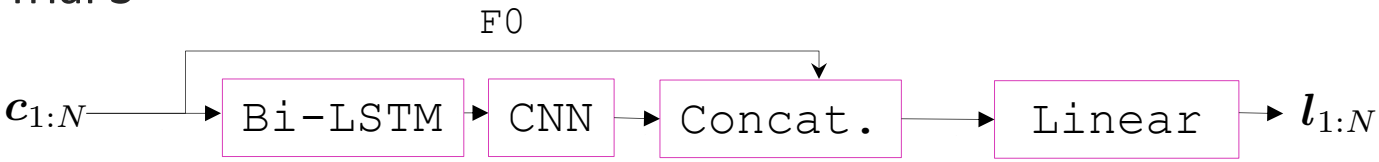
Conditional network

- Architecture of the conditional network

- Trial1

- Trial2

- Trial 3



APPENDIX - WAVENET

Conditional network

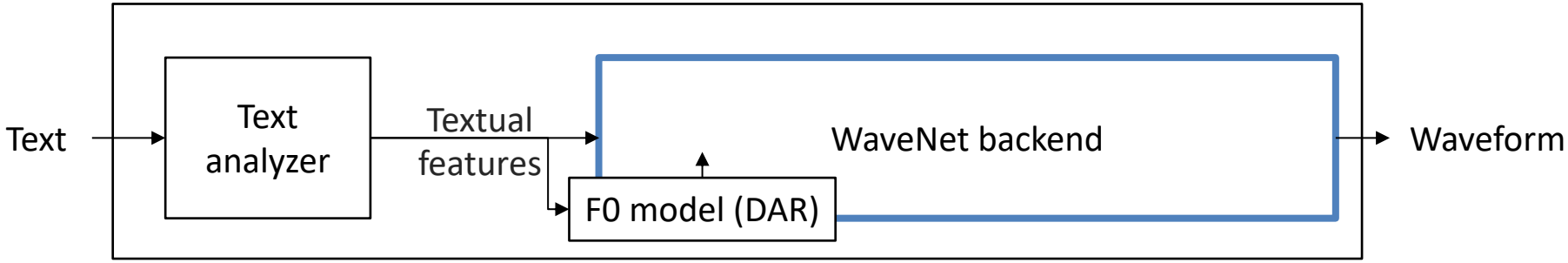
- Experiments on WaveNet Vocoder
  - Given generated MGC/F0

	sample1	sample2	sample3
Natural			
Trial 1 None			
Trial 2 LSTM+CNN			
Trial 3 LSTM+CNN+skip-F0			

APPENDIX - WAVENET

WaveNet backend

- Architecture

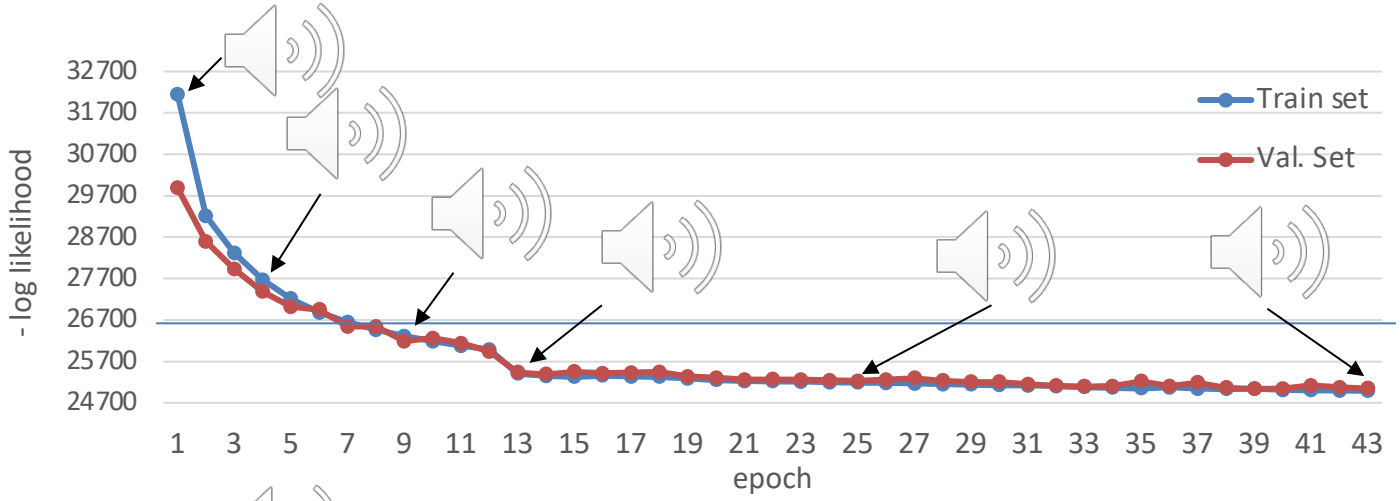


- WaveNet-backend only uses random sampling

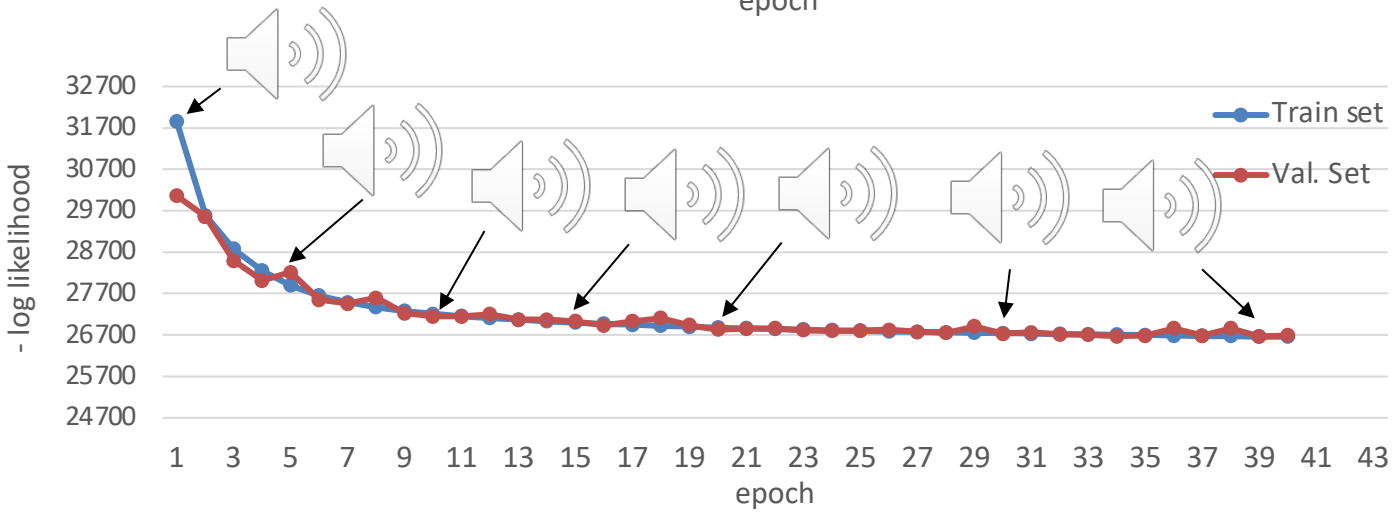
	sample1	sample2	sample3	sample4	sample5
Natural					
WaveNet-vocoder					
WaveNet-backend					

APPENDIX - WAVENET

Learning curve



WaveNet vocoder

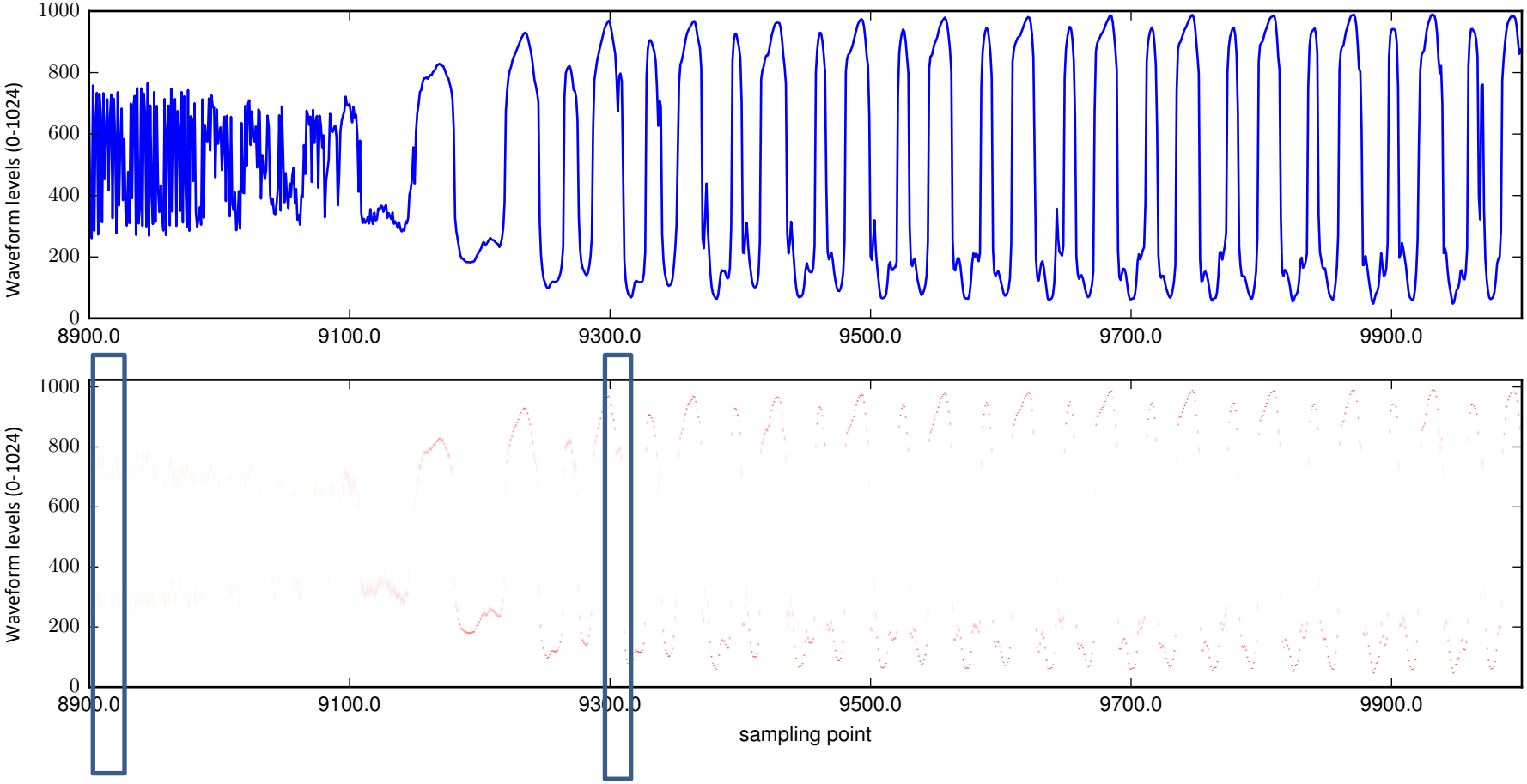


WaveNet backend

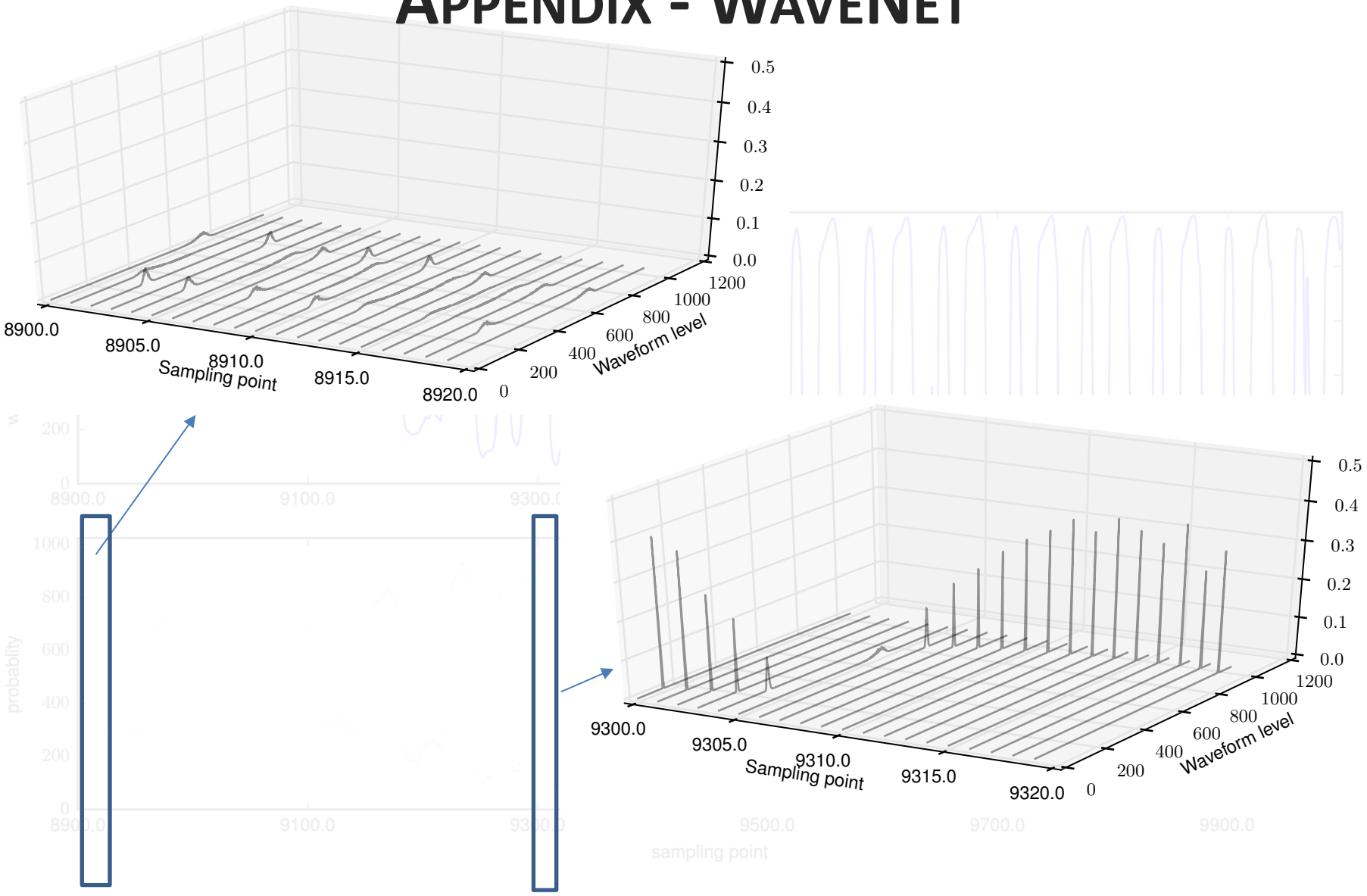
- We trained WaveNet backend for more than 100 epochs

APPENDIX - WAVENET

Generation method



APPENDIX - WAVENET

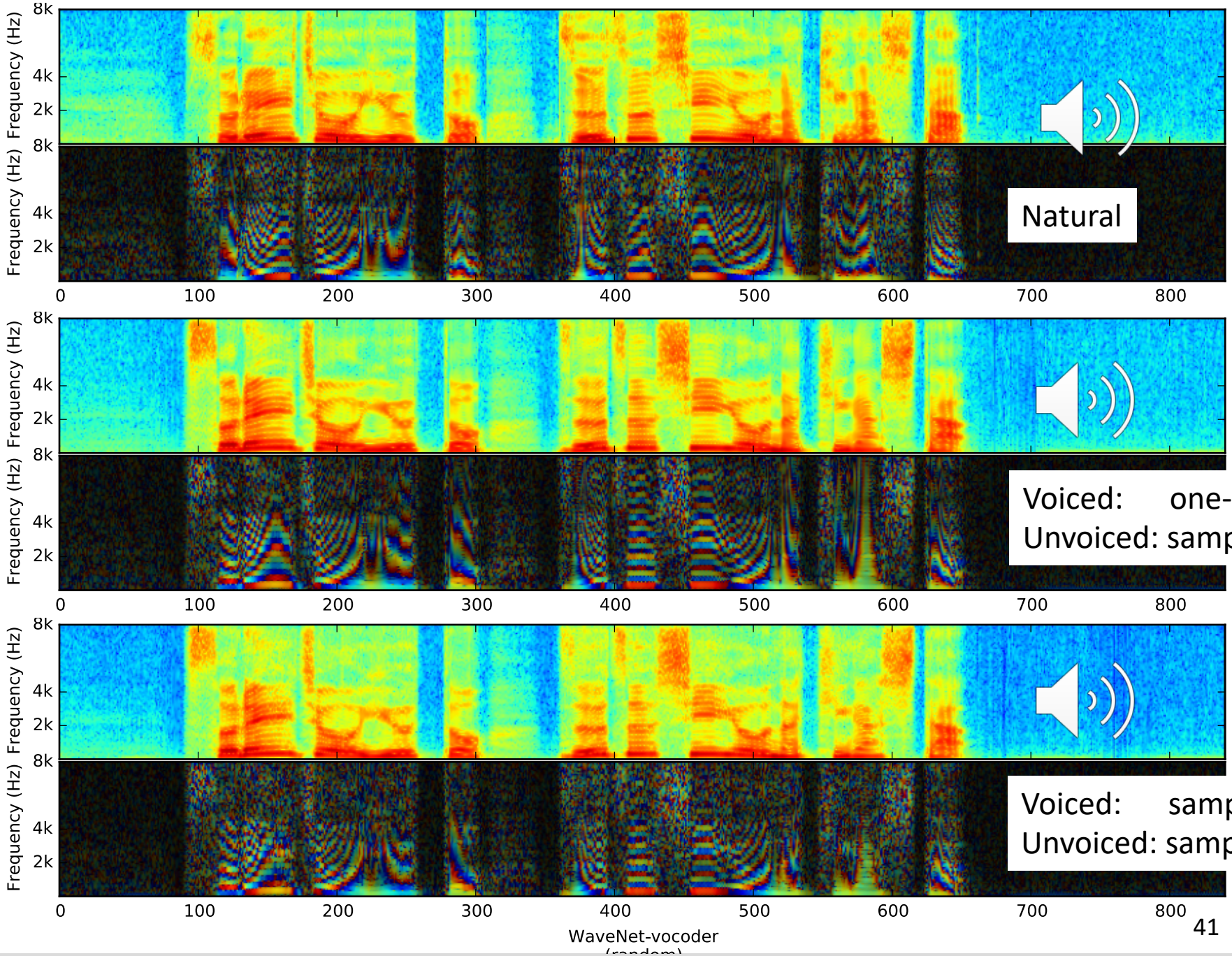


APPENDIX - WAVENET

Generation method

- One-best + random sampling
  - Given generated MGC/F0
  - Mix: voiced region: one-best; unvoiced region: random sampling
  - Random: random sampling all the time

	sample1	sample2	sample3
Natural			
Mix			
Random			


Frequency (Hz)
8k
4k
2k
0
100
200
300
400
500
600
700
800
Natural
Voiced: one-best
Unvoiced: sampling
Voiced: sampling
Unvoiced: sampling
WaveNet-vocoder

APPENDIX - WAVENET

Generation method

- One-best + random sampling

- Mix:
  - voiced region: one-best
  - unvoiced region: random sampling
- Mix2:
  - 75% voiced frames: one-best
  - else: random sampling
- …
- Mix4:
  - 25% voiced frames: one-best
  - else: random sampling

INVESTIGATION

Generation method

- One-best + random sampling

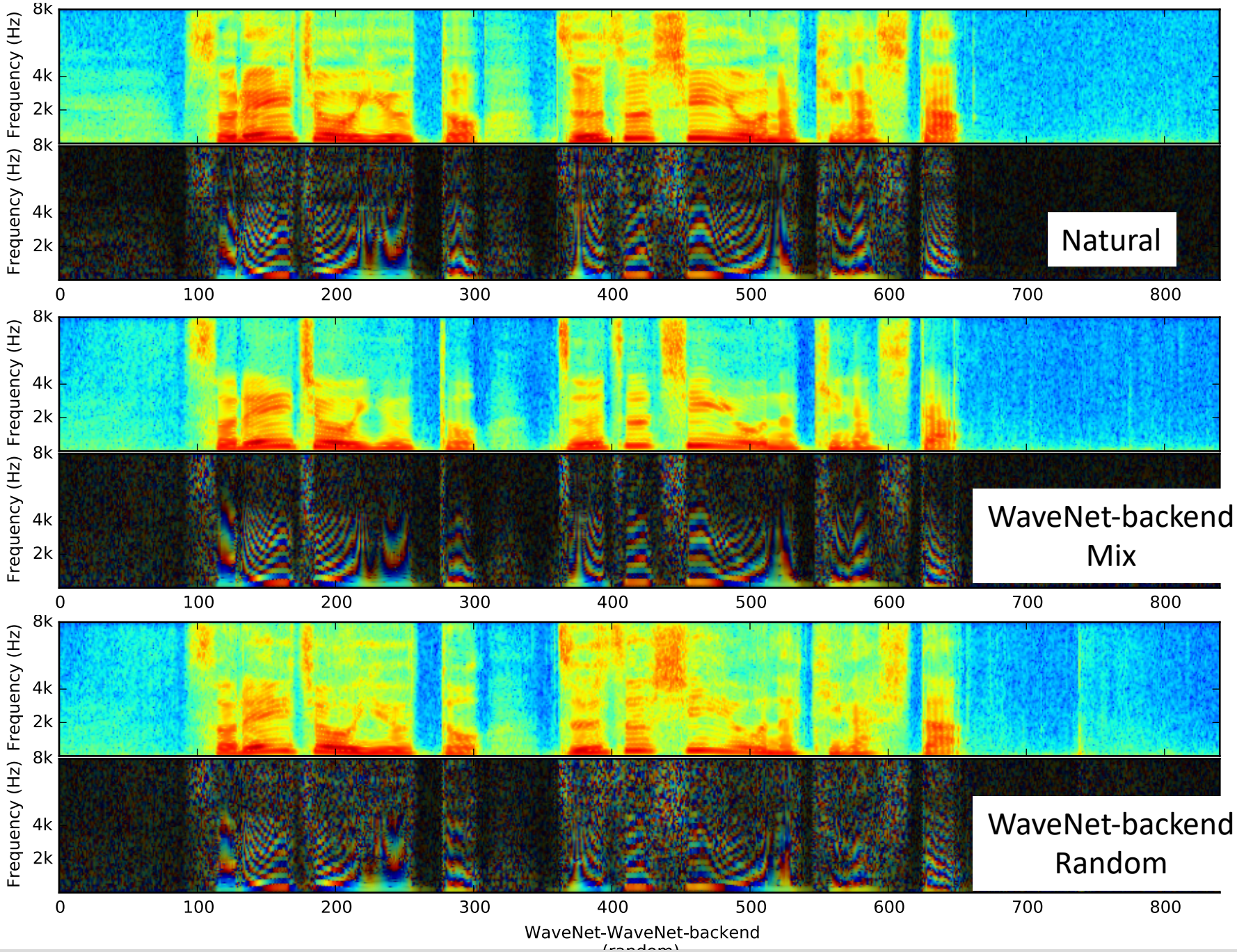
	sample1	sample2	sample3	
Natural				
Mix				100%
Mix2				75%
Mix3				50%
Mix4				20%
Random				00%

APPENDIX - WAVENET

Generation method

- One-best + random sampling

		sample1	sample2	sample3
Natural				
WaveNet backend	Mix			
	Random			
WaveNet vocoder	Mix			


Frequency (Hz)
8k
4k
2k
Natural
WaveNet-backend
Mix
WaveNet-backend
Random
0
100
200
300
400
500
600
700
800
WaveNet-WaveNet-backend

APPENDIX - WAVENET

WaveNet-vocoder + other acoustic models

	sample1	sample2	sample3
Natural			
SAR + DAR			
SGA + DAR			
RGA + DAR			
RNN + DAR			
extended SAR + DAR			

APPENDIX - WAVENET

WaveNet-vocoder: training using generated MGC

		sample1	sample2	sample3
	Natural			
	WaveNet-Backend			
WaveNet-Vocoder	Trained on natural MGC			
	Trained on generated MGC Epoch 35			
	Trained on generated MGC Epoch 45			
	Trained on generated MGC Epoch 55			

APPENDIX – ACOUSTIC MODELS

General comparison

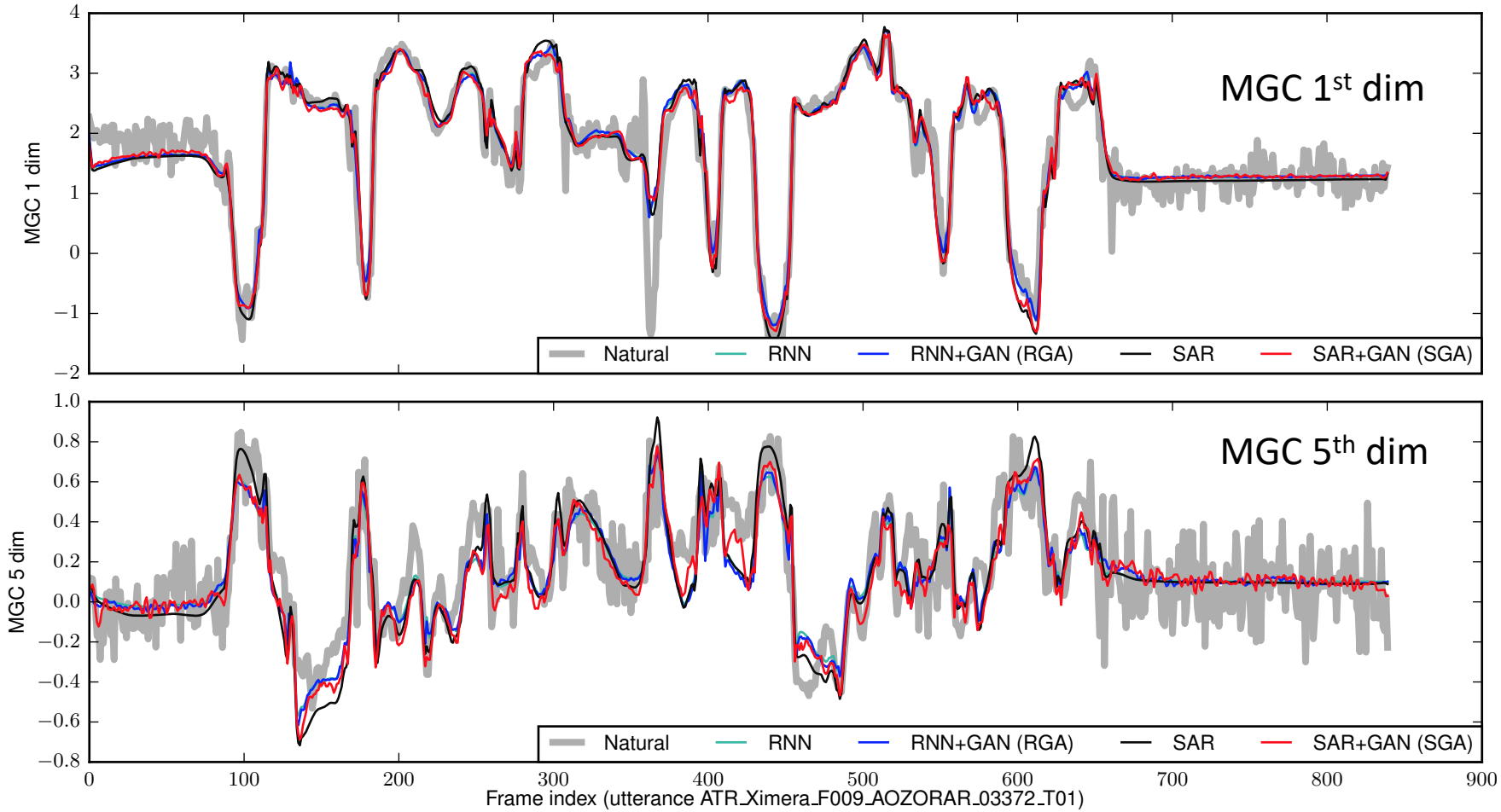
SAR & DAR

SAR extension

APPENDIX – ACOUSTIC MODELS

General comparison

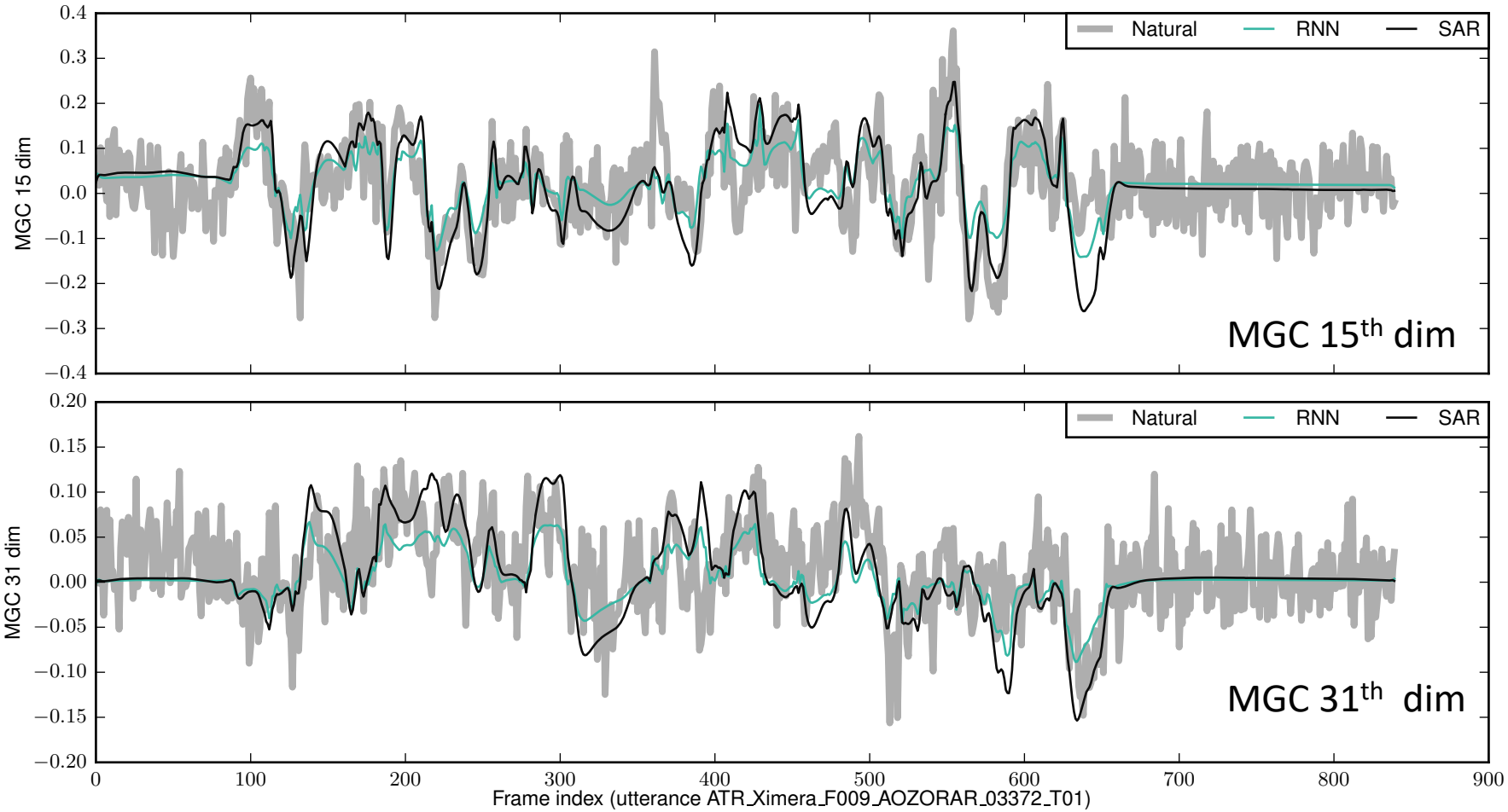
- Generated trajectories



APPENDIX – ACOUSTIC MODELS

General comparison

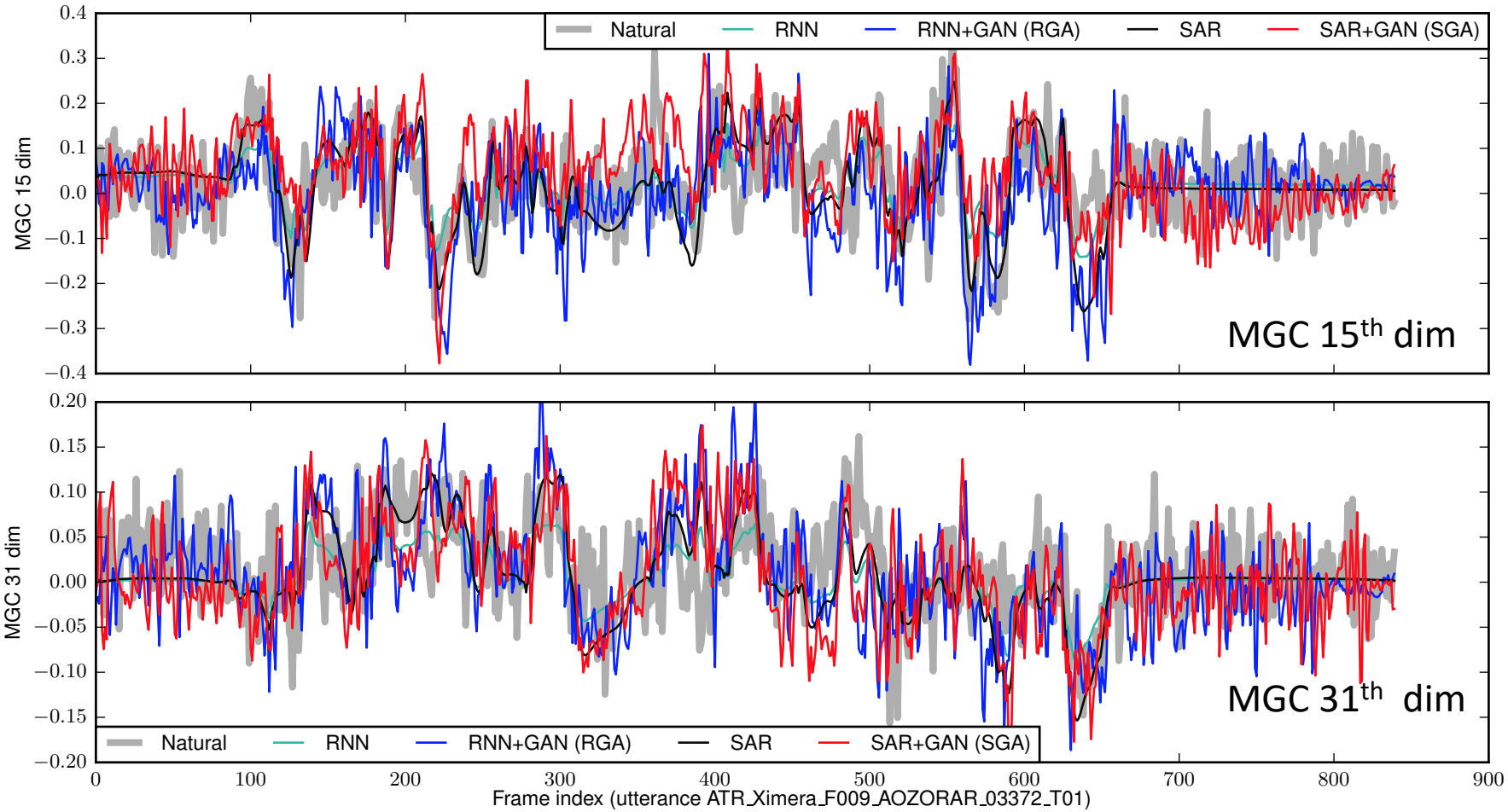
- Generated trajectories



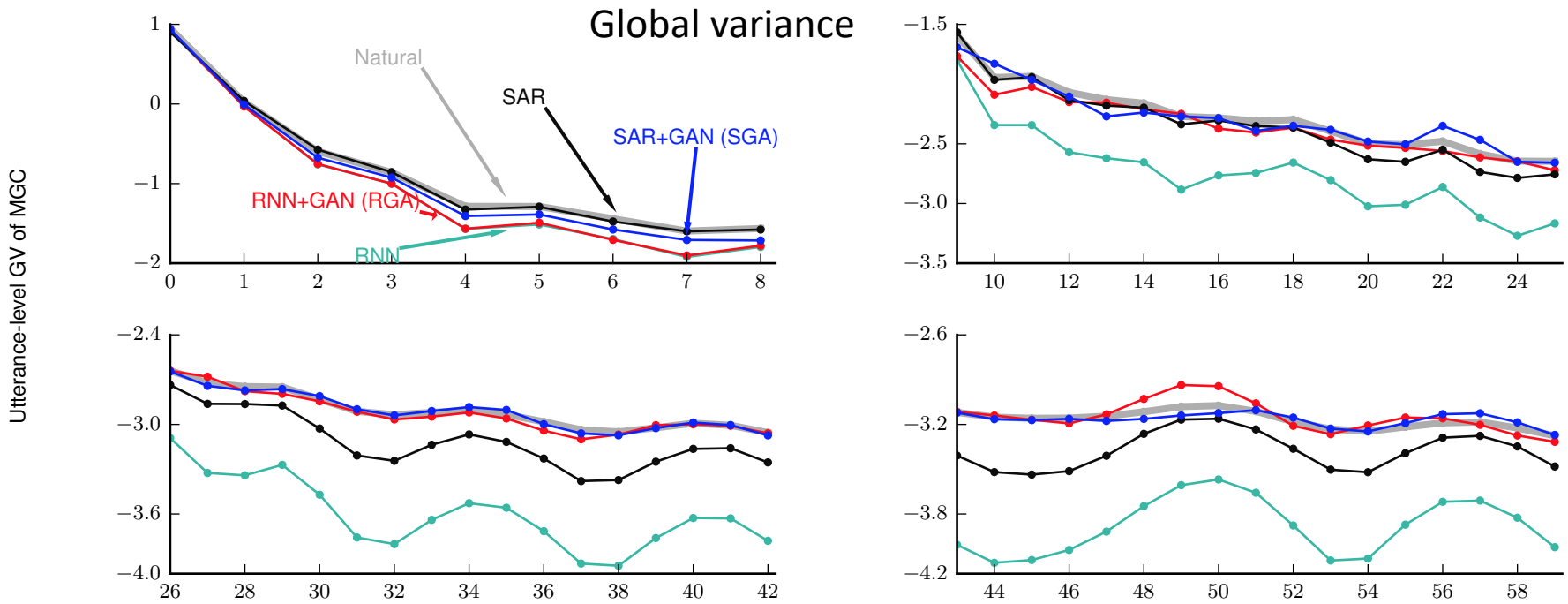
APPENDIX – ACOUSTIC MODELS

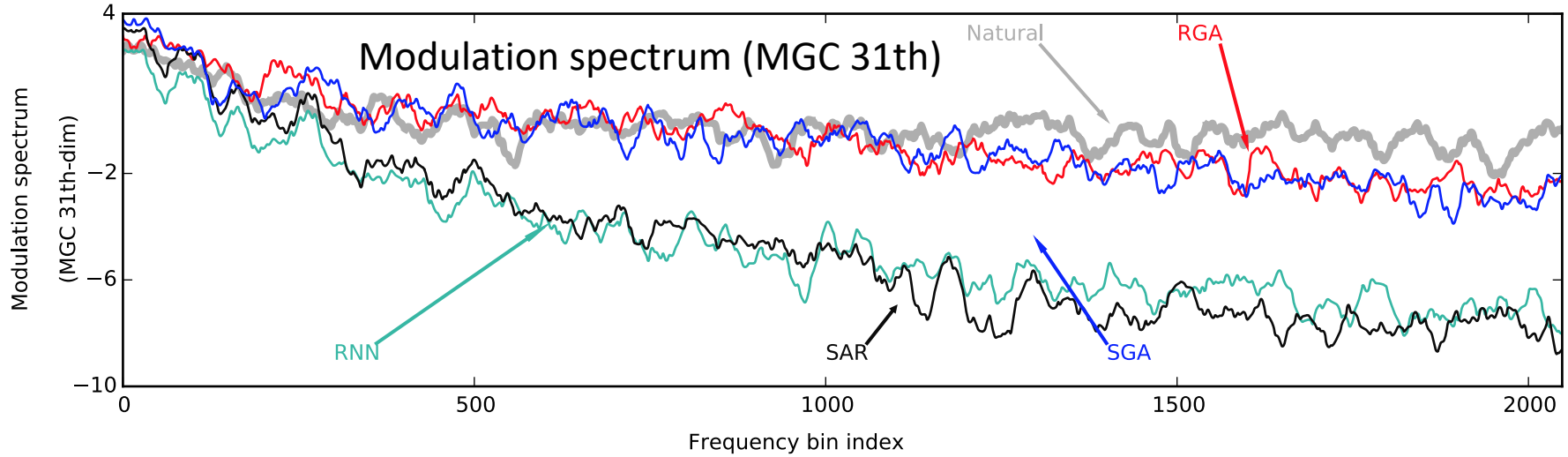
General comparison

- Generated trajectories



APPENDIX – ACOUSTIC MODELS



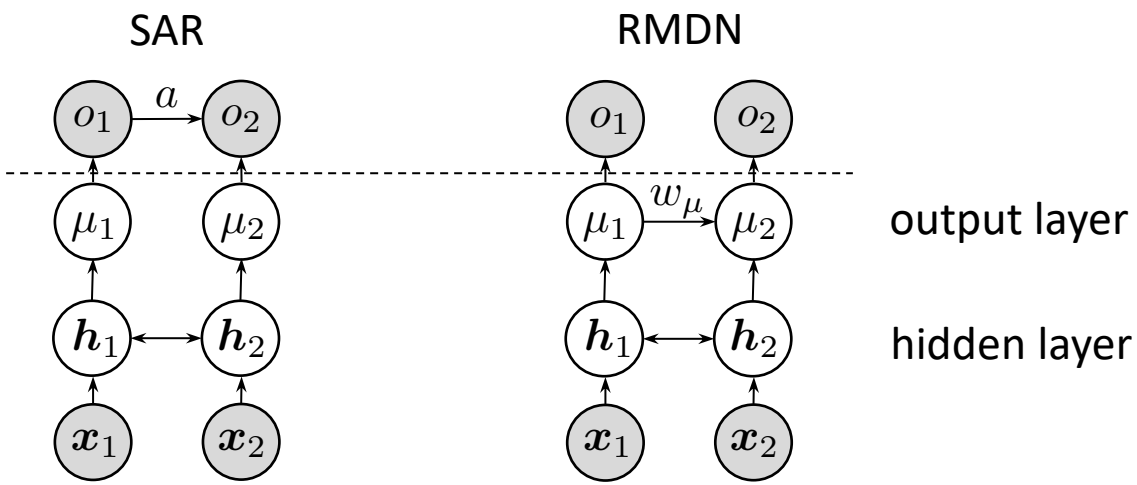


APPENDIX – ACOUSTIC MODELS

SAR & DAR

❑ Toy SAR example

- SAR versus RMDN with a recurrent output layer [11]



- Assume $o_t \in \mathbb{R}$ and $\Sigma_t = 1$, linear activation function

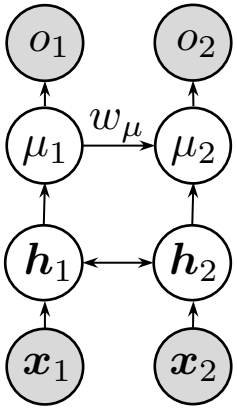
[11] H. Zen and H. Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In Proc. ICASSP, pages 4470–4474, 2015.

APPENDIX – ACOUSTIC MODELS

SAR & DAR

- Toy SAR example

RMDN

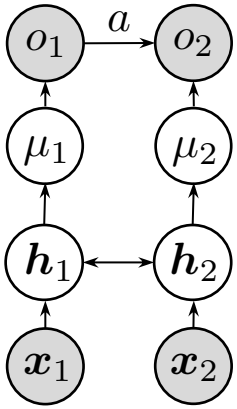


$$p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \tilde{\mu}_2 + w_\mu \mu_1, 1)$$

$$\mu_1 = \boldsymbol{w}^\top \boldsymbol{h}_1 + b$$

$$\mu_2 = \boldsymbol{w}^\top \boldsymbol{h}_2 + b + w_\mu \mu_1 = \tilde{\mu}_2 + w_\mu \mu_1$$

SAR



$$p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \mu_2 + a o_1, 1)$$

$$\mu_1 = \boldsymbol{w}^\top \boldsymbol{h}_1 + b$$

$$\mu_2 = \boldsymbol{w}^\top \boldsymbol{h}_2 + b$$

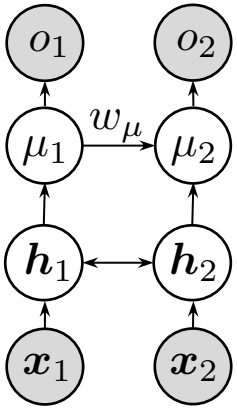
[11] H. Zen and H. Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In Proc. ICASSP, pages 4470–4474, 2015.

APPENDIX – ACOUSTIC MODELS

SAR & DAR

- Toy SAR example

RMDN

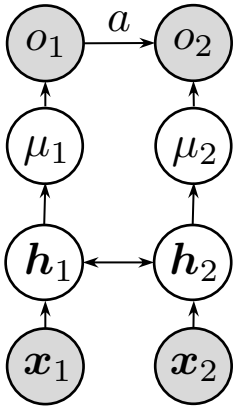


$$p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \tilde{\mu}_2 + w_\mu \mu_1, 1)$$
$$= \frac{1}{2\pi} \exp(-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{o} - \boldsymbol{\mu}))$$

$$\boldsymbol{o} = [o_1, o_2]^\top \qquad \boldsymbol{\mu} = [\mu_1, \tilde{\mu}_2 + w_\mu \mu_1]^\top \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Dependency between μ_t or o_t ?

SAR



$$p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \mu_2 + a o_1, 1)$$
$$= \frac{1}{2\pi} \exp(-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{o} - \boldsymbol{\mu}))$$

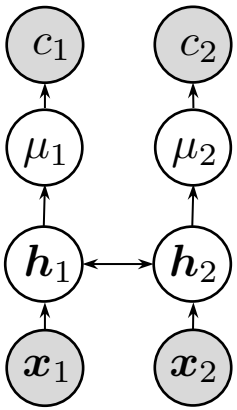
$$\boldsymbol{o} = [o_1, o_2]^\top \qquad \boldsymbol{\mu} = [\mu_1, \mu_2 + a\mu_1]^\top \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & a \\ a & 1 + a^2 \end{bmatrix}$$

APPENDIX – ACOUSTIC MODELS

SAR & DAR

- Toy SAR example

RMDN

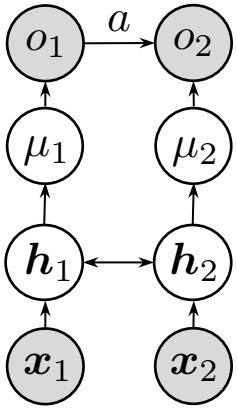


$$p(\boldsymbol{o}) = p(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_{\boldsymbol{c}}, \boldsymbol{\Sigma}_{\boldsymbol{c}})$$

$$\boldsymbol{\mu}_c = [\mu_1, \mu_2]^\top \qquad \boldsymbol{\Sigma}_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} o_1 \\ o_2 - a o_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix} \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} = \boldsymbol{A}\boldsymbol{o}$$

SAR



$$p(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{o}}, \boldsymbol{\Sigma}_{\boldsymbol{o}})$$

$$\boldsymbol{o} = [o_1, o_2]^\top \qquad \boldsymbol{\mu}_{\boldsymbol{o}} = [\mu_1, \mu_2 + a\mu_1]^\top \qquad \boldsymbol{\Sigma}_{\boldsymbol{o}} = \begin{bmatrix} 1 & a \\ a & 1 + a^2 \end{bmatrix}$$

APPENDIX – ACOUSTIC MODELS

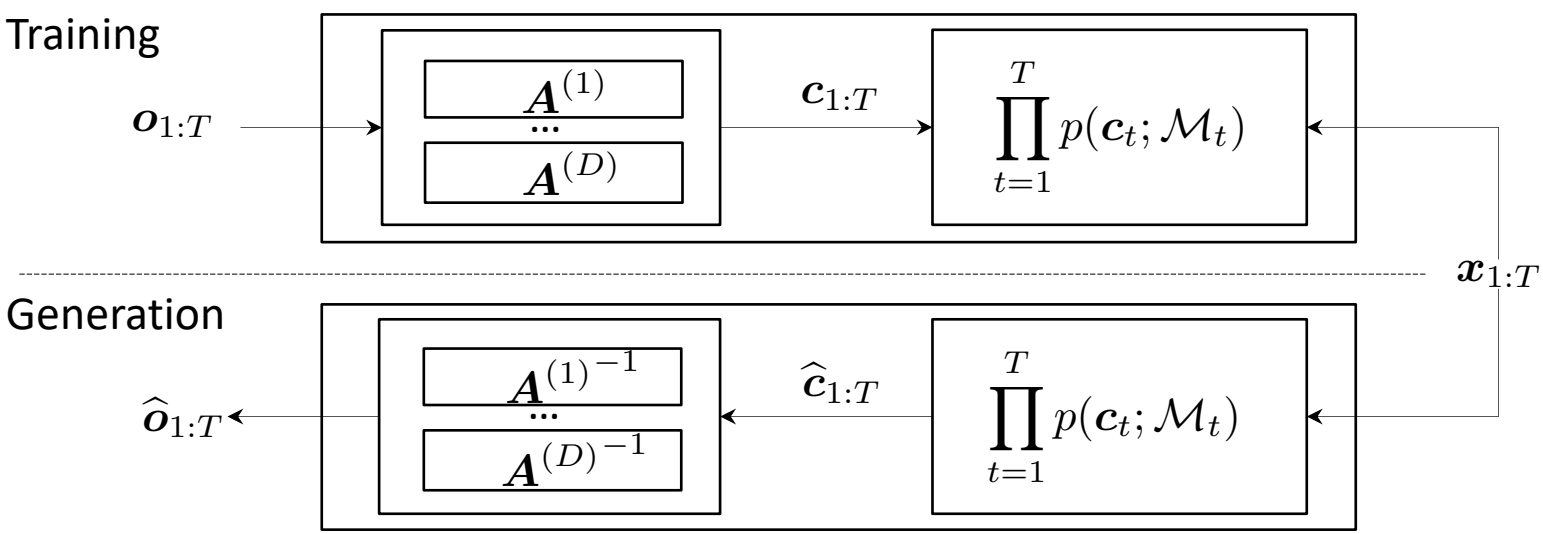
SAR & DAR

- SAR: invertible linear feature/model transformation

- For $\boldsymbol{o}_{1:T} \in \mathbb{R}^{D \times T}$

$$\boldsymbol{o}_{1:T} = \begin{bmatrix} \boldsymbol{o}_{1:T,1} \\ \cdots \\ \boldsymbol{o}_{1:T,D} \end{bmatrix} \begin{matrix} \longrightarrow \\ \longrightarrow \\ \longrightarrow \end{matrix} \begin{matrix} \boldsymbol{A}^{(1)} \\ \cdots \\ \boldsymbol{A}^{(D)} \end{matrix} \begin{matrix} \longrightarrow \\ \longrightarrow \\ \longrightarrow \end{matrix} \begin{bmatrix} \boldsymbol{c}_{1:T,1} \\ \cdots \\ \boldsymbol{c}_{1:T,D} \end{bmatrix}$$

- SAR is equivalent to:



APPENDIX – ACOUSTIC MODELS

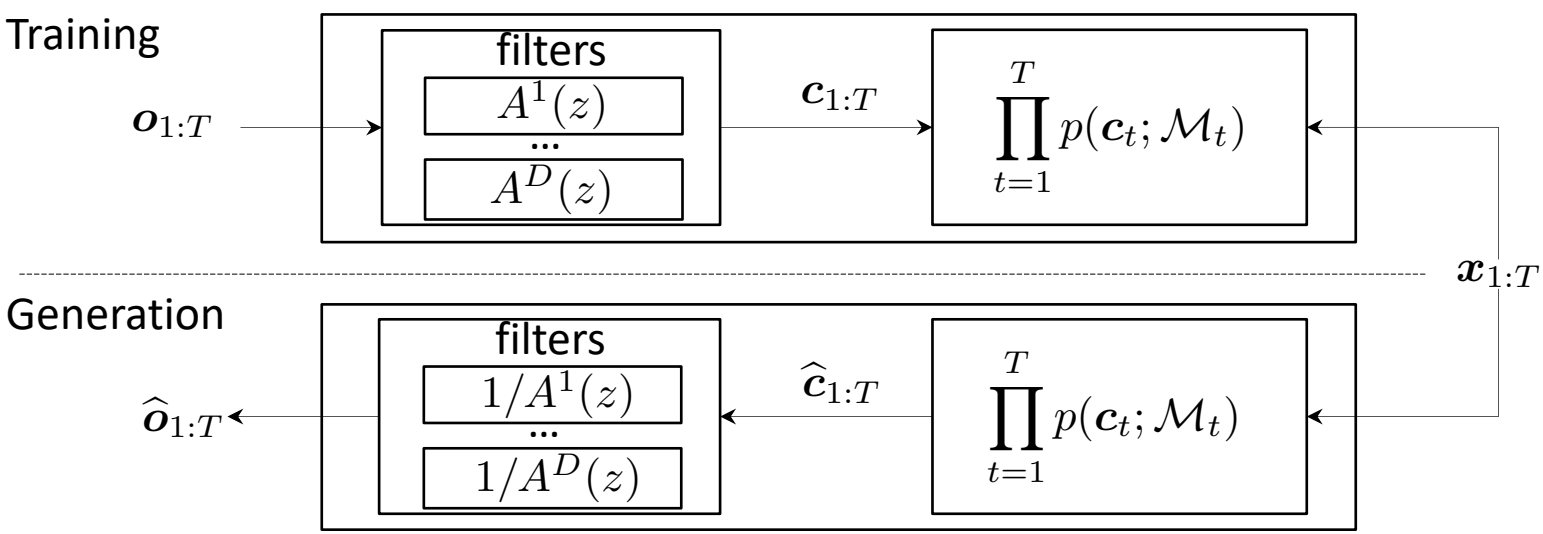
SAR & DAR

- SAR: invertible linear feature/model transformation

- For $\boldsymbol{o}_{1:T} \in \mathbb{R}^{D \times T}$

$$\boldsymbol{o}_{1:T} = \begin{bmatrix} \boldsymbol{o}_{1:T,1} \\ \cdots \\ \boldsymbol{o}_{1:T,D} \end{bmatrix} \begin{matrix} \longrightarrow \boxed{\text{filter 1}} \longrightarrow \\ \longrightarrow \boxed{\cdots} \longrightarrow \\ \longrightarrow \boxed{\text{filter D}} \longrightarrow \end{matrix} \begin{bmatrix} \boldsymbol{c}_{1:T,1} \\ \cdots \\ \boldsymbol{c}_{1:T,D} \end{bmatrix}$$

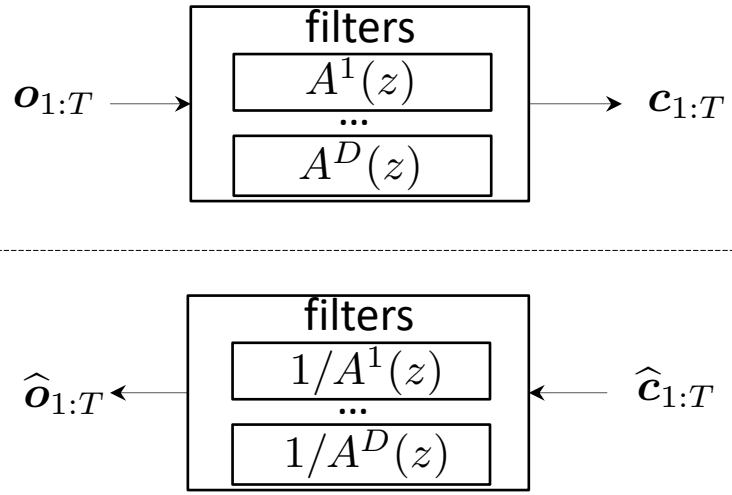
- SAR is equivalent to:

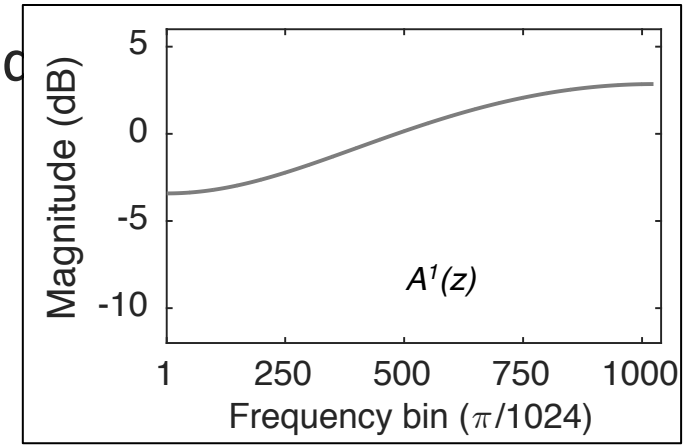


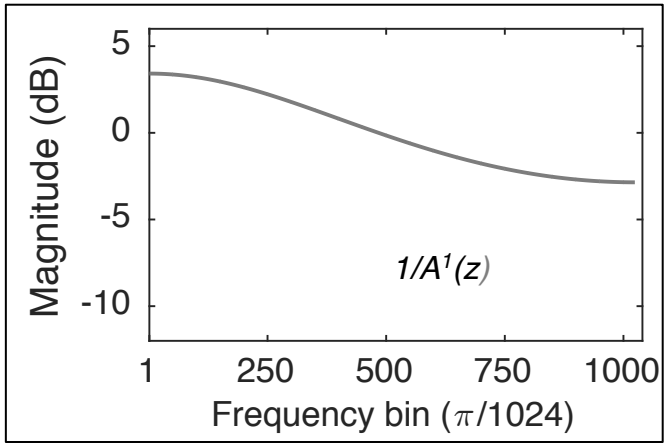
APPENDIX – ACOUSTIC MODELS

SAR & DAR

- SAR: invertible linear feature/mod





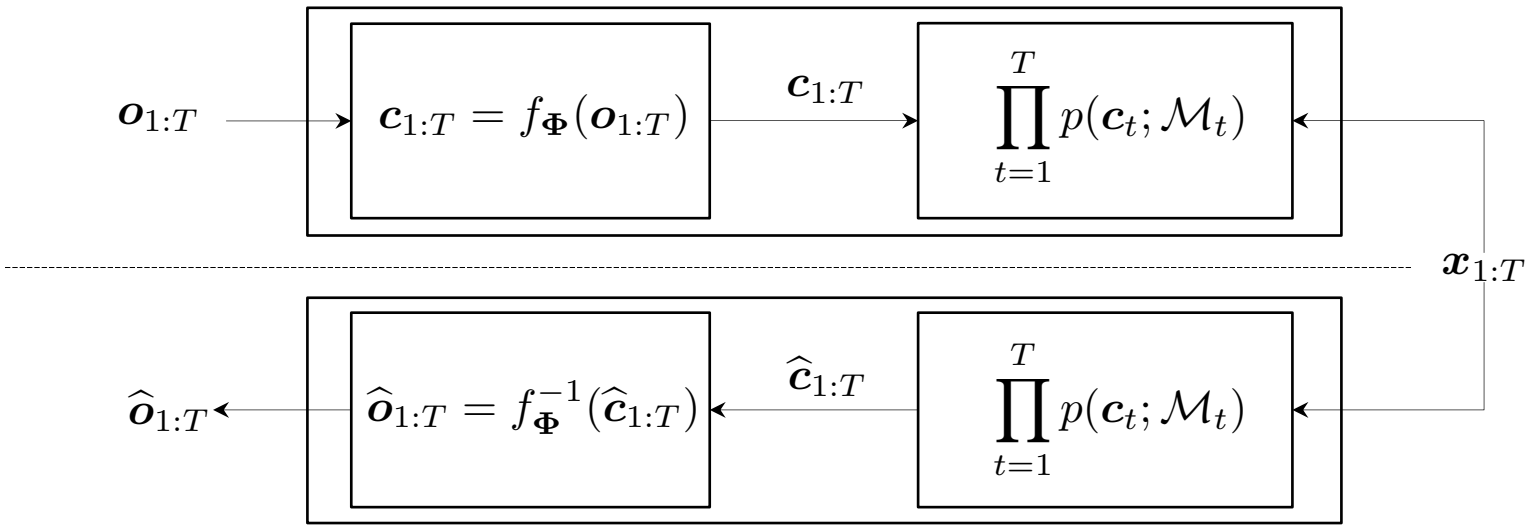


- Only due to $1/A^d(z)$?
- Due to $\{A^d(z), 1/A^d(z)\}$, less mismatch between $\boldsymbol{c}_{1:T}$ and RMDN

APPENDIX – ACOUSTIC MODELS

SAR extension: normalizing flow

- Basic idea



$$p_o(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}) = p_c(\boldsymbol{c}_{1:T}|\boldsymbol{x}_{1:T}) \left| \det \frac{\partial \boldsymbol{c}_{1:T}}{\partial \boldsymbol{o}_{1:T}} \right|$$

- Jacobian matrix must be simple
- f (.) must be invertible

[13] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In International Conference on Machine Learning, pages 1530–1538, 2015.
[14] 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.

APPENDIX – ACOUSTIC MODELS

SAR extension: normalizing flow

- Basic idea

$$p_o(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}) = p_c(\boldsymbol{c}_{1:T}|\boldsymbol{x}_{1:T}) \left| \det \frac{\partial \boldsymbol{c}_{1:T}}{\partial \boldsymbol{o}_{1:T}} \right|$$

	SAR	AR Flow
Transform	$\boldsymbol{c}_t = \boldsymbol{o}_t - \sum_{k=1}^{K} \boldsymbol{a}_k \odot \boldsymbol{o}_{t-k}$	$\boldsymbol{c}_t = \boldsymbol{o}_t - \sum_{k=1}^{K} f^{(k)}(\boldsymbol{o}_{1:t-k}) \odot \boldsymbol{o}_{t-k}$
De-transform	$\widehat{\boldsymbol{o}}_t = \boldsymbol{c}_t + \sum_{k=1}^{K} \boldsymbol{a}_k \odot \widehat{\boldsymbol{o}}_{t-k}$	$\widehat{\boldsymbol{o}}_t = \boldsymbol{c}_t + \sum_{k=1}^{K} f^{(k)}(\widehat{\boldsymbol{o}}_{1:t-k}) \odot \widehat{\boldsymbol{o}}_{t-k}$

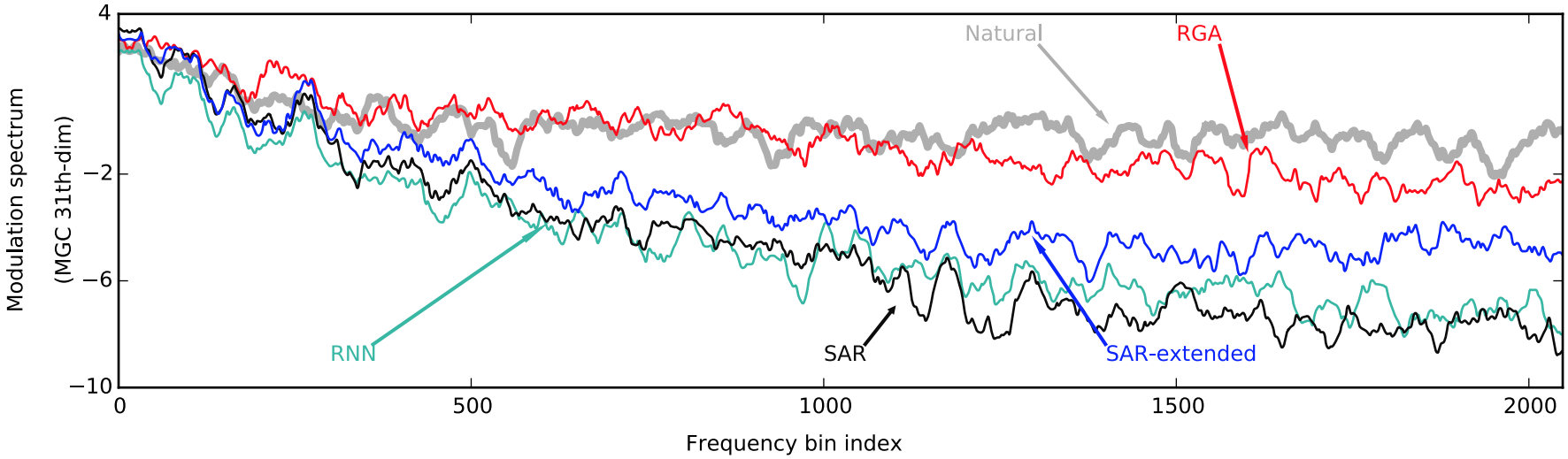
$$\boldsymbol{\mu_t} = \mathrm{RNN}(o_{1:t-1})$$

- Simple for SAR: $\left| \det \frac{\partial \boldsymbol{c}_{1:T}}{\partial \boldsymbol{o}_{1:T}} \right| = 1$

APPENDIX – ACOUSTIC MODELS

SAR extension

- SAR can be extended



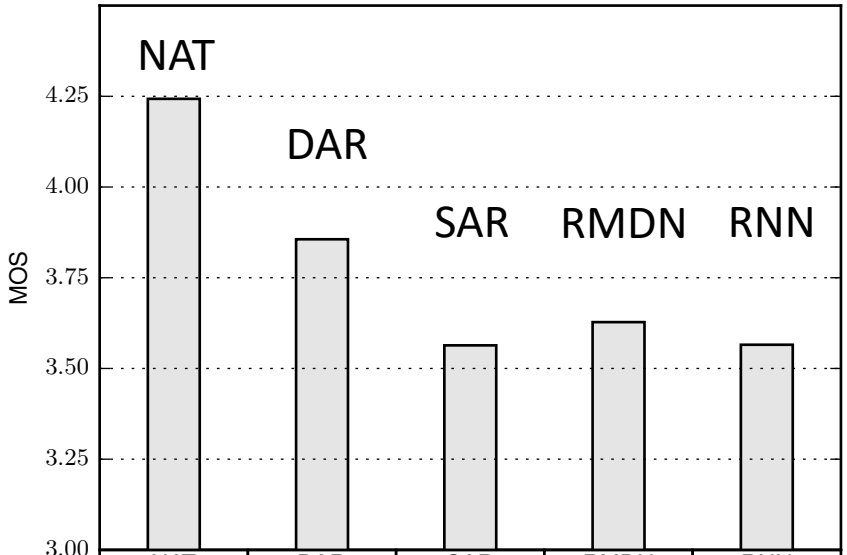
More details: http://tonywangx.github.io/pdfs/talk.pdf

APPENDIX – ACOUSTIC MODELS

DAR

- Same autoregressive principle
- But noninvertible nonlinear

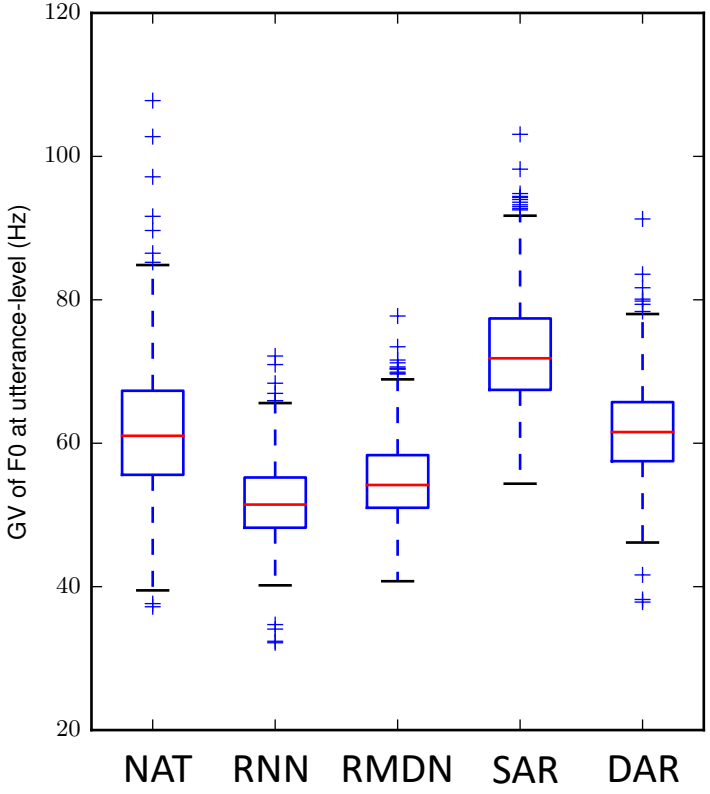
MOS score



p-value

	NAT	DAR	SAR	RMDN	RNN
NAT		<1e-30	<1.0e-30	<1.0e-30	<1.0e-30
DAR	<1e-30		1.6e-28	6.3e-19	2.4e-30
SAR	<1e-30	1.6e-28		0.015	0.949
RMDN	<1e-30	6.3e-19	0.015		0.014
RNN	<1e-30	2.4e-30	0.949	0.014	

F0 GV



[7] X. Wang, S. Takaki, and J. Yamagishi. Autoregressive neural F0 model for statistical parametric speech synthesis. IEEE Transactions on Audio, Speech and Language Processing. (Accepted)