Slides by Xin Wang
National Institute of Informatics

Note: Natural Japanese speech data belonging to ATR Ximera corpus are deleted in this public available version

Fundamental Frequency Modeling for Neural-Network-based Statistical Parametric Speech Synthesis

Xin Wang

ID: 20151706

2018-07-13

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

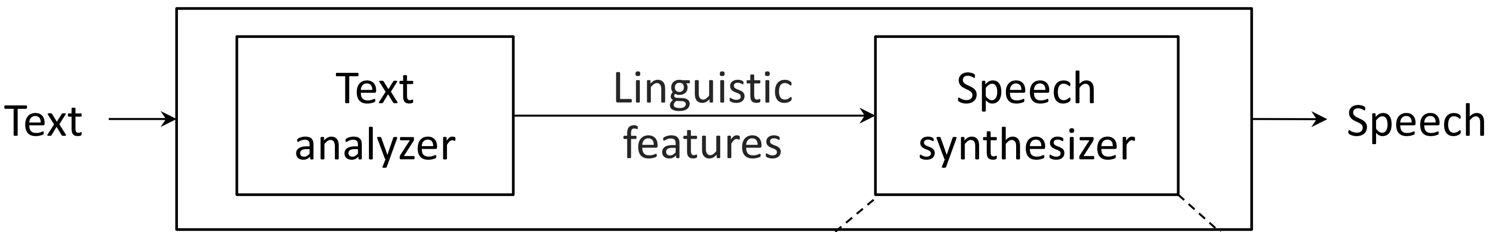
CONTENTS

★ New results / updated explanation

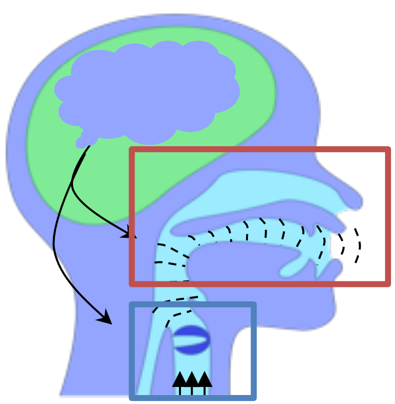
INTRODUCTION

Text-to-speech (TTS)

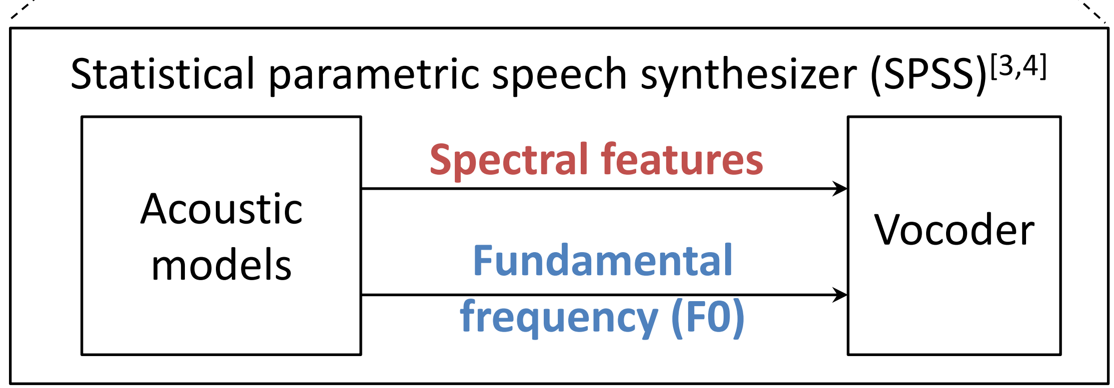
- TTS pipeline [1,2]



- Synthesizer



(c.f. HTS Slides, by HTS Working Group)

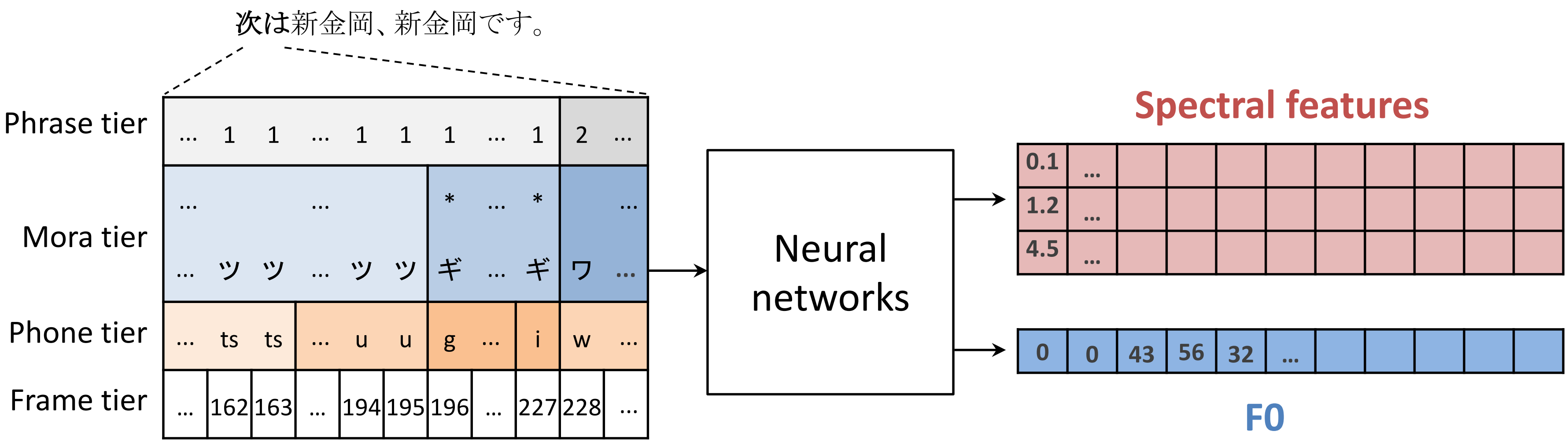


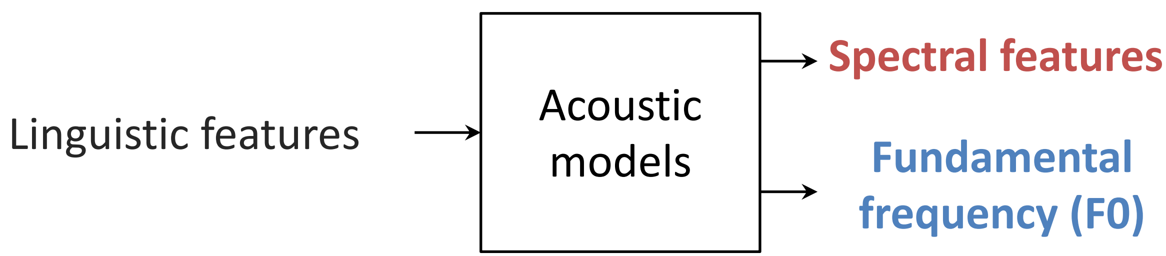
[1] Taylor, P. (2009). Text-to-Speech Synthesis.
[2] Dutoit, T. (1997). An Introduction to Text-to-speech Synthesis.
[3] Tokuda, K., et al., (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*, *101*(5), 1234–1252.
[4] Zen, H., et al. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*, 1039–1064.

Introduction

Text-to-speech (TTS)

- Neural-network-based acoustic models [5,6,7]



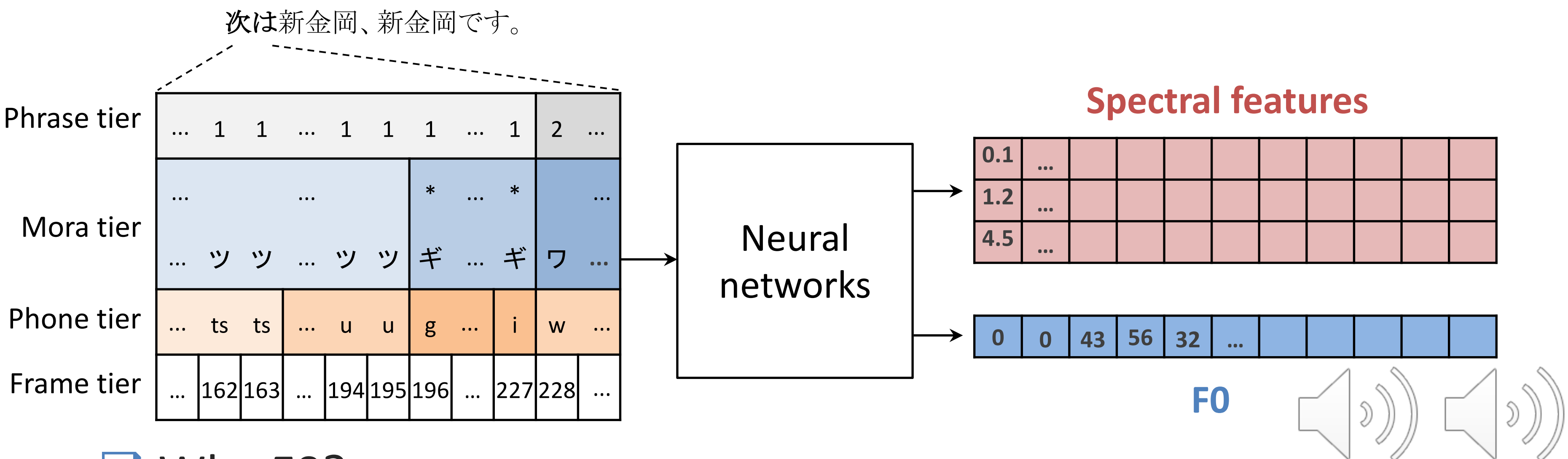


[5] H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In Proc. ICASSP, pages 7962–7966, 2013.
[6] Z. H. Ling, et al. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. IEEE Signal Processing Magazine, 32(3):35–52, 2015.
[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proc. Interspeech, pages 1964–1968, 2014.

INTRODUCTION

Topic

- ❑ Neural F0 modeling for TTS



- ❑ Why F0?

Speaker A: Who made the marmalade.

Speaker B: Marianna made the marmalade.

Speaker A: Bob made the marmalade.

Speaker B: (No,) Marianna made the marmalade.

Speaker B: Marianna made the marmalade.

Speaker B: Marianna made the marmalade.

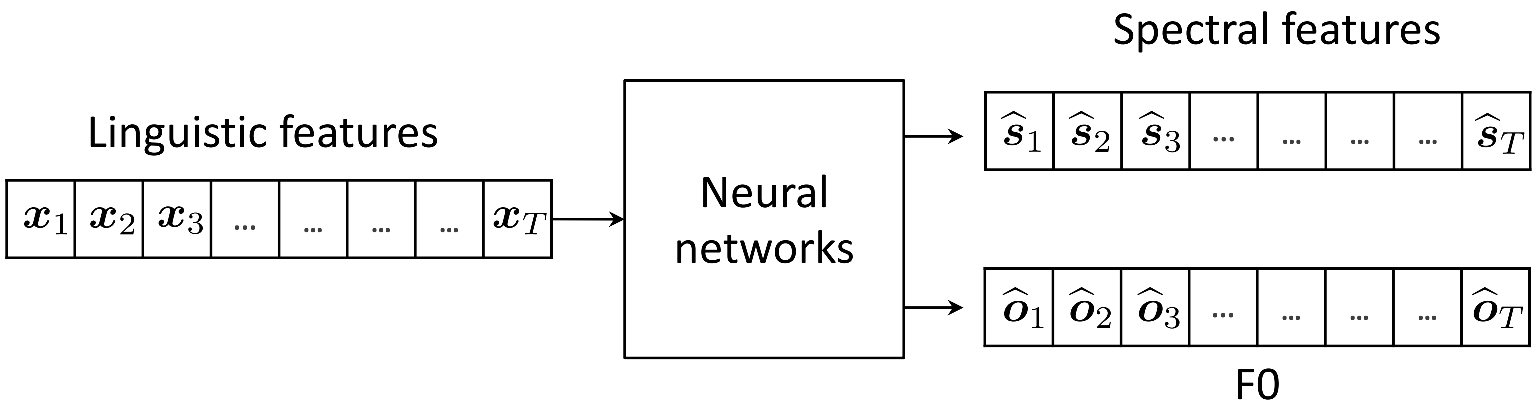
Speaker B: Marianna made the marmalade.

[8] Nanette Veilleux, et al. 6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI. January IAP 2006. https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

INTRODUCTION

Topic

- Issues to be addressed



$$p(\underline{\boldsymbol{o}_{1:T}}, \boldsymbol{s}_{1:T} | \underline{\boldsymbol{x}_{1:T}}; \boldsymbol{\Theta}) = \underline{\prod_{t=1}^{T}} \mathcal{N}(\underline{[\boldsymbol{o}_t, \boldsymbol{s}_t]}; \underline{\mathrm{Network}_{\boldsymbol{\Theta}}(\boldsymbol{x}_{1:T}, t),} \beta \boldsymbol{I})$$

F0 features [9]

Linguistic features [10]

Issue 3: efficient enough?

Issue 1: joint modeling?

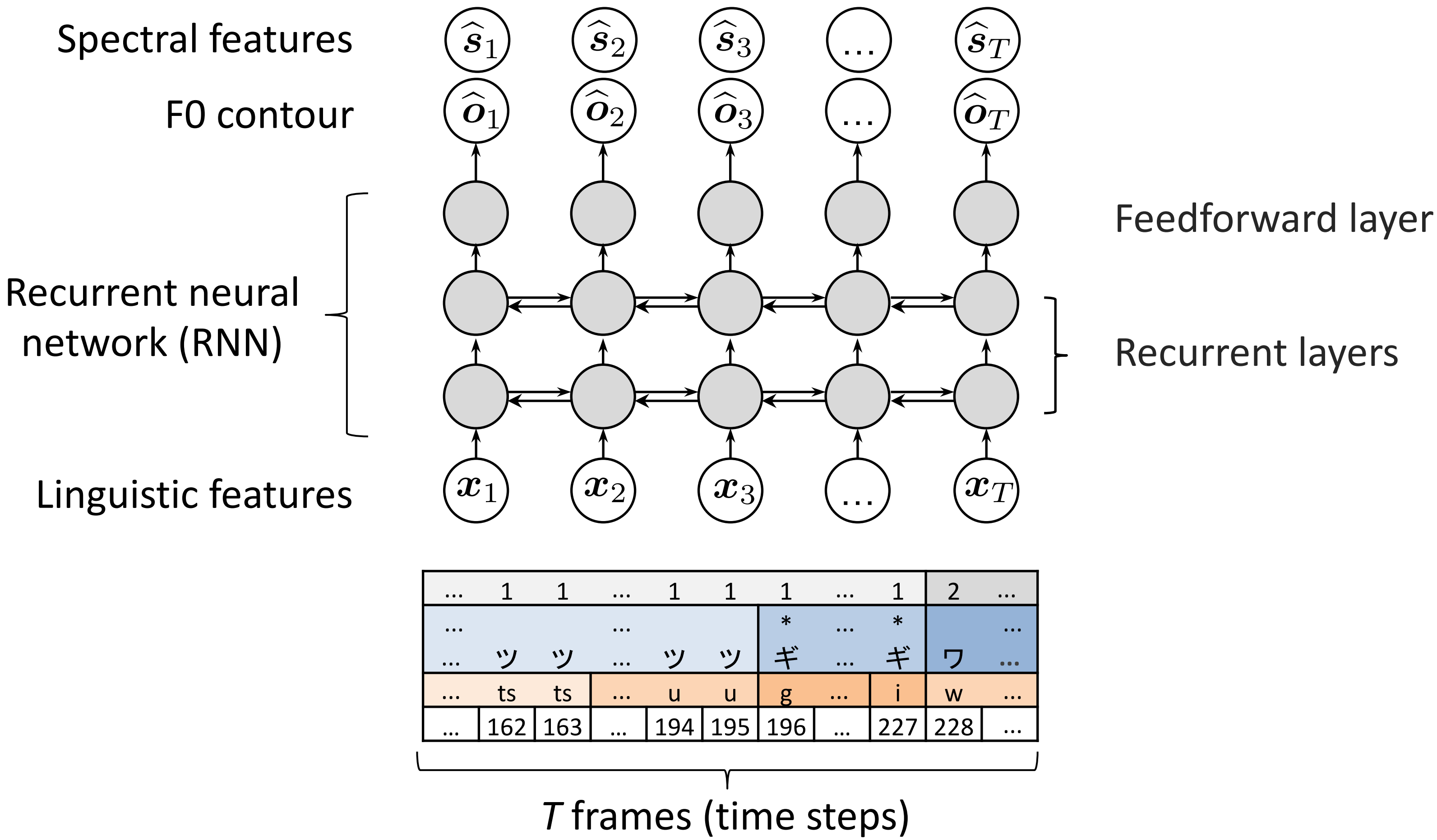
Issue 2: temporal dependency?

[9] M. S. Ribeiro. Suprasegmental representations for the modeling of fundamental frequency in statistical parametric speech synthesis. PhD thesis, The University of Edinburgh, 2018.

[10] J. Hirschberg. Pitch accent in context predicting intonational prominence from text. Artificial Intelligence, 63(1):305–340, 1993.

Thesis outline

- Conventional approaches [7] (Table 3.1 in thesis)



[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proc. Interspeech, pages 1964–1968, 2014.

Introduction

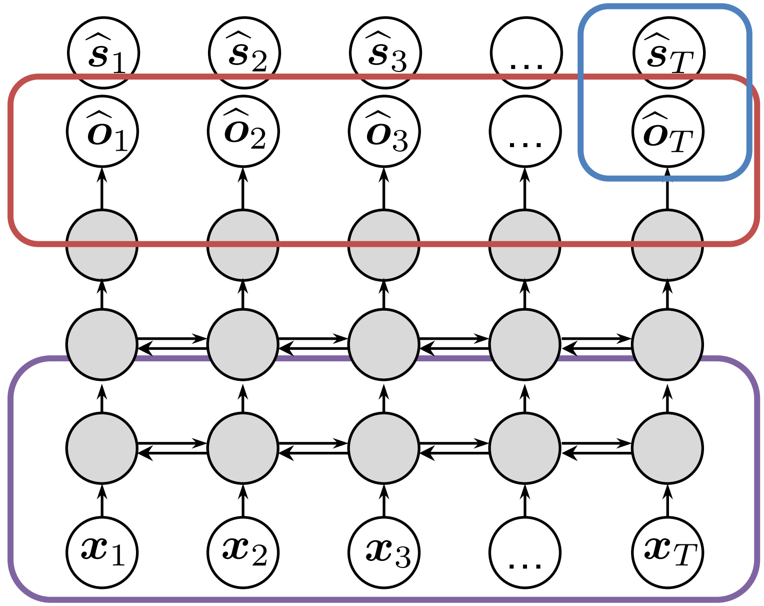
Thesis outline

- Three issues

Issue 2:
Temporal dependency?

Issue 1:
Joint modeling?

Issue 3:
Frame-by-frame
processing is efficient?



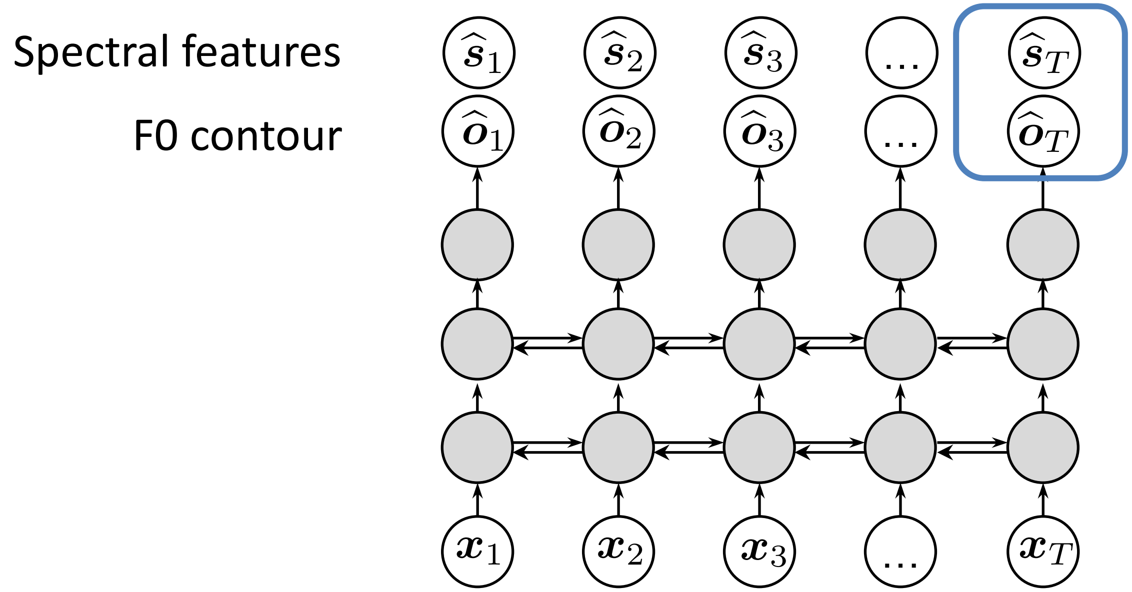
...	1	1	...	1	1	1	...	1	2	...
...			...			*	...	*		...
...	ツ	ツ	...	ツ	ツ	ギ	...	ギ	ワ	...
...	ts	ts	...	u	u	g	...	i	w	...
...	162	163	...	194	195	196	...	227	228	...

T frames

INTRODUCTION

Thesis outline (Chapter 4)

❑ On issue 1: Joint modeling of F0 and spectral features?



- Investigation using highway networks
  - Spectral features are prioritized
  - Different input/hidden features for F0 and spectral

Novel analysis

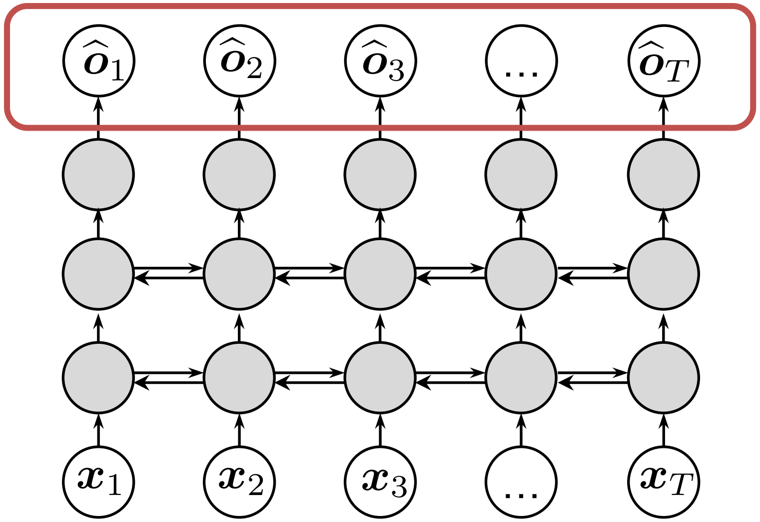
× Sub-optimal for F0 modeling

✓ Only F0 as target

INTRODUCTION

Thesis outline (Chapter 5-6)

❑ On issue 2: Temporal dependency?



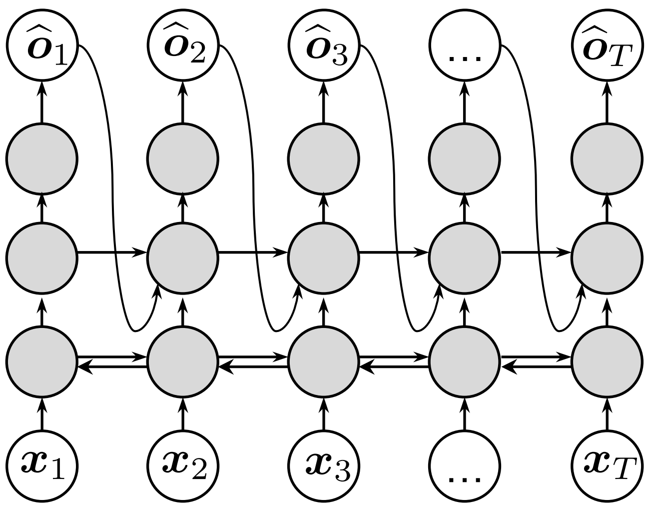
- Evidence from random sampling
- × RNN ignores temporal dependency

INTRODUCTION

Thesis outline (Chapter 5-6)

❑ On issue 2: Temporal dependency?

DAR

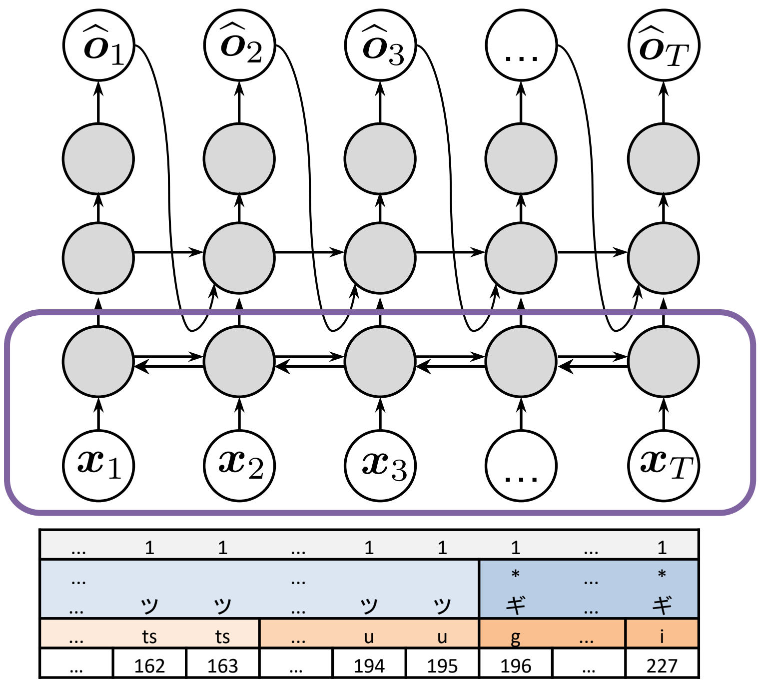


- ✓ Shallow autoregressive model (SAR)
  - Short-term dependency
- ✓ Deep autoregressive model (DAR)
  - Longer dependency & best results & random sampling

Novel models & interpretations

Thesis outline (Chapter 7)

- On issue 3: Frame-by-frame processing?



× Inefficient

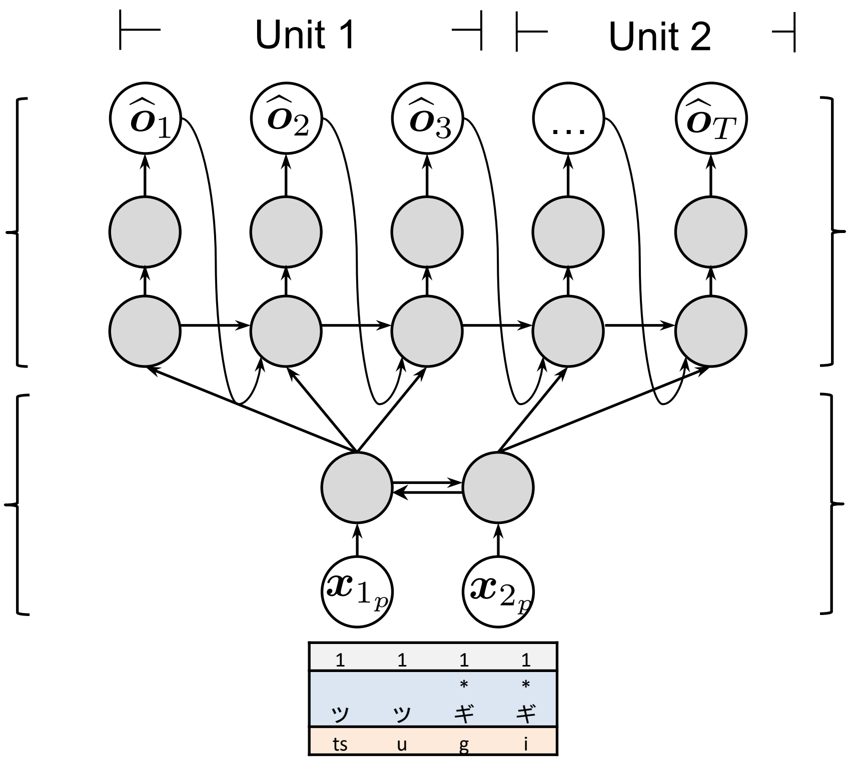
INTRODUCTION

Thesis outline (Chapter 7)

- On issue 3: Frame-by-frame processing?

Stage 1:
F0 contour modeling

Stage 2:
Linguistic linking



Unsupervised
learning

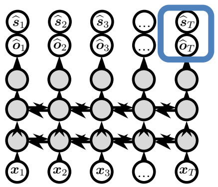
Fast supervised
learning

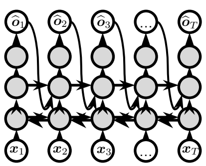
✓ Two-stage model: efficient & interpretable & multi-level   Novel model

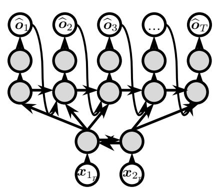
INTRODUCTION

Thesis outline

Chapter 1-3	Preliminary	
Chapter 4	Issue 1 Joint modeling for F0?	Highway networks Tools for analysis
Chapter 5 Chapter 6	Issue 2 Temporal dependency?	SAR & extended SAR DAR & techniques
Chapter 7	Issue 3 Frame-by-frame?	Two-stage F0 model stage1: frame-by-frame stage2: unit-by-unit
Chapter 8	Summary	







INTRODUCTION

Updated results

Chapter 1-3	Preliminary		
Chapter 4	Issue 1 Joint modeling for F0?		
Chapter 5 Chapter 6	Issue 2 Temporal dependency?	SAR + log area ratio Additional listening tests	★
Chapter 7	Issue 3 Frame-by-frame?	Listening test	★
Chapter 8	Summary		

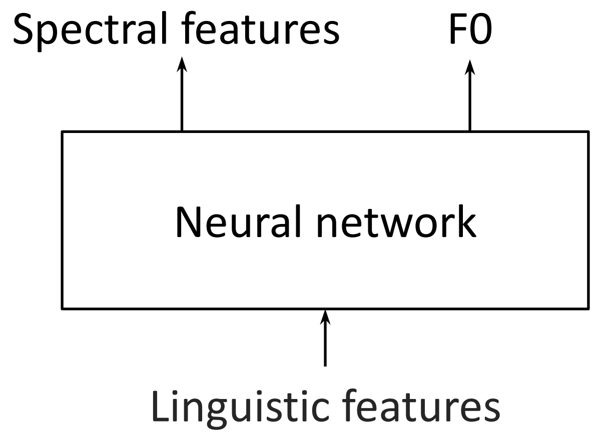
Contents

- Introduction
- Issue 1: joint modeling of F0 and spectral features
- Issues and methods
- Summary

ISSUE 1: JOINT LEARNING FOR F0?

Motivation

- Common approach [5, 7, 11]



- Joint (multi-task) learning
  - ? Beneficial for both targets
  - ? Sharing hidden features

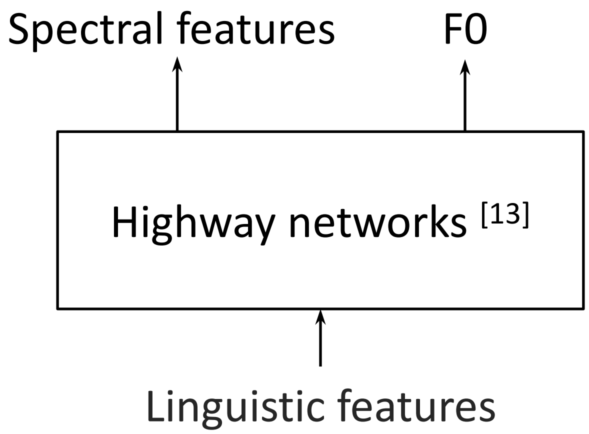
? True or not?
More evidence?

- Empirical results against joint learning [5, 12]

[5] H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In Proc. ICASSP, pages 7962–7966, 2013.
[7] Y. Fan, Y. Qian, F. Xie, and F. K. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proc. Interspeech, pages 1964–1968, 2014.
[11] H. Zen and A. Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In Proc. ICASSP, pages 3844–3848, 2014.
[12] S. Kang and H. Meng. Statistical parametric speech synthesis using weighted multi-distribution deep belief network. In Proc. Interspeech, pages 1959–1963, 2014.

ISSUE 1: JOINT LEARNING FOR F0?

Method



- Joint (multi-task) learning
  - ? Beneficial for both targets
  - ? Sharing hidden features

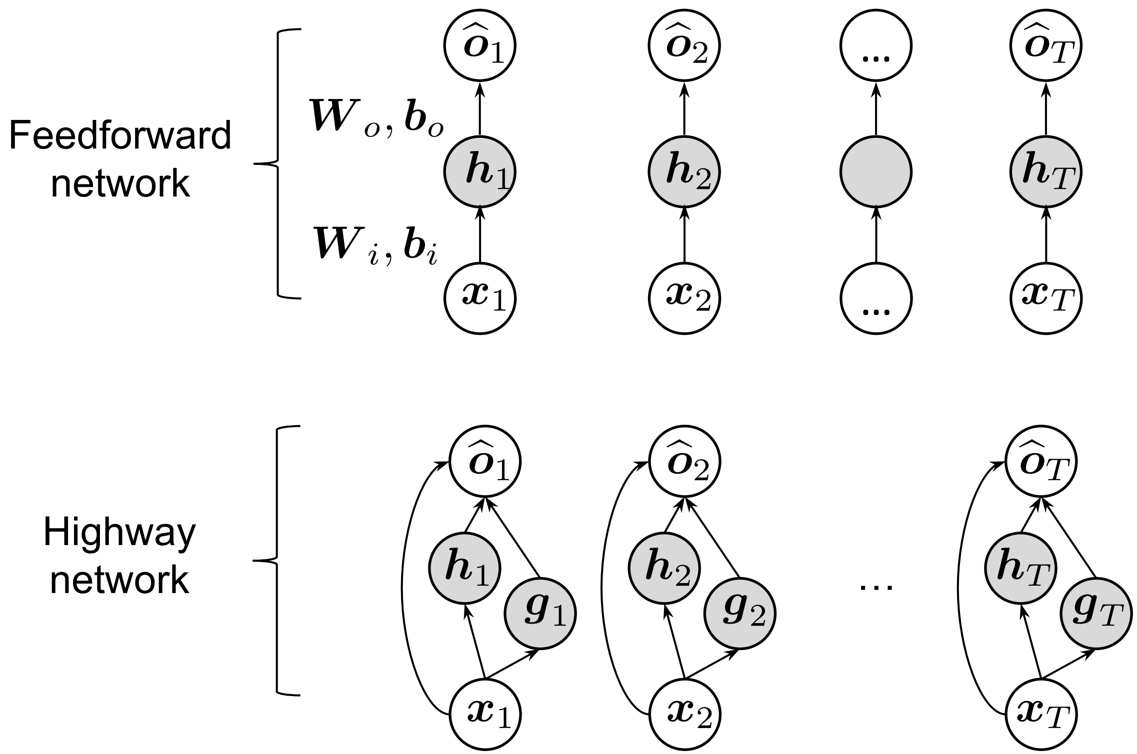
- Model and tools:
  - Highway network [13]
  - Histogram & sensitivity tools

[13] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. In Proc. Deep Learning Workshop, 2015.

ISSUE 1: JOINT LEARNING FOR F0?

Method

- ❑ Definition of highway network



$$\widehat{\boldsymbol{o}}_t = \boldsymbol{W}_o \boldsymbol{h_t} + \boldsymbol{b}_o.$$

$$\boldsymbol{h}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{b}_i)$$

$$\widehat{\boldsymbol{o}}_t = \boldsymbol{g}_t \odot \boldsymbol{h}_t + (\boldsymbol{1} - \boldsymbol{g}_t) \odot \boldsymbol{x}_t$$

$$\boldsymbol{h}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{b}_i)$$

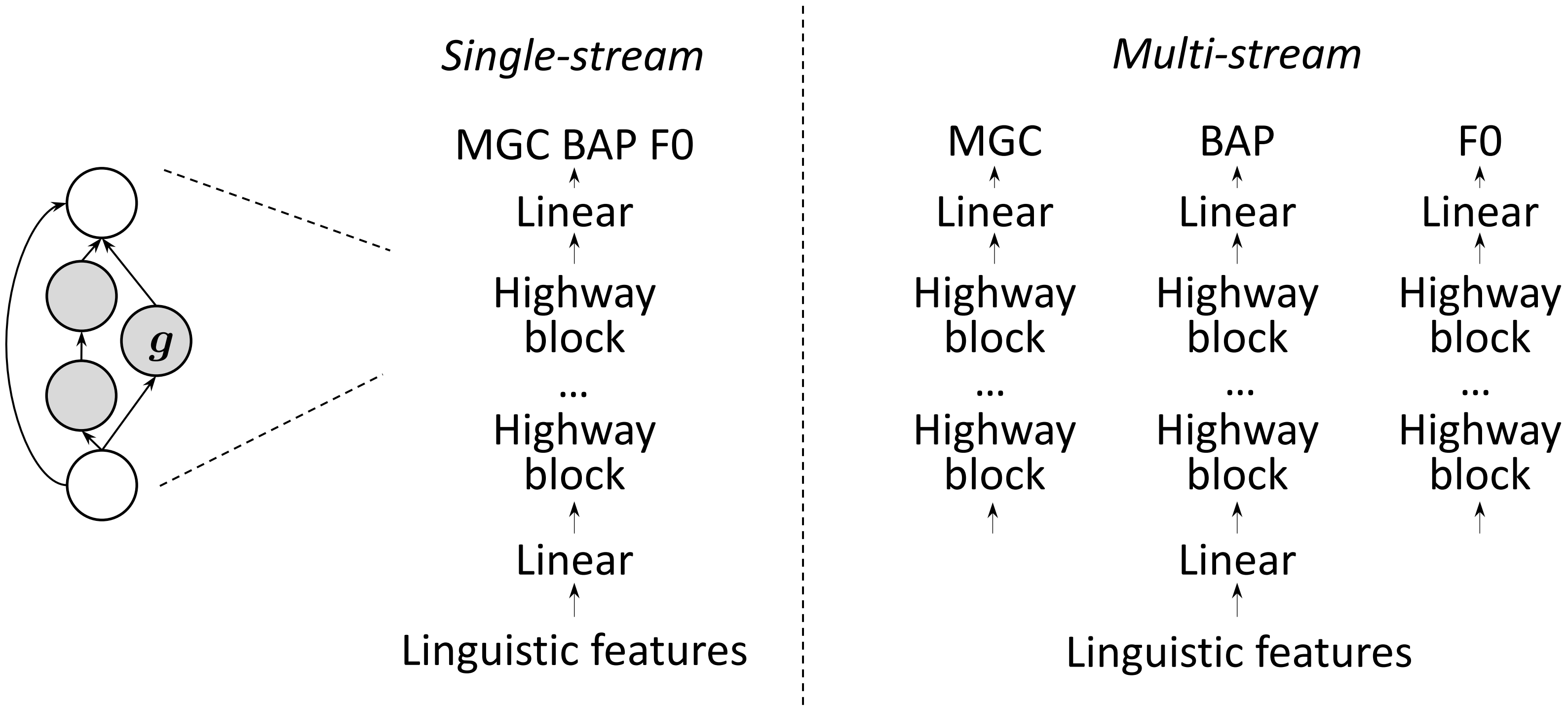
$$\boxed{\boldsymbol{g}_t = \mathrm{sigmoid}(\boldsymbol{W}_g \boldsymbol{x}_t + \boldsymbol{b}_g)}$$

Highway gate vector

ISSUE 1: JOINT LEARNING FOR F0?

Method

- Highway network for acoustic modeling



- Spectral features
  - MGC: Mel-generalized cepstral (MGC) coefficients [14]
  - BAP: band aperiodicity coefficients [15, 16]

[14] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis a unified approach. In Proc. ICSLP, pages 1043–1046, 1994.

[15] H. Kawahara, J. Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, 2001.

[16] H. Zen and T. Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In Proc. Interspeech, pages 93–96, 2005.

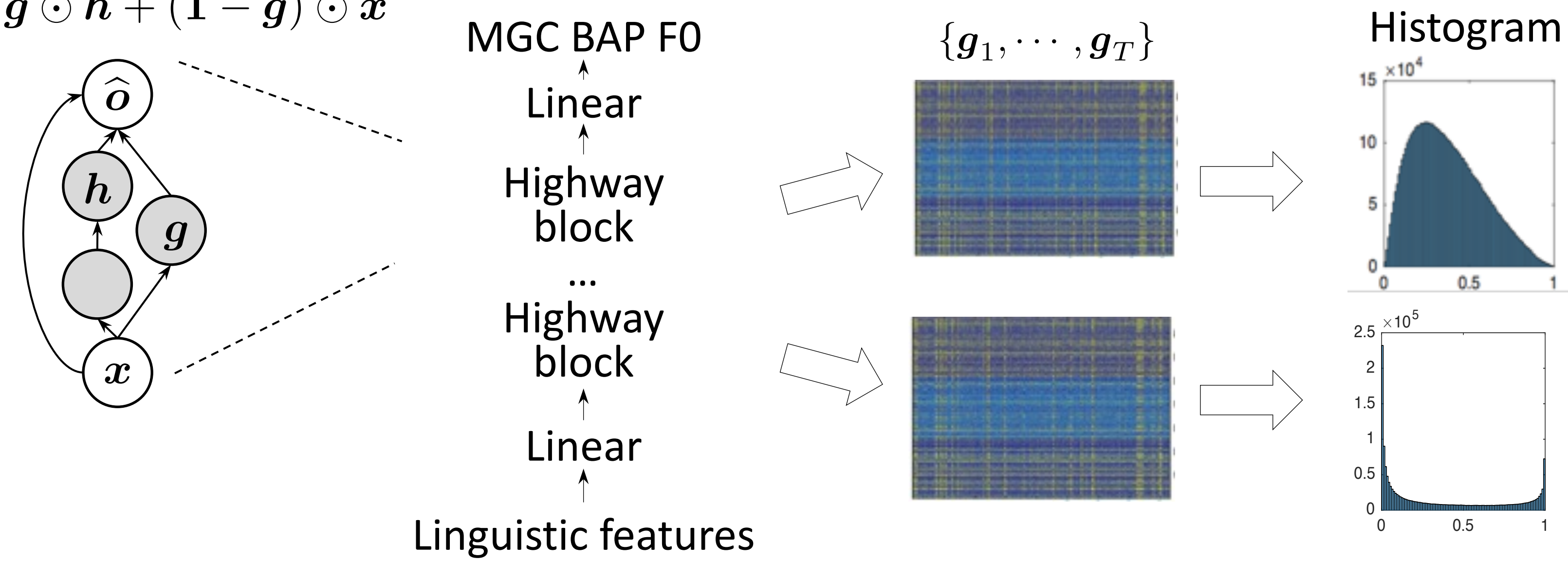
Issue 1: Joint learning for F0?

Method

- Analysis tools

1. Histogram of gate vectors $\boldsymbol{g}$

$$\widehat{\boldsymbol{o}} = \boldsymbol{g} \odot \boldsymbol{h} + (\boldsymbol{1} - \boldsymbol{g}) \odot \boldsymbol{x}$$



$\boldsymbol{g} \approx \boldsymbol{1}$ Non-linear transformation

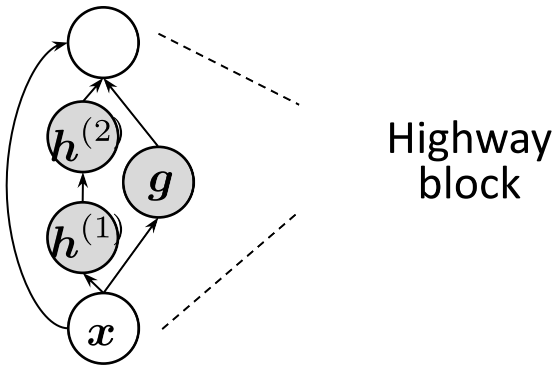
$\boldsymbol{g} \approx \boldsymbol{0}$ No transformation

2. Sensitivity of $\boldsymbol{g}$ to different linguistic features (Sec. 4.3.2)

ISSUE 1: JOINT LEARNING FOR F0?

Experiments

- Configuration
  - Data: English, 16 hours
  - Feature: MGC, BAP, F0 (Interpolated F0 + voicing (U/V))
  - Metric: Root mean square error (RMSE); Correlation coefficients (CORR)
- Three models:
  - Single-stream feedforward network
  - Single-stream highway network
  - Multi-stream highway network

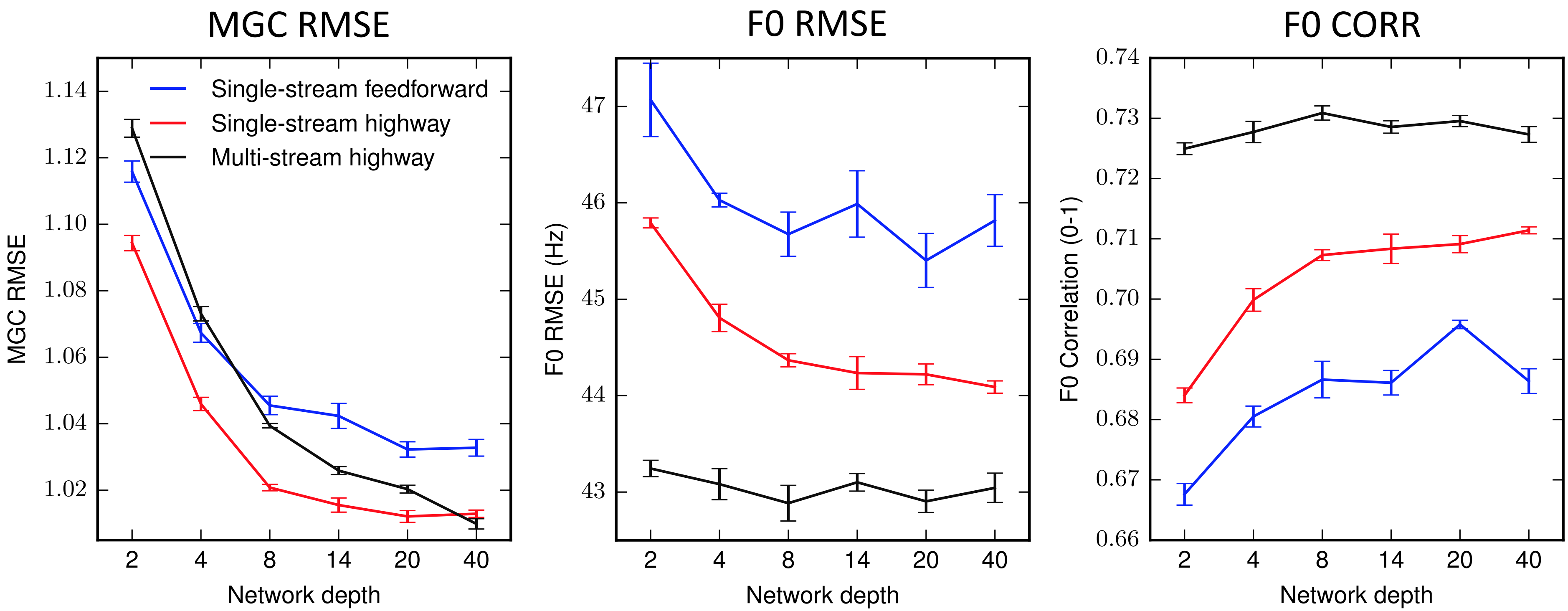


- Two tests:
  1. Fixed layer width, varying network depth
  2. Fixed network depth, varying layer width

ISSUE 1: JOINT LEARNING FOR F0?

Experiments

- Objective results: increasing network depth

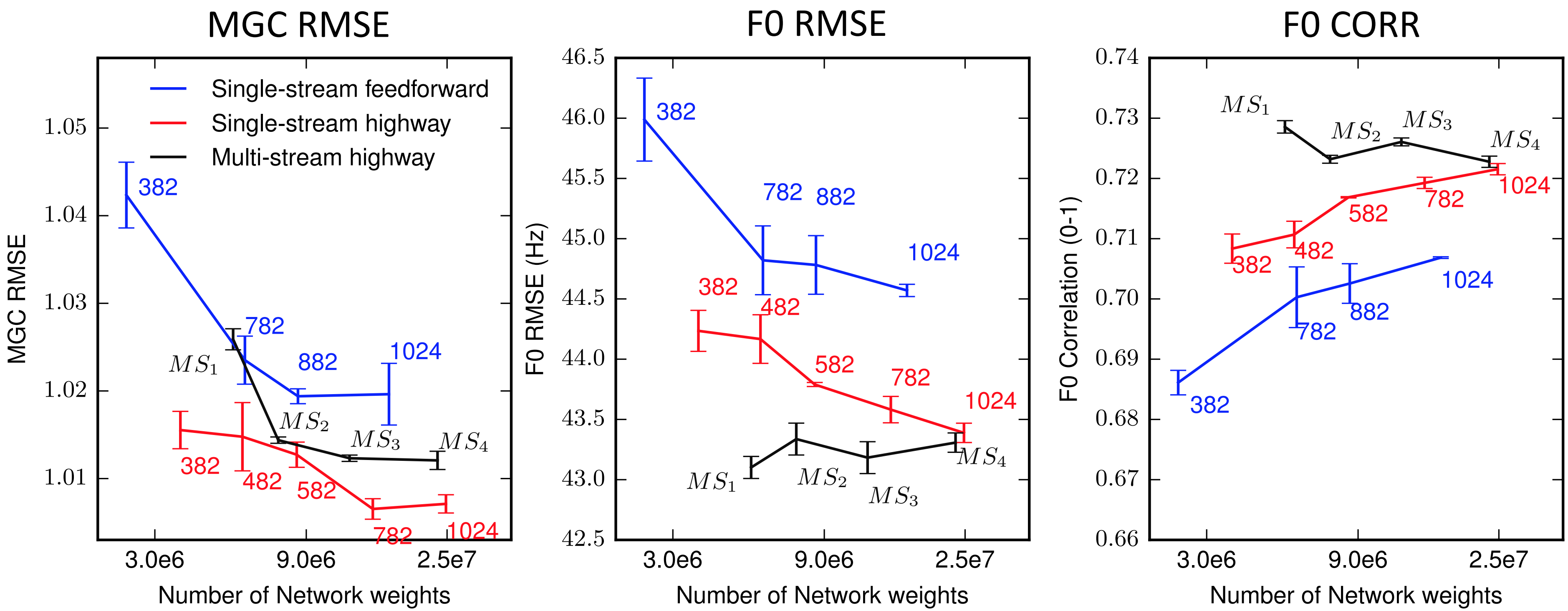


- Network depth: Number of tanh-based transformation layers
- Single-stream network prioritizes MGC?

Issue 1: Joint learning for F0?

Experiments

❑ Objective results: increasing width (depth = 14)



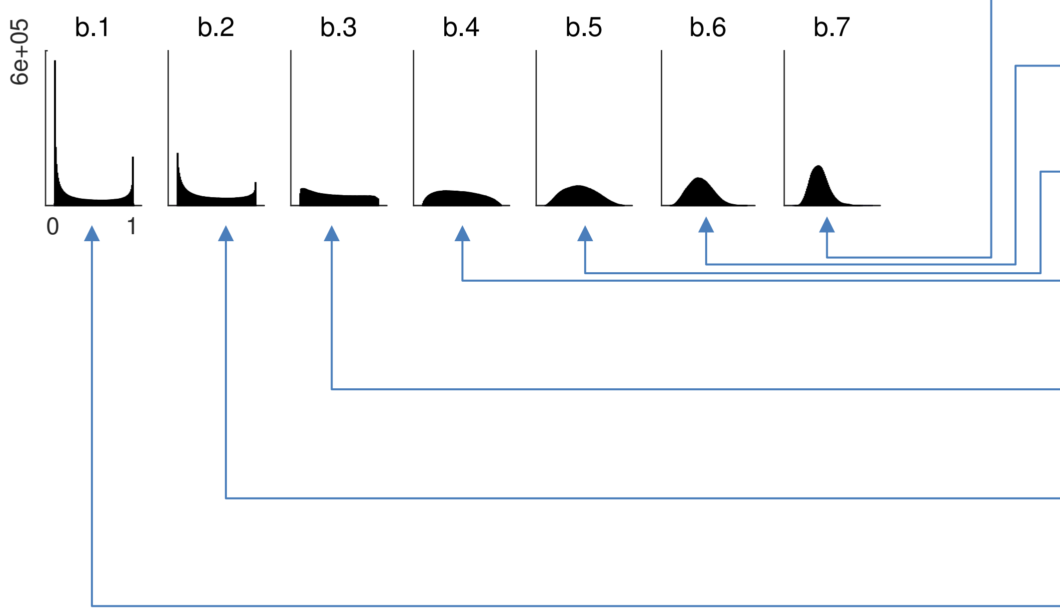
- MS_1 : [MGC 256] – [F0 256]
- MS_2 : [MGC 382] – [F0 256]
- MS_3 : [MGC 512] – [F0 382]
- MS_4 : [MGC 768] – [F0 512]

- Single-stream network prioritizes MGC?

ISSUE 1: JOINT LEARNING FOR F0?

Experiments

- Histogram of g
  - Multi-stream highway (depth 14)

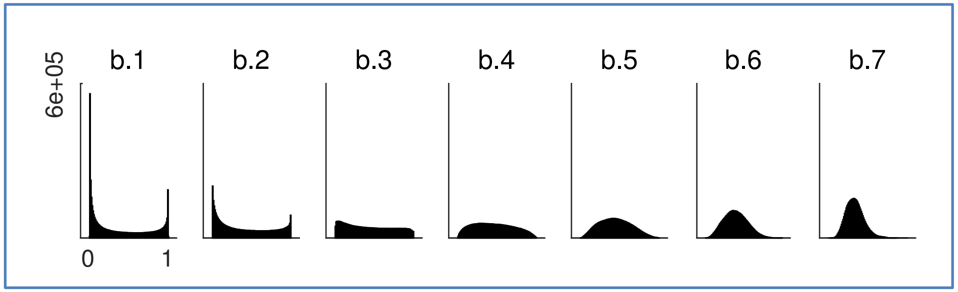


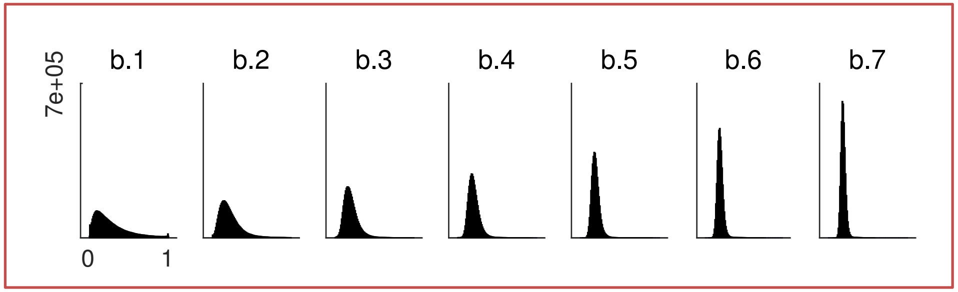
  - $g \approx 1$ Non-linear transformation
    $g \approx 0$ No transformation

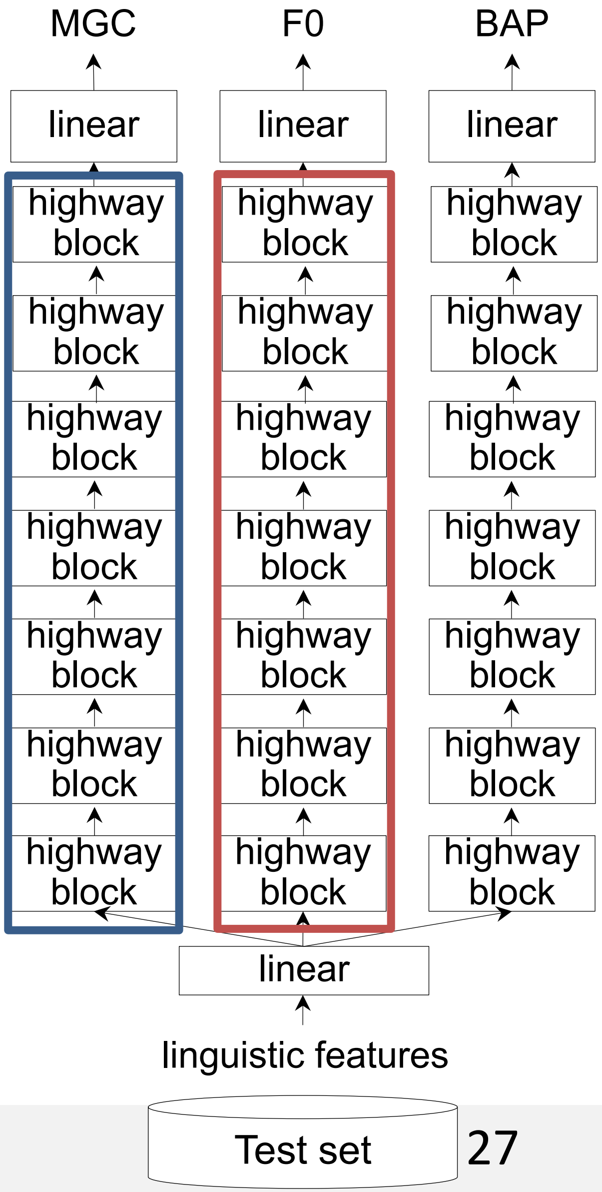
Issue 1: Joint learning for F0?

Experiments

- Histogram of g
  - Multi-stream highway (depth 14, 7 blocks)





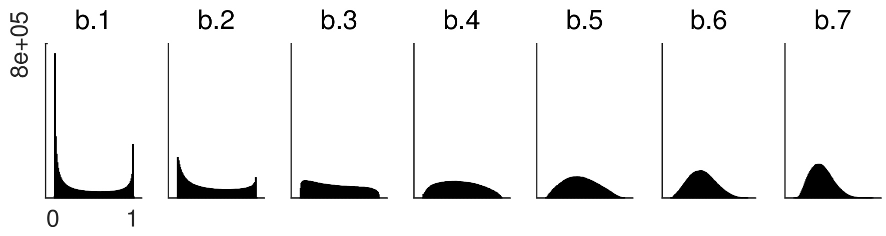


  - $g \approx 1$ Non-linear transformation
    $g \approx 0$ No transformation
  - Different hidden features for MGC and F0

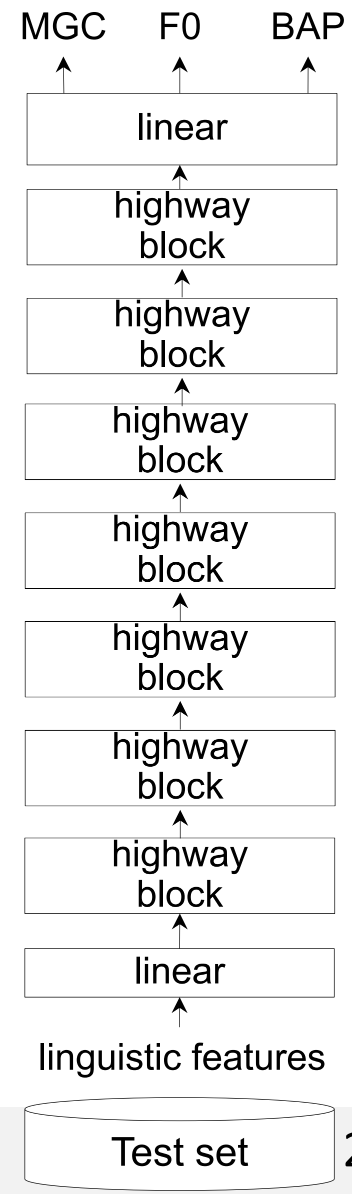
Issue 1: Joint learning for F0?

Experiments

- Histogram of g
  - Single-stream highway (depth 14, 7 blocks)



  - Similar to MGC sub-net in multi-stream highway
  - Single-stream network prioritizes MGC?



Issue 1: Joint learning for F0?

Summary

- ❑ Answer to issue 1

Joint (multi-task) learning	Negative evidence
? Beneficial for both F0 and spectral features	• Joint learning (single-stream network) prioritizes spectral features
? They share hidden features	• They use different hidden features
	• They use different input features (Sec.4.4.3)
	• Results on English and Japanese corpora (Sec.4.5)

NOT for the sake of F0 modeling!

- ❑ Only F0 modeling in the following chapters
- ❑ F0 is useful for MGC modeling? How to do? (slides appendix) ★

CONTENTS

Issue 2: Temporal Dependency?

Motivation

- Baseline RNN model [17]

$$\widehat{\boldsymbol{o}}_{1:T} = \{\widehat{\boldsymbol{o}}_1, \cdots, \widehat{\boldsymbol{o}}_T\}$$

F0 contour

$\uparrow$

Recurrent neural network (RNN)

$\uparrow$

Linguistic features

$$\boldsymbol{x}_{1:T} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$$

- *T*: number of frames (time steps)

[17] R. Fernandez, et. al. Prosody contour prediction with long short- term memory, bi-directional, deep recurrent neural networks. In Proc. Interspeech, pages 2268–2272, 2014.

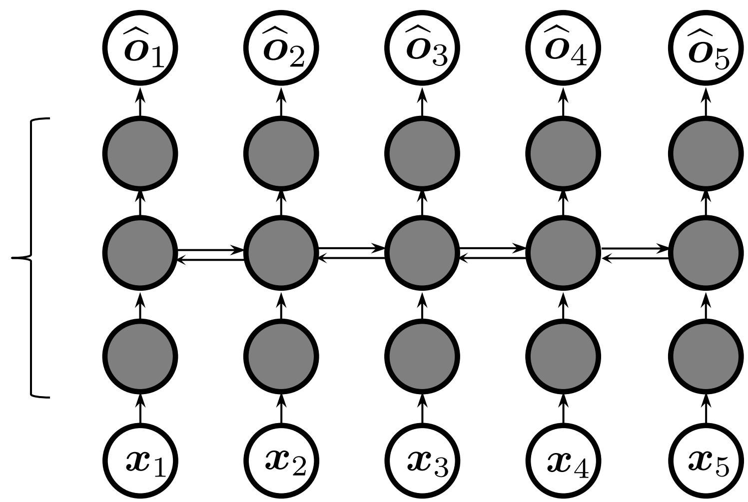
ISSUE 2: TEMPORAL DEPENDENCY?

Motivation

❑ Baseline RNN model

$$\widehat{\boldsymbol{o}}_{1:T} = \{\widehat{\boldsymbol{o}}_1, \cdots, \widehat{\boldsymbol{o}}_T\}$$

$\mathcal{H}_{\boldsymbol{\Theta}}^{(RNN)}(\cdot)$



$$\widehat{\boldsymbol{o}}_t = \mathcal{H}_{\boldsymbol{\Theta}}^{(RNN)}(\boldsymbol{x}_{1:T}, t)$$

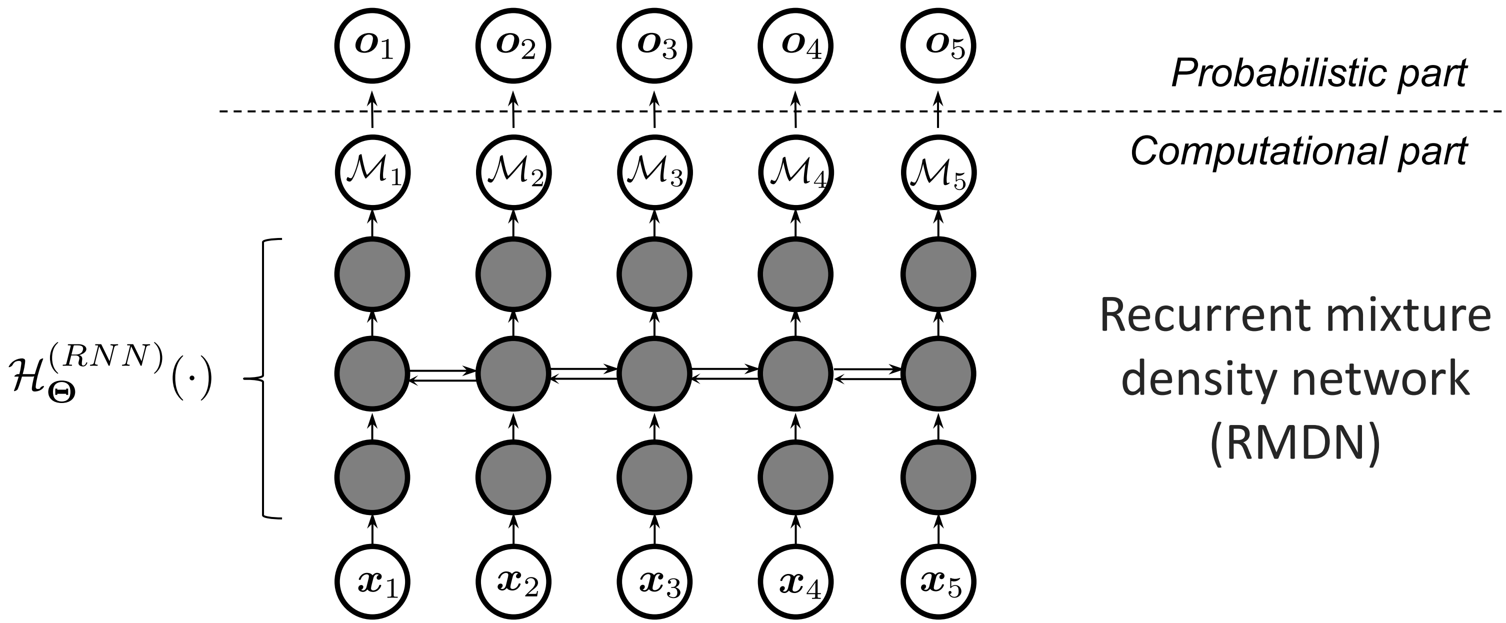
$$\boldsymbol{x}_{1:T} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$$

- Learn the correlation between $\boldsymbol{o}_{t_1}$ and $\boldsymbol{o}_{t_2}$, $t_2 \neq t_1$ **?**

Issue 2: Temporal Dependency?

Motivation

- Baseline RNN model



Recurrent mixture density network (RMDN)

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

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

$$\widehat{\boldsymbol{o}}_t = \arg\max_{\boldsymbol{o}_t} p(\boldsymbol{o}_t|\boldsymbol{x}_{1:T};\boldsymbol{\Theta}^*) = \boldsymbol{\mu}_t$$ Mean-based generation

[18] C. M. Bishop. Mixture Density Networks. Technical report, Aston University, 2004.
[19] C. M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995.
[20] M. Schuster. Better generative models for sequential data problems: Bidirectional recurrent mixture density networks. In Proc. NIPS, pages 589–595, 1999.

Issue 2: temporal dependency?

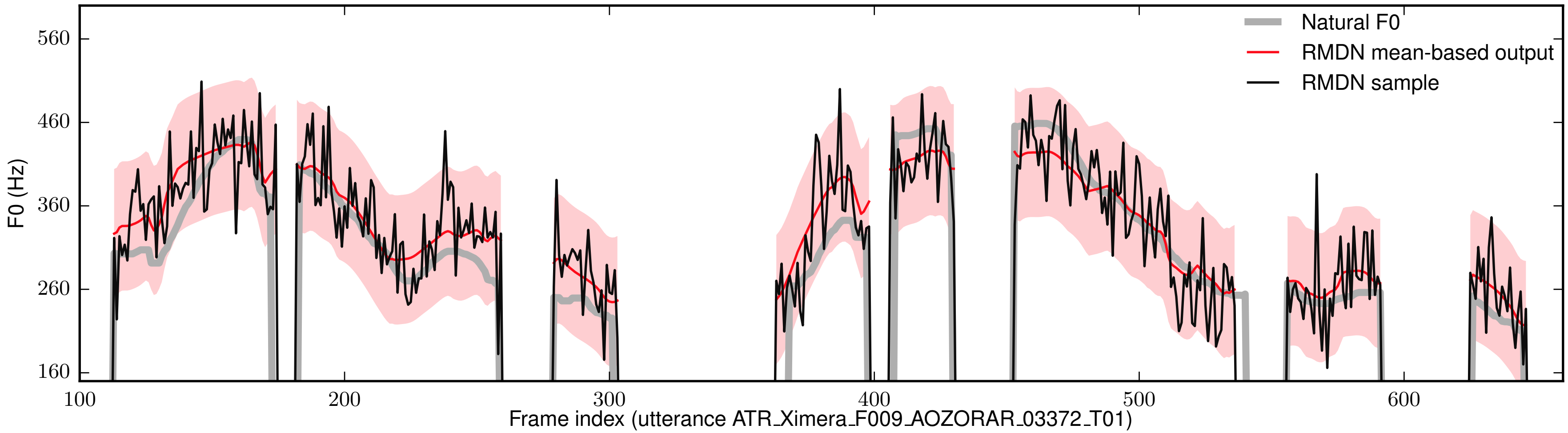
Motivation

❑ Initial answer

Temporal dependency is ignored by RNN/RMDN

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

- Evidence from random sampling



ISSUE 2: TEMPORAL DEPENDENCY?

Motivation

- Initial answer

Temporal dependency is ignored by RNN/RMDN

- Better models?

RNN/RMDN

$$p(\boldsymbol{o}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t)$$

→

Autoregressive (AR) [23] models

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

- Shallow AR models (SAR) & extended SAR
- Deep AR models (DAR)

[23] B. Frey. Graphical Models for Machine Learning and Digital Communication. A Bradford book. Bradford book, 1998.

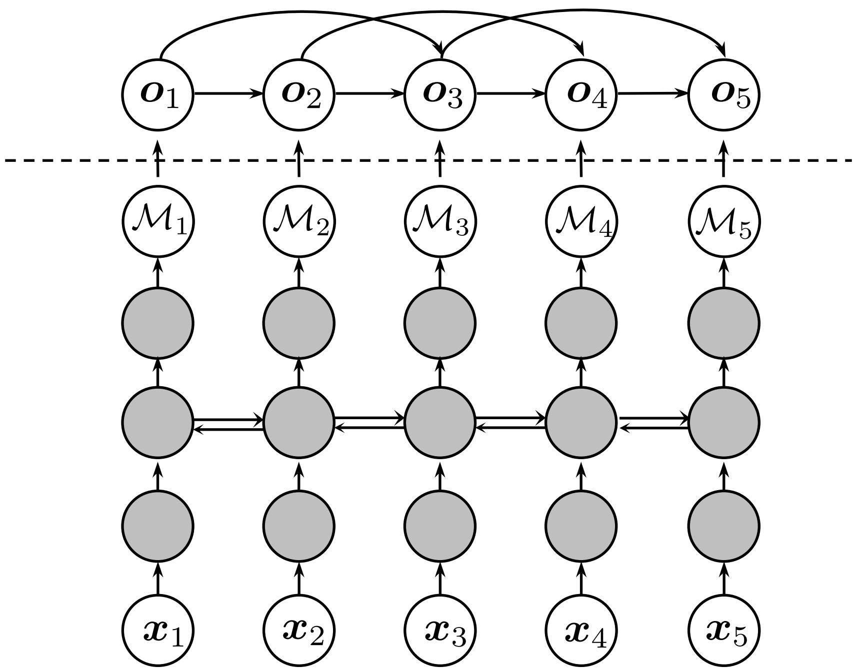
CONTENTS

- Introduction
- Issue 1: joint modeling of F0 and spectral features
- Issue 2: temporal dependency modeling of F0 contours
  - SAR and extension
  - DAR
- Issues and methods
- Summary

ISSUE 2: TEMPORAL DEPENDENCY?

SAR

- Definition

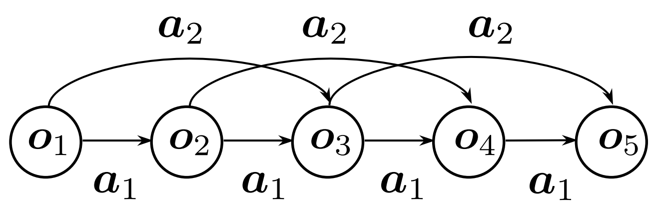


$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\underline{\boldsymbol{o}_{t-K:t-1}}, \boldsymbol{x}_{1:T})$$

ISSUE 2: TEMPORAL DEPENDENCY?

SAR

❑ Definition



K = 2

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

$$= \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t + f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{t-K:t-1}), \boldsymbol{\Sigma}_t)$$

$$f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{t-K:t-1}) = \sum_{k=1}^{K} \boldsymbol{a}_k \odot \boldsymbol{o}_{t-k} + \boldsymbol{b}$$

Trainable parameters

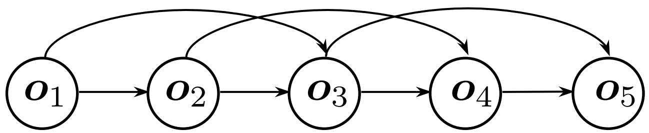
- Time invariant $\boldsymbol{\Psi} = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_K, \boldsymbol{b}\}$
- *K*: hyper-parameter

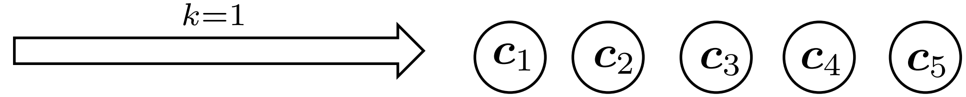
Issue 2: Temporal Dependency?

SAR

- Interpretation 1 (Sec. 5.3)

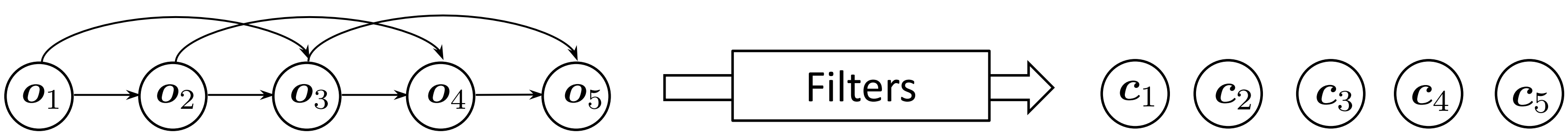
$$\boldsymbol{c}_t = \boldsymbol{o}_t - \sum_{k=1}^{K} \boldsymbol{a}_k \odot \boldsymbol{o}_{t-k}$$





$$\begin{aligned} p(\boldsymbol{o}_{1:T} | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) &= \prod_{t=1}^{T} p(\boldsymbol{o}_t | \boldsymbol{o}_{t-K:t-1}, \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) \\ &= \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t + f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{t-K:t-1}), \boldsymbol{\Sigma}_t) \\ &= \prod_{t=1}^{T} p_c(\boldsymbol{c}_t | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}) \end{aligned}$$

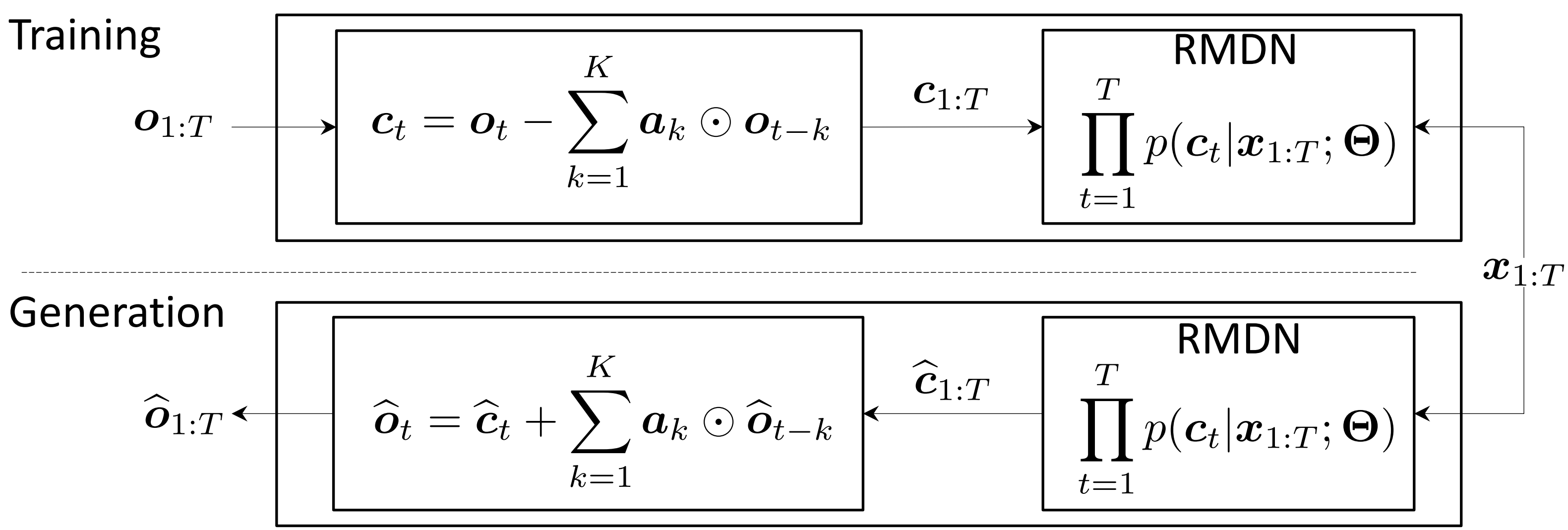
- Interpretation 2 (Sec. 5.3)



ISSUE 2: TEMPORAL DEPENDENCY?

SAR

- Interpretation (Sec. 5.3)
  - SAR = Linear transformation + RMDN

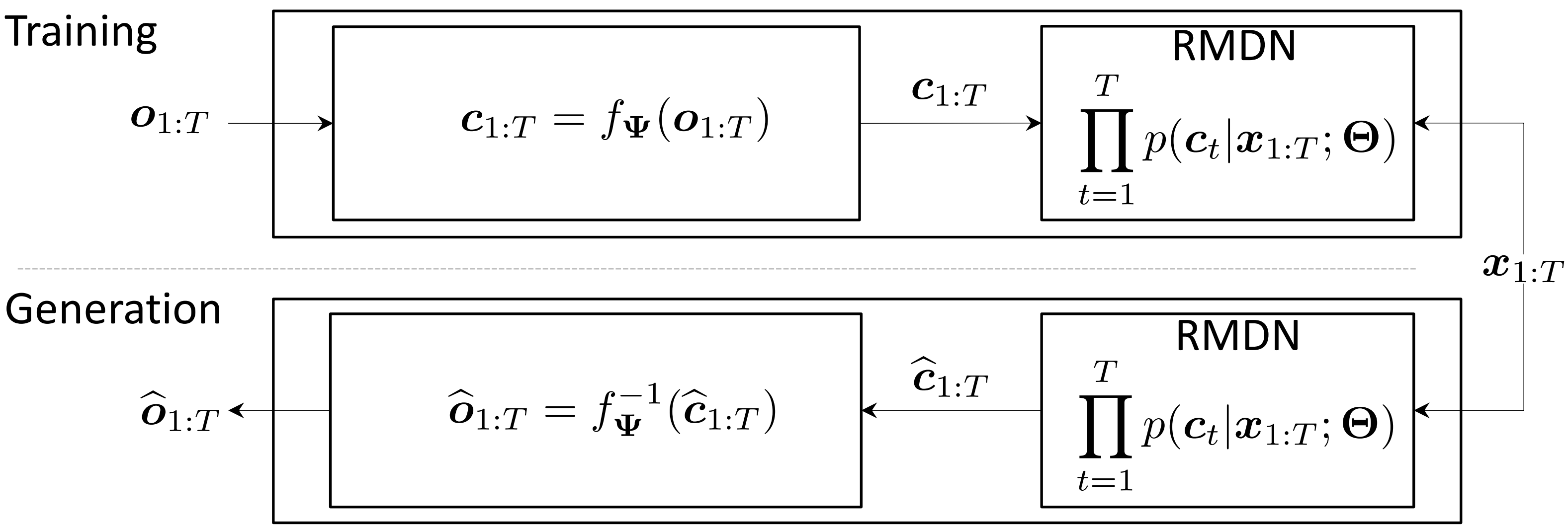


Issue 2: temporal dependency?

Extended SAR (eSAR)

❑ Motivation

- SAR -> Non-linear transformation + RMDN?



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

- Yes, if $f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T})$ is invertible and `simple'

ISSUE 2: TEMPORAL DEPENDENCY?

Extended SAR (eSAR)

❑ Definition

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

	SAR	eSAR
$\boldsymbol{c}_{1:T} = f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T})$	$\boldsymbol{c}_t = \boldsymbol{o}_t - \sum_{k=1}^{K} \boldsymbol{a}_k \odot \boldsymbol{o}_{t-k}$	$\boldsymbol{c}_t = \boldsymbol{o}_t - \mathrm{RNN}_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:t-1}, t)$
$\widehat{\boldsymbol{o}}_{1:T} = f_{\boldsymbol{\Psi}}^{-1}(\widehat{\boldsymbol{c}}_{1:T})$	$\widehat{\boldsymbol{o}}_t = \widehat{\boldsymbol{c}}_t + \sum_{k=1}^{K} \boldsymbol{a}_k \odot \widehat{\boldsymbol{o}}_{t-k}$	$\widehat{\boldsymbol{o}}_t = \widehat{\boldsymbol{c}}_t + \mathrm{RNN}_{\boldsymbol{\Psi}}(\widehat{\boldsymbol{o}}_{1:t-1}, t)$
$\det\frac{\partial f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T})}{\partial \boldsymbol{o}_{1:T}}$	$\det\frac{\partial f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T})}{\partial \boldsymbol{o}_{1:T}} = 1$	$\det\frac{\partial f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T})}{\partial \boldsymbol{o}_{1:T}} = 1$

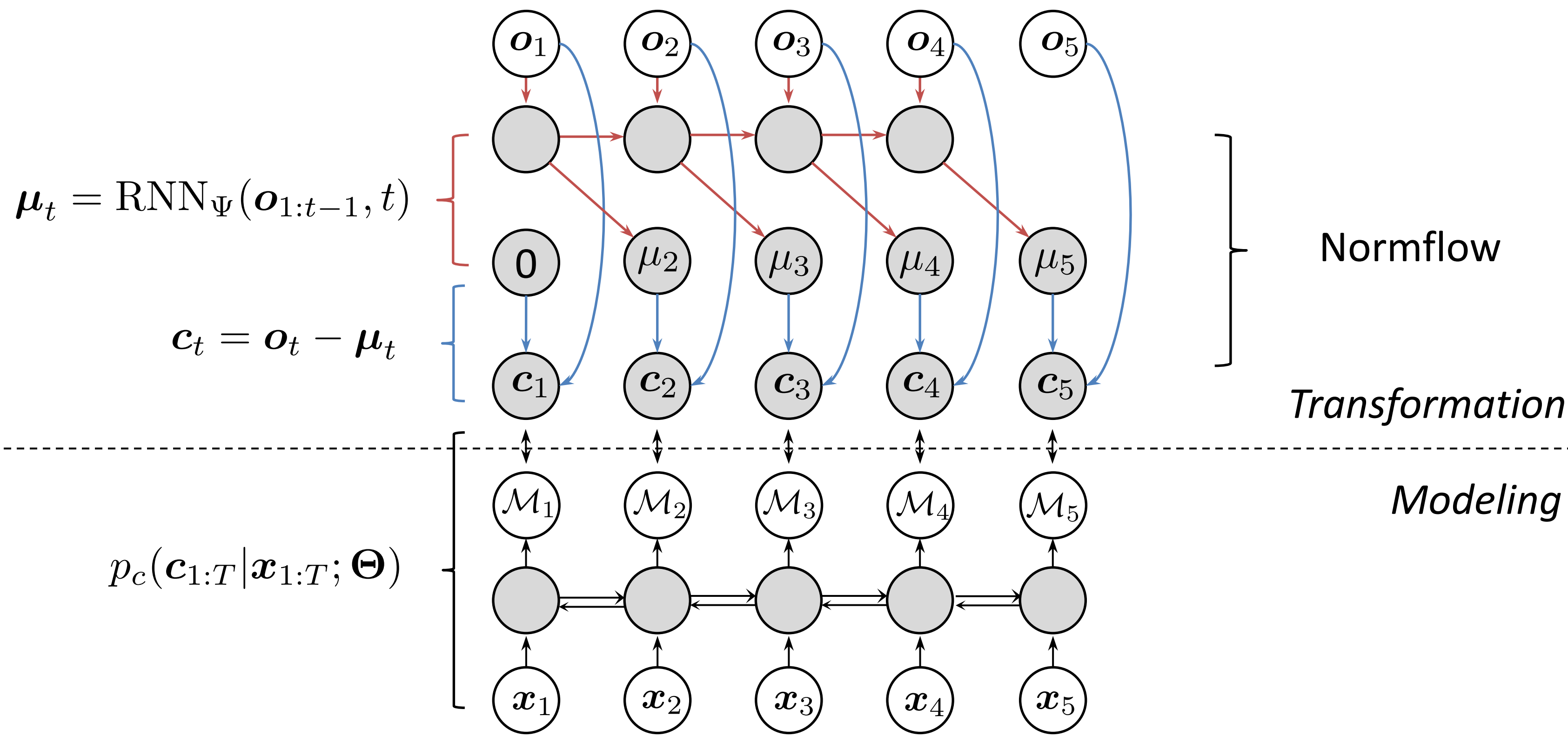
- Volume-preserving [24]

[24] J. M. Tomczak and M. Welling. Improving variational auto-encoders using convex combination linear inverse autoregressive flow. In Proc. Benelearn, pages 162–194, 2017.

ISSUE 2: TEMPORAL DEPENDENCY?

Extended SAR (eSAR)

- Implementation

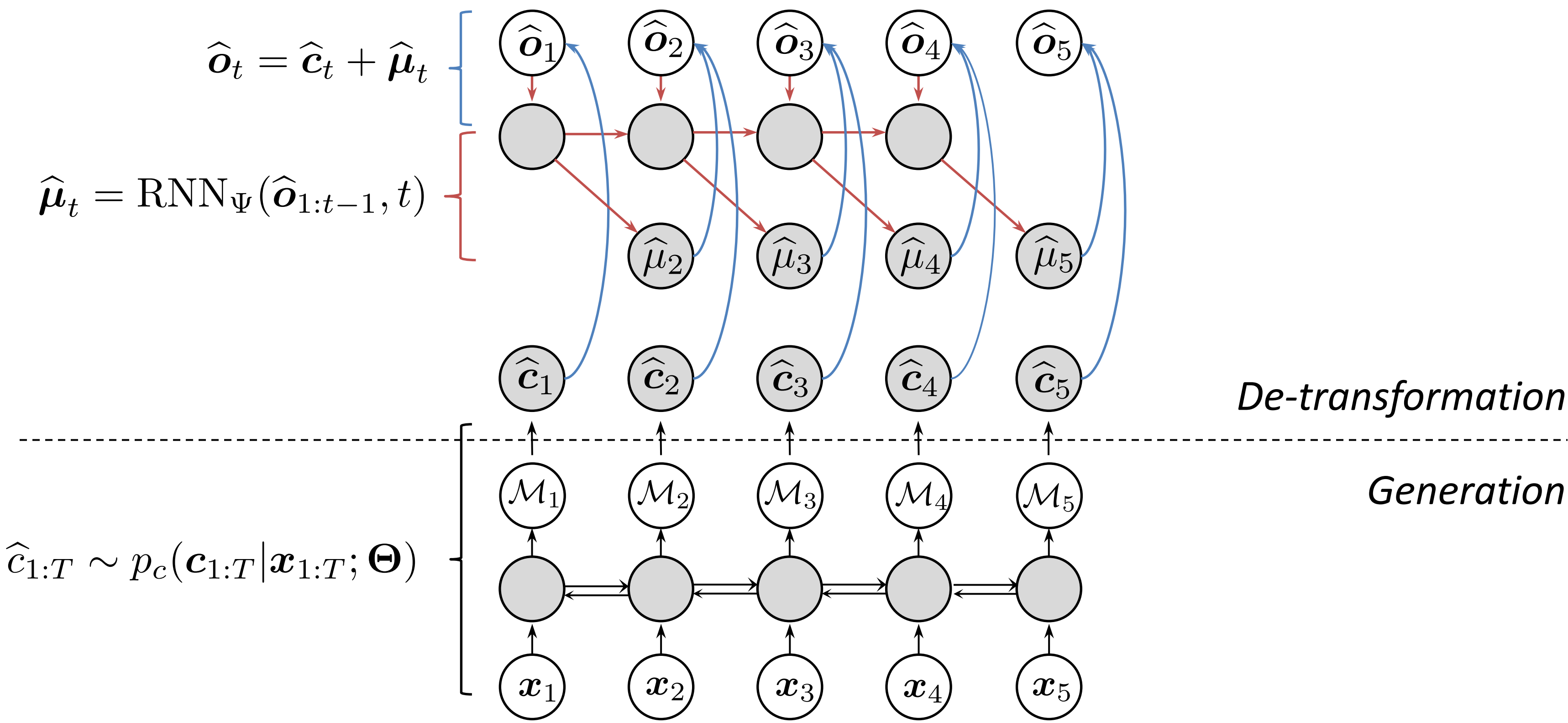


$$p_o(\boldsymbol{o}_{1:T} | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = p_c(\boldsymbol{c}_{1:T} = f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T}) | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta})$$

ISSUE 2: TEMPORAL DEPENDENCY?

Extended SAR (eSAR)

- Implementation

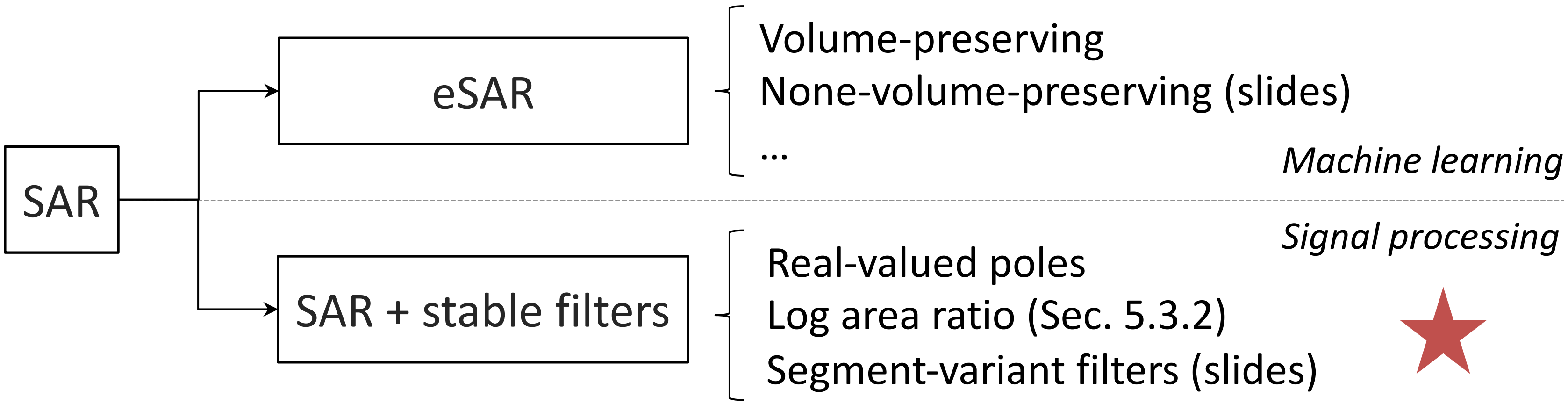


$$p_o(\boldsymbol{o}_{1:T} | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}, \boldsymbol{\Psi}) = p_c(\boldsymbol{c}_{1:T} = f_{\boldsymbol{\Psi}}(\boldsymbol{o}_{1:T}) | \boldsymbol{x}_{1:T}; \boldsymbol{\Theta})$$

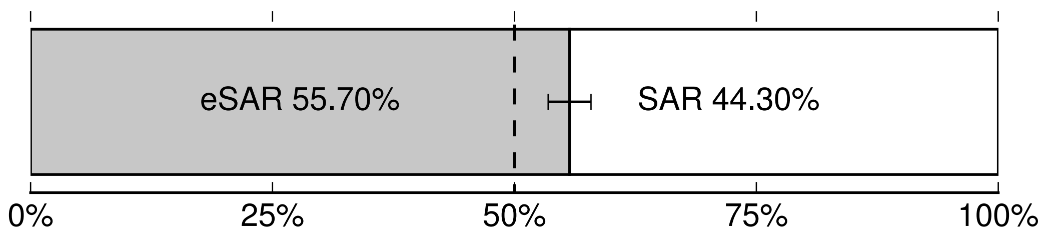
ISSUE 2: TEMPORAL DEPENDENCY?

Summary of SAR

- Theoretically appealing



- Performance for F0 modeling (model details later)



- Performance for MGC modeling
  - Better than RNN/RMDN [25, 26]

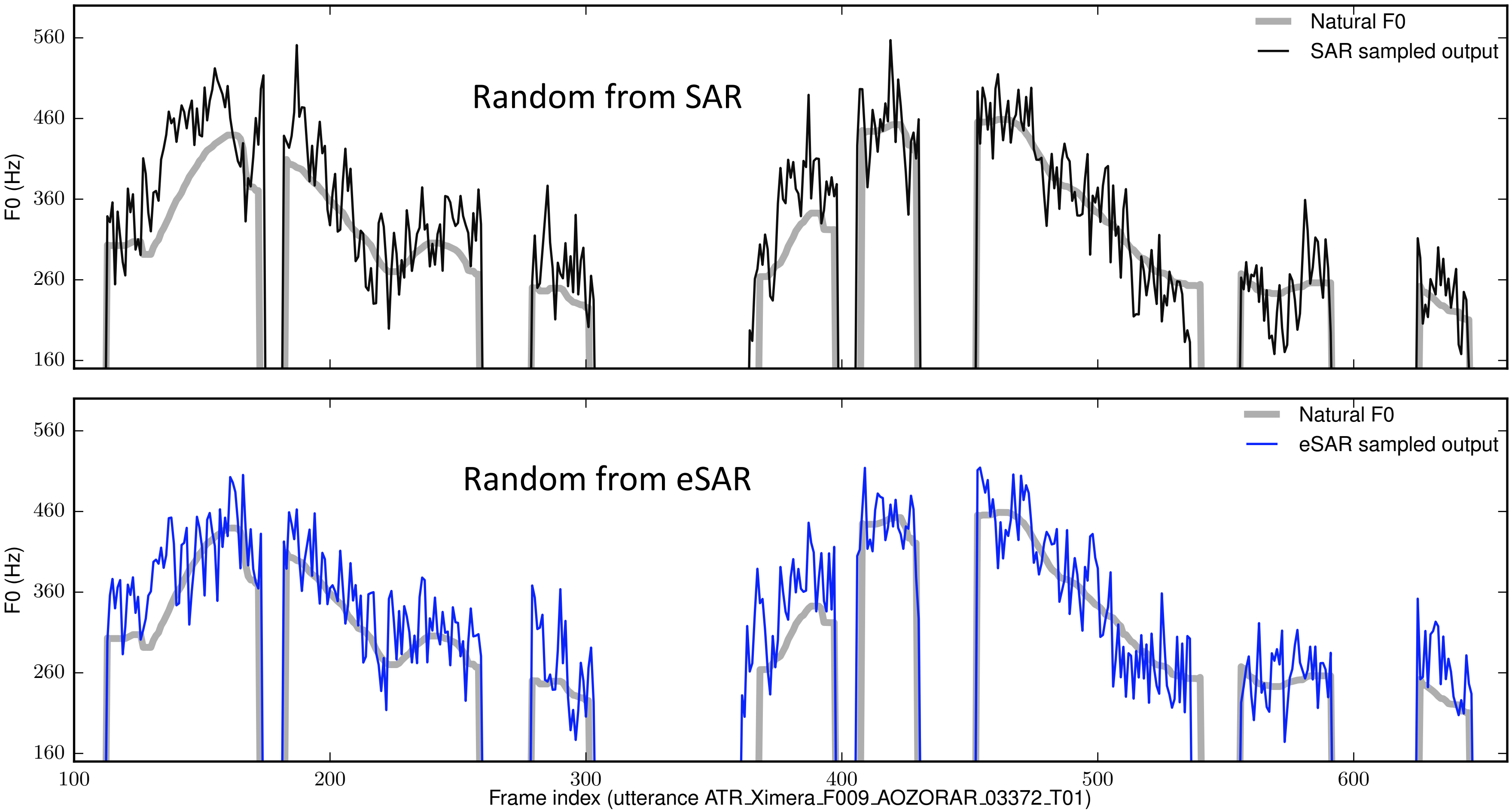
[25] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi. A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In Proc. ICASSP, pages 4804–4808, 2018.
[26] X. Wang, S. Takaki, and J. Yamagishi. An autoregressive recurrent mixture density network for para- metric speech synthesis. In Proc. ICASSP, pages 4895–4899, 2017.

CONTENTS

ISSUE 2: TEMPORAL DEPENDENCY?

DAR

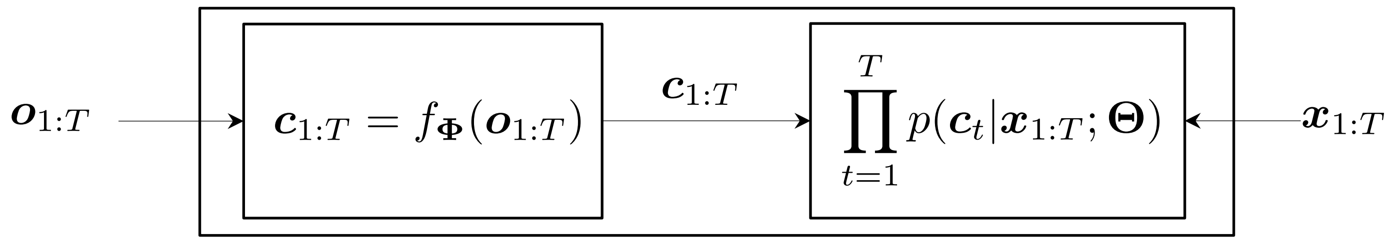
- Motivation
  - Are SAR and eSAR sufficiently good?



ISSUE 2: TEMPORAL DEPENDENCY?

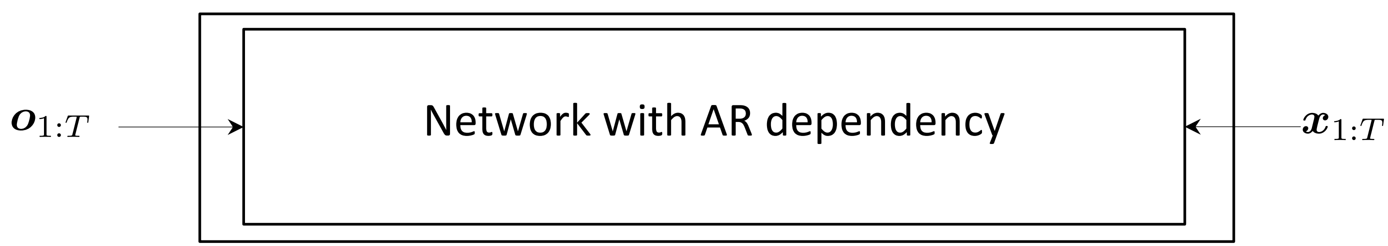
DAR

- Motivation
  - Random sampling on SAR and eSAR?



  - SAR: linear $f_{\boldsymbol{\Phi}}(\cdot)$ (Sec. 6.1)
  - eSAR: non-linear but a special form

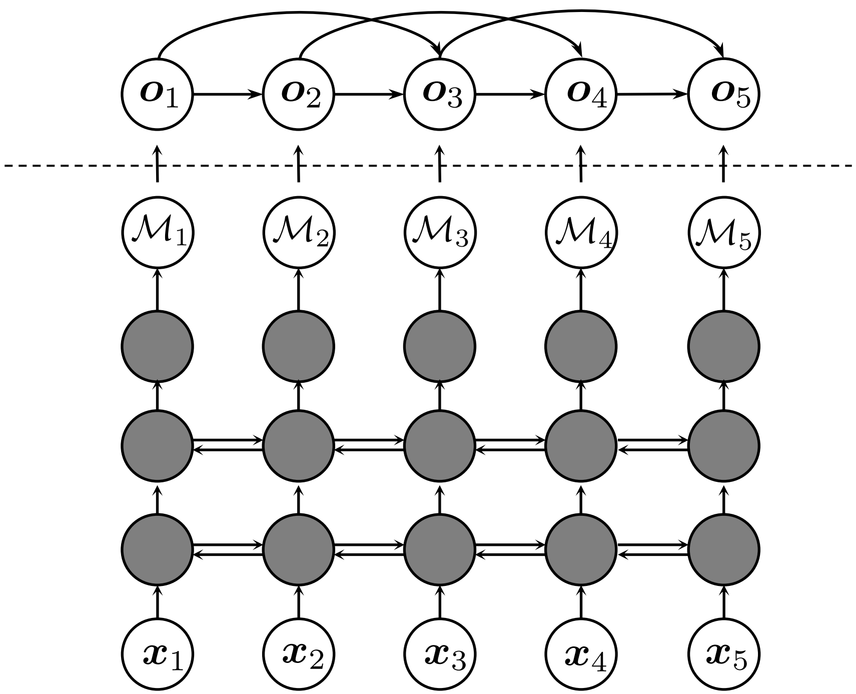
- Non-linear and non-invertible AR transformation?



ISSUE 2: TEMPORAL DEPENDENCY?

DAR

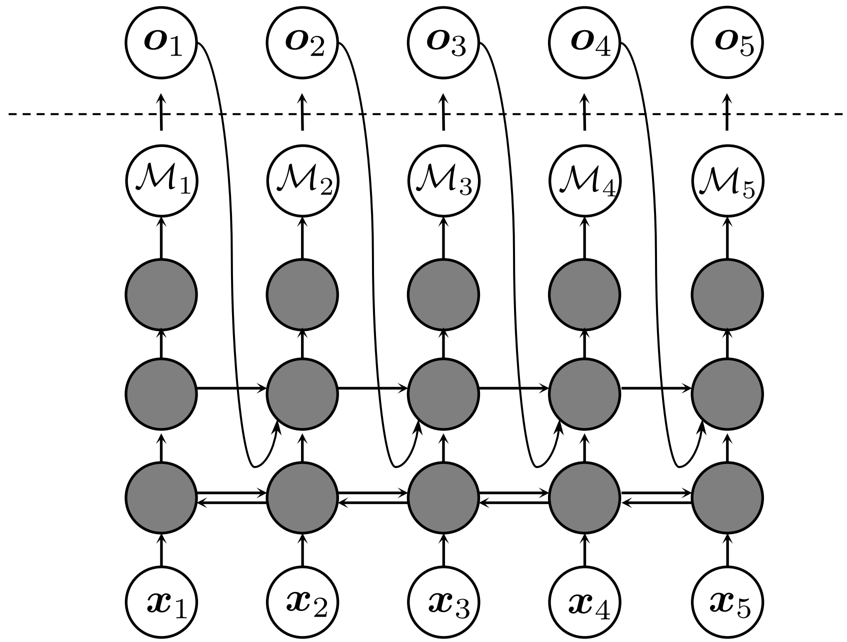
- Definition



ISSUE 2: TEMPORAL DEPENDENCY?

DAR

❑ Definition



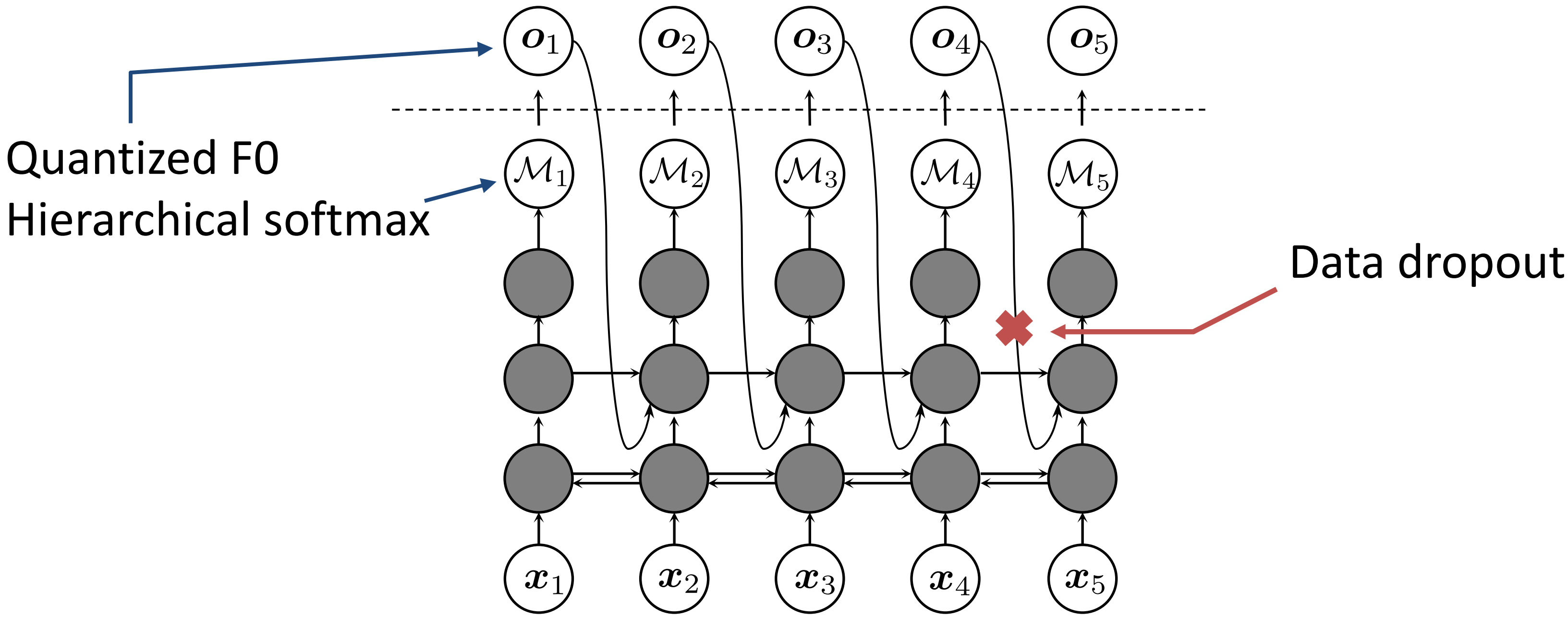
$$p(\boldsymbol{o}_{1:T}|\boldsymbol{x}_{1:T};\boldsymbol{\Theta}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t|\boldsymbol{o}_{1:t-1}, \boldsymbol{x}_{1:T};\boldsymbol{\Theta})$$

- DAR is more general than SAR (Sec 6.2.2 & slides, toy examples)
  1. Longer-time dependency
  2. Non-linear dependency

ISSUE 2: TEMPORAL DEPENDENCY?

DAR

- Implementation (Sec. 6.3)



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

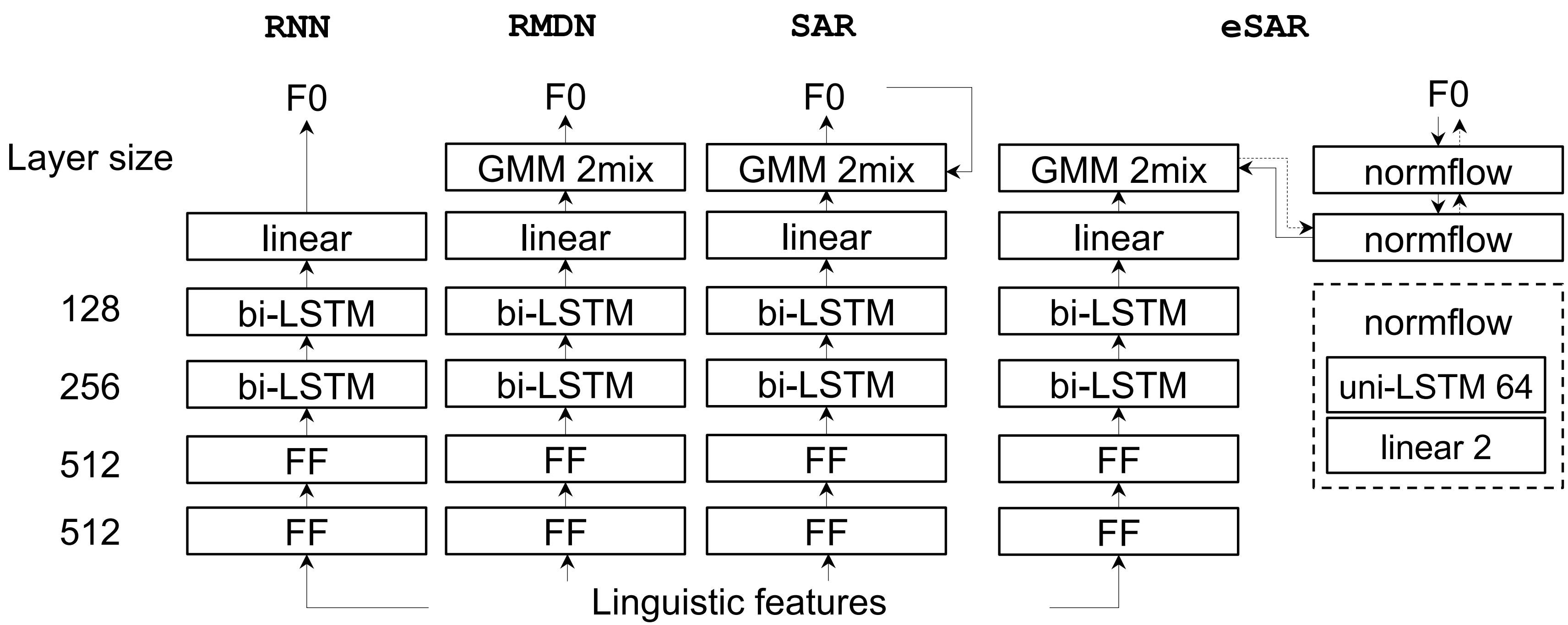
ISSUE 2: TEMPORAL DEPENDENCY?

Experiment

❑ Data and features

- Data: Japanese, 48 hours
- Feature: F0 (interpolated F0 value 1 dim + U/V 1 dim)

❑ Models

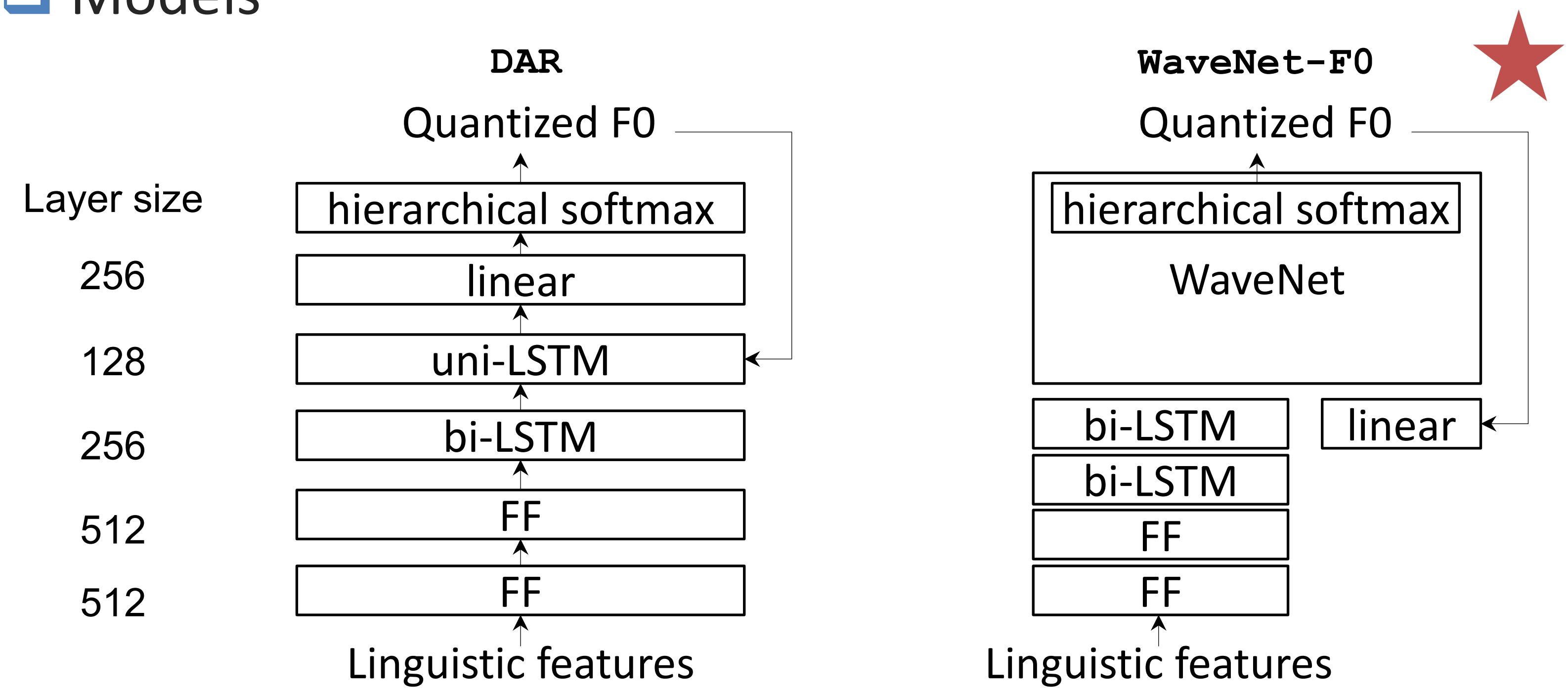


❖ FF: feedforward with tanh-activation function

ISSUE 2: TEMPORAL DEPENDENCY?

Experiment

- Data and features
  - Data: Japanese, 48 hours
  - Feature: F0 (interpolated F0 value 1 dim + U/V 1 dim)
  - Feature: quantized F0 (256 quantization bins)
- Models

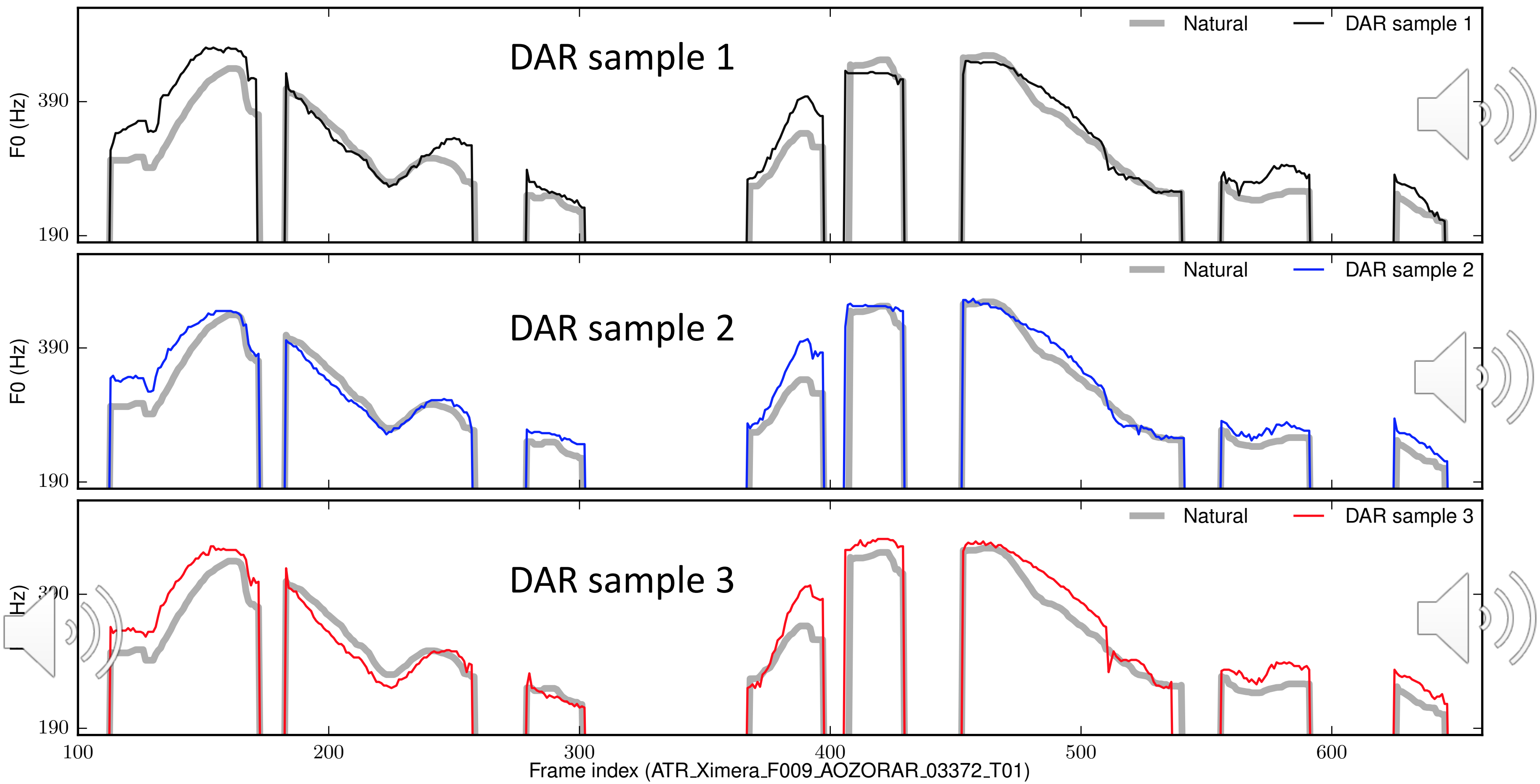


❖ FF: feedforward with tanh-activation function

Issue 2: temporal dependency?

Experiment

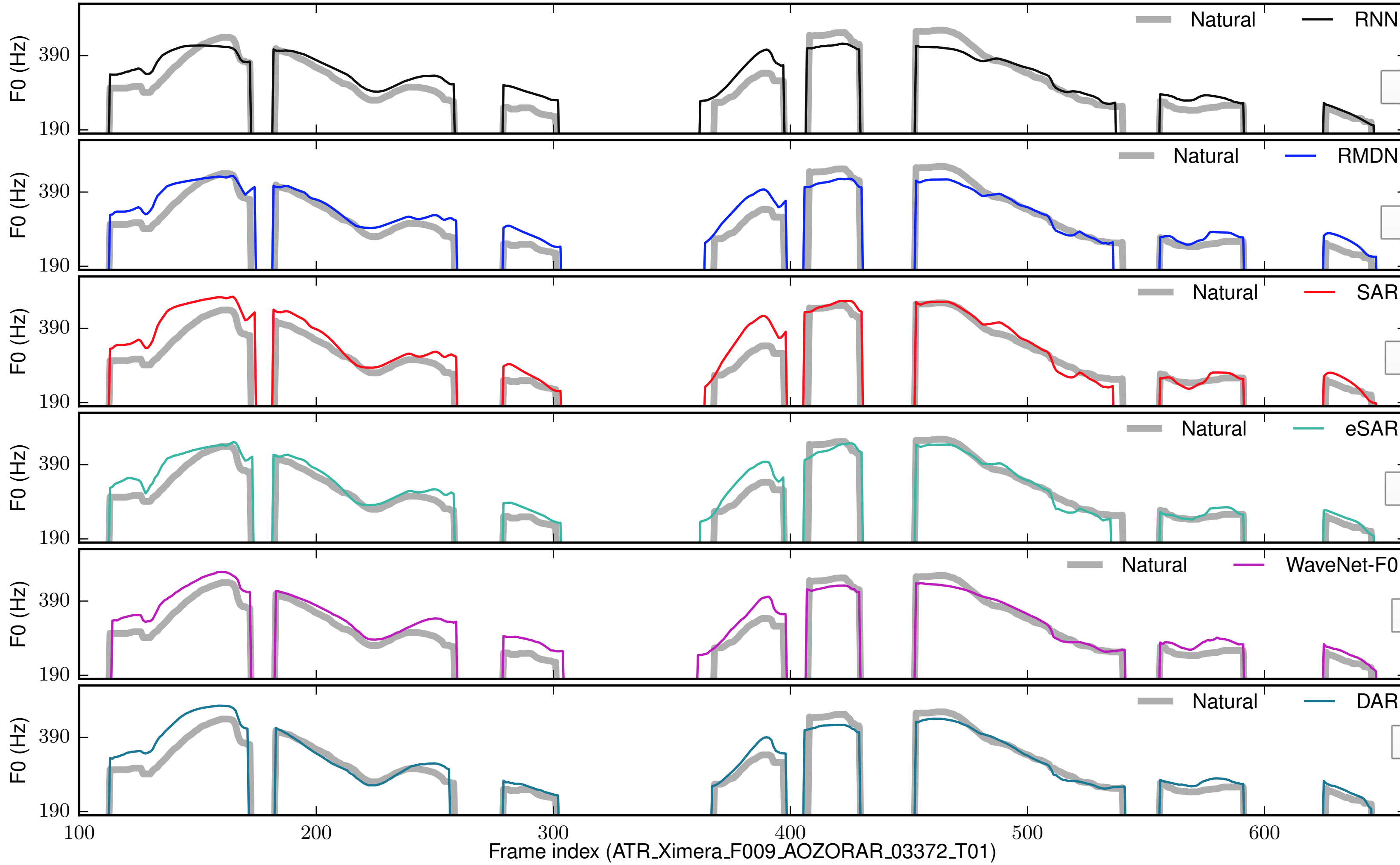
- Random sampling



ISSUE 2: TEMPORAL DEPENDENCY?

Experiment

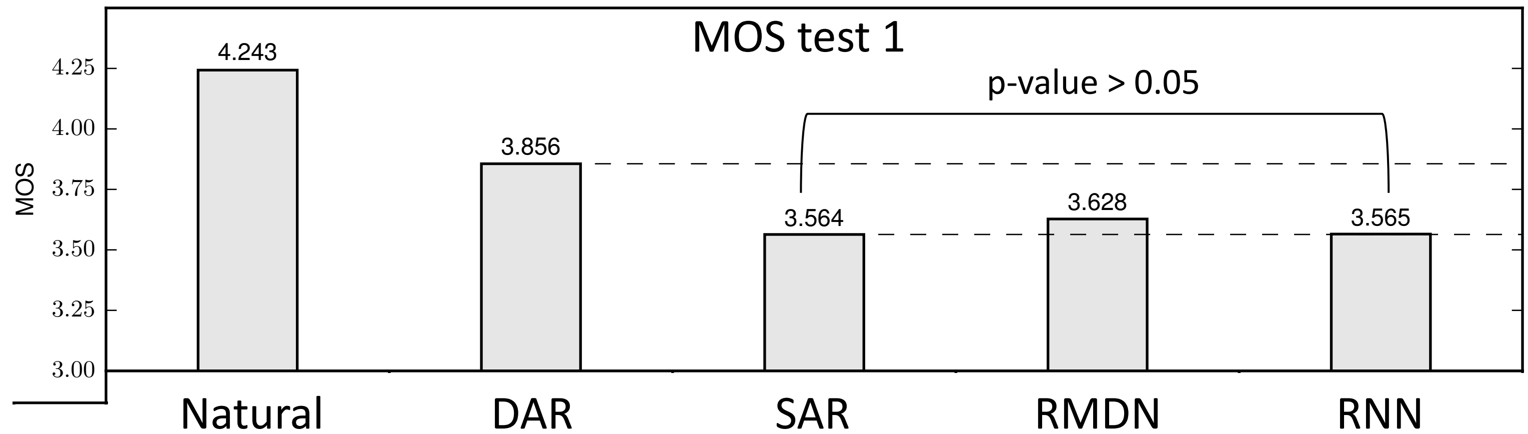
- Mean-based generation

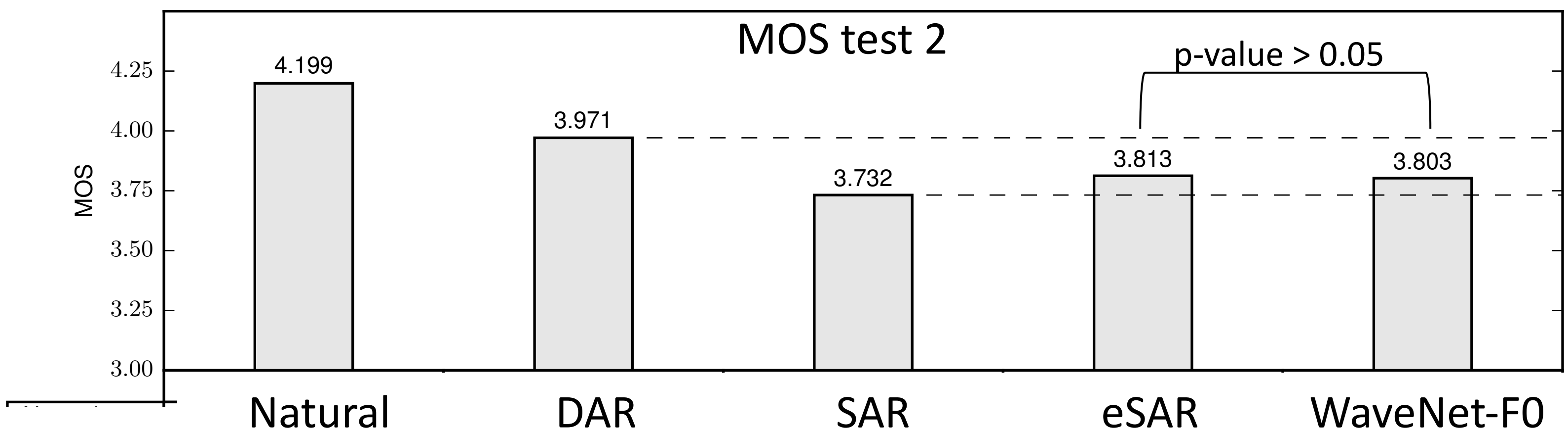


ISSUE 2: TEMPORAL DEPENDENCY?

Experiment

- Mean-based generation





- 500 test utterances, >1000 sets of scores

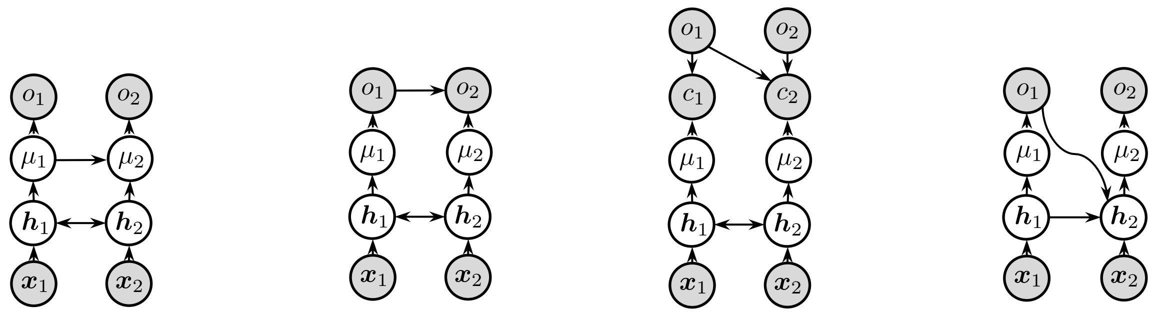
ISSUE 2: TEMPORAL DEPENDENCY?

Summary

- Full answer to issue 2

Temporal dependency is ignored by RNN/RMDN!

But better model can be defined



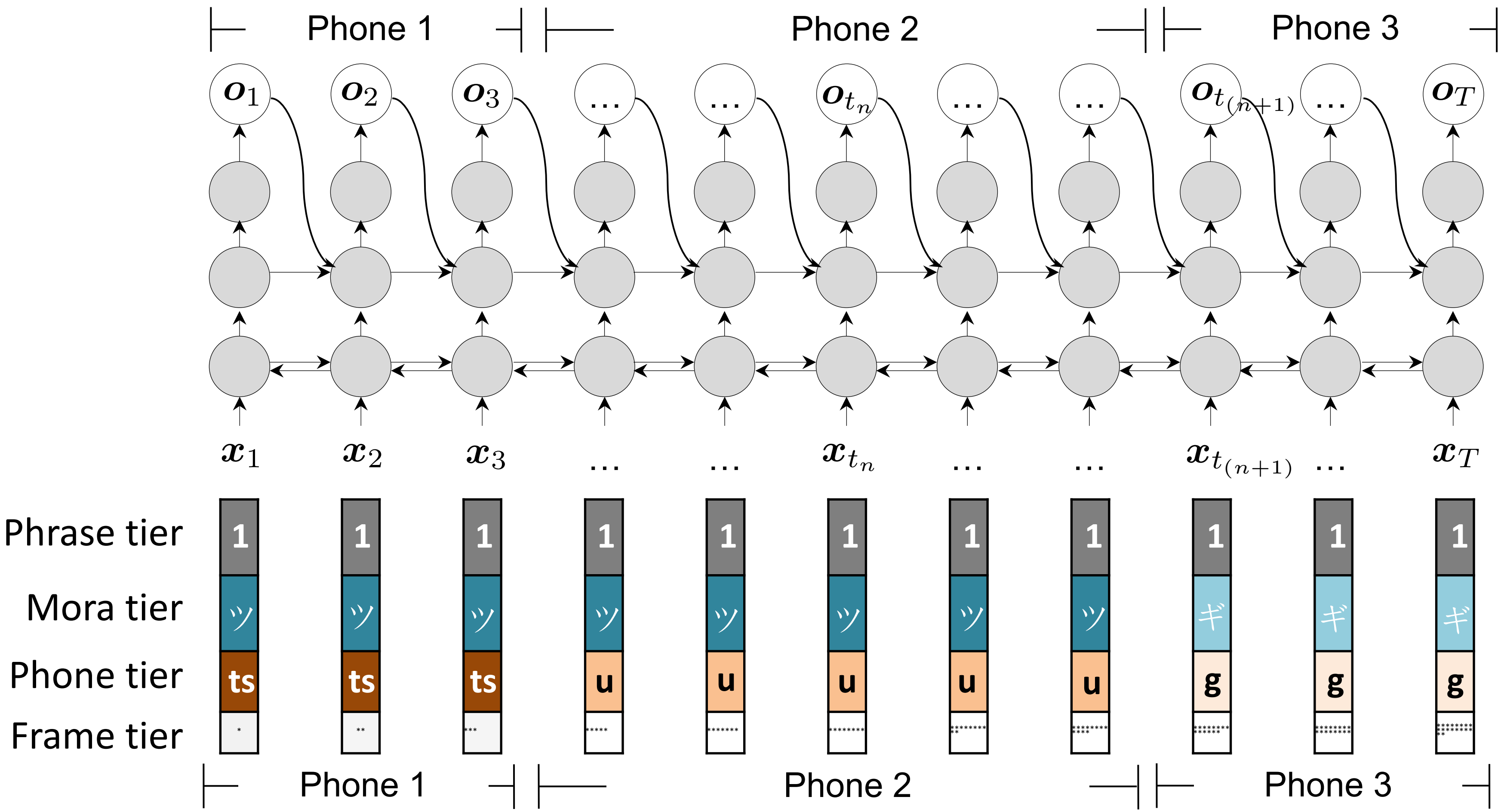
	RMDN	SAR	eSAR	DAR
AR linear?	-	Linear	Non-linear (constrained)	Non-linear
AR time span	-	$t-K:t-1$	$1:t-1$	$1:t-1$
Tractable?	-	Yes	Somewhat	No
Sampling?	No	No	No	Yes

CONTENTS

Issue 3: Frame-by-frame processing?

Motivation

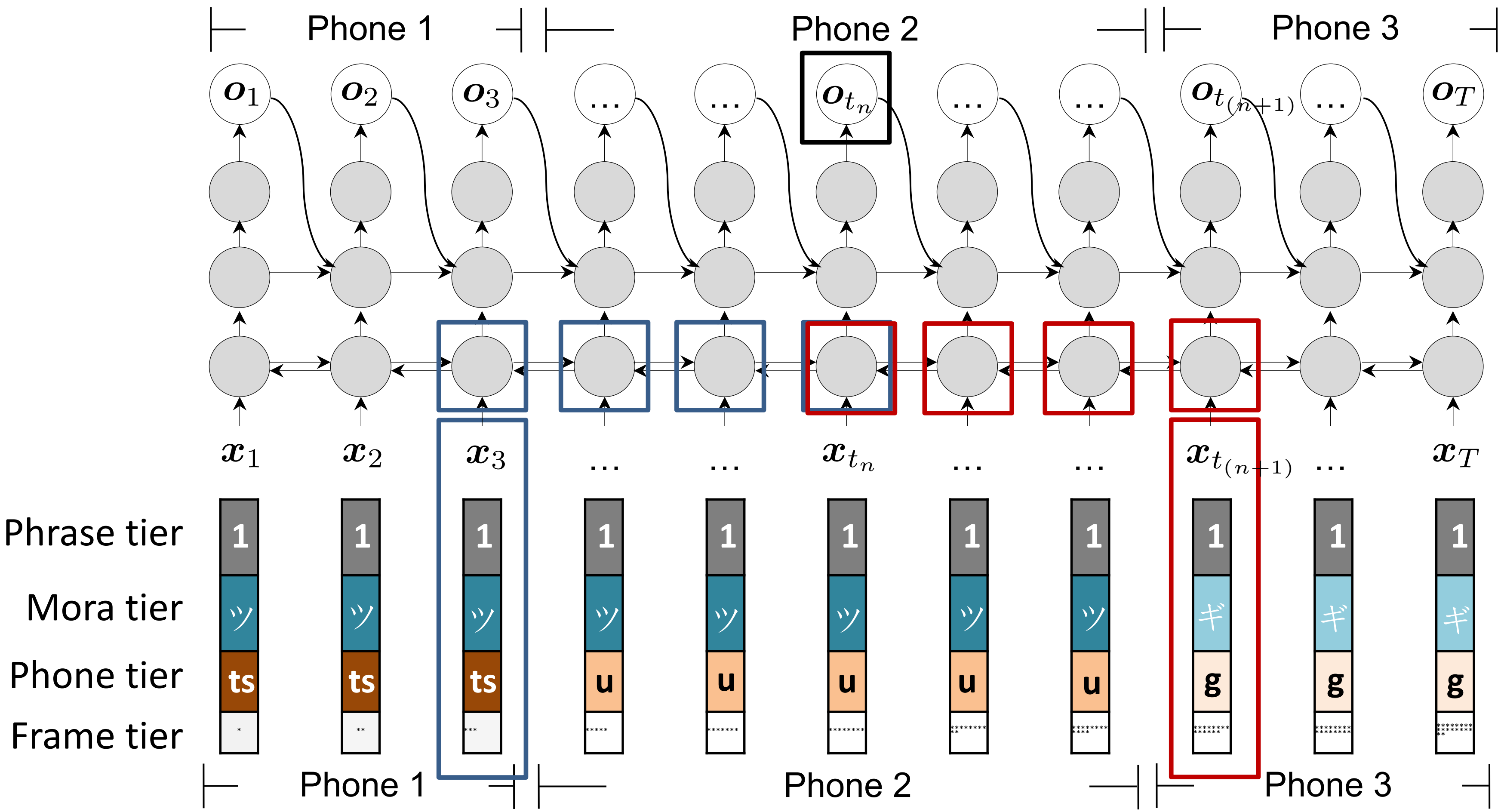
- Inefficient processing



Issue 3: Frame-by-frame processing?

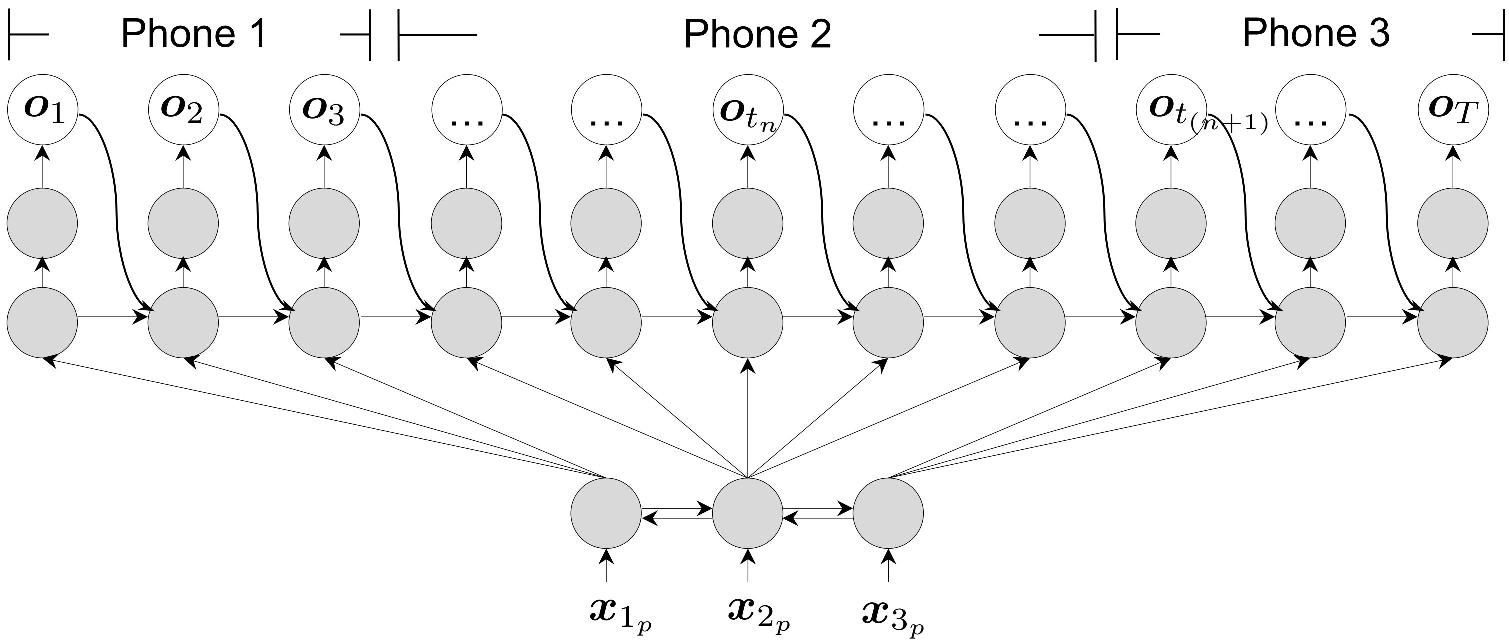
Motivation

- Inefficient processing

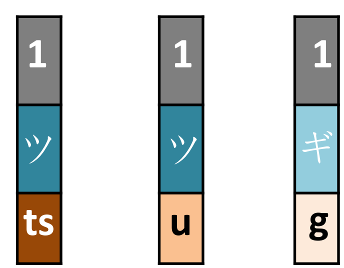


Motivation

- More efficient processing?



- Given duration of each unit

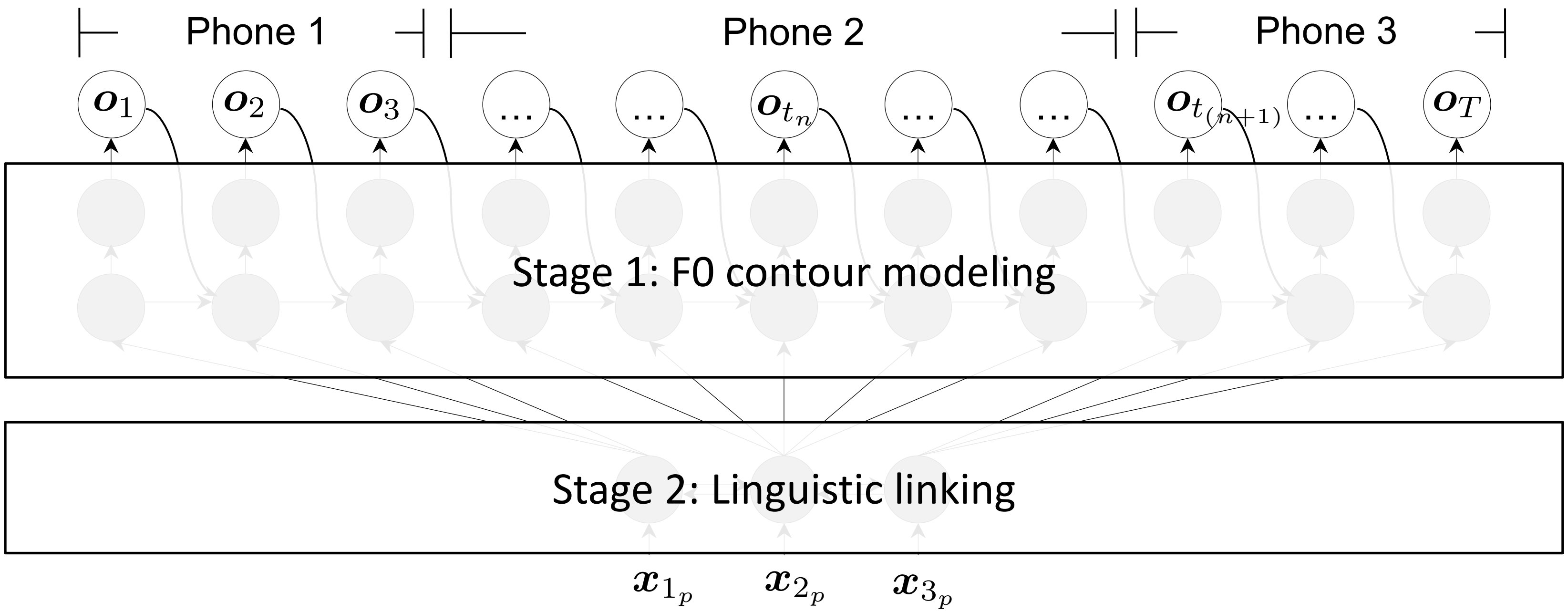


How to do?

Issue 3: Frame-by-frame processing?

Method

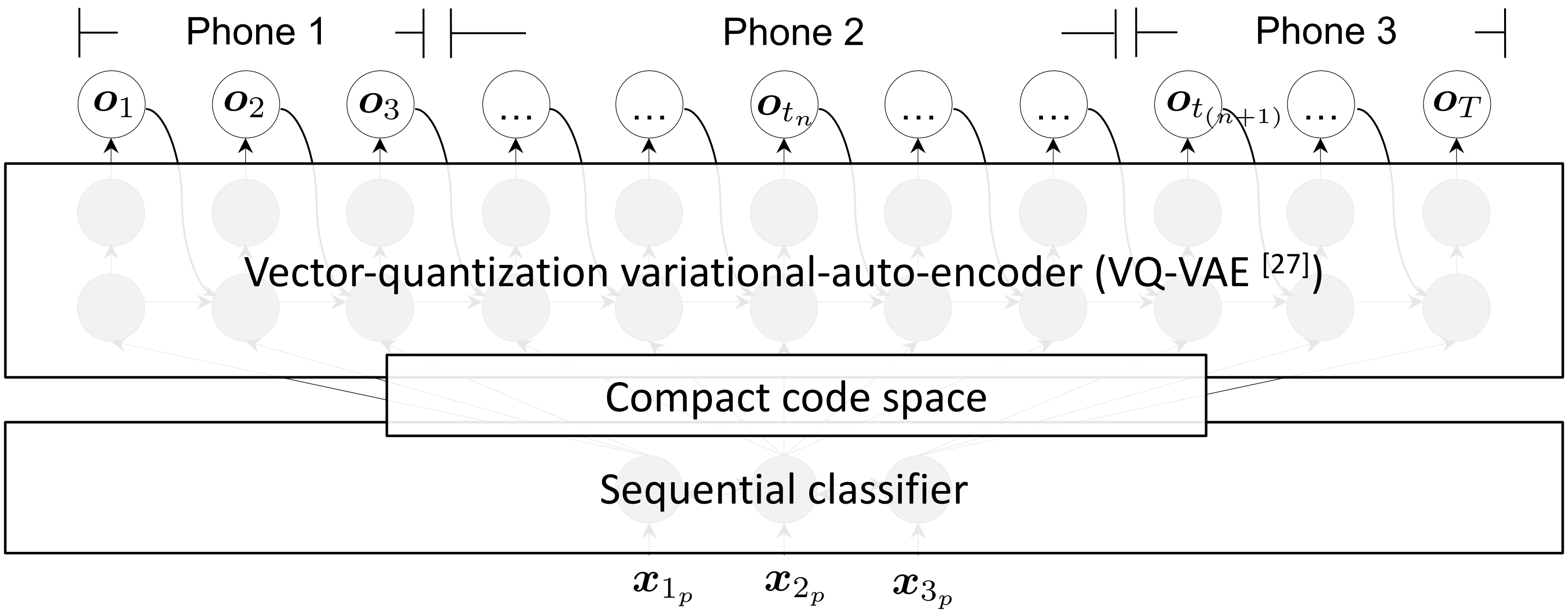
- Two-stage F0 modeling



ISSUE 3: FRAME-BY-FRAME PROCESSING?

Method

- Two-stage F0 modeling



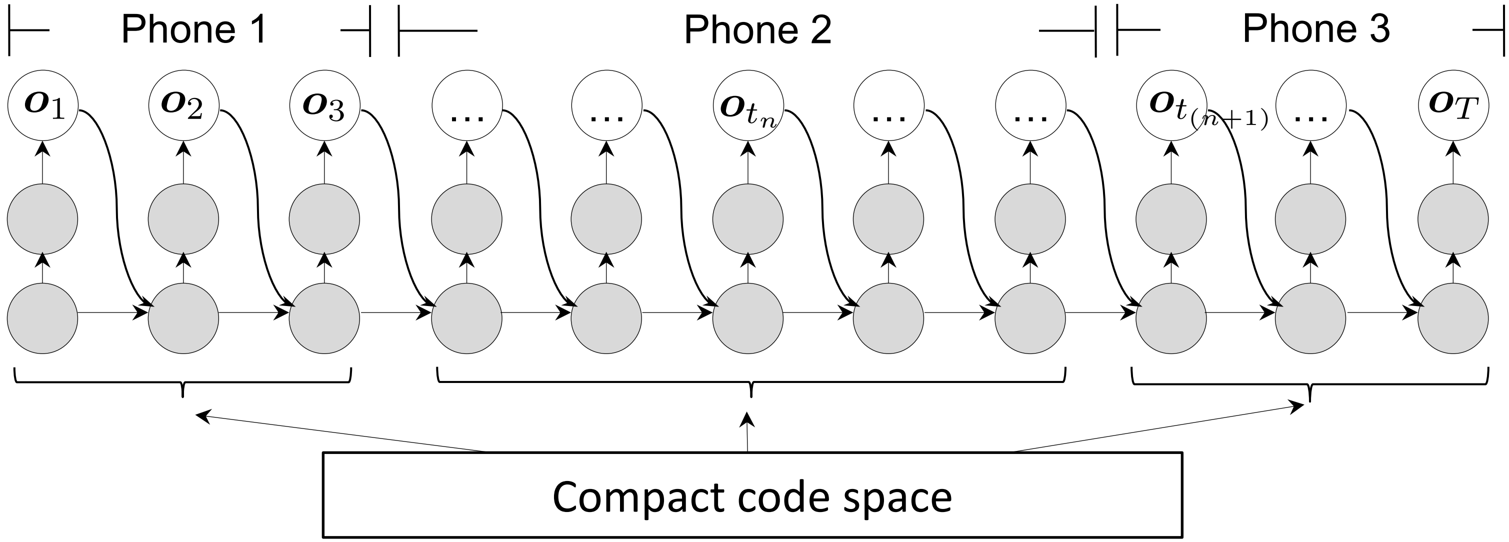
- Revisit linguistic approaches [28, 29]

[27] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In Proc. NIPS, page to appear, 2017.
[28] K. E. Dusterhoff, A. W. Black, and P. A. Taylor. Using decision trees within the Tilt intonation model to predict F0 contours. Proc. Eurospeech, pages 1627–1630, 1999.
[29] K. Hirose, K. Sato, Y. Asano, and N. Minematsu. Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis. Speech communication, 46(3):385–404, 2005.

ISSUE 3: FRAME-BY-FRAME PROCESSING?

Stage 1: F0 contour modeling

❑ VQ-VAE



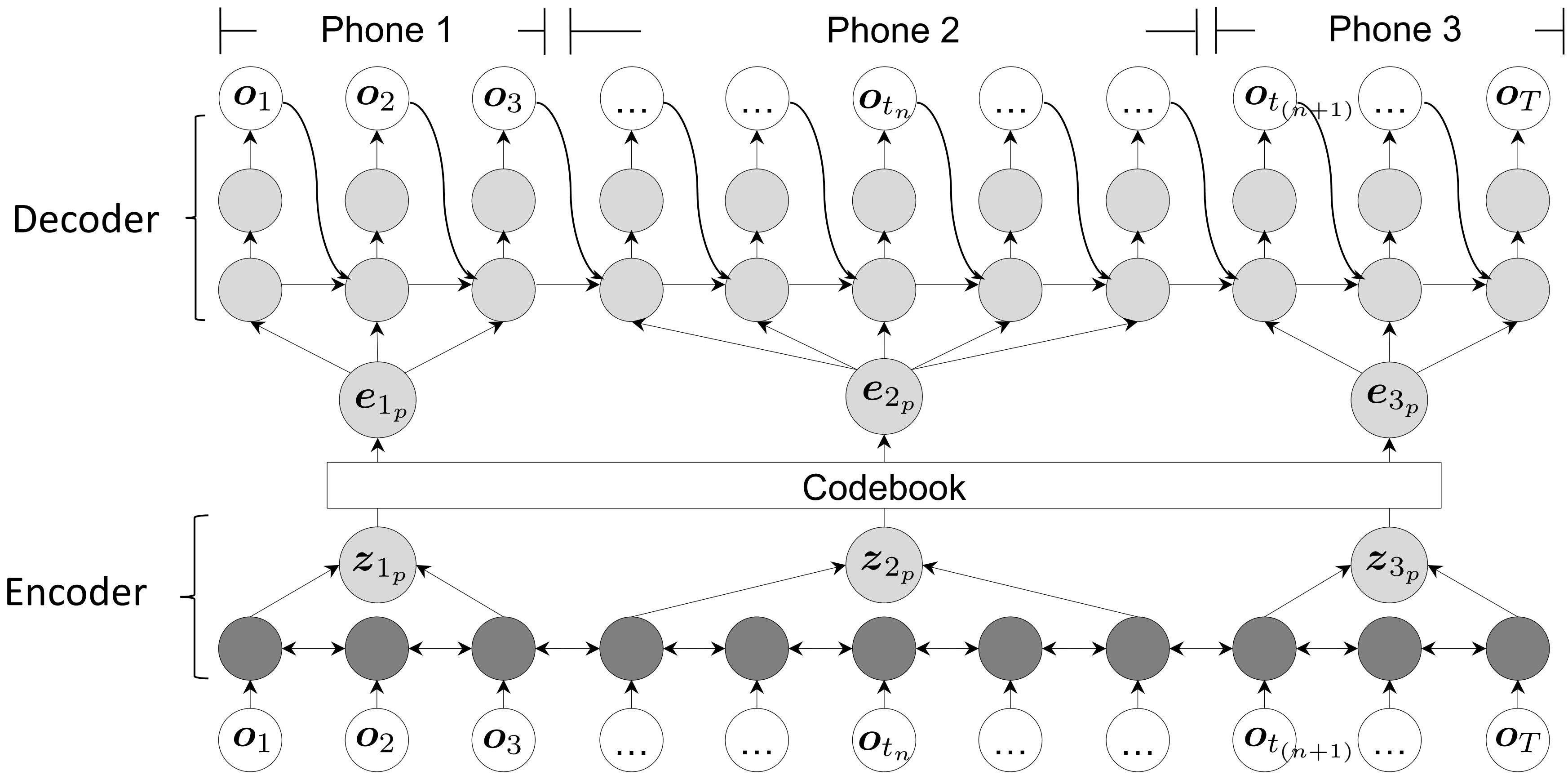
- Unsupervised learning
- One code for varied-length unit
- Multiple linguistic levels

Goals

Issue 3: Frame-by-frame processing?

Stage 1: F0 contour modeling

- VQ-VAE

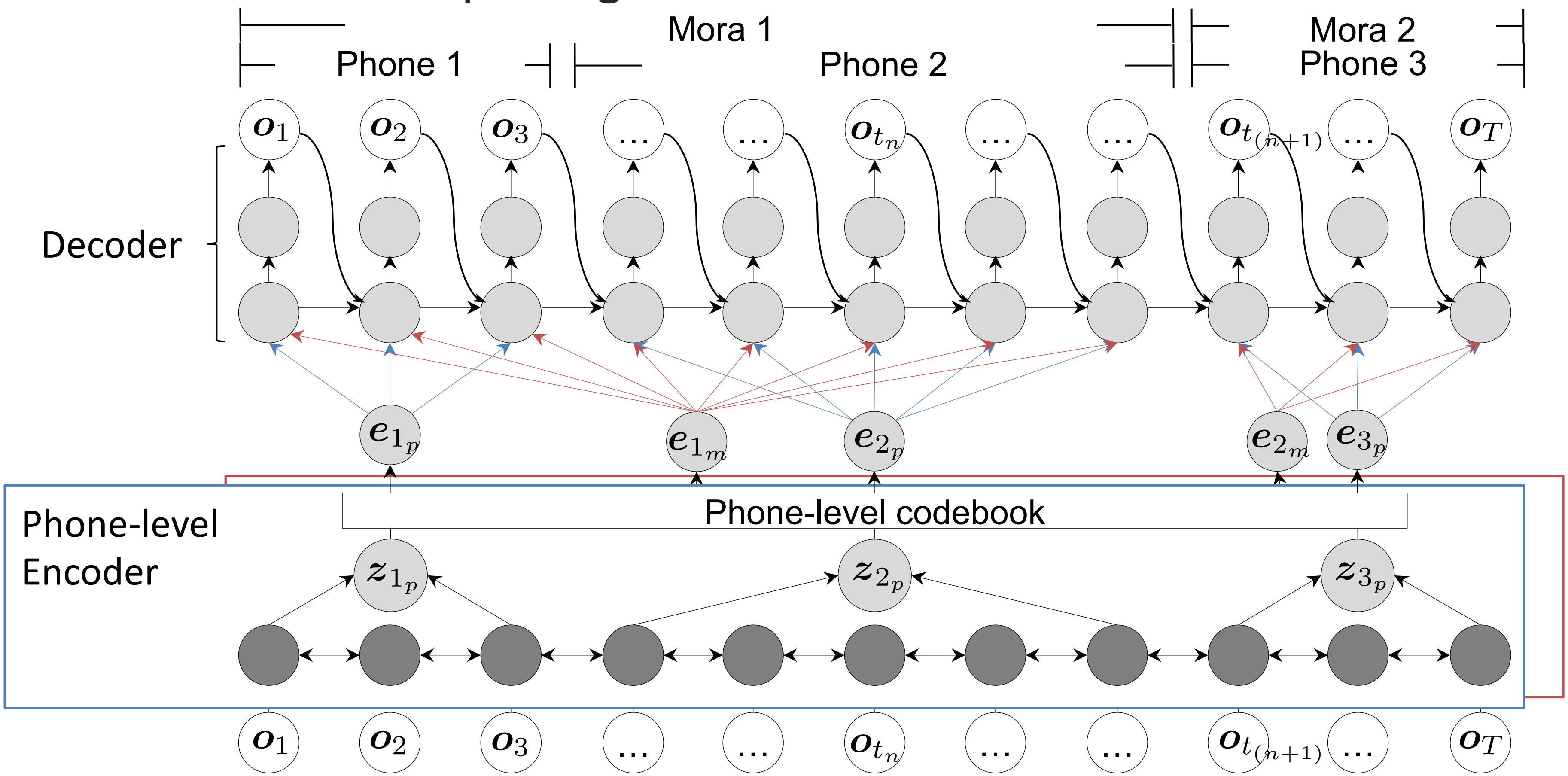


- One code per phone

ISSUE 3: FRAME-BY-FRAME PROCESSING?

Stage 1: F0 contour modeling

- VQ-VAE multiple linguistic tiers

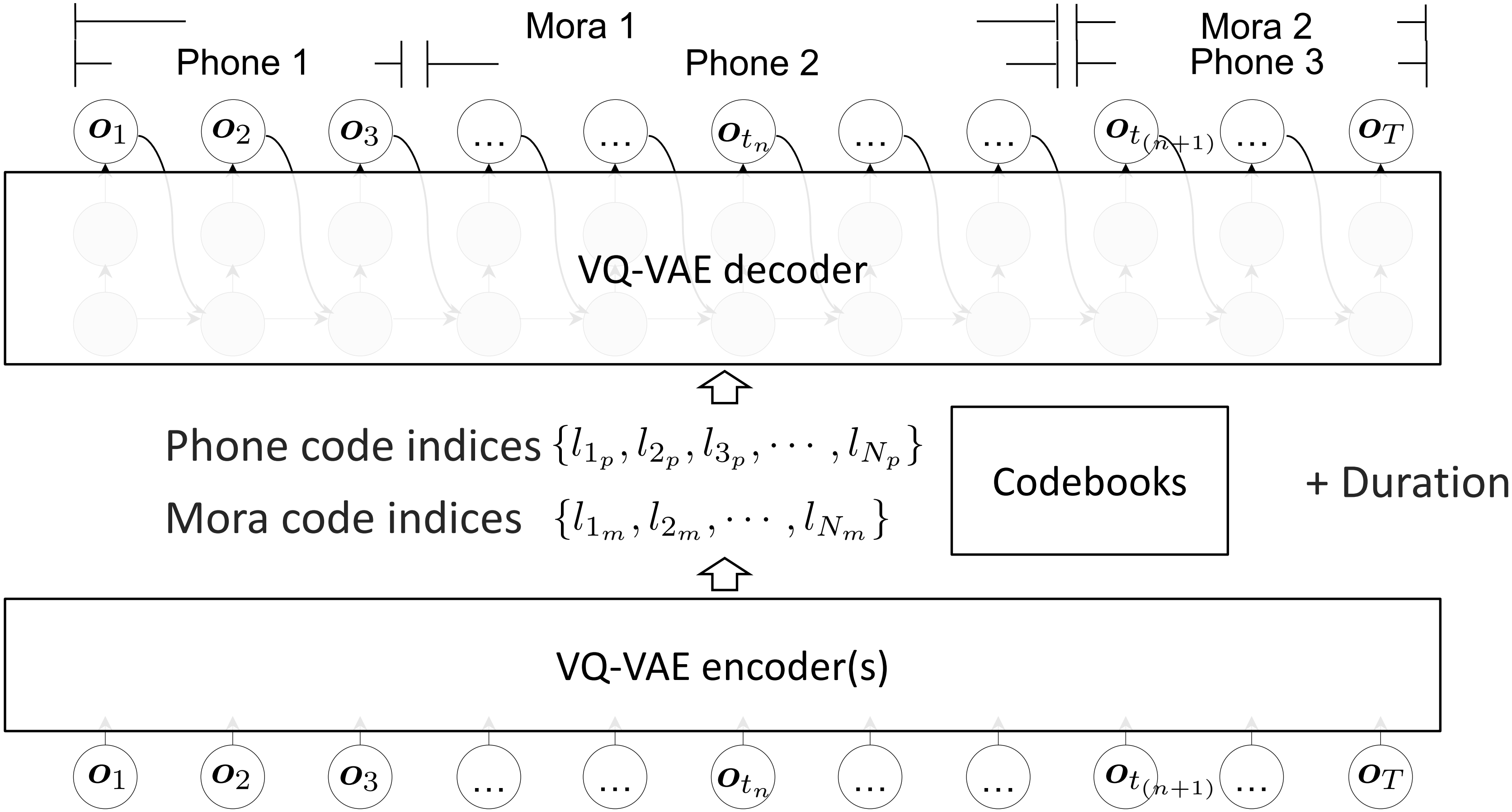


- One code per phone & one code per mora

ISSUE 3: FRAME-BY-FRAME PROCESSING?

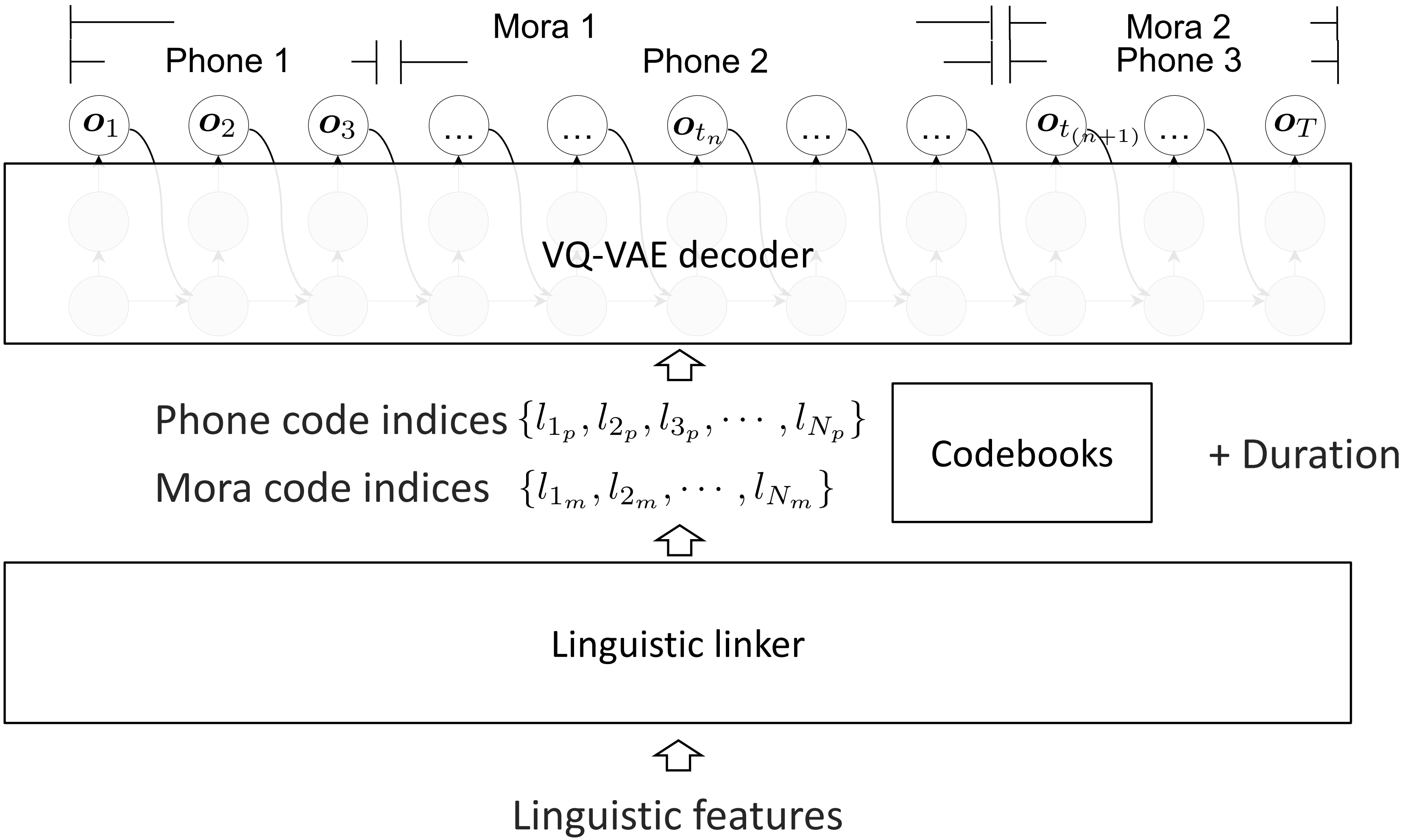
Stage 1: F0 contour modeling

- VQ-VAE multiple linguistic tiers



Issue 3: Frame-by-frame processing?

Stage 2: Linguistic linking



ISSUE 3: FRAME-BY-FRAME PROCESSING?

Stage 2: Linguistic linking

Phone code indices $\{l_{1_p}, l_{2_p}, l_{3_p}, \cdots, l_{N_p}\}$

Mora code indices $\{l_{1_m}, l_{2_m}, \cdots, l_{N_m}\}$

⇧

RNN-based sequential classifier

(Sec. 7.2.2)

- Clockwork RNN [30]
- Highway [31]
- AR & feedback links
- Dropout [32]

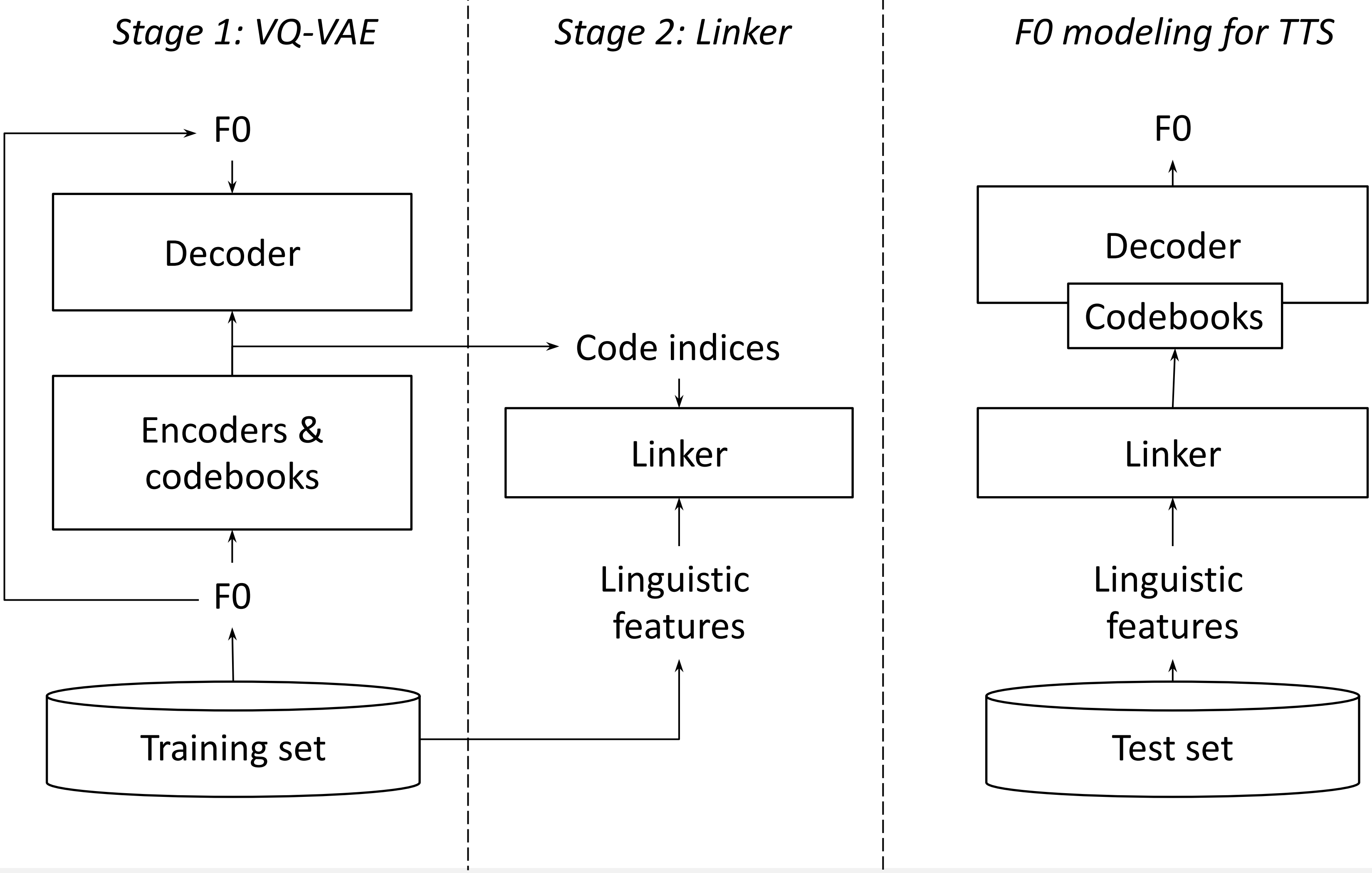
⇧

Linguistic features

[30] J. Koutnik, K. Greff, F. Gomez, and J. Schmidhuber. A Clockwork RNN. In Proc. ICML, pages 1863– 1871, 2014.
[31] K. Greff, R. K. Srivastava, and J. Schmidhuber. Highway and residual networks learn unrolled iterative estimation. In Proc. ICLR, 2017.
[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958, 2014.

ISSUE 3: FRAME-BY-FRAME PROCESSING?

F0 modeling for TTS



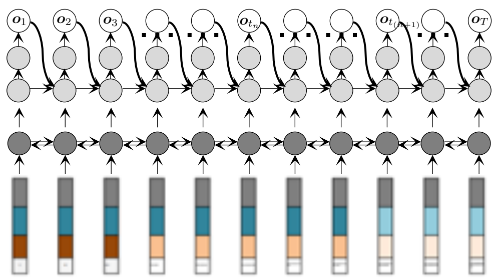
ISSUE 3: FRAME-BY-FRAME PROCESSING?

Experiments on stage 1 (sec.7.3.2)

Experiments on whole model

- Models

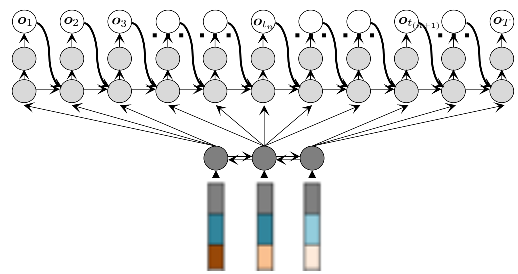
Model 1



VQ-VAE decoder (phone-level)

Frame-by-frame Linker

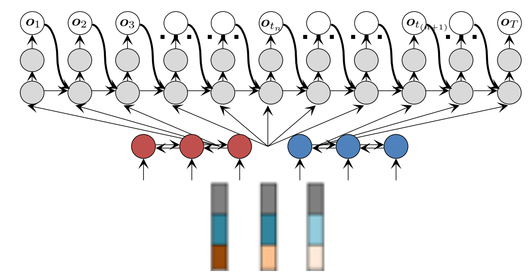
Model 2



VQ-VAE decoder (phone-level)

Phone-by-phone Linker

Model 3



VQ-VAE decoder (phone-mora-level)

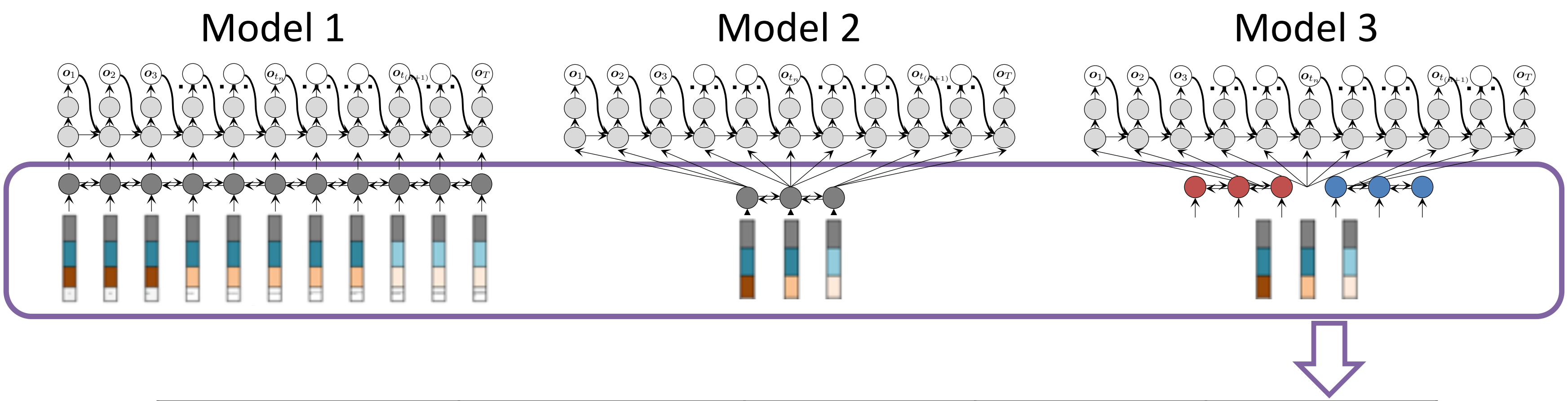
Phone-by-phone Linker + Mora lock

- Given natural duration

Issue 3: Frame-by-frame processing?

Experiments

- Objective results

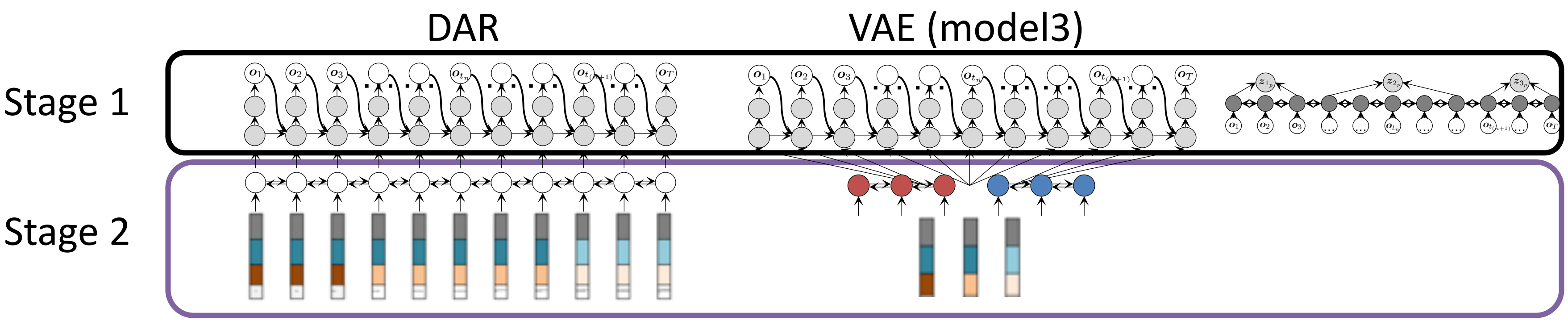


	RMSE	CORR	U/V	Time cost (s/epoch)
Model 1	34.3	0.839	7.96%	1300
Model 2	27.1	0.906	6.36%	54
Model 3	**25.5**	**0.916**	**4.87%**	**65**

ISSUE 3: FRAME-BY-FRAME PROCESSING?

VAE vs DAR

❑ Objective results



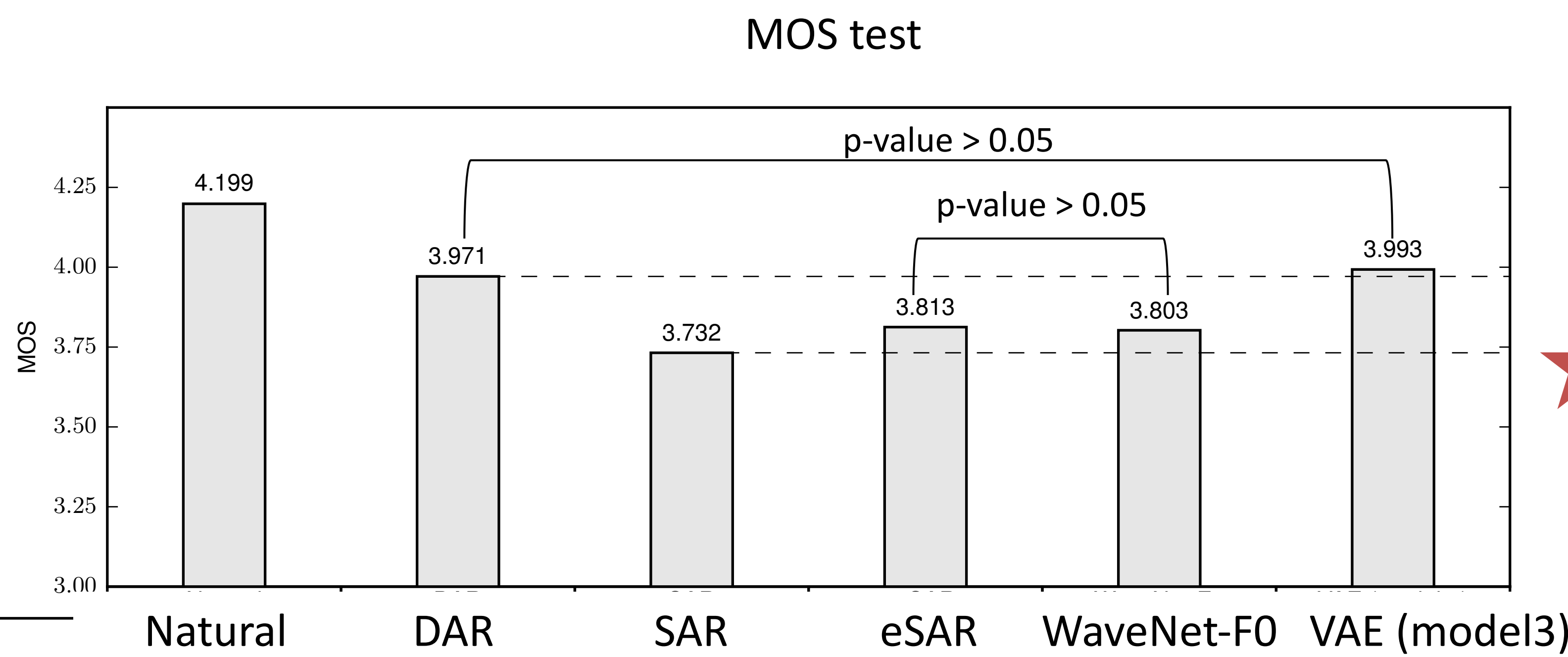
	RMSE	CORR	U/V	Number of parameters (m)			Time cost (s/epoch)		
				Stage 1	Stage 2	Sum	Stage 1	Stage 2	Sum
DAR	28.3	0.903	3.46%	0.36	1.11	1.48	~700	~1300	2000
VAE model3	25.5	0.916	4.87%	0.44	0.67	1.11	1500	65	1565

- VQ-VAE encoder needs time and memory ★

ISSUE 3: FRAME-BY-FRAME PROCESSING?

VAE vs DAR

- Subjective test



- 500 test utterances, >1000 sets of scores

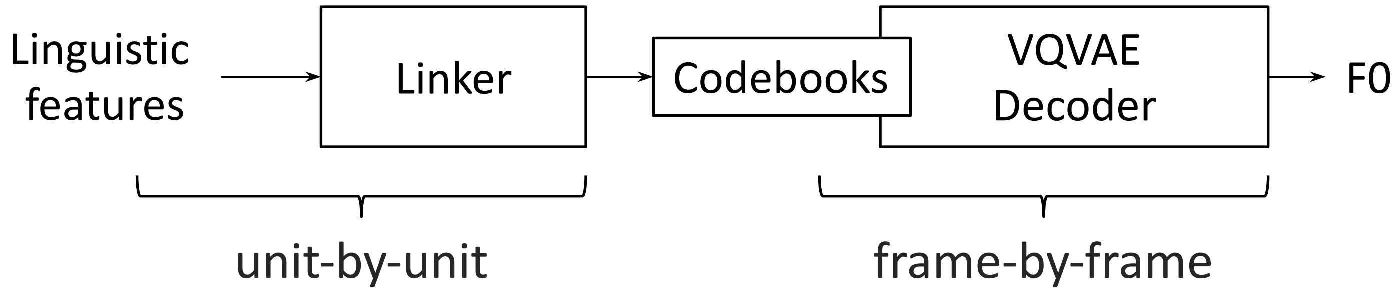
ISSUE 3: FRAME-BY-FRAME PROCESSING?

Summary

- Answer to issue 3

It could be more efficient

- How



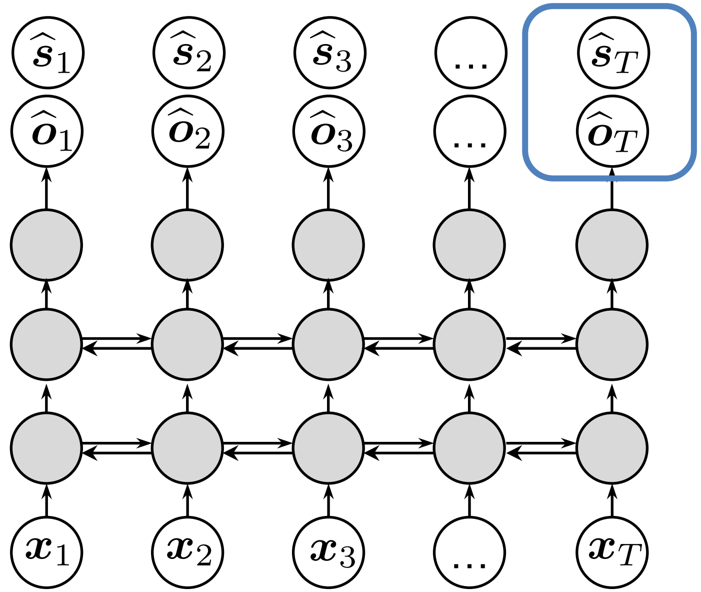
- Results
  - Multiple linguistic tiers
  - More efficient than DAR: smaller + faster + F0 CORR > 0.91
  - Interpretable latent code spaces (Sec. 7.3.2)
  - Random sampling OK (slides)

Contents

- Introduction
- Issue 1: joint modeling of F0 and spectral features
- Issue 2: temporal dependency modeling of F0 contours
- Issue 3: frame-by-frame processing
- Conclusion

Conclusion

Summary



Preliminary → Issue 1: Joint modeling for F0?

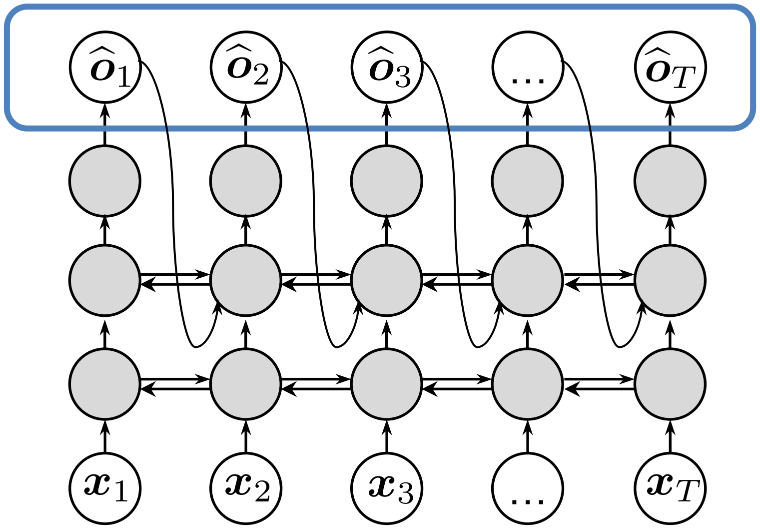
Chapter 1-3 | Chapter 4

- ? Joint modeling of F0 and spectral?
- × Sub-optimal for F0 modeling
- Methods:
  - Highway networks
  - Histogram + sensitivity analysis
- Results:
  - Spectral features are prioritized
  - Different input/hidden features for F0 and spectral features

Conclusion

Summary

DAR



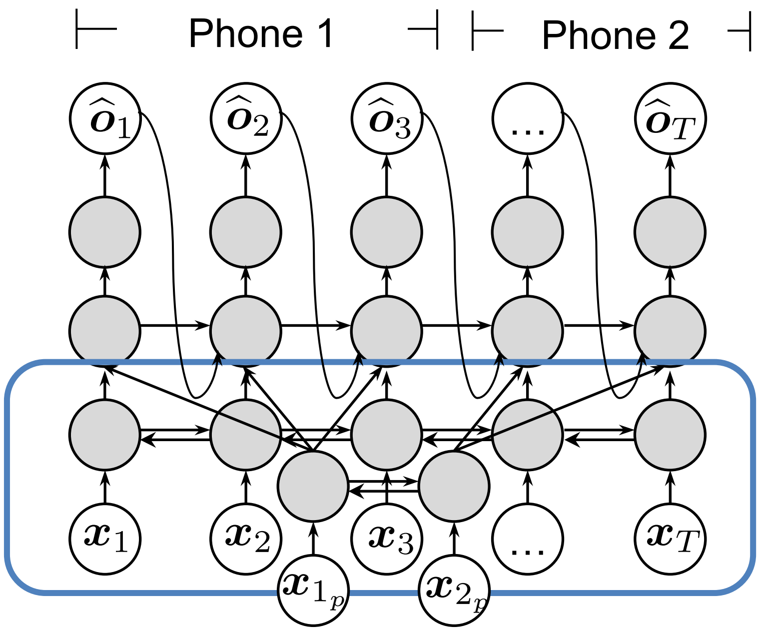
- ? Temporal dependency in RNN/RMDN?
- × Ignored by RNN/RMDN
- Models:
  - SAR : tractable dependency
  - DAR : non-linear + longer dependency
- Results:
  - DAR: F0 CORR > 0.90, MOS score
  - DAR supports random sampling!

Preliminary → Issue 1: Joint modeling for F0? → Issue 2: Temporal dependency?

Chapter 1-3 | Chapter 4 | Chapter 5-6

CONCLUSION

Summary



- ? Frame-by-frame processing
- × Inefficient
- Two-stage F0 model:
  - F0 contour coding: VQ-VAE + DAR
  - Linguistic linking: unit-by-unit classifier
- Results:
  - F0 CORR > 0.91
  - More efficient (smaller, faster)

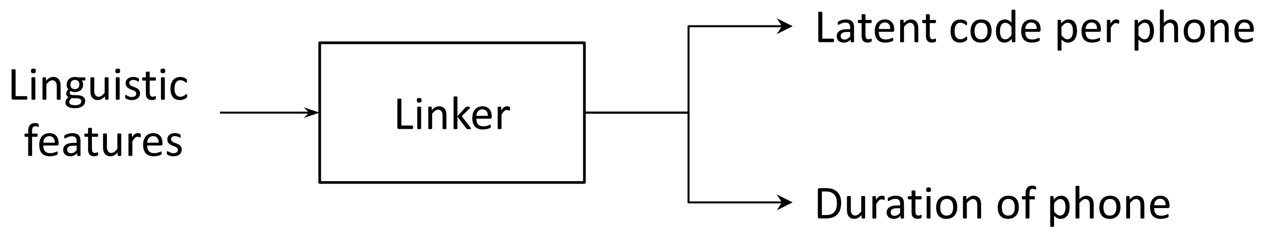
Preliminary → Issue 1: Joint modeling for F0? → Issue 2: Temporal dependency? → Issue 3: Frame-by-frame? → Summary

Chapter 1-3 | Chapter 4 | Chapter 5-6 | Chapter 7 | Chapter 8

CONCLUSION

New topics?

- Towards complete prosody modeling
  - Not only F0 but also duration
  - Easy for joint modeling



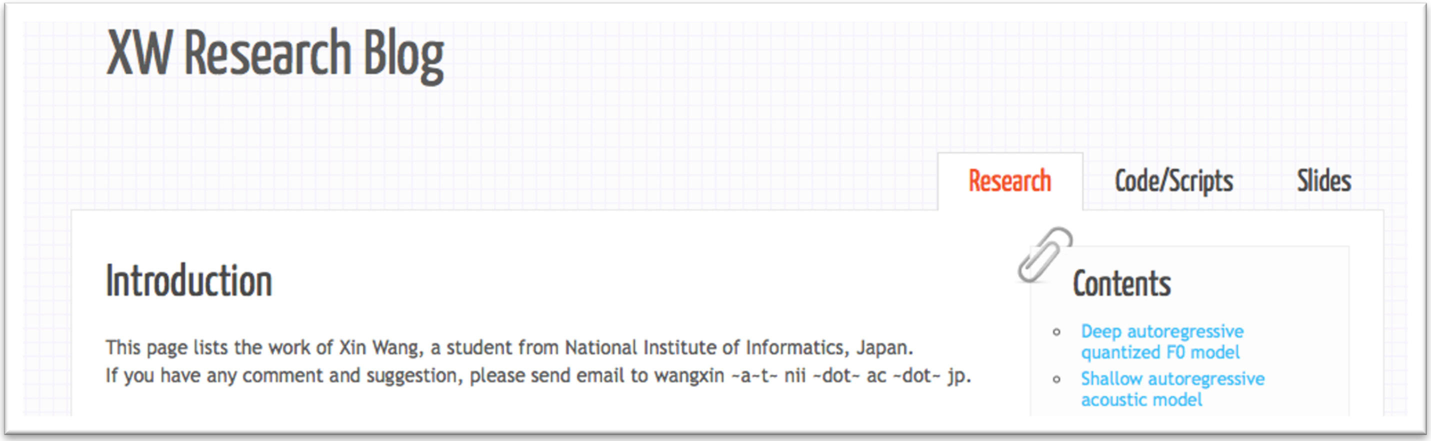
- Waveform modeling
  - F0 and waveform are 1 dimensional signals
  - Signal processing methods available

    SAR + log area ratio + segment-variant filters

Thank you for your attention

Q & A

Codes, scripts, slides: tonywangx.github.io



THESIS OUTLINE

Neural F0 modeling

Chapter 1-3 Introduction, TTS, neural networks	
Chapter 4 Investigation using highway network	*Speech synthesis workshop 2016,* *Speech Communication, Vol 96, Feb. 2018, pp1-9*
Chapter 5 Shallow autoregressive F0 model (SAR) Extended SAR Chapter 6 Deep autoregressive F0 model (DAR)	*ICASSP 2017* *Interspeech 2017* *IEEE Trans. ASLP, accepted, 2018*
Chapter 7 Variational auto-encoder (VAE) + DAR	*Prepare a journal paper*
Chapter 8 Summary and application of F0 model	*ICASSP 2018*

✓ Meet basic requirement on publication: 3 journals + 6 conferences

Work during Ph.D.

- Work not included in Ph.D. thesis
  - Word embedding as input features (IEICE 2018, Interpseech2016)
  - Impact of training data size (SSW 2016)
- Collaborated work
  - DAR using manually-annotated / corrupted data (Interspeech 2018)

    Provide DAR, SAR, and WaveNet-vocoder
  - Voice cloning using found data (Odyssey 2018)

    Provide WaveNet-vocoder, and acoustic models
  - Cyborg speech: multilingual speech synthesis (ICASSP 2018)

    Provide acoustic models
  - Speech synthesis from MFCC (ICASSP 2018)

    Provide DAR on MFCC
  - Controllable speech synthesis (Interspeech 2017)

    Provide acoustic models