ICASSP 2017
New Orleans, USA

An Autoregressive Recurrent Mixture Density Network for Parametric Speech Synthesis

Xin WANG, Shinji TAKAKI, Junichi YAMAGISHI
National Institute of Informatics, Japan

2017-03-07

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

ABBREVIATION

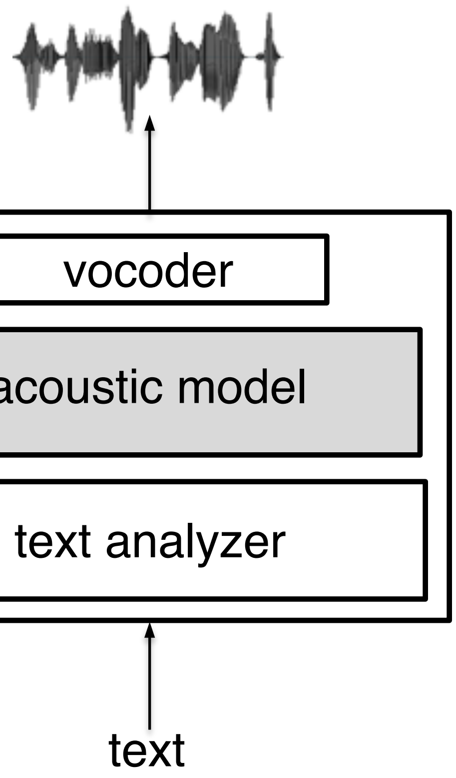
GMM	Gaussian mixture model
NN	Neural network
RNN	Recurrent neural network
MDN	Mixture density network
RMDN	Recurrent mixture density network
AR	Autoregressive model
AR-RMDN	Autoregressive RMDN **(proposed model)**

CONTENTS

Introduction

Text-to-Speech

- Based on parametric speech synthesis

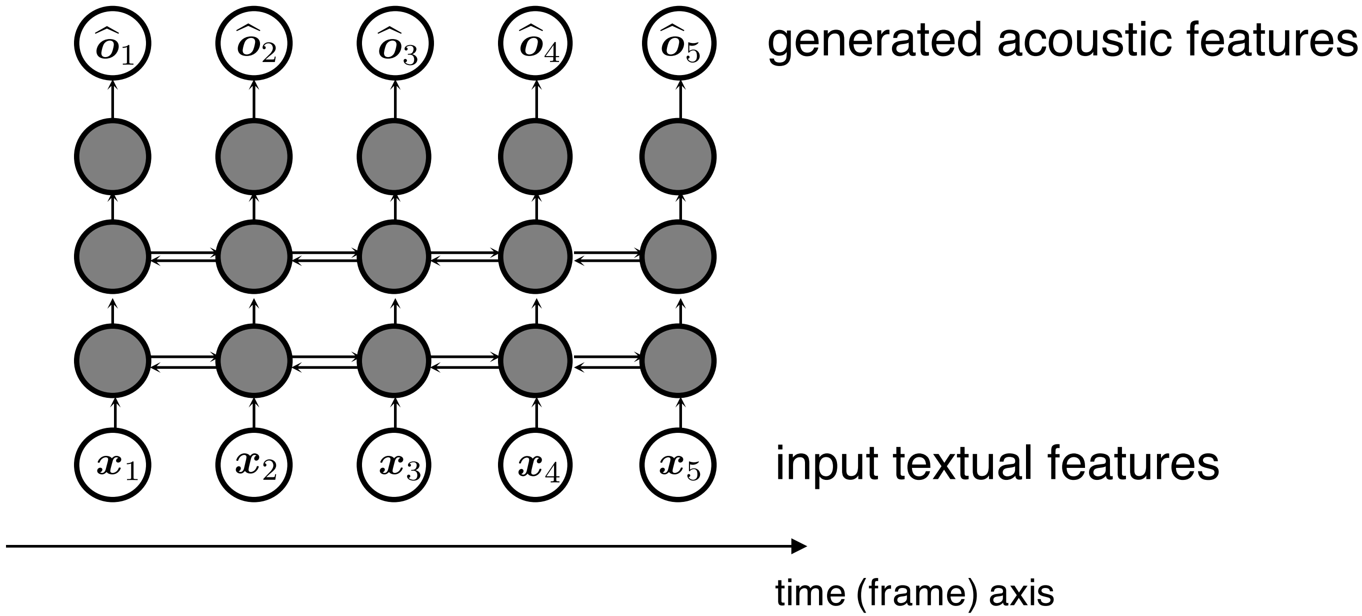


- this work: on acoustic models based on neural networks

INTRODUCTION

Acoustic models based on neural networks

- RNN [1]



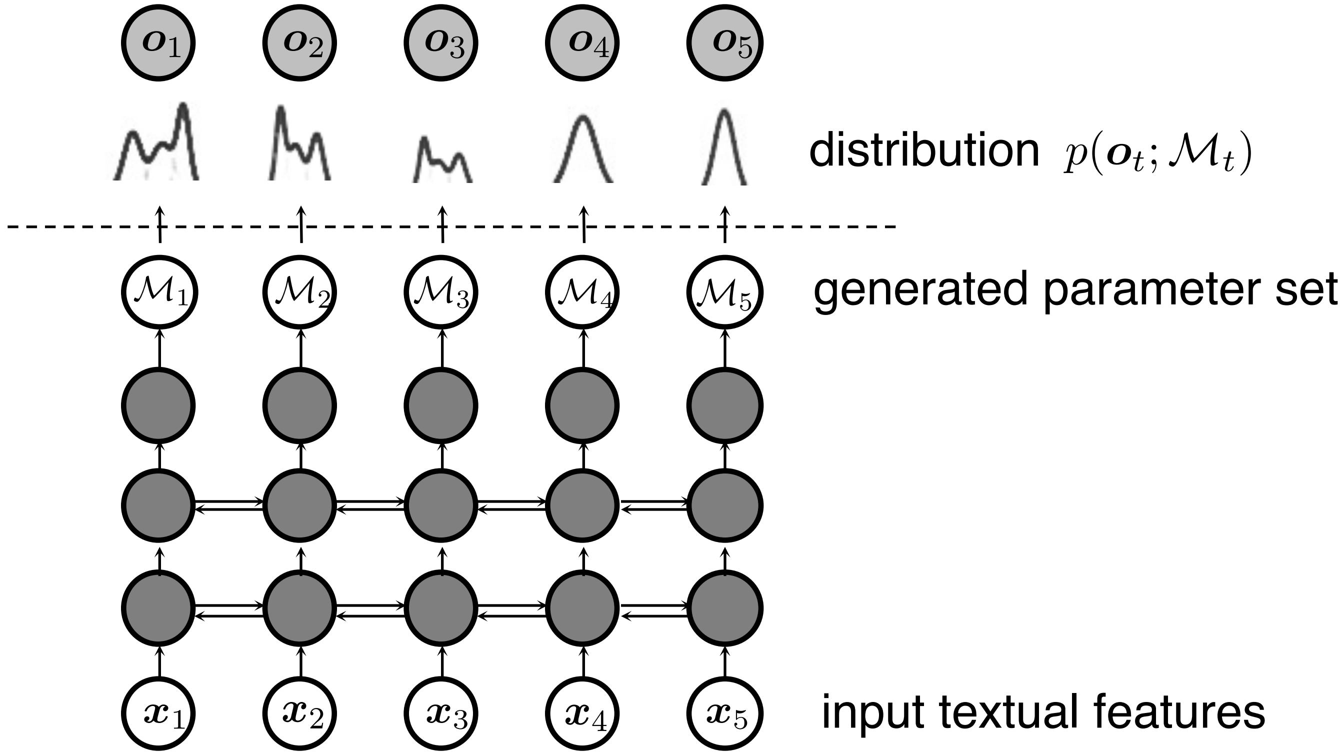
- output of RNN: $\widehat{o}_t$ (spectral features, F0 ...)

[1] Fan, Y., Qian, Y., Xie, F.-L., & Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proc. INTERSPEECH* (pp. 1964–1968).

INTRODUCTION

Acoustic models based on neural networks

- RMDN [2]



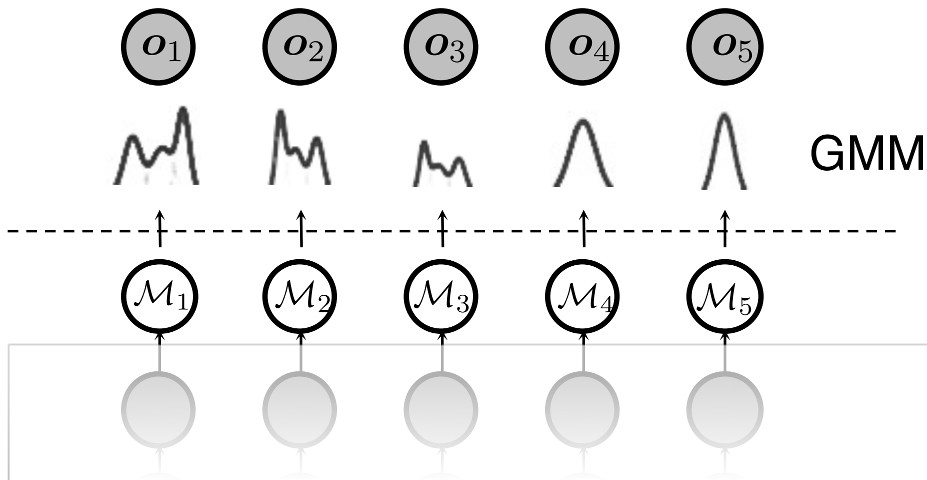
- output of RMDN: $\mathcal{M}_t$ for $p(\boldsymbol{o}_t; \mathcal{M}_t)$

[2] Schuster, M. (1999). Better Generative Models for Sequential Data Problems: Bidirectional Recurrent Mixture Density Networks. In *Proc. NIPS* (pp. 589–595).

INTRODUCTION

Acoustic models based on neural networks

- RMDN [3,4]



  - RMDN using GMM

$$\mathcal{M}_t = \{\omega_t^1, \cdots, \omega_t^M, \boldsymbol{\mu}_t^1, \cdots, \boldsymbol{\mu}_t^M, \boldsymbol{\Sigma}_t^1, \cdots, \boldsymbol{\Sigma}_t^M\}$$

$$p(\boldsymbol{o}_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t; \mathcal{M}_t) = \prod_{t=1}^{T} \sum_{m=1}^{M} \omega_t^m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m).$$

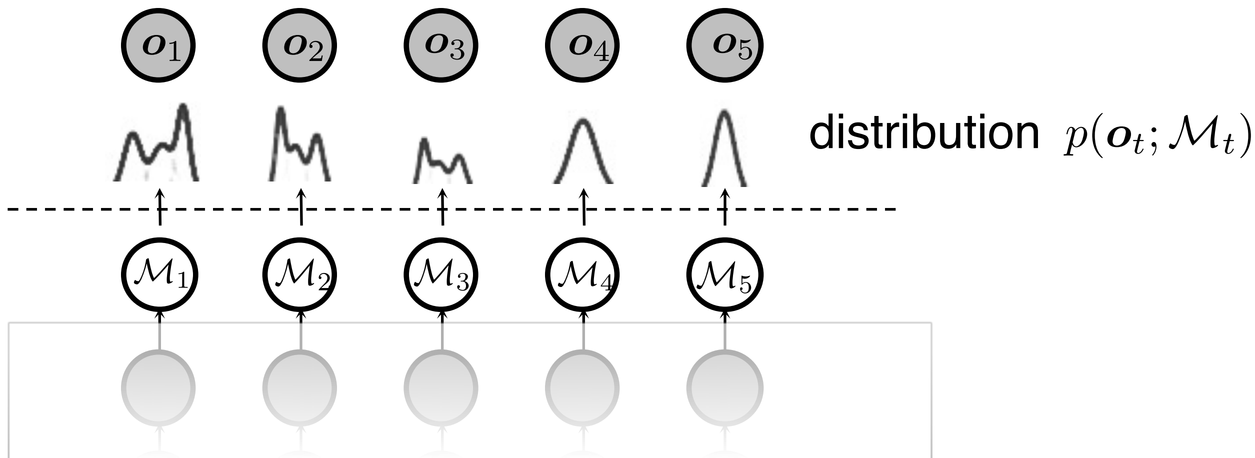
  - generate $\hat{\boldsymbol{o}}_t$ from GMM using MLPG [5]

[3] Bishop, C. M. (2004). *Mixture Density Networks*. Retrieved from http://eprints.aston.ac.uk/373/
[4] Schuster, M. (1999). Better Generative Models for Sequential Data Problems: Bidirectional Recurrent Mixture Density Networks. In *Proc. NIPS* (pp. 589–595).
[5] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318.

CONTENTS

MOTIVATION

Conventional RMDN

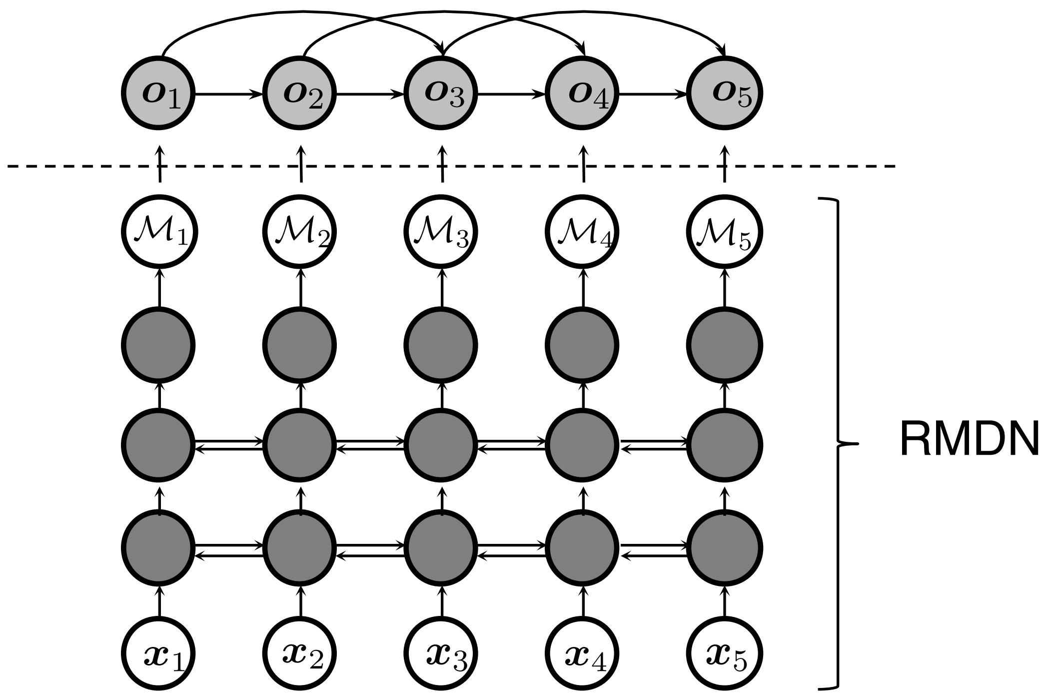


- Independence assumption: $p(\boldsymbol{o}_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t; \mathcal{M}_t)$
- Alternative?
  - the idea of Autoregressive HMM (for speech synthesis [6])
  - ...

[6] Shannon, M., Zen, H., & Byrne, W. (2013). Autoregressive models for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(3), 587–597.

DEFINITION

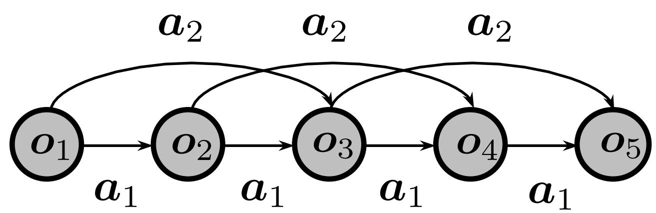
Proposed model: AR + RMDN



- AR assumption: $p(\boldsymbol{o}_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{o}_t | \underline{\boldsymbol{o}_{t-K:t-1}}; \mathcal{M}_t)$

Definition

AR-RMDN using GMM



- baseline RMDN: $$p(\boldsymbol{o}_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^{T} \sum_{m=1}^{M} \omega_t^m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t^m, \boldsymbol{\Sigma}_t^m)$$
- AR-RMDN: $$p(\boldsymbol{o}_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^{T} \sum_{m=1}^{M} \omega_t^m \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t^m + \underline{f(\boldsymbol{o}_{t-K:t-1})}, \boldsymbol{\Sigma}_t^m)$$

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

  - $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_K, \boldsymbol{b}\}$
    1. time invariant (context-independent)
    2. joint training with RMDN using back-propagation

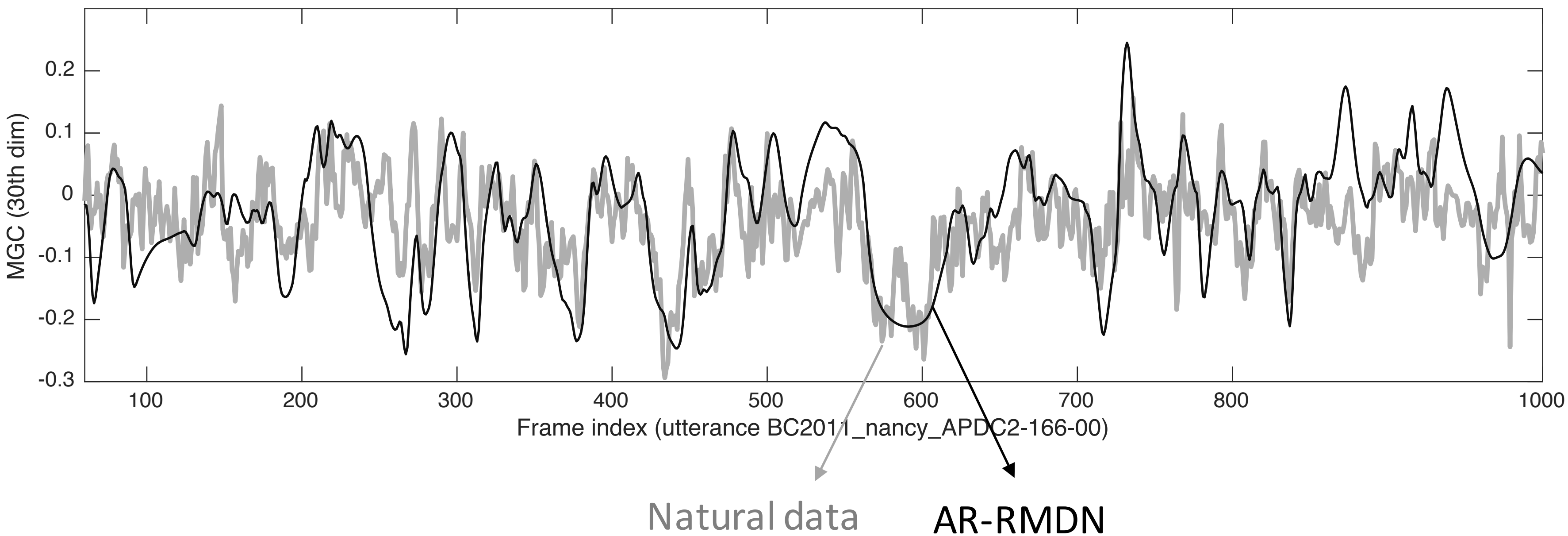
CONTENTS

- Introduction
- Model definition, interpretation, implementation
- Experiments and results
- Conclusion

BEFORE INTERPRETATION

A glimpse of the results

Generated trajectories of the 30th dimension of mel-generalized cepstrum coefficients

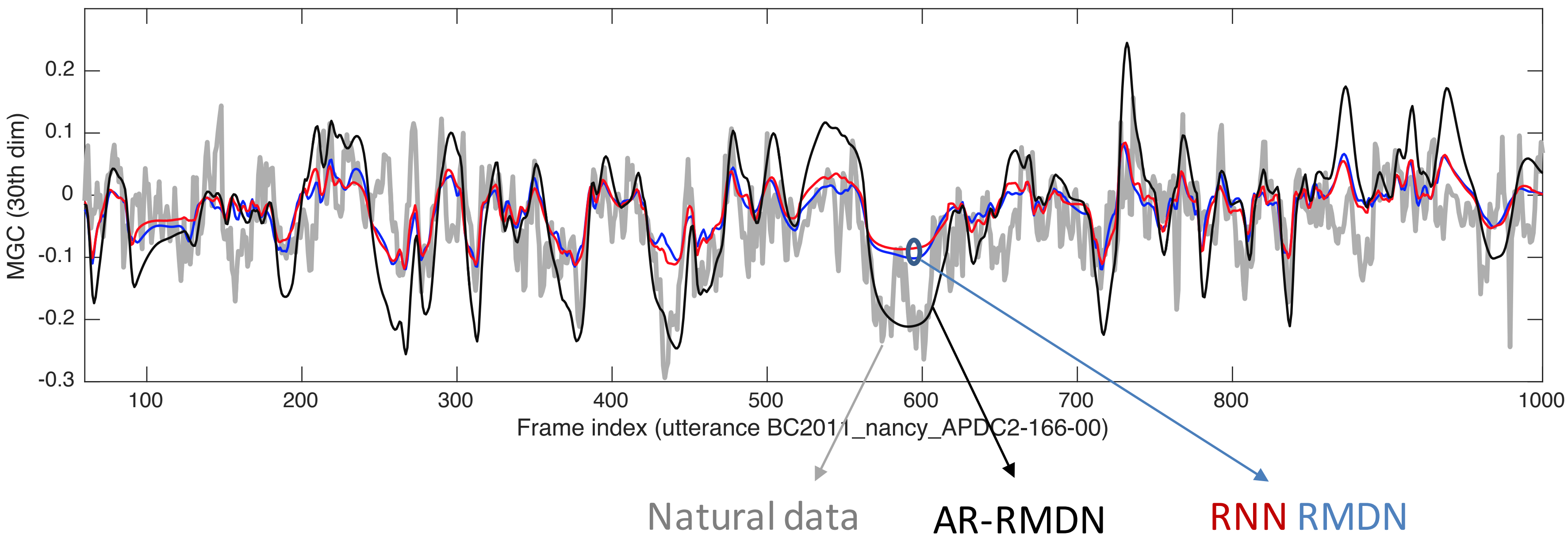


- smooth?

BEFORE INTERPRETATION

A glimpse of the results

Generated trajectories of the 30th dimension of mel-generalized cepstrum coefficients



- smooth?
- larger dynamic range?

INTERPRETATION

Signals and filters in AR-RMDN

- Simple case: 1st order AR

$$f(\boldsymbol{o}_{t-1}) = \boldsymbol{a} \odot \boldsymbol{o}_{t-1} + \boldsymbol{b}, \text{where } \boldsymbol{a} = [a_1, \cdots, a_D]^\top$$

  - for GMM in AR-RMDN

$$p(\boldsymbol{o}_t|\boldsymbol{o}_{t-1}; \mathcal{M}_t) = \sum_{m=1}^{M} w_t^m \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{t,d}^{m2}}} \exp(-\frac{(\boxed{o_{t,d} - f(o_{t-1,d})} - \mu_{t,d}^m)^2}{2{\sigma_{t,d}^m}^2})$$

$$o_{t,d} - f(o_{t-1,d}) = \underbrace{\boxed{o_{t,d} - a_d o_{t-1,d}}}_{c_{t,d}} - b_d$$

$d \in [1, D]$, dimension index

$t \in [1, T]$, frame index

Interpretation

Signals and filters in AR-RMDN

- Simple case: 1st order AR:

$$c_{t,d} = o_{t,d} - a_d o_{t-1,d}$$

$d \in [1, D]$, dimension index

$t \in [1, T]$, frame index

- if $T = 2$

$$c_{1,d} = o_{1,d}$$
$$c_{2,d} = o_{2,d} - a_d o_{1,d}$$

$$\Downarrow$$

$$\begin{bmatrix} 1 & 0 \\ -a_d & 1 \end{bmatrix} \cdot \begin{bmatrix} o_{1,d} \\ o_{2,d} \end{bmatrix} = \begin{bmatrix} c_{1,d} \\ c_{2,d} \end{bmatrix}$$

INTERPRETATION

Signals and filters in AR-RMDN

- Simple case: 1st order AR:

$$c_{t,d} = o_{t,d} - a_d o_{t-1,d}$$

$d \in [1, D]$, dimension index

$t \in [1, T]$, frame index

- if $T = 3$

$$c_{1,d} = o_{1,d}$$
$$c_{2,d} = o_{2,d} - a_d o_{1,d}$$
$$c_{3,d} = o_{3,d} - a_d o_{2,d}$$

$$\Downarrow$$

$$\begin{bmatrix} 1 & 0 & 0 \\ -a_d & 1 & 0 \\ 0 & -a_d & 1 \end{bmatrix} \cdot \begin{bmatrix} o_{1,d} \\ o_{2,d} \\ o_{3,d} \end{bmatrix} = \begin{bmatrix} c_{1,d} \\ c_{2,d} \\ c_{3,d} \end{bmatrix}$$

INTERPRETATION

Signals and filters in AR-RMDN

- Simple case: 1st order AR:

$$c_{t,d} = o_{t,d} - a_d o_{t-1,d}$$

$d \in [1, D]$, dimension index

$t \in [1, T]$, frame index

  - in general

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_d & 1 & 0 & \cdots & 0 \\ 0 & -a_d & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & 0 & -a_d & 1 \end{bmatrix} \cdot \begin{bmatrix} o_{1,d} \\ o_{2,d} \\ o_{3,d} \\ \vdots \\ o_{T,d} \end{bmatrix} = \begin{bmatrix} c_{1,d} \\ c_{2,d} \\ c_{3,d} \\ \vdots \\ c_{T,d} \end{bmatrix}$$

INTERPRETATION

Signals and filters in AR-RMDN

- Simple case: 1^{st} order AR:
  - a filtering processing

$d \in [1, D]$, dimension index

$$\boldsymbol{o}_{1:T,d} \longrightarrow \boxed{A^d(z) = 1 - a_d z^{-1}} \longrightarrow \boldsymbol{c}_{1:T,d}$$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_d & 1 & 0 & \cdots & 0 \\ 0 & -a_d & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & 0 & -a_d & 1 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} o_{1,d} \\ o_{2,d} \\ o_{3,d} \\ \vdots \\ o_{T,d} \end{bmatrix}}_{\boldsymbol{o}_{1:T,d}} = \underbrace{\begin{bmatrix} c_{1,d} \\ c_{2,d} \\ c_{3,d} \\ \vdots \\ c_{T,d} \end{bmatrix}}_{\boldsymbol{c}_{1:T,d}}$$

- $A^d(z)$ denotes the filter in z-domain

INTERPRETATION

Signals and filters in AR-RMDN

- Simple case: 1st order AR:
  - a filtering processing

$d \in [1, D]$, dimension index

$$\boldsymbol{o}_{1:T,d} \longleftarrow \boxed{H^d(z) = \frac{1}{A^d(z)}} \longleftarrow \boldsymbol{c}_{1:T,d}$$

$$\Updownarrow$$

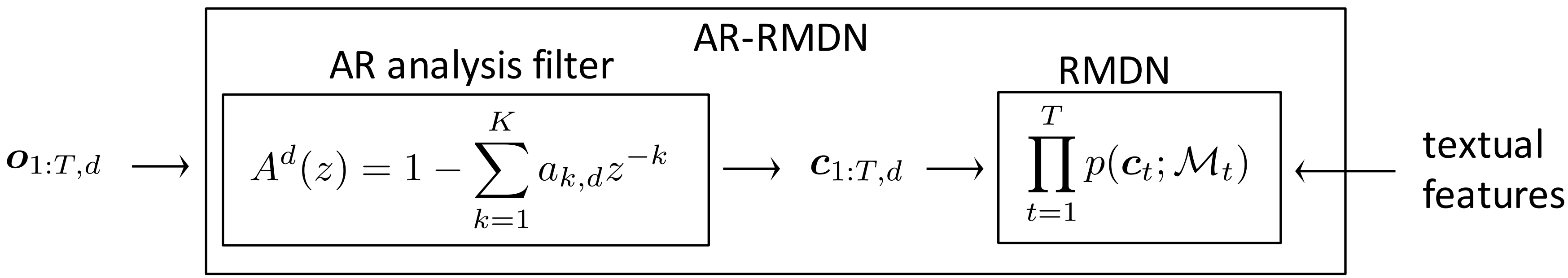
$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_d & 1 & 0 & \cdots & 0 \\ 0 & -a_d & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & 0 & -a_d & 1 \end{bmatrix}^{-1} \cdot \underbrace{\begin{bmatrix} c_{1,d} \\ c_{2,d} \\ c_{3,d} \\ \vdots \\ c_{T,d} \end{bmatrix}}_{\boldsymbol{c}_{1:T,d}} = \underbrace{\begin{bmatrix} o_{1,d} \\ o_{2,d} \\ o_{3,d} \\ \vdots \\ o_{T,d} \end{bmatrix}}_{\boldsymbol{o}_{1:T,d}}$$

- $H^d(z)$ denotes the filter in z-domain

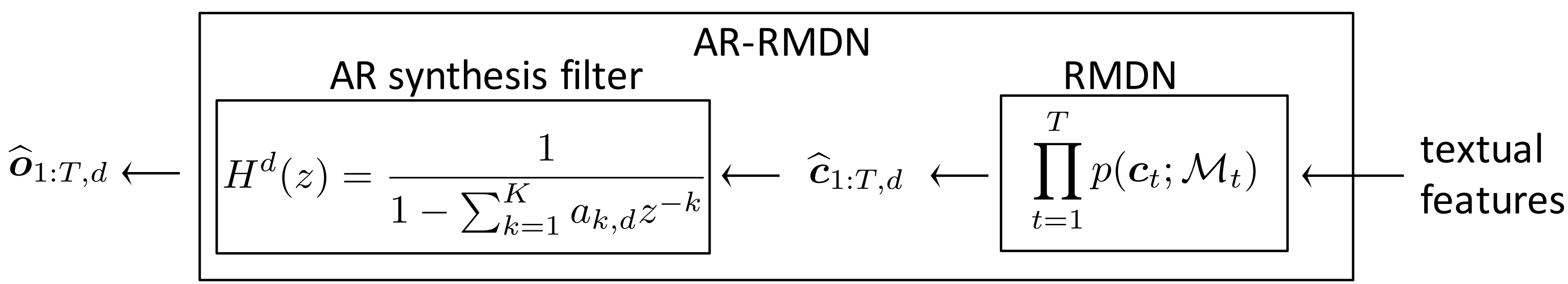
INTERPRETATION

Signals and filters in AR-RMDN

- With high order AR
  - joint training of AR filter and RMDN using back-propagation:



  - generation:



$d \in [1, D]$, dimension index

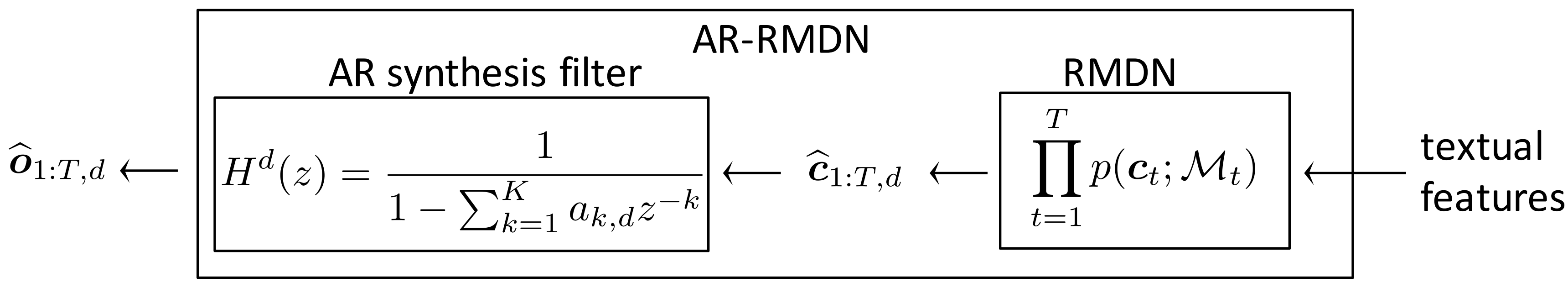
CONTENTS

- Introduction
- Model definition, interpretation, implementation
- Experiments and results
- Conclusion

IMPLEMENTATION

Stability of the filter

- $H^d(z)$ must be stable

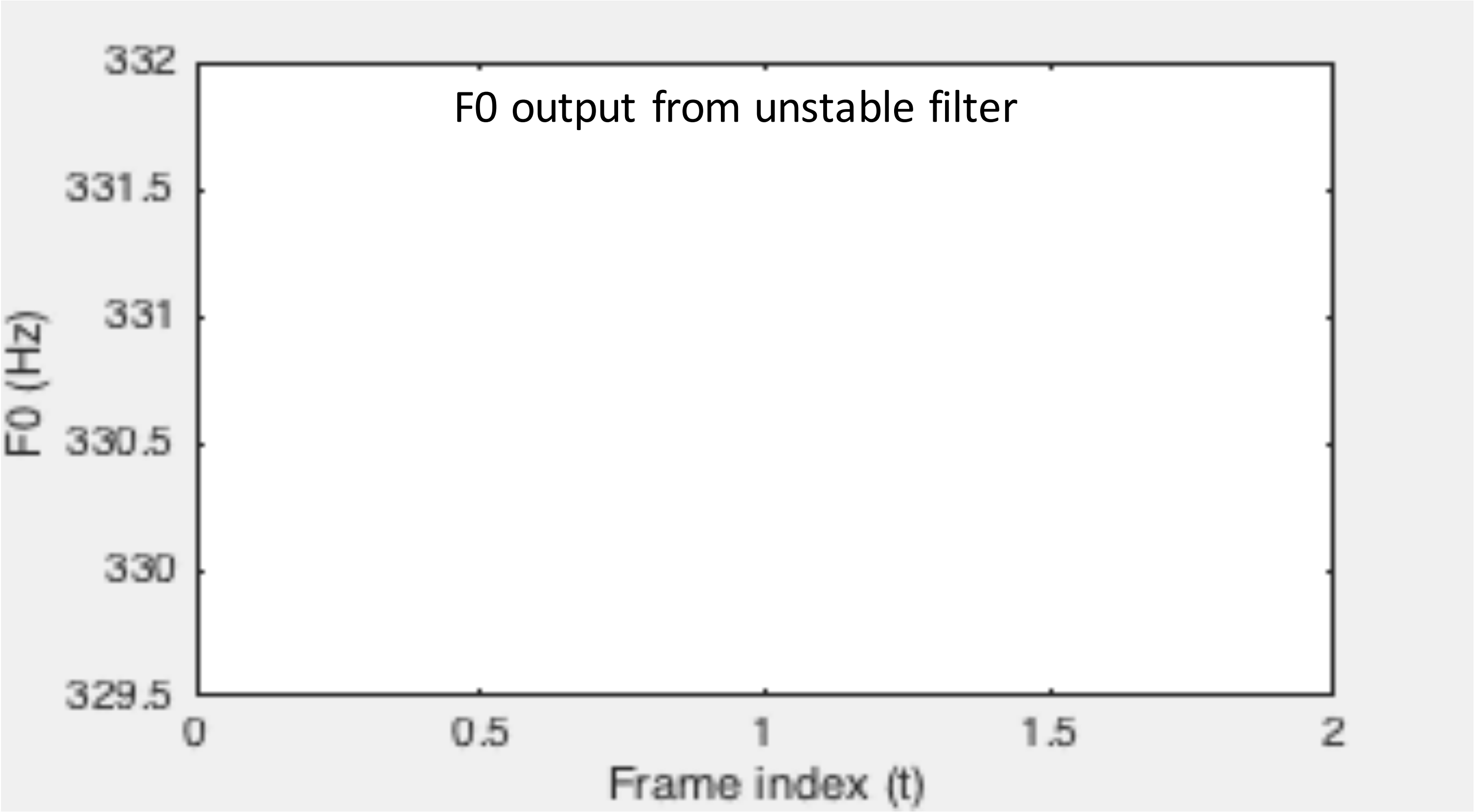


infinite impulse response filter

Implementation

Stability of the filter

- $H^d(z)$ must be stable
  - unstable $H^d(z)$ ------> $\widehat{\boldsymbol{o}}_{1:T,d}$ will be infinitely large !



Implementation

Stability of the filter

- General requirement:
  - $H^d(z)$'s poles are inside the unit circle [7]

- A simple trick (with a strong constraint):

$$H(z) = \frac{1}{1 - \sum_{k=1}^{K} a_k z^{-k}} = \prod_{k=1}^{K} \frac{1}{1 - \alpha_k z^{-1}} \qquad \alpha_k \in \mathbb{R}$$

  - requirement: $\alpha_k \in (-1, 1)$
  - how to: $\alpha_k = \tanh(\hat{\alpha}_k)$

[7] Oppenheim, A. V, Schafer, R. W., & Buck, J. R. (1999). *Discrete-time Signal Processing (2nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

CONTENTS

DATA

Corpus

Name	Size	Usage
Blizzard Challenge 2011 corpus [8]	About 12,000 utterances 16 hours	Validation set: 500 utterances Test set: 500 utterances Training set: the rest

Features

	Feature and description	Dimension
Input features $\boldsymbol{x}_t$	Phoneme sequence, stress, pitch accent... (extracted using Flite[9])	382
Target features $\boldsymbol{o}_t$	Mel-generalized cepstrum coefficients (MGC)	60
	Continuous F0 trajectory + Unvoiced/voiced flag	1 + 1
	Band aperiodicities (BAP)	25

[8] King, S., & Karaiskos, V. (2011). The Blizzard Challenge 2011. Retrieved from http://festvox.org/blizzard/bc2011/summary_Blizzard2011.pdf
[9] HTS Working Group. (2014). The English TTS System "Flite+hts_engine." Retrieved from http://hts-engine.sourceforge.net/

SYSTEMS

	Description		Configuration
RNN	without	$\Delta \boldsymbol{o}_t, \Delta^2 \boldsymbol{o}_t$	(1) feedforward tanh [512] (2) feedforward tanh [512] (3) Bi-LSTM [256] (4) Bi-LSTM [256] (5) feedforward linear output [259]
RMDN	without	$\Delta \boldsymbol{o}_t, \Delta^2 \boldsymbol{o}_t$	Same RNN + MDN: 1. GMM (2 mix) for MGC 2. GMM (2 mix) for F0 3. GMM (1 mix) for BAP 4. Binary distribution for U/V
RNN+MLPG [5]	with	$\Delta \boldsymbol{o}_t, \Delta^2 \boldsymbol{o}_t$	Same as RNN
AR-RMDN	without	$\Delta \boldsymbol{o}_t, \Delta^2 \boldsymbol{o}_t$	Same as RMDN, with AR 1. GMM (**2 mix, 1st order AR**) for MGC 2. GMM (**2 mix, 2nd order AR**) for F0 3. GMM (1 mix) for BAP 4. Binary distribution for U/V

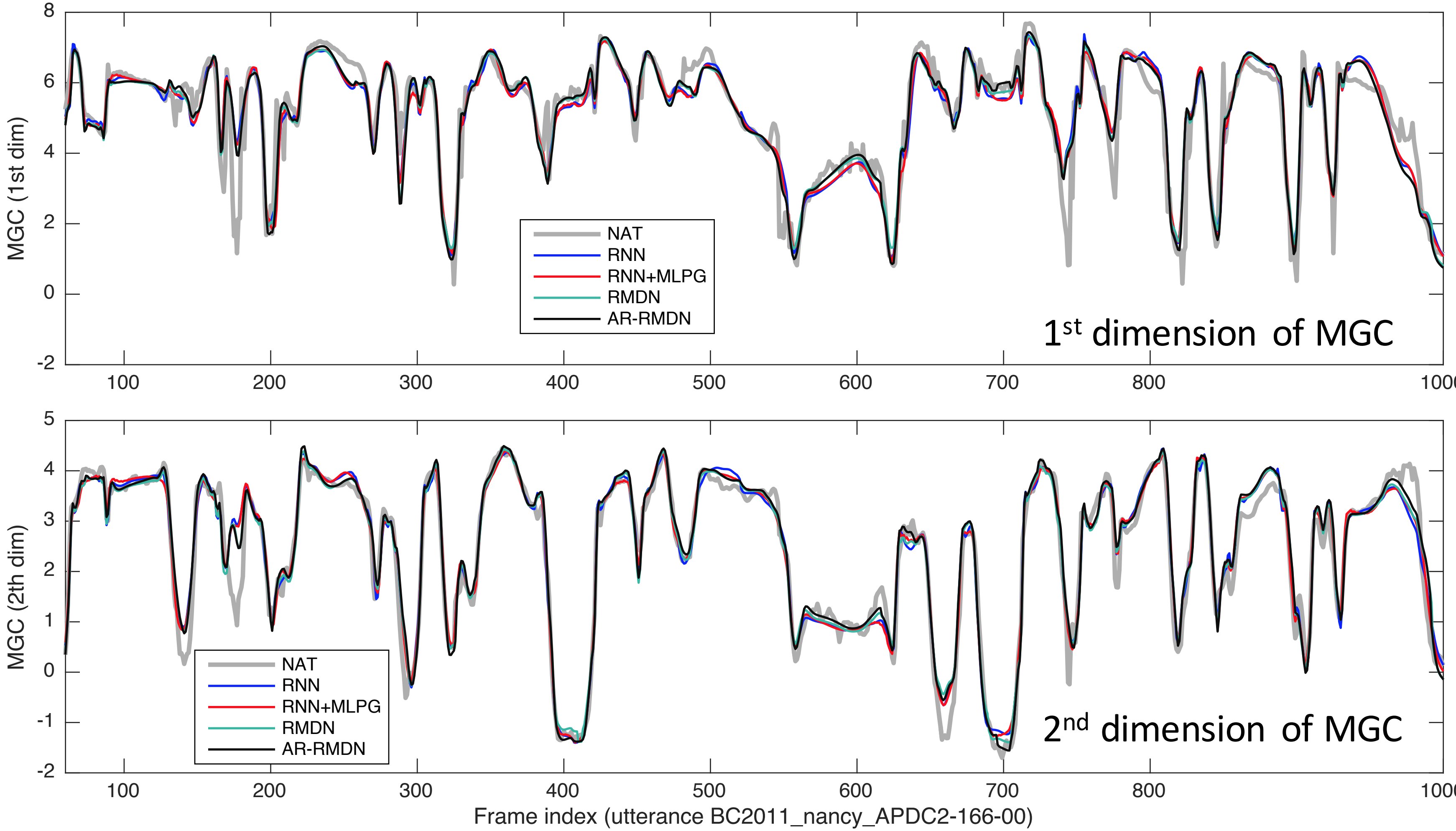
number of AR parameters: 1 * 60 + 60 + 2 * 1 + 1 = **123**

[5] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP* (Vol. 3, pp. 1315–1318).

EXPERIMENTS

Results

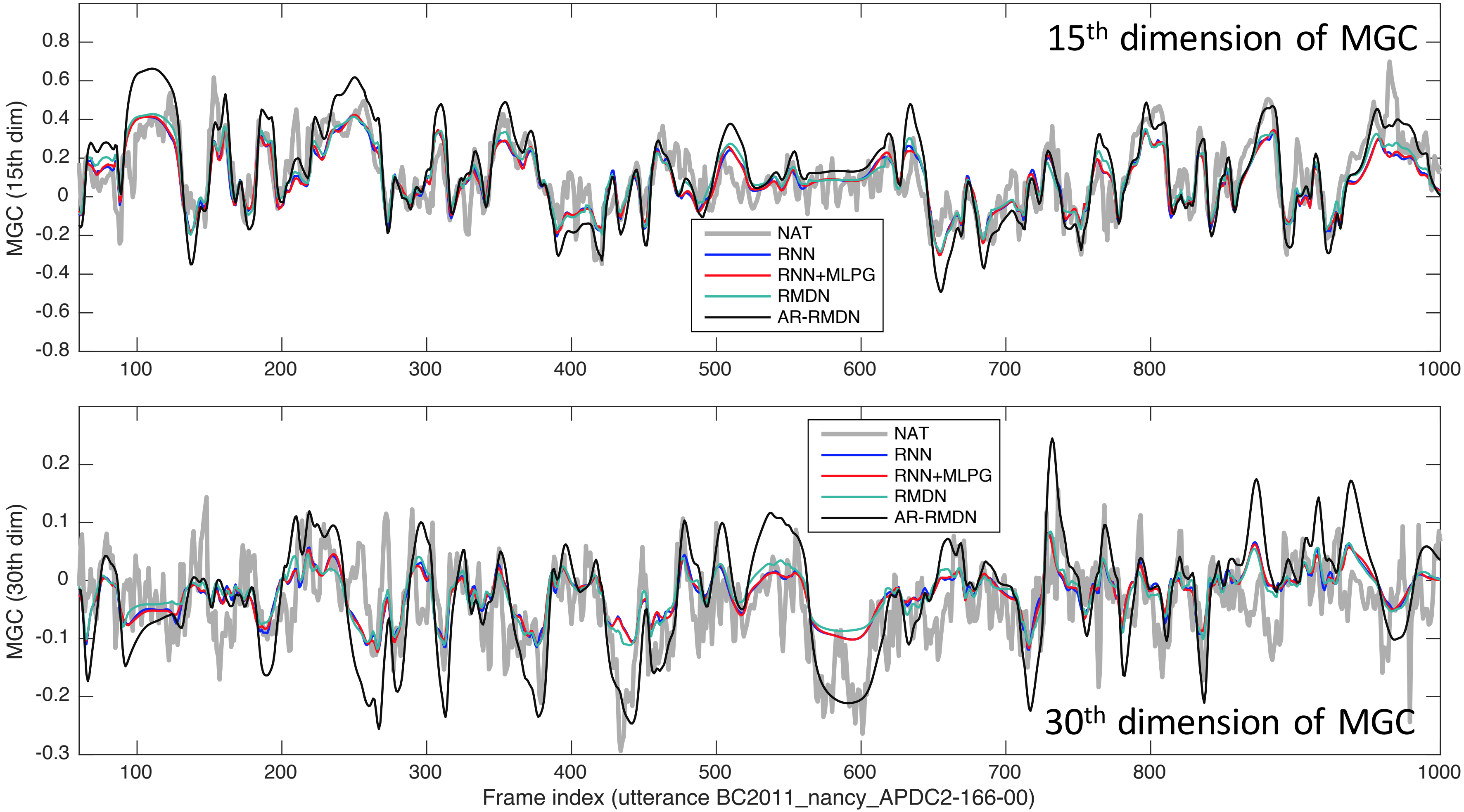
- Trajectory of the generated MGC:



EXPERIMENTS

Results

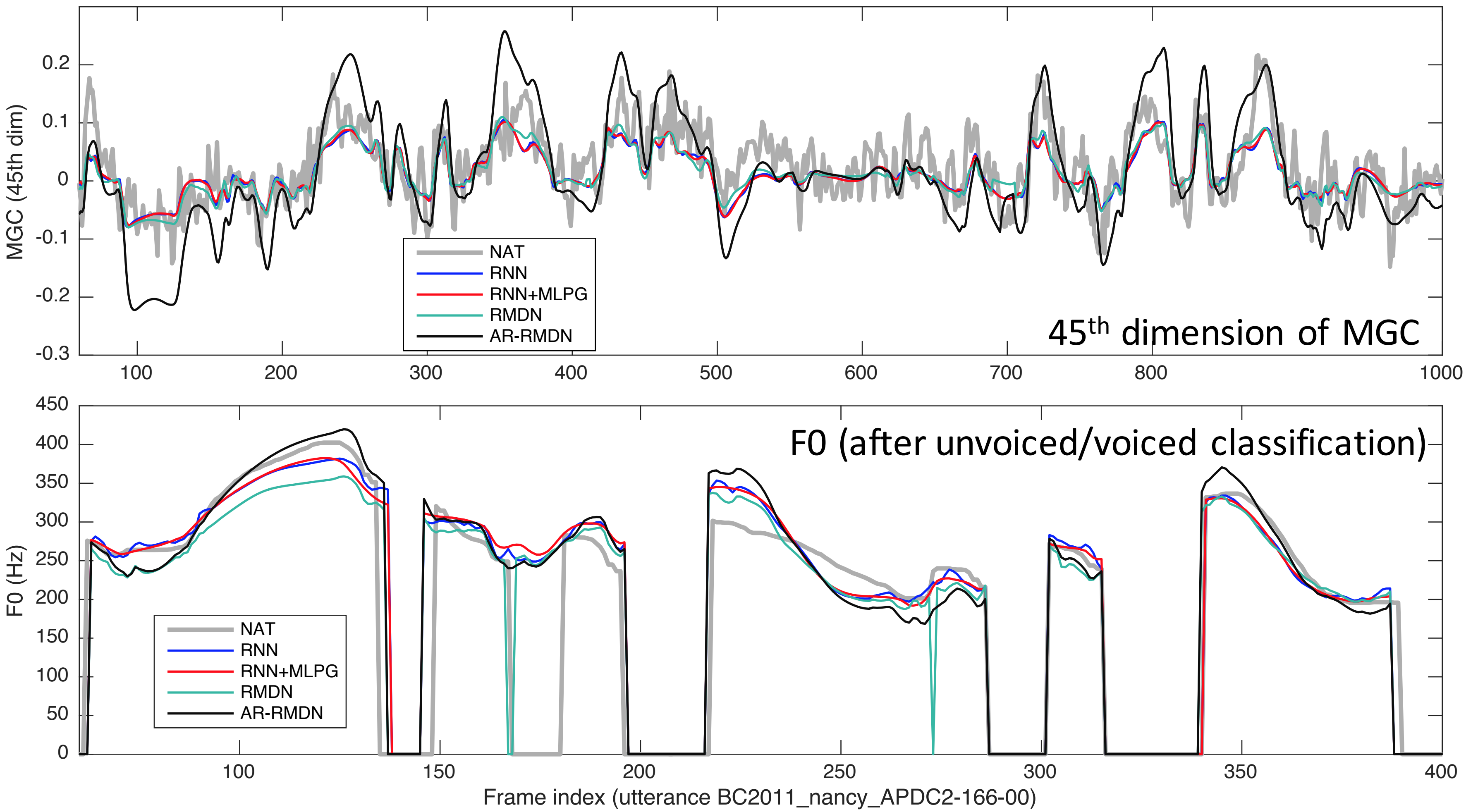
- Trajectory of the generated MGC:



Experiments

Results

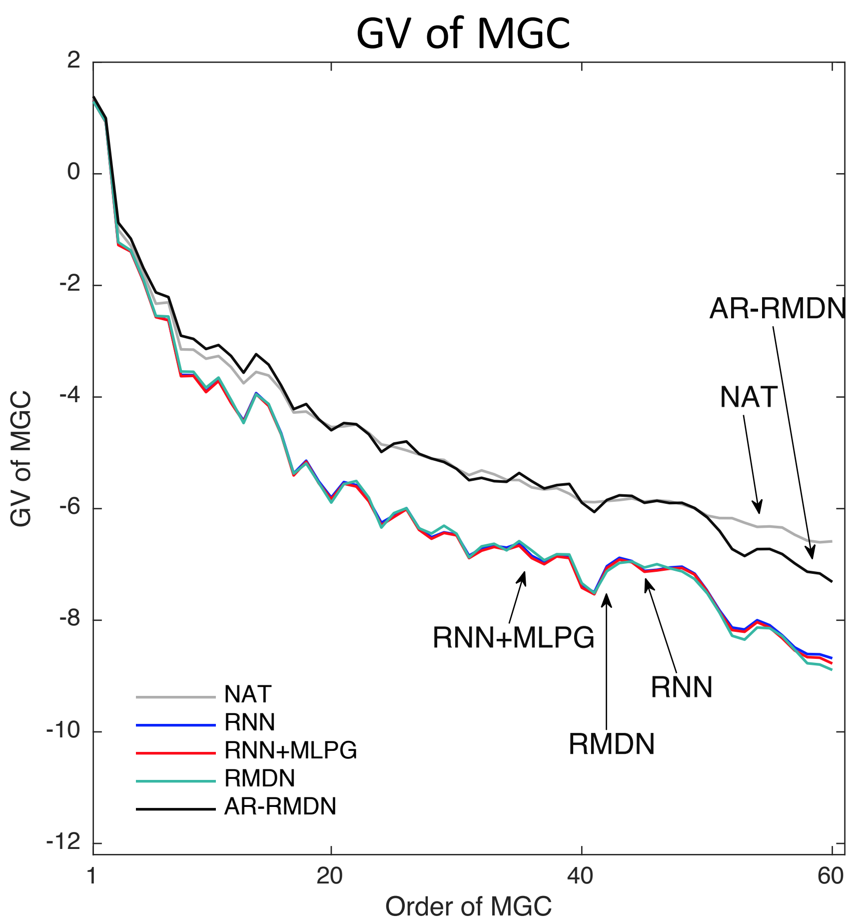
- Trajectory of the generated MGC and F0:

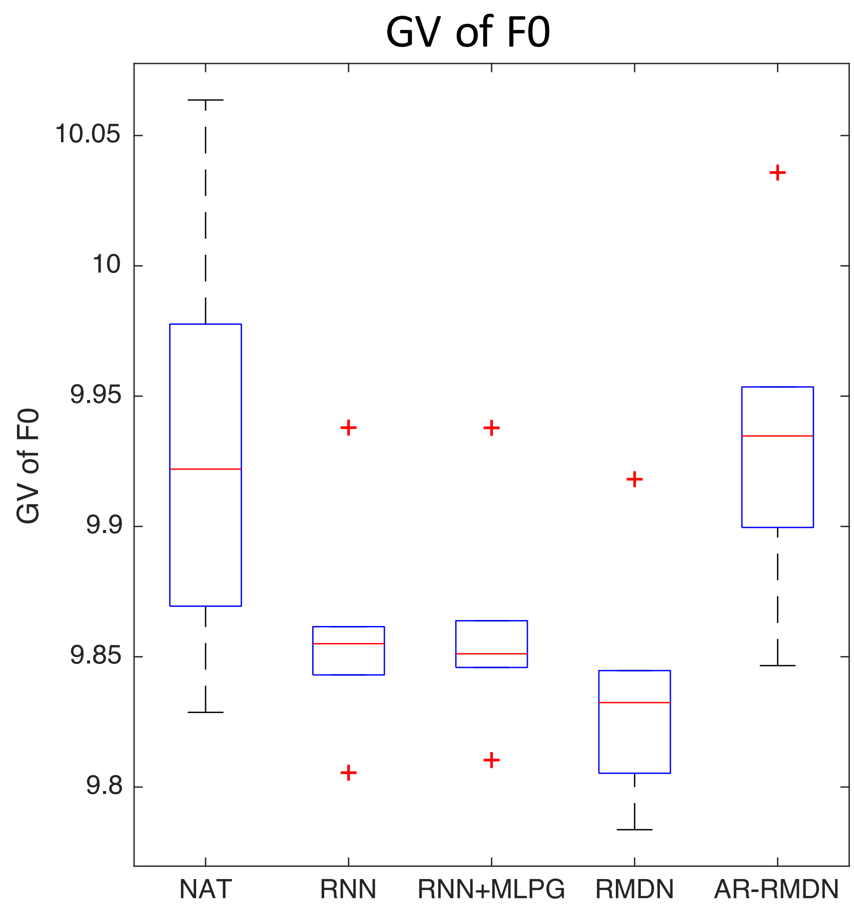


EXPERIMENTS

Analysis

- Global variance [10] of the generated MGC and F0 trajectories





[10] Toda, T., & Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for {HMM}-based speech synthesis. *IEICE Transactions on Information and Systems*, *90*(5), 816–824.

EXPERIMENTS

Analysis

- A larger dynamic range (higher GV)?

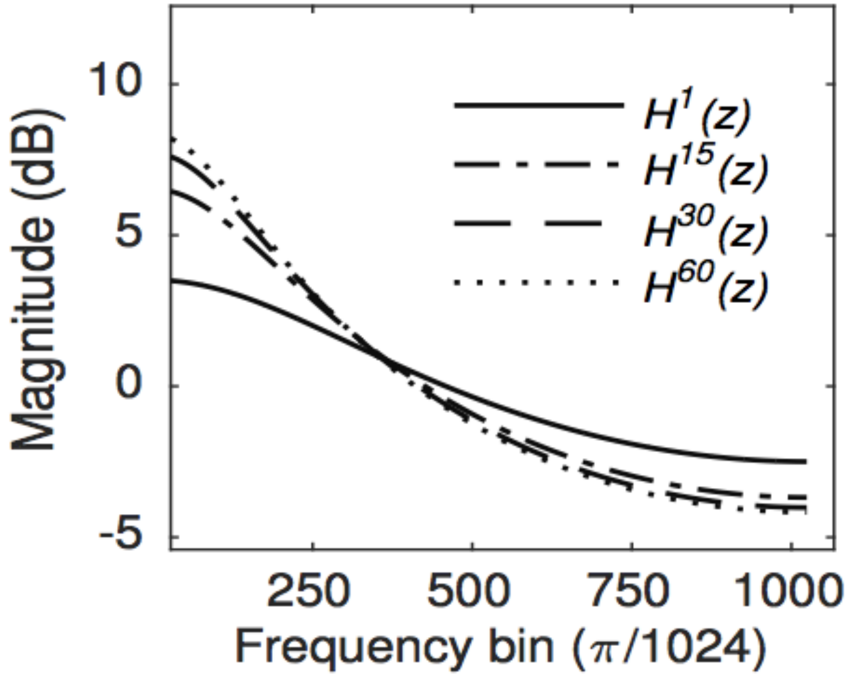
AR-RMDN

AR synthesis filter

RMDN part

$$\widehat{\boldsymbol{o}}_{1:T,d} \longleftarrow \boxed{H^d(z) = \frac{1}{1 - \sum_{k=1}^{K} a_{k,d} z^{-k}}} \longleftarrow \widehat{\boldsymbol{c}}_{1:T,d} \longleftarrow \boxed{\prod_{t=1}^{T} p(\boldsymbol{c}_t; \mathcal{M}_t)}$$

Frequency response of $H^d(z)$ for MGC



$d \in [1, D]$, dimension index

EXPERIMENTS

Results

- Samples:

	RNN	RNN+MLPG	RMDN	AR-RMDN	Natural
w/o formant enhancement					
with formant enhancement					

- formant enhancement was used in subjective evaluation

Other samples : here or tonywangx.github.io

CONTENTS

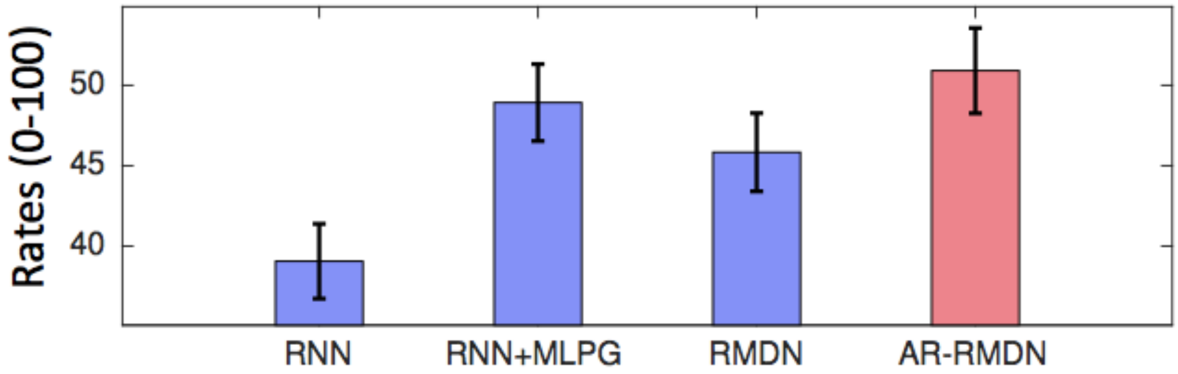
CONCLUSION

- Model:
  - AR-RMDN : a pair of AR filter + RMDN

- Results:
  - AR synthesis filter: a '**low-pass**' filter
  - AR analysis filter: a '**high-pass**' filter (see pp.45-46)

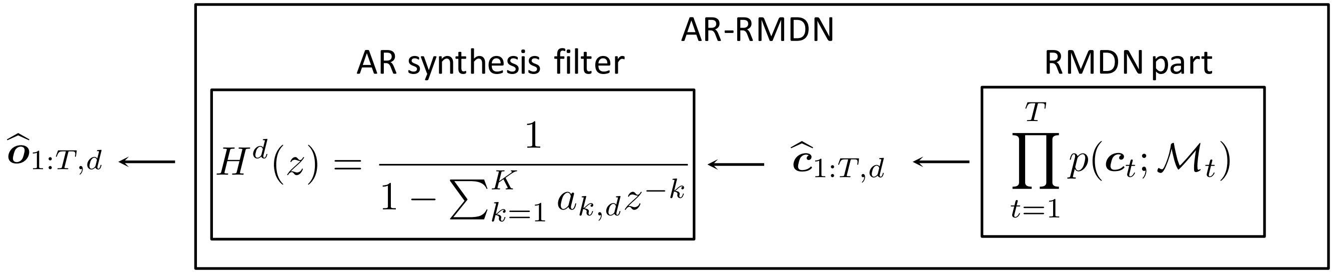
  - generated trajectories with a larger dynamic range
  - better perceived quality

Subjective evaluation (MUSHRA test)



Recent work

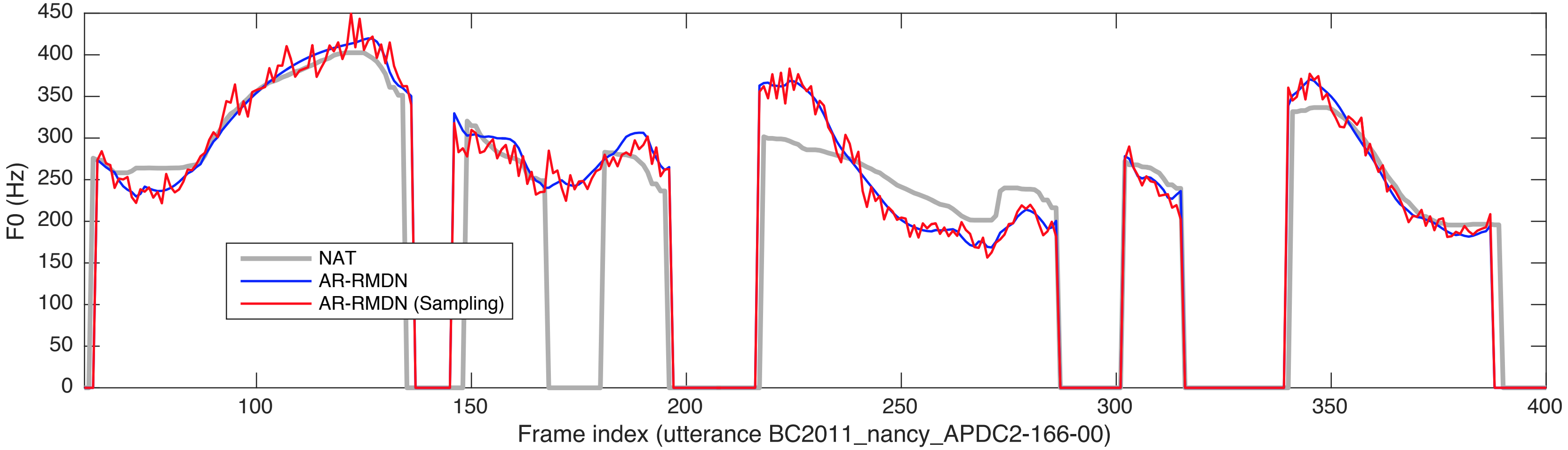
- Inverse AR filter with complex poles



  - using sigmoid & tanh function
  - please see attached slides pp.58

Future work

- Sampling from the model



  - AR is still weak for temporal correlation ?

Thank you for your attention

Q & A

- Toolkit, scripts, slides, and samples:

tonywangx.github.io