先端ニューラルネットワークに基づく音声合成

Deep and Shallow Autoregressive Neural Networks for Statistical Parametric Speech Synthesis

Abstract

- Fundamental frequency (F0) determines the pitch of sound, it conveys the linguistic and para-linguistic information of speech.
- Normal deep neural networks are imperfect for F0 modeling:
  - they generate dull and boring F0
  - they generate the same F0 for the same utterance
- SAR and DAR are improved neural networks for F0 modeling
  - they can generate F0 with natural variation
  - they can generate different F0s for the same utterance

Abbreviation

GMM	Gaussian mixture model
RNN	Recurrent neural network
MDN	Mixture density network
RMDN	Recurrent MDN
GV	Global variance
AR	Autoregressive
SAR	Shallow AR neural network
DAR	Deep AR neural network

Towards deep/shallow autoregressive neural network

- RMDN

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

Output

GMM parameter

hidden layers

Input

✗ independence across time

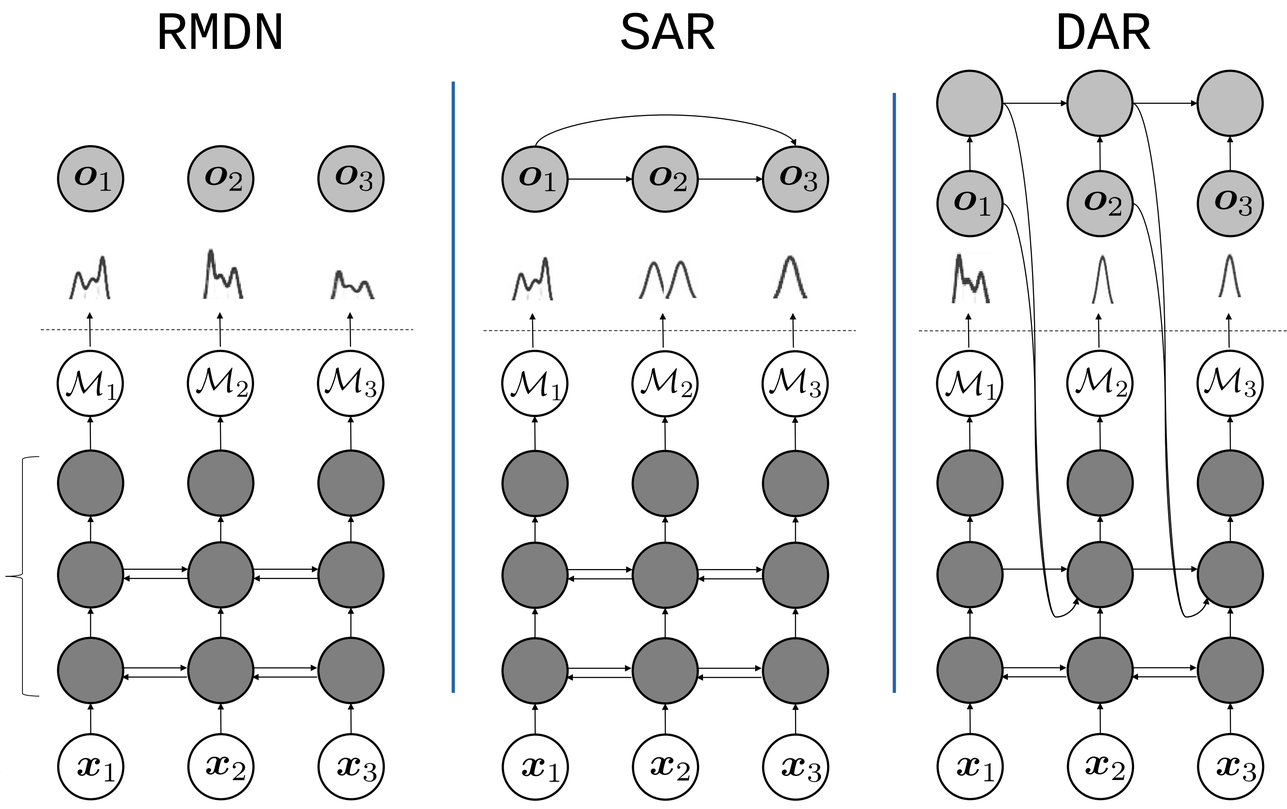
- Combing AR models with RMDN

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

  - SAR: linear AR dependence (*K* steps before)

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

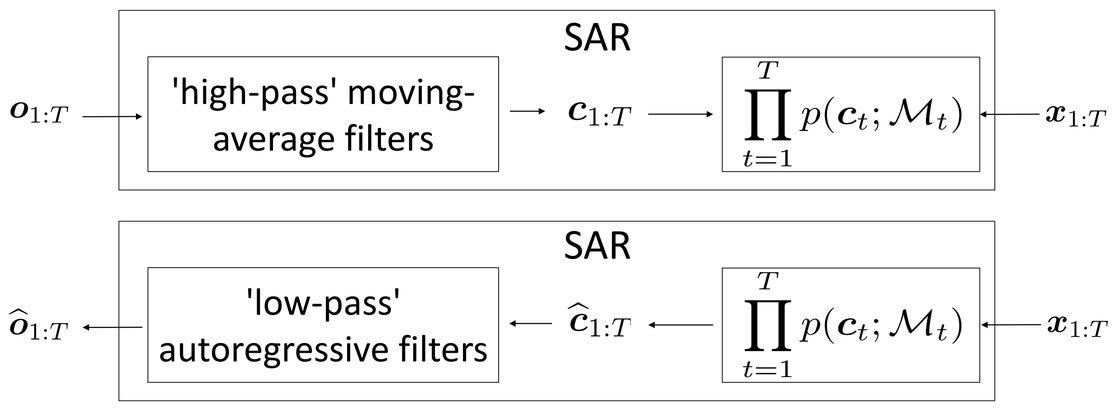
  - DAR: non-linear AR dependence (by RNN)



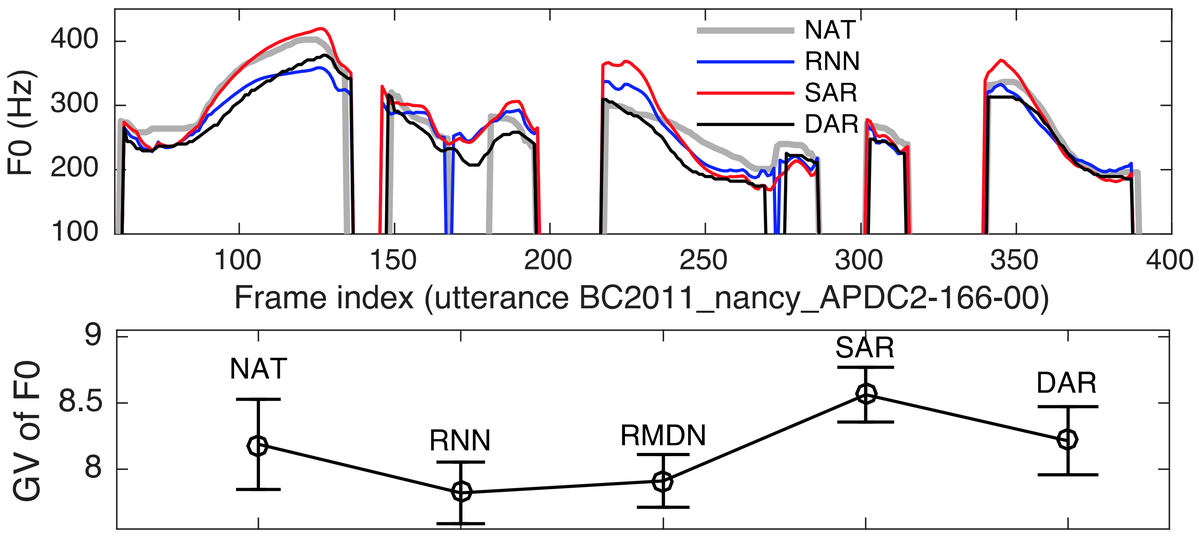
Experiment, analysis and interpretation

Results of SAR

- SAR = linear filters + RMDN

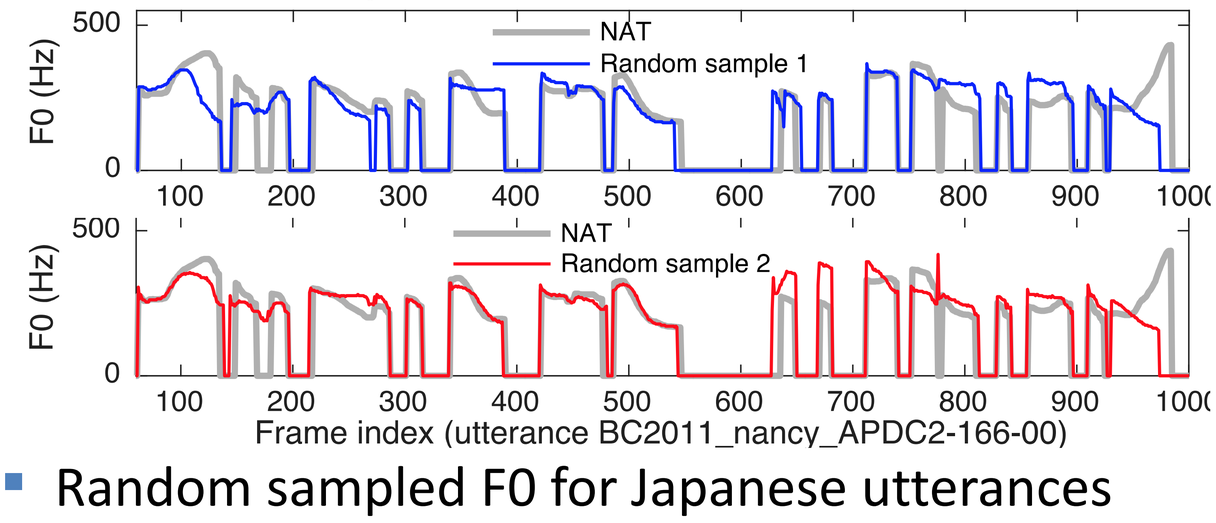


- Generated F0 by SAR is less over-smoothed

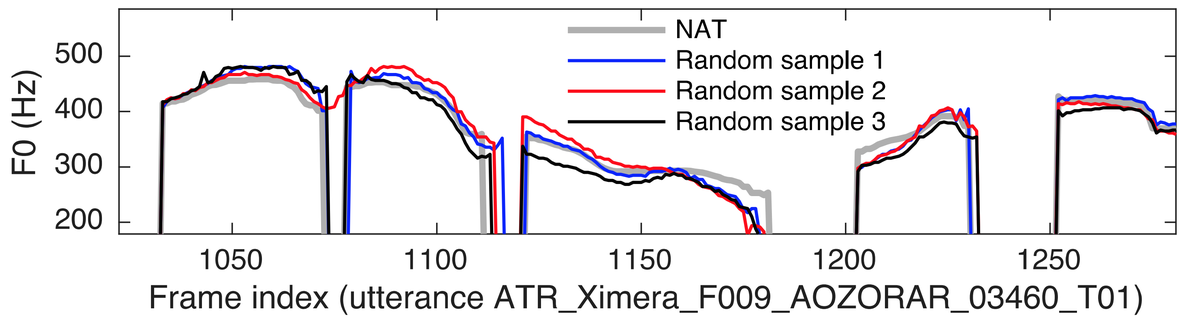


Results of DAR

- Random sampled F0s show natural variation



- Random sampled F0 for Japanese utterances



Future work

➢ SAR/DAR will incorporate rich linguistic and para-linguistic information that influence F0

連絡先：ワン　シン　／国立情報学研究所 総合研究大学院大学複合科学研究科
Email：wangxin@nii.ac.jp