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

National Institute of Informatics, Japan & Aalto University, Finland
2018-04-17

contact: wangxin@nii.ac.jp
we welcome critical comments, suggestions, and discussion
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 \([1]\)

  ![Diagram of conventional TTS pipeline]

  - Text → Front-end → Linguistic features → Back-end → Speech

- SPSS back-end \([2,3]\)

  ![Diagram of SPSS back-end]

  - Linguistic features → Acoustic models → MGC & BAP → F0 → Waveform generators → Speech

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

---

INTRODUCTION

Topic of this work

Better modules for SPSS back-end?

- Recurrent neural networks (RNNs)
- Autoregressive (AR) models
- General adversarial network (GAN)

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

WORLD vocoder
+ Phrase recovery
WaveNet-based vocoder

Acoustic models

Linguistic features

MGC & BAP

F0

Waveform generators

Speech
CONTENTS

- Introduction
- Models and modules
- Experiments
- Summary
MODELS & METHODS

Speech waveforms

WORLD

MGC & BAP & F0

RNN

Linguistic features

Waveform generators

Acoustic models
Models & Methods

Acoustic models

- Baseline RNN

- Sequence of linguistic features \( \{x_1, \cdots, x_t, \cdots \} \)
- Sequence of generated acoustic features \( \{\hat{o}_1, \cdots, \hat{o}_t, \cdots \} \)
Acoustic models

- Baseline RNN

\[
M_t = \{\mu_t\}, \quad \text{where} \quad \mu_t = \mathcal{H}_{\Theta}^{(RNN)}(x_{1:T}, t)
\]

\[
\hat{o}_t = \mu_t
\]

Neural network

\[
\mathcal{H}_{\Theta}^{(RNN)}(\cdot)
\]

Probabilistic models

\[
p(o_{1:T} | x_{1:T}; \Theta) = \prod_{t=1}^{T} p(o_t | x_{1:T}; \Theta) = \prod_{t=1}^{T} \mathcal{N}(o_t; \mu_t, I)
\]

MODELS & METHODS

Acoustic models

- Baseline RNN

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

\[
p(o_{1:T}|x_{1:T}; \Theta) = \prod_{t=1}^{T} p(o_t|x_{1:T}; \Theta) = \prod_{t=1}^{T} \mathcal{N}(o_t; \mu_t, I)
\]
Acoustic models

- Shallow AR (SAR)

Alternative interpretation: trainable filter + RNN

\[ p(o_{1:T}|x_{1:T}; \Theta, \Psi) = \prod_{t=1}^{T} p(o_t|o_{t-K:t-1}, x_{1:T}; \Theta) = \prod_{t=1}^{T} \mathcal{N}(o_t; \mu_t + f_\Psi(o_{t-K:t-1}), I) \]

Acoustic models

- Deep AR (DAR)

$$p(o_{1:T} | x_{1:T}; \Theta) = \prod_{t=1}^{T} p(o_t | o_{1:t-1}, x_{1:T}; \Theta)$$

- Only for (quantized) F0 modeling

Acoustic models

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

Models & Methods

Acoustic features (natural) → Discriminator

Residual generator → Acoustic features (generated)

Acoustic model

Linguistic features

MODELS & METHODS

Acoustic models

Speech waveforms

WORLD

Waveform generators

F0

MGC

Acoustic models

DAR

SAR

RNN

Linguistic features

• BAP is not shown
Models & Methods

Acoustic models

- SAR-Wo
- SGA-Wo
- RGA-Wo
- RNN-Wo

Waveform generators

- WORLD
- GAN
- DAR
- SAR
- RNN

Linguistic features

- F0
- MGC

- 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

MODELS & METHODS

Waveform generators

- WaveNet-based vocoder [12, 13]

How to generate (search) a waveform:

1. **Exploration**: sampling
2. **Exploitation**: picking one-best

---


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

WaveNet

Phrase recovery

PML

WORLD

GAN

DAR

SAR

RNN

F0

minimum phase

SAR-Wa

SAR-Pr

SAR-Pm

SAR-Wo

SGA-Wo

RGA-Wo

RNN-Wo

Acoustic models

Waveform generators

Linguistic features
EXPERIMENTS

Configuration

Data

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATR Ximera F009 voice [14]</td>
<td>~30,000 utterances 48 hours</td>
<td>Sampling rate: 48kHz Japanese, neutral style, reading</td>
</tr>
</tbody>
</table>

- Recording period: over 1 year

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>Phone identity, prosodic tags ...</td>
</tr>
<tr>
<td>Acoustic</td>
<td>MGC</td>
</tr>
<tr>
<td></td>
<td>BAP</td>
</tr>
<tr>
<td></td>
<td>F0</td>
</tr>
</tbody>
</table>

Front-end: Open JTalk [15]

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 $\Delta, \Delta^2$, nor formant enhancement
- Sampling rate: 48kHz & 16kHz, except SAR-Wa at 16kHz (10 bits, $\mu$-law)
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>Natural</th>
<th>Abs-Pm</th>
<th>Abs-Wo</th>
<th>SAR-Wa</th>
<th>SAR-Pr</th>
<th>SAR-Pm</th>
<th>SAR-Wo</th>
<th>SGA-Wo</th>
<th>RGA-Wo</th>
<th>RNN-Wo</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>48kHz</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>16kHz</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PML</strong></td>
<td><strong>WORLD</strong></td>
<td>WaveNet</td>
<td>+PhraseRec</td>
<td><strong>WORLD</strong></td>
<td><strong>PML</strong></td>
<td><strong>WORLD</strong></td>
<td><strong>PML</strong></td>
<td><strong>WORLD</strong></td>
<td><strong>PML</strong></td>
<td><strong>WORLD</strong></td>
</tr>
<tr>
<td>natural</td>
<td>natural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>GAN</td>
<td>SAR</td>
<td>GAN</td>
<td>RNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Natural* and *Abs-Pm* indicate natural and manipulated natural data, respectively. *Abs-Wo* indicates manipulated data with a world model. *SAR-Wa* and *SAR-Pr* indicate Sar model trained with world or phrase-recognition data, respectively. *SAR-Pm* indicates Sar model trained with manipulated data. *SGA-Wo* and *RGA-Wo* indicate models trained with SGA or RGA data, respectively. *RNN-Wo* indicates models trained with RNN data.
EXPERIMENTS

Quality scores

<table>
<thead>
<tr>
<th></th>
<th>16kHz</th>
<th>48kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>Abs-Pm</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Abs-Wo</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>SAR-Wa</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>SAR-Pr</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>SAR-Pm</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>SAR-Wo</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>SGA-Wo</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>RGA-Wo</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>RNN-Wo</td>
<td>-0.5</td>
<td></td>
</tr>
</tbody>
</table>

PML  WORLD  WaveNet  +PhraseRec WORLD  PML  WORLD

natural  natural  SAR  GAN  SAR  GAN  RNN  RNN

24
EXPERIMENTS

Similarity scores

![Bar chart showing MOS scores for different conditions.

- Natural
- Abs-Pm
- Abs-Wo
- SAR-Wa
- SAR-Pr
- SAR-Pm
- SAR-Wo
- SGA-Wo
- RGA-Wo
- RNN-Wo

- PML
- WORLD
- WaveNet
- +PhraseRec
- PML
- WORLD

- natural
- natural
- SAR
- GAN
- SAR
- GAN
- RNN
- RNN

- 16kHz
- 48kHz]
CONTENTS

- Introduction
- Models and methods
- Experiments
- Summary
SUMMARY

Plug and test

- SAR:
  - Avoid conditional independence assumption
  - Alleviate the over-smoothing effect  

WaveNet

- Generation method: one-best generation + random sampling
- Less distortion of harmonics

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

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 - WAVE NET

Conditional network

WaveNet-Backend

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

Learning curve

Generation method

Generation with other acoustic models

Training based on generated features

APPENDIX - WAVE NET

Structure

\[ P(o_t|o_{t-R:t-1}, c_{1:N}) \]

\[ \text{Post-processing network} \]

\[ \text{Wavenet block 1} \]

\[ \text{Wavenet block 2} \]

\[ \text{Wavenet block M} \]

Clock rate: 16kHz

\[ \text{Conditional feature network} \]

Clock rate: 200Hz (frame shift = 5ms)
Conditional network

- **Architecture of the conditional network**

  - Trial 1
    - $F_0$, $MGC$
    - $c_{1:N}$
    - Bi-LSTM
    - CNN
    - Concat.
    - Linear
    - $l_{1:N}$

  - Trial 2
    - $c_{1:N}$
    - Bi-LSTM
    - CNN
    - Linear
    - $l_{1:N}$

  - Trial 3
    - $c_{1:N}$
    - Bi-LSTM
    - CNN
    - Concat.
    - Linear
    - $l_{1:N}$
**APPENDIX - WAVE_NET**

**Conditional network**

- Experiments on WaveNet Vocoder
  - Given generated MGC/F0

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>Trial 1</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>None</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>Trial 2</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>LSTM+CNN</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>Trial 3</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
<tr>
<td>LSTM+CNN+skip-F0</td>
<td><img src="sample1" alt="Audio" /></td>
<td><img src="sample2" alt="Audio" /></td>
<td><img src="sample3" alt="Audio" /></td>
</tr>
</tbody>
</table>
APPENDIX - WAVE_NET

WaveNet backend

- Architecture

WaveNet-backend only uses random sampling

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th>sample4</th>
<th>sample5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td><img src="image1.png" alt="Sample1" /></td>
<td><img src="image2.png" alt="Sample2" /></td>
<td><img src="image3.png" alt="Sample3" /></td>
<td><img src="image4.png" alt="Sample4" /></td>
<td><img src="image5.png" alt="Sample5" /></td>
</tr>
<tr>
<td>WaveNet-vocoder</td>
<td><img src="image1.png" alt="Sample1" /></td>
<td><img src="image2.png" alt="Sample2" /></td>
<td><img src="image3.png" alt="Sample3" /></td>
<td><img src="image4.png" alt="Sample4" /></td>
<td><img src="image5.png" alt="Sample5" /></td>
</tr>
<tr>
<td>WaveNet-backend</td>
<td><img src="image1.png" alt="Sample1" /></td>
<td><img src="image2.png" alt="Sample2" /></td>
<td><img src="image3.png" alt="Sample3" /></td>
<td><img src="image4.png" alt="Sample4" /></td>
<td><img src="image5.png" alt="Sample5" /></td>
</tr>
</tbody>
</table>
We trained WaveNet backend for more than 100 epochs.
APPENDIX - WAVENET

Generation method
APPENDIX - WAVE NET

<table>
<thead>
<tr>
<th>Sampling point</th>
<th>Waveform level</th>
</tr>
</thead>
<tbody>
<tr>
<td>8900.0</td>
<td>0</td>
</tr>
<tr>
<td>8905.0</td>
<td>200</td>
</tr>
<tr>
<td>8910.0</td>
<td>400</td>
</tr>
<tr>
<td>8915.0</td>
<td>600</td>
</tr>
<tr>
<td>8920.0</td>
<td>800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling point</th>
<th>Waveform level</th>
</tr>
</thead>
<tbody>
<tr>
<td>9300.0</td>
<td>0</td>
</tr>
<tr>
<td>9305.0</td>
<td>200</td>
</tr>
<tr>
<td>9310.0</td>
<td>400</td>
</tr>
<tr>
<td>9315.0</td>
<td>600</td>
</tr>
<tr>
<td>9320.0</td>
<td>800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sampling point</th>
<th>Waveform level</th>
</tr>
</thead>
<tbody>
<tr>
<td>9300.0</td>
<td>0</td>
</tr>
<tr>
<td>9305.0</td>
<td>200</td>
</tr>
<tr>
<td>9310.0</td>
<td>400</td>
</tr>
<tr>
<td>9315.0</td>
<td>600</td>
</tr>
<tr>
<td>9320.0</td>
<td>800</td>
</tr>
</tbody>
</table>
APPENDIX - WAVE NET

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

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>Mix</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>Random</td>
<td>🎧</td>
<td>🎧</td>
<td>🎧</td>
</tr>
</tbody>
</table>
Voiced:

Unvoiced:
testing

Voiced:
testing

Unvoiced:
testing
APPENDIX - WAVE NET

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

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td></td>
</tr>
<tr>
<td>Mix</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td>100%</td>
</tr>
<tr>
<td>Mix2</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td>75%</td>
</tr>
<tr>
<td>Mix3</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td>50%</td>
</tr>
<tr>
<td>Mix4</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td>20%</td>
</tr>
<tr>
<td>Random</td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td><img src="#" alt="Sound" /></td>
<td>00%</td>
</tr>
</tbody>
</table>
# Appendix - WaveNet

## Generation method

- One-best + random sampling

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
</tr>
<tr>
<td>WaveNet backend</td>
<td>Mix</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
</tr>
<tr>
<td>Random</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
</tr>
<tr>
<td>WaveNet vocoder</td>
<td>Mix</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound1.png" alt="Sound" /></td>
</tr>
</tbody>
</table>
## Appendix - WaveNet

WaveNet-vocoder + other acoustic models

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
<tr>
<td>SAR + DAR</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
<tr>
<td>SGA + DAR</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
<tr>
<td>RGA + DAR</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
<tr>
<td>RNN + DAR</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
<tr>
<td>extended SAR + DAR</td>
<td><img src="sound1.png" alt="Sound" /></td>
<td><img src="sound2.png" alt="Sound" /></td>
<td><img src="sound3.png" alt="Sound" /></td>
</tr>
</tbody>
</table>
**APPENDIX - WAVENET**

WaveNet-vocoder: training using generated MGC

<table>
<thead>
<tr>
<th></th>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Natural</strong></td>
<td><img src="image1" alt="Sound" /></td>
<td><img src="image2" alt="Sound" /></td>
<td><img src="image3" alt="Sound" /></td>
</tr>
<tr>
<td><strong>WaveNet-Backend</strong></td>
<td><img src="image4" alt="Sound" /></td>
<td><img src="image5" alt="Sound" /></td>
<td><img src="image6" alt="Sound" /></td>
</tr>
<tr>
<td><strong>Trained on natural MGC</strong></td>
<td><img src="image7" alt="Sound" /></td>
<td><img src="image8" alt="Sound" /></td>
<td><img src="image9" alt="Sound" /></td>
</tr>
<tr>
<td><strong>Trained on generated MGC Epoch 35</strong></td>
<td><img src="image10" alt="Sound" /></td>
<td><img src="image11" alt="Sound" /></td>
<td><img src="image12" alt="Sound" /></td>
</tr>
<tr>
<td><strong>Trained on generated MGC Epoch 45</strong></td>
<td><img src="image13" alt="Sound" /></td>
<td><img src="image14" alt="Sound" /></td>
<td><img src="image15" alt="Sound" /></td>
</tr>
<tr>
<td><strong>Trained on generated MGC Epoch 55</strong></td>
<td><img src="image16" alt="Sound" /></td>
<td><img src="image17" alt="Sound" /></td>
<td><img src="image18" alt="Sound" /></td>
</tr>
</tbody>
</table>
APPENDIX — ACOUSTIC MODELS

General comparison

SAR & DAR

SAR extension

More details: http://tonywangx.github.io/pdfs/talk.pdf
APPENDIX – ACOUSTIC MODELS

General comparison

- Generated trajectories

Diagram showing comparison of generated trajectories for MGC 1st and 5th dimensions.
APPENDIX – ACOUSTIC MODELS

General comparison

- Generated trajectories

---

Frame index (utterance ATR_Ximera_F009_AOZORAR_03372_T01)

![MGC 15\textsuperscript{th} dim](chart1)

![MGC 31\textsuperscript{th} dim](chart2)

<table>
<thead>
<tr>
<th>Natural</th>
<th>RNN</th>
<th>SAR</th>
</tr>
</thead>
</table>

**MGC 15\textsuperscript{th} dim**

**MGC 31\textsuperscript{th} dim**
APPENDIX – ACOUSTIC MODELS

General comparison

- Generated trajectories

![Graphs showing generated trajectories for different acoustic models.](image-url)
APPENDIX — ACOUSTIC MODELS

Global variance

Modulation spectrum (MGC 31th)
APPENDIX — ACOUSTIC MODELS

SAR & DAR

- Toy SAR example

  - SAR versus RMDN with a recurrent output layer \(^{[11]}\)

\[
RMDN: \ h_{t} \rightarrow \ h_{t-1} \rightarrow \ \mu_{1} \rightarrow \ \mu_{2} \rightarrow \ o_{2}
\]

\[
SAR: \ x_{t} \rightarrow \ h_{t} \rightarrow \ \mu_{1} \rightarrow \ \mu_{2} \rightarrow \ o_{2}
\]

- Assume \( o_{t} \in \mathbb{R} \) and \( \Sigma_{t} = 1 \), linear activation function

APPENDIX — ACOUSTIC MODELS

SAR & DAR

Toy SAR example

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

\[ \mu_1 = w^T h_1 + b \]

\[ \mu_2 = w^T h_2 + b + w_{\mu}\mu_1 = \tilde{\mu}_2 + w_{\mu}\mu_1 \]

SAR

\[ p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \mu_2 + ao_1, 1) \]

\[ \mu_1 = w^T h_1 + b \]

\[ \mu_2 = w^T h_2 + b \]

**APPENDIX — ACOUSTIC MODELS**

**SAR & DAR**

- Toy SAR example

![Diagram of SAR & DAR](image)

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

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(o - \mu)^\top \Sigma^{-1}(o - \mu)\right)$$

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

Dependency between $\mu_t$ or $o_t$?

- SAR

$$p(o_{1:2}) = \mathcal{N}(o_1; \mu_1, 1)\mathcal{N}(o_2; \mu_2 + ao_1, 1)$$

$$= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(o - \mu)^\top \Sigma^{-1}(o - \mu)\right)$$

$$o = [o_1, o_2]^\top \quad \mu = [\mu_1, \mu_2 + a\mu_1]^\top \quad \Sigma = \begin{bmatrix} 1 & a \\ a & 1 + a^2 \end{bmatrix}$$
APPENDIX — ACOUSTIC MODELS

SAR & DAR

- Toy SAR example

\[ p(o) = p(c) = \mathcal{N}(c; \mu_c, \Sigma_c) \]

\[ \mu_c = [\mu_1, \mu_2]^\top \]
\[ \Sigma_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

\[ c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} o_1 \\ o_2 - ao_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix} \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} = Ao \]

\[ p(o) = \mathcal{N}(o; \mu_o, \Sigma_o) \]

\[ o = [o_1, o_2]^\top \]
\[ \mu_o = [\mu_1, \mu_2 + a\mu_1]^\top \]
\[ \Sigma_o = \begin{bmatrix} 1 & a \\ a & 1 + a^2 \end{bmatrix} \]
APPENDIX — ACOUSTIC MODELS

SAR & DAR

- SAR: invertible linear feature/model transformation

- For \( o_{1:T} \in \mathbb{R}^{D \times T} \)

\[
\begin{bmatrix}
    o_{1:T,1} \\
    \vdots \\
    o_{1:T,D}
\end{bmatrix} \xrightarrow{A^{(1)}} \begin{bmatrix}
    \cdots \\
    \vdots \\
    \cdots
\end{bmatrix} \xrightarrow{A^{(D)}} \begin{bmatrix}
    c_{1:T,1} \\
    \cdots \\
    c_{1:T,D}
\end{bmatrix}
\]

- SAR is equivalent to:

![Diagram showing the training and generation process for SAR and DAR](image)
**APPENDIX — ACOUSTIC MODELS**

**SAR & DAR**

- **SAR**: invertible linear feature/model transformation

  - For \( o_{1:T} \in \mathbb{R}^{D \times T} \)

\[
o_{1:T} = \begin{bmatrix} o_{1:T,1} \\
\vdots \\
o_{1:T,D} \end{bmatrix} \xrightarrow{\text{filter 1}} \begin{bmatrix} c_{1:T,1} \\
\vdots \\
c_{1:T,D} \end{bmatrix}
\]

- **SAR** is equivalent to:

  - Training
    \[ o_{1:T} \xrightarrow{\text{filters}} c_{1:T} \]
    \[ c_{1:T} \xrightarrow{\prod_{t=1}^{T} p(c_t; \mathcal{M}_t)} x_{1:T} \]

  - Generation
    \[ \hat{o}_{1:T} \xrightarrow{\text{filters}} \hat{c}_{1:T} \]
    \[ \hat{c}_{1:T} \xrightarrow{\prod_{t=1}^{T} p(c_t; \mathcal{M}_t)} \]
APPENDIX — ACOUSTIC MODELS

SAR & DAR

- SAR: invertible linear feature/model transformation

\[ o_{1:T} \rightarrow \begin{array}{c}
\text{filters} \\
A_1^1(z) \\
\ldots \\
A_D^D(z)
\end{array} \rightarrow c_{1:T} \]

\[ 1/A_1^1(z) \\
\ldots \\
1/A_D^D(z) \]

\[ \hat{o}_{1:T} \leftarrow \begin{array}{c}
\text{filters} \\
1/A_1^1(z) \\
\ldots \\
1/A_D^D(z)
\end{array} \leftarrow \hat{c}_{1:T} \]

- Only due to \( 1/A^d(z) \)?
- Due to \( \{A^d(z), 1/A^d(z)\} \), less mismatch between \( c_{1:T} \) and RMDN
**APPENDIX — ACOUSTIC MODELS**

**SAR extension: normalizing flow**

- **Basic idea**

\[ o_{1:T} \rightarrow c_{1:T} = f_{\Phi}(o_{1:T}) \rightarrow c_{1:T} \rightarrow \prod_{t=1}^{T} p(c_t; M_t) \]

\[ \hat{o}_{1:T} \leftarrow \hat{c}_{1:T} = f_{\Phi}^{-1}(\hat{c}_{1:T}) \leftarrow \hat{c}_{1:T} \rightarrow \prod_{t=1}^{T} p(c_t; M_t) \]

\[
p_{o}(o_{1:T}|x_{1:T}) = p_{c}(c_{1:T}|x_{1:T}) \left| \det \frac{\partial c_{1:T}}{\partial o_{1:T}} \right|
\]

- Jacobian matrix must be simple
- \( f(\cdot) \) must be invertible

---


**APPENDIX — ACOUSTIC MODELS**

**SAR extension: normalizing flow**

- **Basic idea**

\[
p_o(o_{1:T} | x_{1:T}) = p_c(c_{1:T} | x_{1:T}) \left| \det \frac{\partial c_{1:T}}{\partial o_{1:T}} \right|
\]

<table>
<thead>
<tr>
<th>SAR</th>
<th>AR Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transform</strong></td>
<td>( c_t = o_t - \sum_{k=1}^{K} a_k \odot o_{t-k} )</td>
</tr>
<tr>
<td><strong>De-transform</strong></td>
<td>( \hat{o}<em>t = c_t + \sum</em>{k=1}^{K} a_k \odot \hat{o}_{t-k} )</td>
</tr>
</tbody>
</table>

- **Simple for SAR:** \( \left| \det \frac{\partial c_{1:T}}{\partial o_{1:T}} \right| = 1 \)

\[\mu_t = \text{RNN}(o_{1:t-1})\]
APPENDIX — ACOUSTIC MODELS

SAR extension

- SAR can be extended

APPENDIX – ACOUSTIC MODELS

DAR

- Same autoregressive principle
- But noninvertible nonlinear

<table>
<thead>
<tr>
<th></th>
<th>NAT</th>
<th>DAR</th>
<th>SAR</th>
<th>RMDN</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>&lt;1e-30</td>
<td>&lt;1.0e-30</td>
<td>&lt;1.0e-30</td>
<td>&lt;1.0e-30</td>
<td></td>
</tr>
<tr>
<td>MOS</td>
<td>4.25</td>
<td>4.00</td>
<td>3.75</td>
<td>3.50</td>
<td>3.25</td>
</tr>
</tbody>
</table>

MOS score

<table>
<thead>
<tr>
<th></th>
<th>NAT</th>
<th>DAR</th>
<th>SAR</th>
<th>RMDN</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>&lt;1e-30</td>
<td>1.6e-28</td>
<td>6.3e-19</td>
<td>2.4e-30</td>
<td></td>
</tr>
<tr>
<td>GV of F0 at utterance-level (Hz)</td>
<td>30</td>
<td>25</td>
<td>50</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

F0 GV