The Blizzard Machine Learning Challenge 2017

Kei Sawada¹, Keiichi Tokuda¹, Simon King², Alan W Black³

¹Nagoya Institute of Technology, ²University of Edinburgh, ³Carnegie Mellon University

https://synsig.org/index.php/Blizzard_Challenge_2017

ASRU 2017 on December 18, 2017
Introduction

- Text-to-speech (TTS) system
  - Technique for generating for artificial speech given input text
  - Evaluation of methods for TTS systems
    - Comparisons are difficult when the training corpus, task, and listening test are different

- Blizzard Challenge [Black & Tokuda; ’05]
  - Better understand and compare research techniques in building corpus-based TTS systems with the same data
  - A lot of time has to be spent on speech-specific tasks
    ⇒ Not attractive to machine learning researchers

- Blizzard Machine Learning Challenge
  - Focus on machine learning problems for speech synthesis
History of TTS system

Rule-based

Formant synthesis

Corpus-based

Concatenative synthesis

Diphone synthesis

Unit selection synthesis

HMM/DNN-driven unit selection synthesis

Statistical speech synthesis

HMM-based speech synthesis

DNN-based speech synthesis

1980s 1990s 2000s 2010s

Heuristic approach ✒ Statistic approach
Statistical speech synthesis

- **Statistical speech synthesis**
  - Mapping to speech waveform from text on the basis of a statistical model

- **HMM-based speech synthesis (’95～)**
  - Context-dependent subword HMMs
  - Regression trees to cluster and tie HMM states

- **DNN-based speech synthesis (’13～)**
  - Replace regression trees with DNN

- **More recent DNN-based speech synthesis (’16～)**
  - Integration of vocoder and acoustic modeling
    - *WaveNet, SampleRNN, etc.*
  - Integration of text analyzer and acoustic modeling
    - *Seq2seq model, Char2Wav, Tacotron, etc.*
Blizzard Challenge

- Evaluations of TTS systems
  - Comparisons are difficult when the training corpus, task, and listening test are different

- Blizzard Challenge [Black, Tokuda, King, et al.]
  - Goal
    - Better understand and compare research techniques in building corpus-based TTS systems
    - Evaluation campaign rather than competition
      ⇒ Purpose of the challenge is to share knowledge
  - Method
    - Participants build voices on a common dataset
    - Organizers evaluate them in a single listening test
  - Annual Blizzard Challenge 2005-2017
    - Need of construct all components for a complete TTS system
    - A lot of time has to be spent on speech-specific tasks
      ⇒ Not attractive to machine learning researchers
Blizzard Machine Learning Challenge 2017

- **Blizzard Machine Learning Challenge**
  - Does not involve speech-specific tasks
  - Allows participants to concentrate on machine learning problem

**Tasks**

- **Text**
  - Text analyzer
  - Linguistic features
  - Acoustic model
  - Acoustic features
  - Vocoder
  - Speech waveform
  - 2017-ES1

- **Text**
  - Text analyzer
  - Linguistic features
  - Acoustic model + Vocoder
  - Acoustic features
  - Speech waveform
  - 2017-ES2

- **Text**
  - Text analyzer + Acoustic model
  - Acoustic features
  - Vocoder
  - Speech waveform

- **Text**
  - Text analyzer + Acoustic model + Vocoder
  - Acoustic features
  - Speech waveform
Tasks

- **2017-ES1**
  - Prediction of acoustic features from linguistic features

- **2017-ES2**
  - Prediction of speech waveforms from linguistic features
Datasets (1/2)

- **Data**
  - Commercial-quality children’s audiobooks from Usborne Publishing Ltd.
  - Same as the Blizzard Challenge 2016
  - 5 hours of speech data

- **Data pruning**
  - Mismatches between speech waveform and text
  - Excessively expressive speech data (e.g. scream, singing voice)
  - Negative effect on acoustic model training
  - Speech data including phoneme alignment errors were pruned
  - 4 hours of speech data (4651 files when divided into sentences)

"I'm king of the jungle," roared Lion.
"I'm going to eat you all up."
"No!" cried the jungle animals.
Datasets (2/2)

- **Speech waveforms (2017-ES2)**
  - 44.1kHz 16 bits monaural Waveform Audio File Format (WAVE)

- **Acoustic features (2017-ES1)**
  - 77-dimensional acoustic features
    - $\log F_0$ (*linearly interpolated values in unvoiced parts*)
    - Voiced and unvoiced information
    - 50-dimensional mel-cepstrum representing spectral envelope
    - 25-dimensional mel-cepstrum representing aperiodicity measures

- **Linguistic features (2017-ES1 and 2017-ES2)**
  - 687-dimensional linguistic features
    - *Forced phoneme alignment* $\Rightarrow$ *Frame-level linguistic features*
    - Normalized to be within $0.0-1.0$ based on minimum and maximum
## Systems

- 7 teams registered and 3 teams submitted
- Pairs of team ID and name are confidential

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>Task</th>
<th>Model</th>
<th>Sampling frequency</th>
<th>Syn. speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Natural speech</td>
<td>–</td>
<td>–</td>
<td>44.1kHz</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Benchmark</td>
<td>2017-ES1</td>
<td>FFNN</td>
<td>44.1kHz</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Benchmark</td>
<td>2017-ES1</td>
<td>FFNN + Trajectory training</td>
<td>44.1kHz</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Submitted</td>
<td>2017-ES1</td>
<td>LSTM</td>
<td>44.1kHz</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Submitted</td>
<td>2017-ES1</td>
<td>LSTM + GAN postfilter</td>
<td>44.1kHz</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Benchmark</td>
<td>2017-ES2</td>
<td>WaveNet</td>
<td>16kHz</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Submitted</td>
<td>2017-ES2</td>
<td>LSTM + WaveNet</td>
<td>22.05kHz</td>
<td></td>
</tr>
</tbody>
</table>
Listening test

- **Design of listening test**
  - The evaluation combined the entries for 2017-ES1 and 2017-ES2 into a single listening test
  - 50 paid native listeners

- **Evaluation criteria**
  - **Naturalness**
    - 5-point mean opinion score (MOS) test
    - 1: completely unnatural – 5: completely natural
  - **Speaker similarity**
    - 5-point MOS test
    - 1: sounds like a different person – 5: sounds like the same person
  - **Intelligibility**
    - Dictation test
    - Word error rate (WER)
    - Semantically unpredictable sentence (SUS)
Result (naturalness)

Mean Opinion Scores (naturalness)  Paid listeners

2017-ES1
2017-ES2
Result (naturalness)

Mean Opinion Scores (naturalness)  Paid listeners

Linguistic features
LSTM
STFT
WaveNet
Speech waveform

Linguistic features
WaveNet
Speech waveform

Original WaveNet
Linguistic features + Log $F_0$
WaveNet
Speech waveform

2017-ES1
2017-ES2
Result (speaker similarity)

Mean Opinion Scores (similarity to original speaker)  Paid listener:

- 2017-ES1
- 2017-ES2
Result (speaker similarity)

Mean Opinion Scores (similarity to original speaker) Paid listeners:

- 2017-ES1
- 2017-ES2

Diagram:

- Linguistic features
- LSTM
- Acoustic features
- GAN
- Acoustic features
- Vocoder
- Speech waveform

Natural speech
LSTM + WaveNet
LSTM + GAN
FFNN + Trajectory
LSTM
WaveNet
FFNN
Result (intelligibility)

Word Error Rate (SUS data)  Paid listeners

2017-ES1
2017-ES2

Good intelligibility

WER (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + WaveNet</td>
<td>198</td>
<td>5</td>
</tr>
<tr>
<td>LSTM + GAN</td>
<td>200</td>
<td>10</td>
</tr>
<tr>
<td>FFNN + Trajectory</td>
<td>199</td>
<td>15</td>
</tr>
<tr>
<td>LSTM</td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td>WaveNet</td>
<td>179</td>
<td>80</td>
</tr>
<tr>
<td>FFNN</td>
<td>199</td>
<td>5</td>
</tr>
</tbody>
</table>
Result (intelligibility)

Word Error Rate (SUS data)  Paid listeners

Linguistic features
WaveNet
Speech waveform

2017-ES1
2017-ES2
Discussion and future plan

- **Recruit machine learning researchers**
  - Lack of advertisement
    - *Difficult to control listening test if there are many participants*
  - Quality confirmation of synthesized speech
    - *Release synthesized speech of benchmark system in advance*
    - *Release training script of benchmark system in advance*
    - *Release simple objective measure*

- **End-to-end speech synthesis**
  - Text → Acoustic feature
  - Text → Speech waveform

Diagram:
- **Text** → **Text analyzer** → **Linguistic features** → **Acoustic model** → **Acoustic features** → **Vocoder** → **Speech waveform**
- ES1, ES2, ES3, ES4
Conclusions

- **Blizzard Machine Learning Challenge 2017**
  - **2017-ES1**
    - *Prediction of acoustic features from linguistic features*
  - **2017-ES2**
    - *Prediction of speech waveform from linguistic features*
  - **Listening test**
    - *Naturalness, speaker similarity, and intelligibility evaluated*
  - **Results**
    - *State-of-the-art machine learning approaches achieved higher scores*