The Blizzard Machine Learning Challenge 2017

Kei Sawada[1], Keiichi Tokuda[1],
Simon King[2], Alan W Black[3]

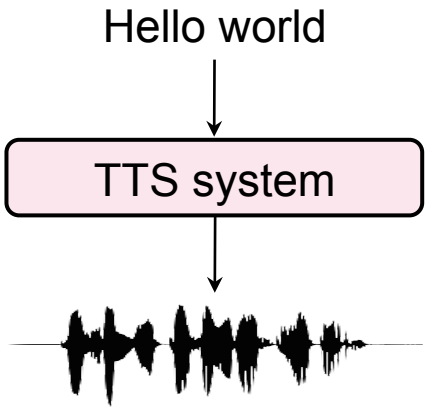
[1]Nagoya Institute of Technology,
[2]University of Edinburgh, [3]Carnegie Mellon University

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*

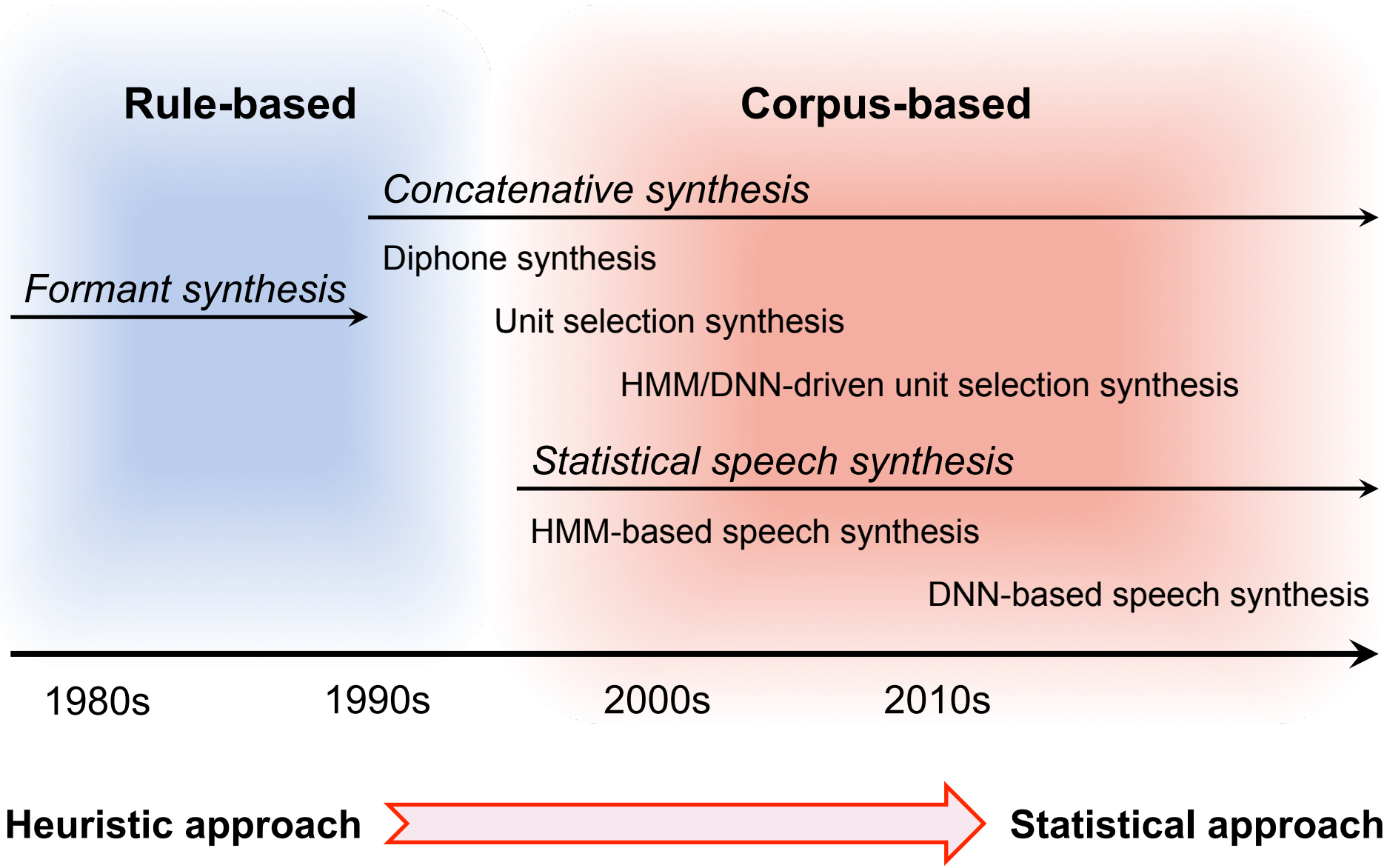


- **Blizzard Challenge [Black & Tokuda; '05]**
  - Better understand and compare research techniques in building corpus-based TTS systems with the same data
  - A lot of time has to be spent on speech-specific tasks

  ⇒ Not attractive to machine learning researchers

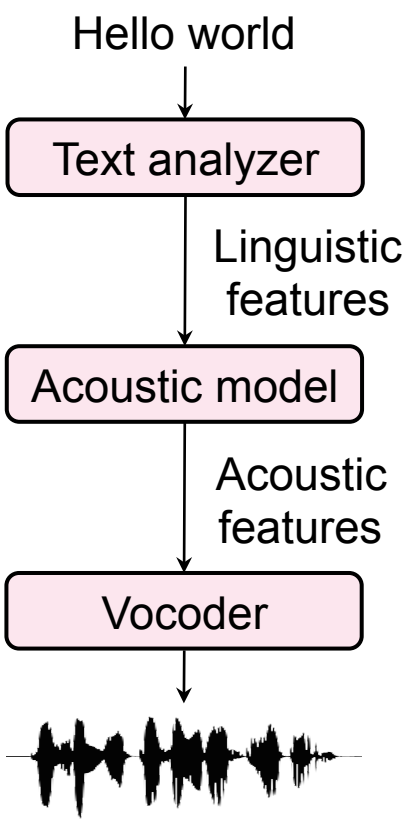
- **Blizzard Machine Learning Challenge**
  - Focus on machine learning problems for speech synthesis

History of TTS system



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*

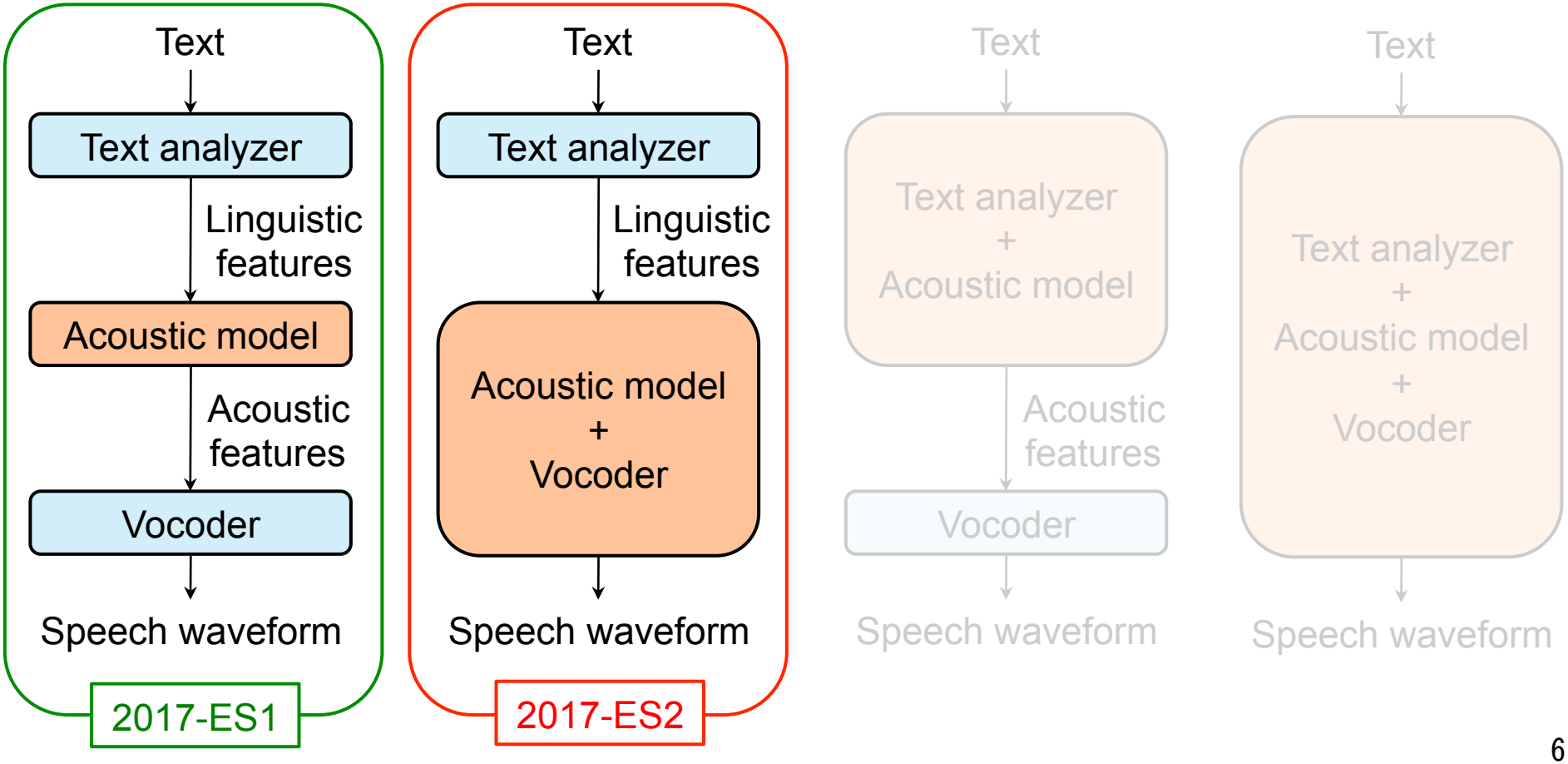
      ⇒ *Purpose of the challenge is to share knowledge*
  - Method
    - *Participants build voices on a common dataset*
    - *Organizers evaluate them in a single listening test*
  - Annual Blizzard Challenge 2005-2017
    - *Need of construct all components for a complete TTS system*
    - *A lot of time 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**

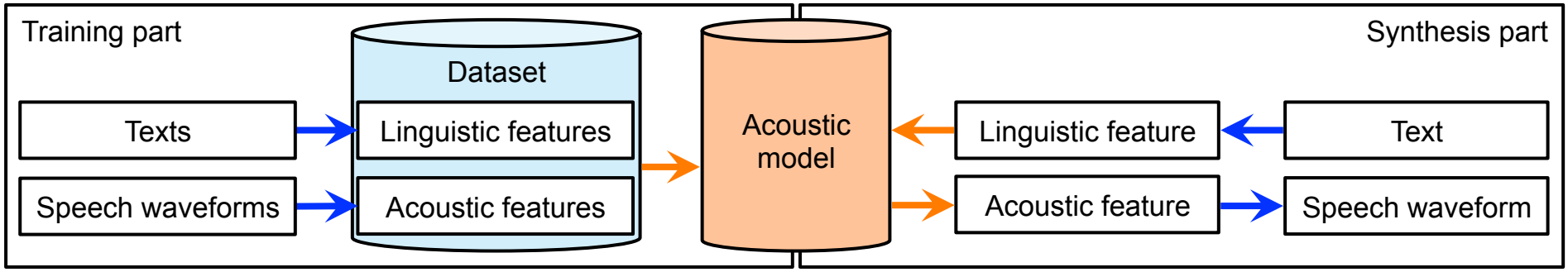


Tasks

Processes of the organizers Processes of the participants

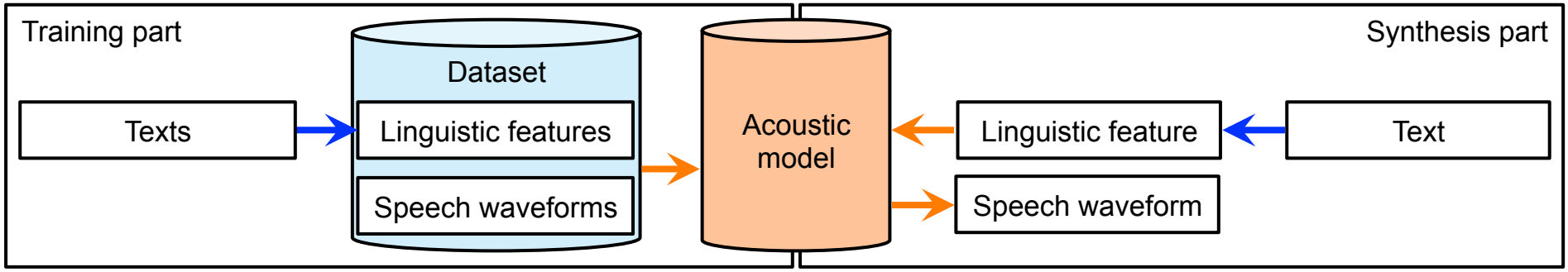
2017-ES1

- 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

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

Character1
Character2
Descriptive part

- **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)

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-dimentional linguistic features
    - *Forced phoneme alignment ⇒ 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**

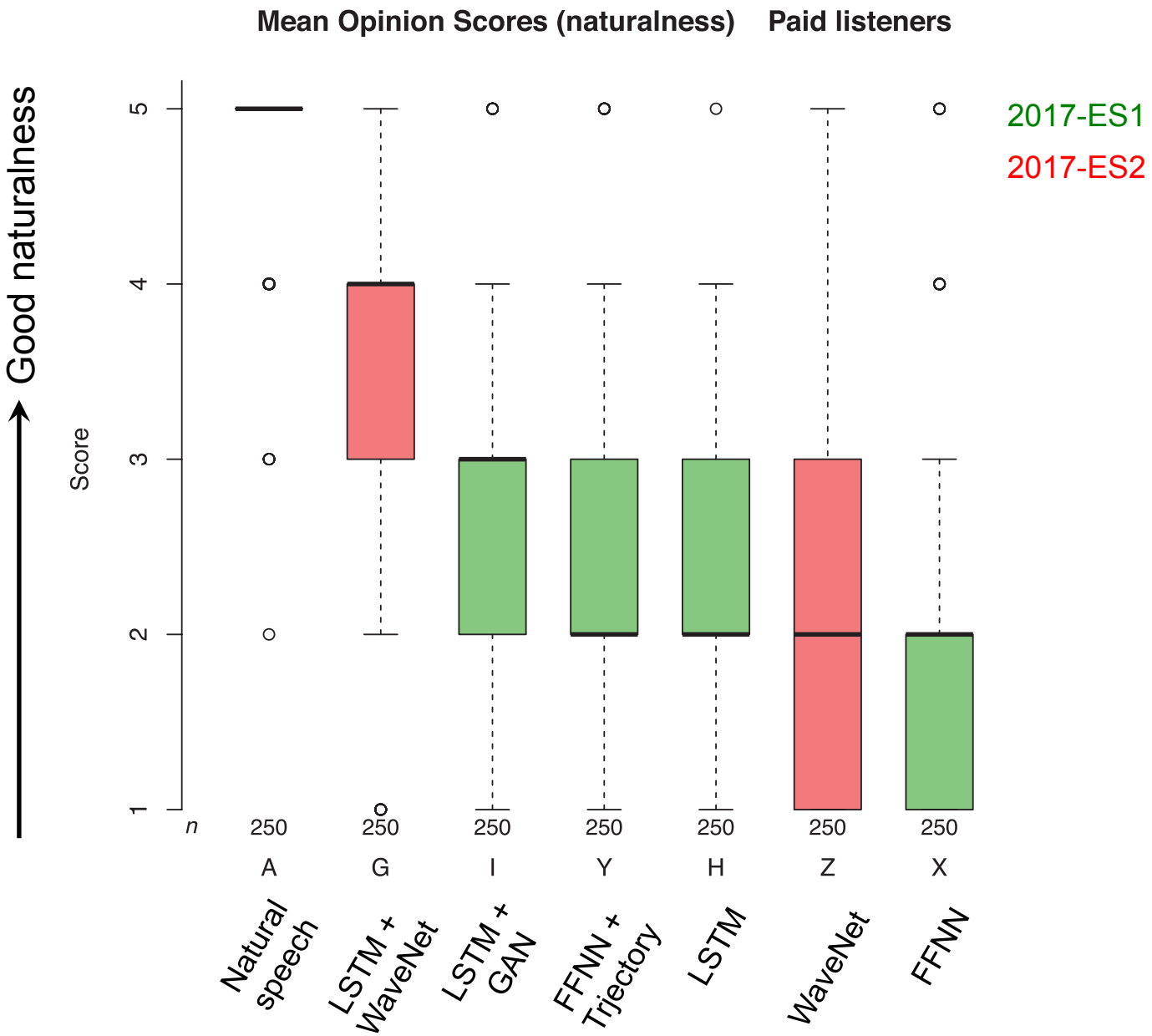
ID	Category	Task	Model	Sampling frequency	Syn. speech
A	Natural speech	–	–	44.1kHz	
X	Benchmark	2017-ES1	FFNN	44.1kHz	
Y	Benchmark	2017-ES1	FFNN + Trajectory training	44.1kHz	
H	Submitted	2017-ES1	LSTM	44.1kHz	
I	Submitted	2017-ES1	LSTM + GAN postfilter	44.1kHz	
Z	Benchmark	2017-ES2	WaveNet	16kHz	
G	Submitted	2017-ES2	LSTM + WaveNet	22.05kHz	

Listening test

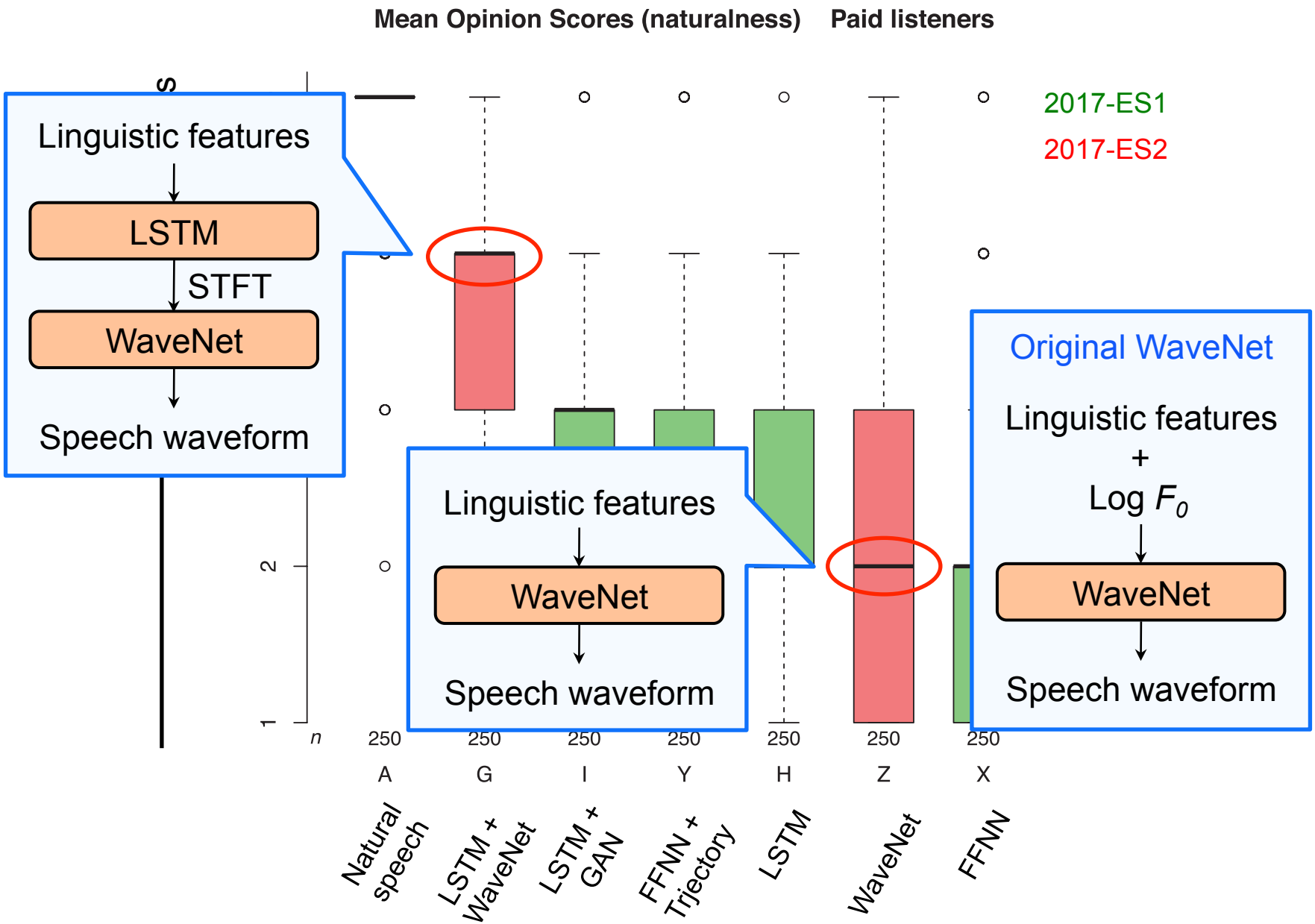
- **Design of listening test**
  - The evaluation combined the entries for 2017-ES1 and 2017-ES2 into a single listening test
  - 50 paid native listeners

- **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)*

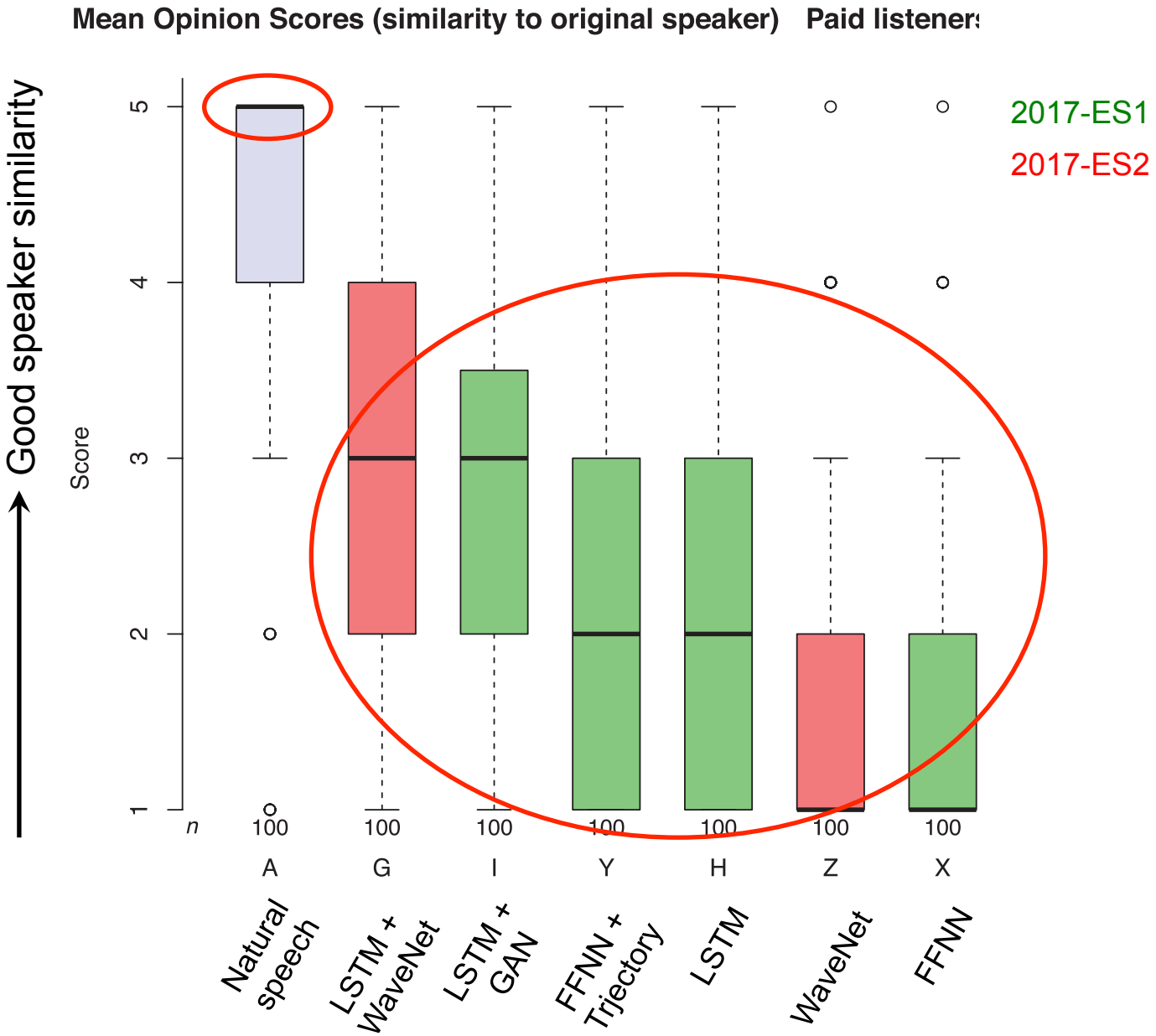
Result (naturalness)



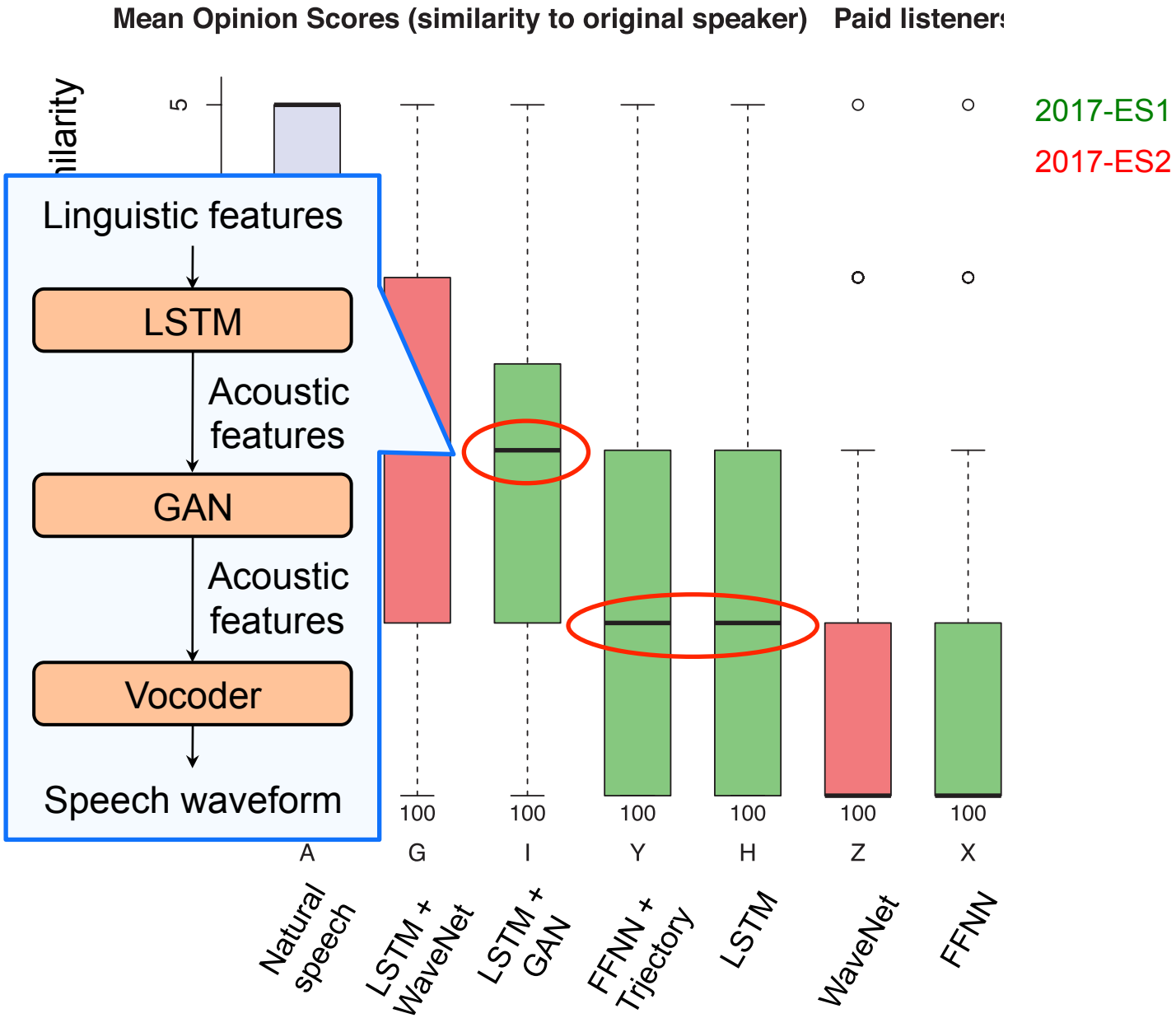
Result (naturalness)



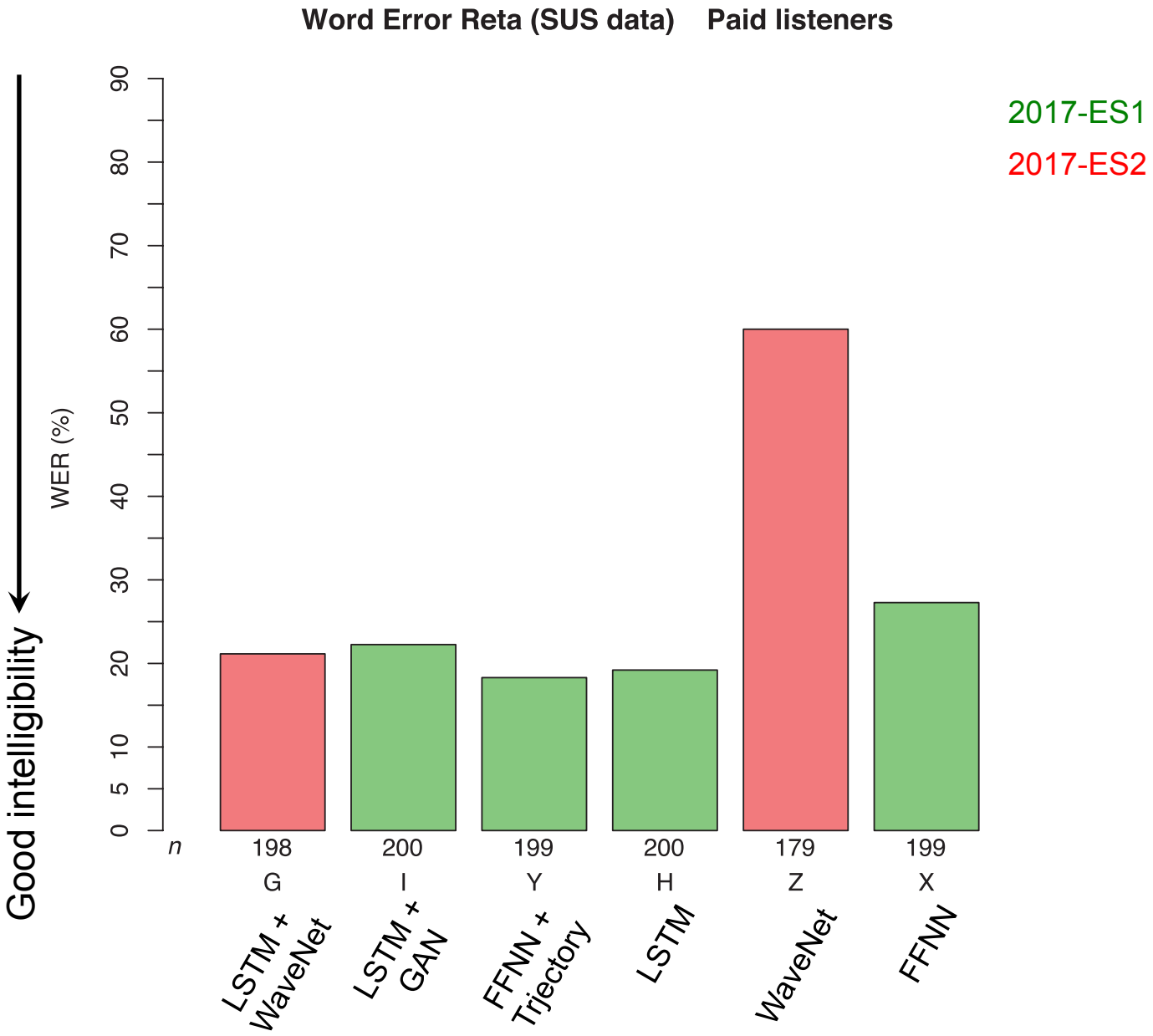
Result (speaker similarity)



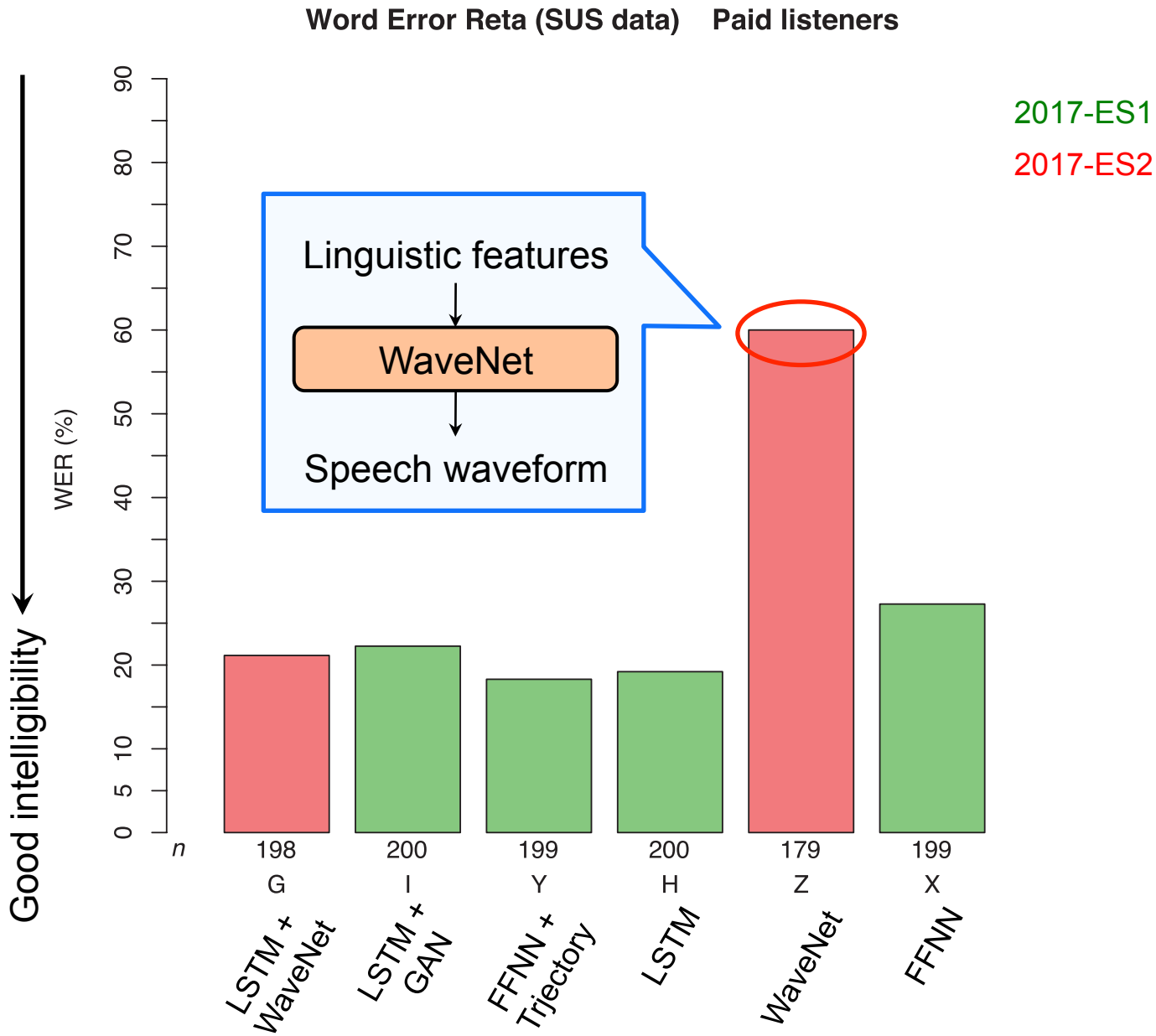
Result (speaker similarity)



Result (intelligibility)



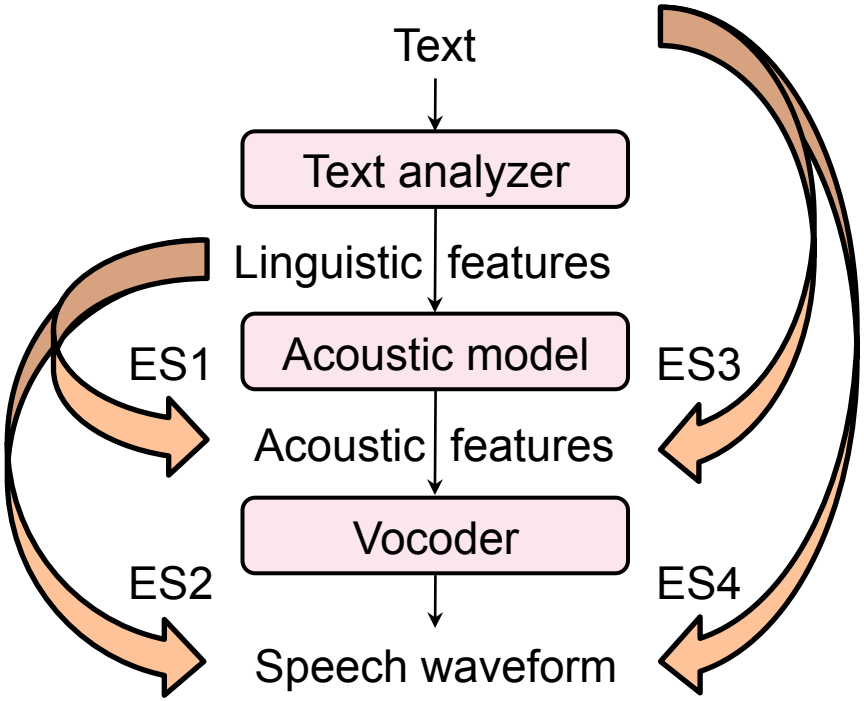
Result (intelligibility)



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



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*