The Blizzard Machine Learning Challenge 2017の概要

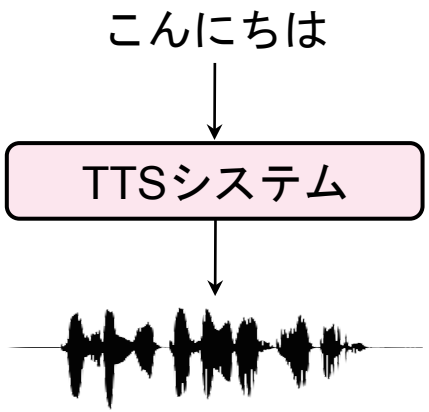
◎沢田慶, 徳田恵一 (名古屋工業大学),
Simon King (エジンバラ大学),
Alan W Black (カーネギーメロン大学)

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

2018年3月15日

Blizzard Challenge

- **テキスト音声合成 (text-to-speech; TTS) システム**
  - HMM・DNNなどの機械学習手法の導入により進展
  - TTSシステムの評価
    - *受聴試験に基づく主観評価実験*
    - *データセット・タスク・受聴試験が異なる主観評価実験の結果を議論することは不適切*

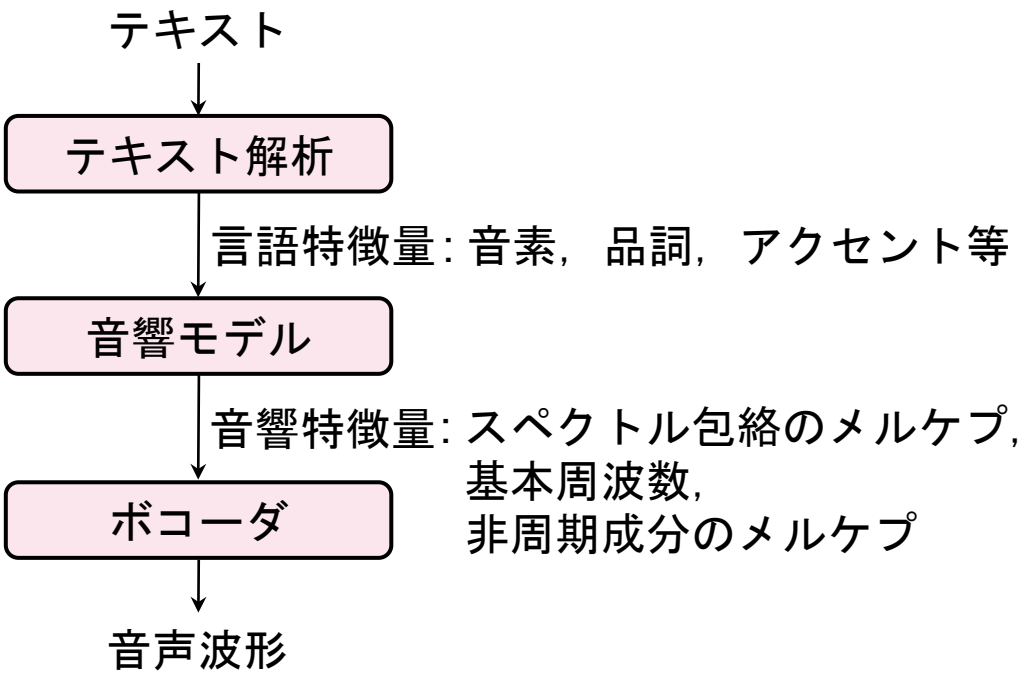


- **Blizzard Challenge [Black & Tokuda; '05]**
  - 目的
    - *コーパスベースのTTSシステムのさらなる理解と手法の比較*
    - *競技会というよりは評価会 ⇒ "チャレンジの目的は知見の共有"*
  - 方法
    - *参加者は共通のデータセットを用いてTTSシステムを構築*
    - *主催者は統一した受聴試験によりTTSシステムを評価*
  - Annual Blizzard Challenge 2005-2017
    - *TTSシステムに関する多くの知見を共有*
    - *高品質な音声を生成するには音声合成に関する多くの知見が必要*
      ⇒ 機械学習分野の研究者が参加することへの妨げに

Blizzard Machine Learning Challenge

- **Blizzard Machine Learning Challenge 2017**
  - 主催者により音声合成に関する知見を必要とする事前処理
  - 参加者は機械学習に基づく音響モデル構築に集中

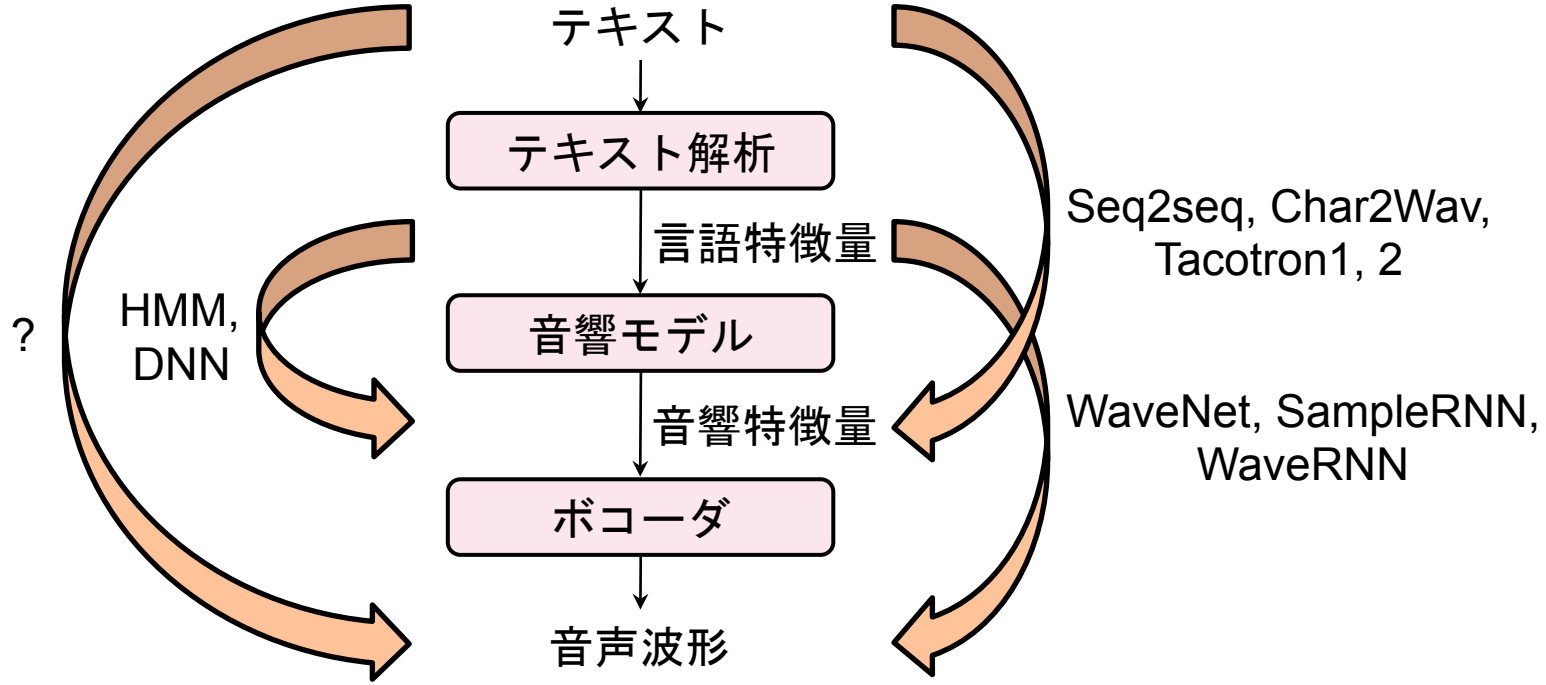
- **統計モデルを用いた音声合成**



Blizzard Machine Learning Challenge

- **Blizzard Machine Learning Challenge 2017**
  - 主催者により音声合成に関する知見を必要とする事前処理
  - 参加者は機械学習に基づく音響モデル構築に集中

- **統計モデルを用いた音声合成**

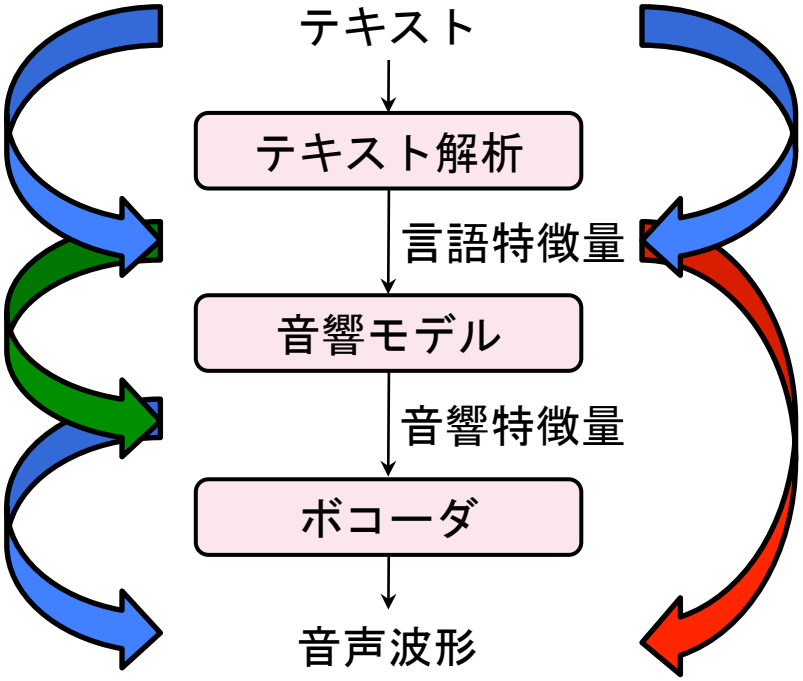


Blizzard Machine Learning Challenge

- **Blizzard Machine Learning Challenge 2017**
  - 主催者により音声合成に関する知見を必要とする事前処理
  - 参加者は機械学習に基づく音響モデル構築に集中

- **タスク**

2017-ES1:
言語特徴量系列から
音響特徴量系列を
予測するタスク



2017-ES2:
言語特徴量系列から
音声波形系列を
予測するタスク

データセット (1/2)

- データ
  - Usborne Publishing Ltd. にて市販されている児童書のオーディオブック
  - Blizzard Challenge 2016と同様のデータ

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

キャラクター1
キャラクター2
地の文

- データの取捨選択
  - 過度に感情表現した音声データ (叫び声や歌声など)
  - 音声データとテキストの不一致
    ⇒ 音響モデル学習に悪影響
  - 音素アライメントエラーが含まれる音声データを削除
  - 5時間 ⇒ 4時間の音声データ (文章単位の4651ファイル)

データセット (2/2)

- **音声波形 (2017-ES2)**
  - 44.1kHz 16 bits モノラルのWaveform Audio File Format (WAVE)

- **音響特徴量 (2017-ES1)**
  - 77次元の音響特徴
    - *Log F_0 (無声部を有声部の値で線形補間)*
    - *有声無声情報*
    - *スペクトル包絡の50次元メルケプストラル表現*
    - *非周期成分の25次メルケプストラル表現*

- **言語特徴量 (2017-ES1, 2017-ES2)**
  - 687次元のフレーム単位の言語特徴 (HTS形式)
    - *強制音素アライメント*
    - *最大値と最小値を用いて0.0-1.0に正規化*

システム

- 7チームが登録して3チームが提出
- チームIDとチーム名のペアを用いた議論は禁止

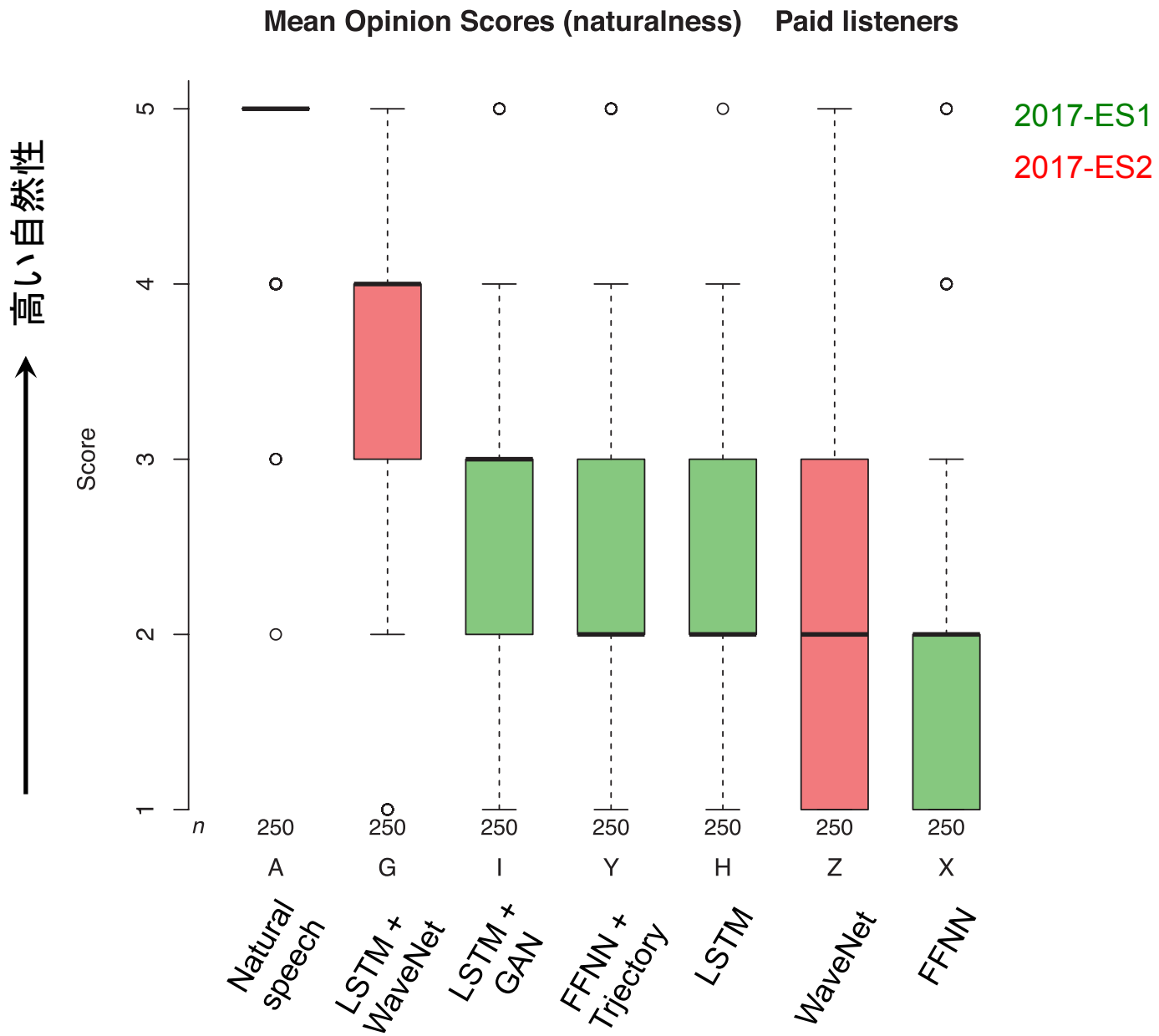
ID	カテゴリ	タスク	手法	サンプリング周波数	合成音声
A	自然音声	–	–	44.1kHz	
X	ベンチマーク (HTS)	2017-ES1	FFNN	44.1kHz	
Y	ベンチマーク	2017-ES1	FFNN + Trajectory training	44.1kHz	
H	参加者	2017-ES1	LSTM	44.1kHz	
I	参加者	2017-ES1	LSTM + GAN postfilter	44.1kHz	
Z	ベンチマーク	2017-ES2	WaveNet	16kHz	
G	参加者	2017-ES2	LSTM + WaveNet	22.05kHz	

受聴試験

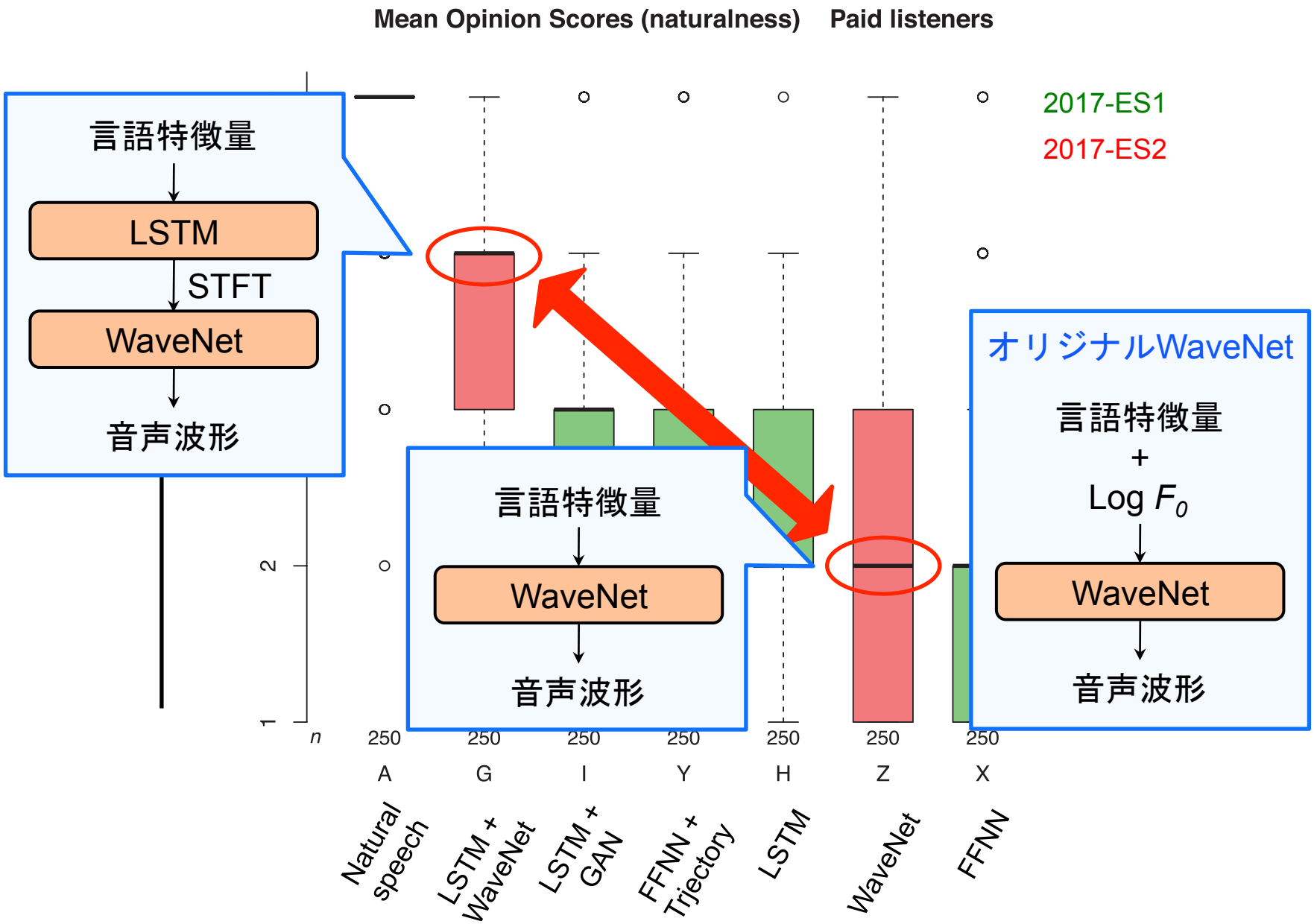
- **受聴試験の設計**
  - 2017-ES1と2017-ES2のシステムを同一の受聴試験で評価
  - 50人の有償被験者 (エジンバラ大学の学生)

- **評価基準**
  - 自然性
    - *5段階 MOS (mean opinion score) テスト*
    - *1: completely unnatural – 5: completely natural*
  - 話者類似性
    - *5段階 MOS テスト*
    - *1: sounds like a different person – 5: sounds like the same person*
  - 明瞭性
    - *無意味文 (SUS; semantically unpredictable sentence) の書き取りテスト*
    - *単語誤り率 (WER; word error rate)*

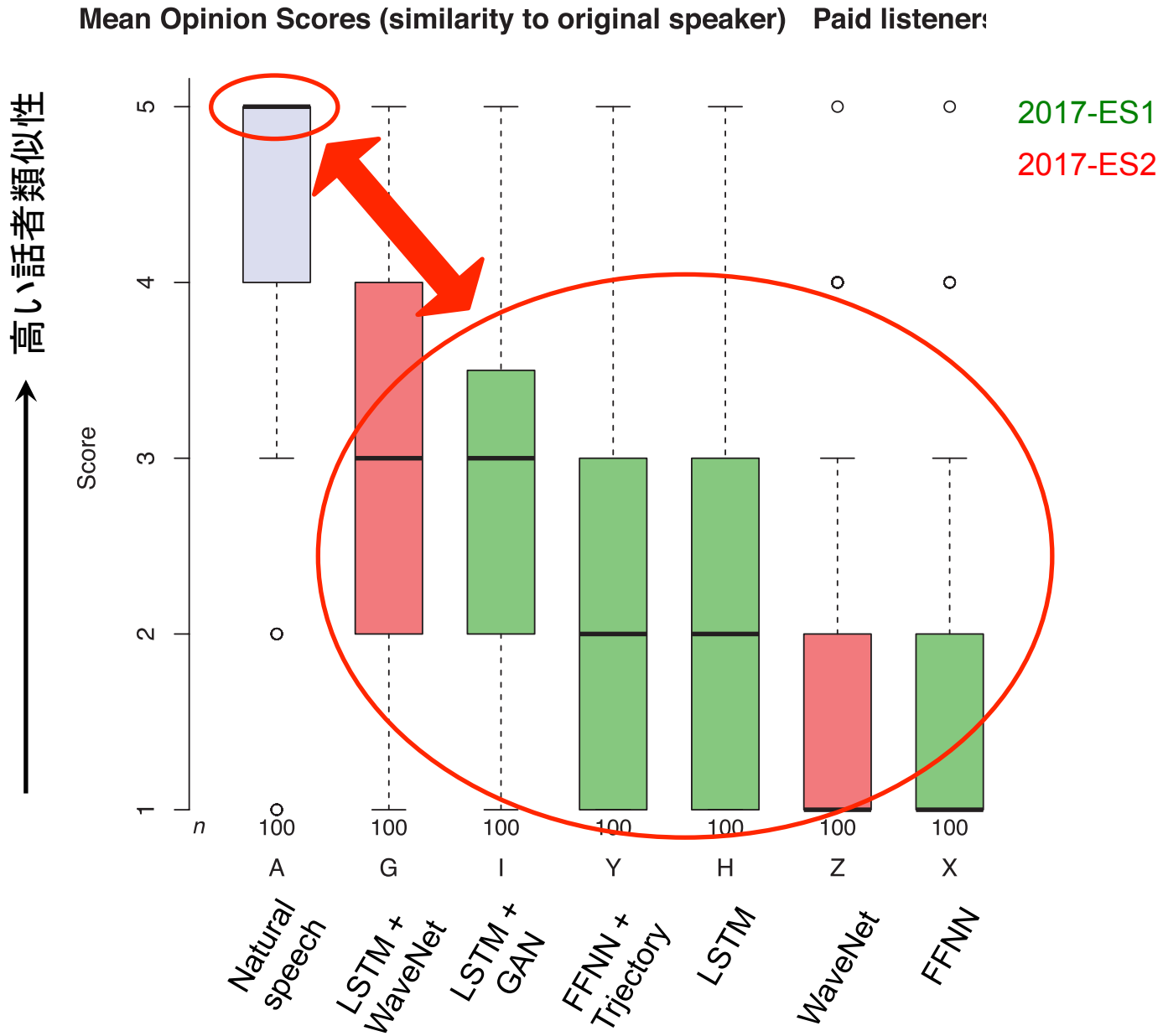
自然性に関する評価結果



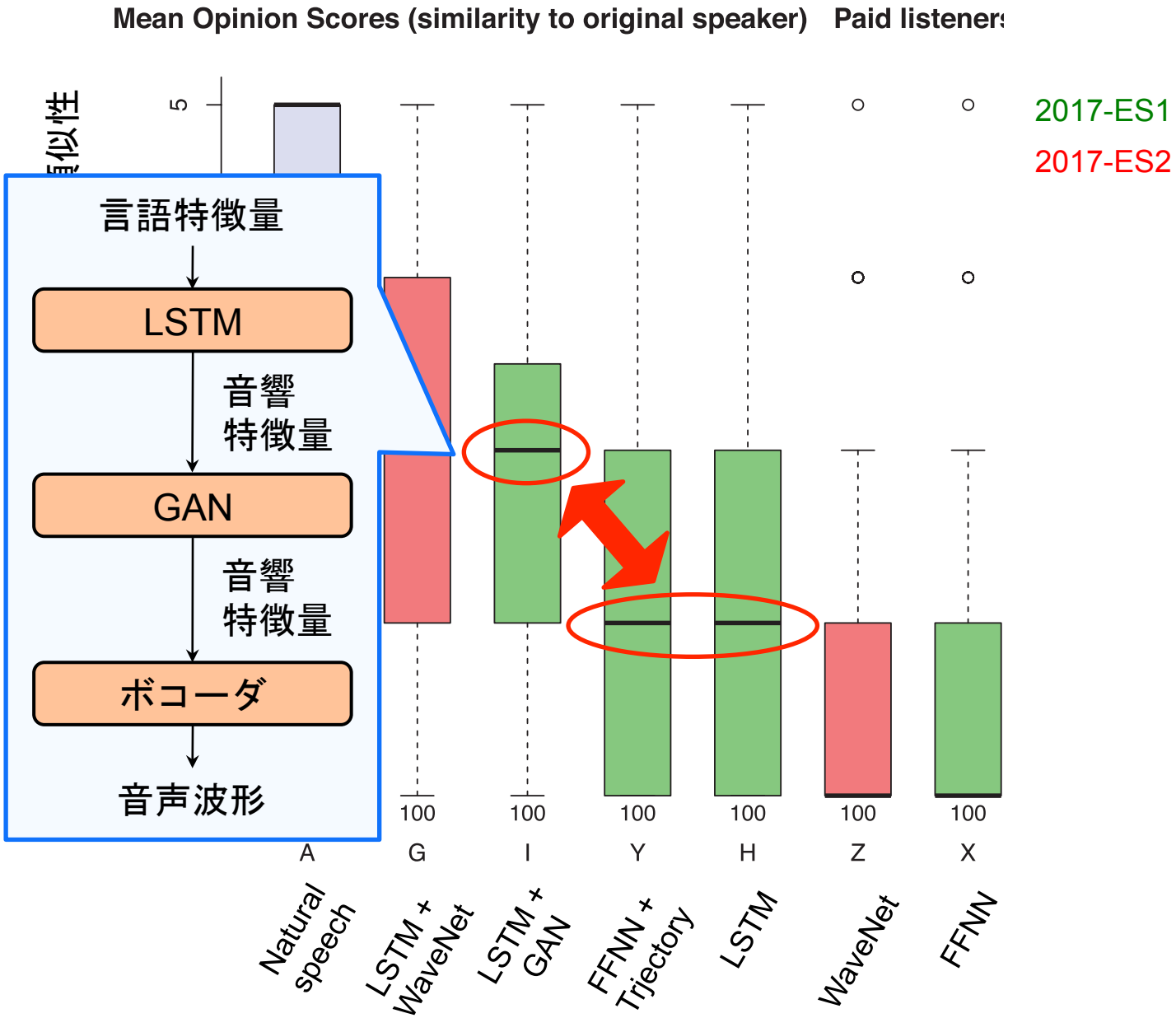
自然性に関する評価結果



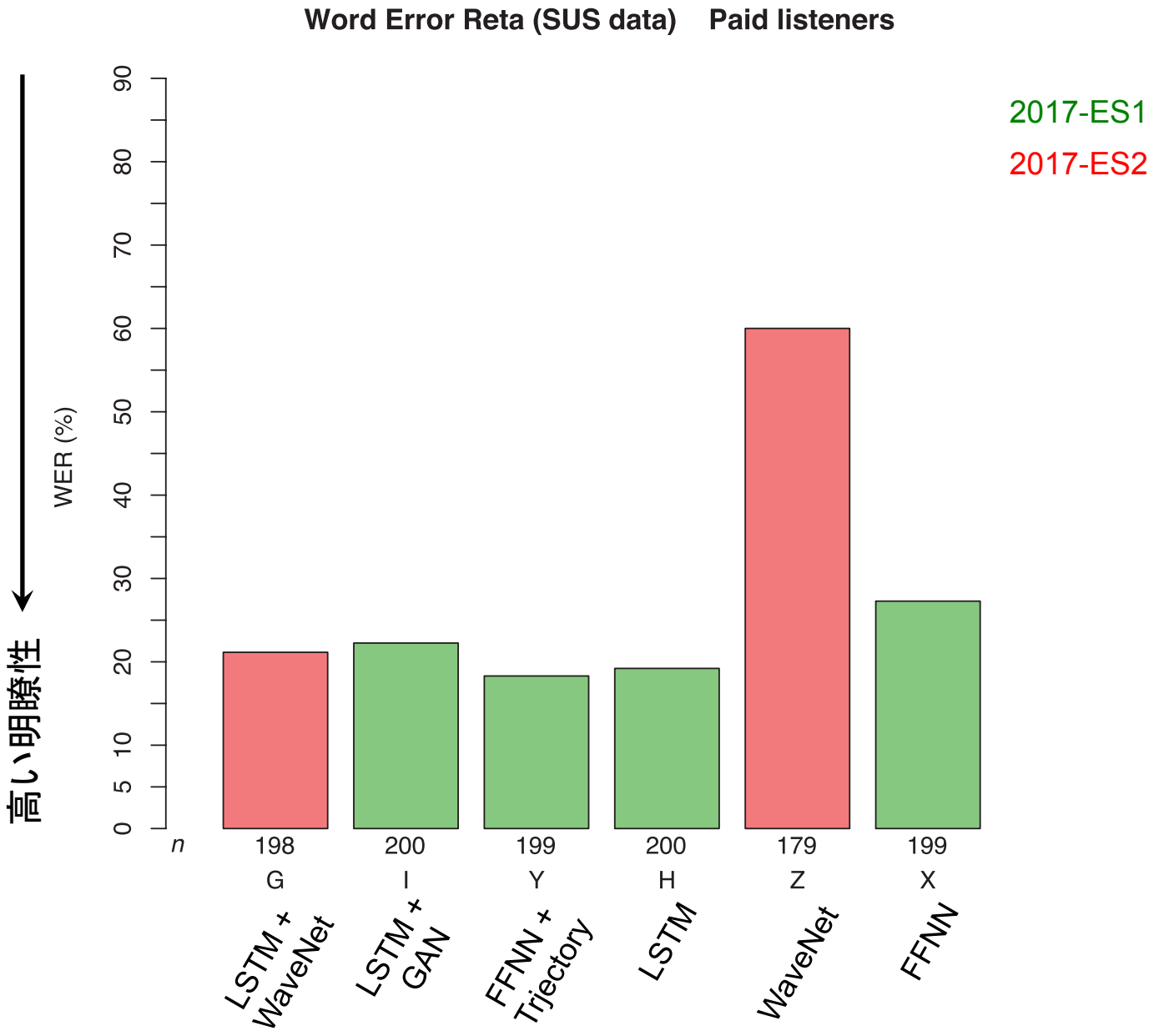
話者類似性に関する評価結果



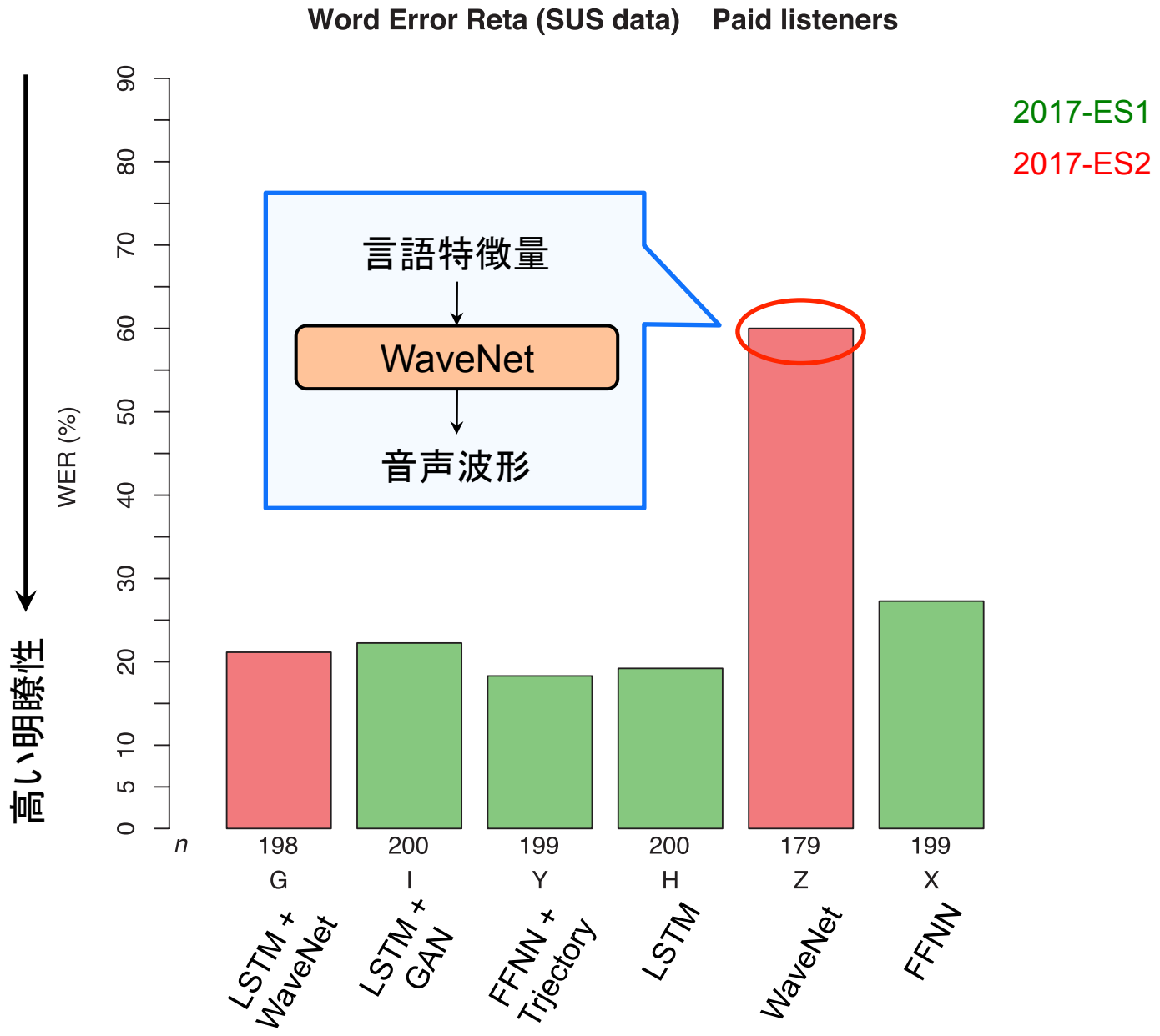
話者類似性に関する評価結果



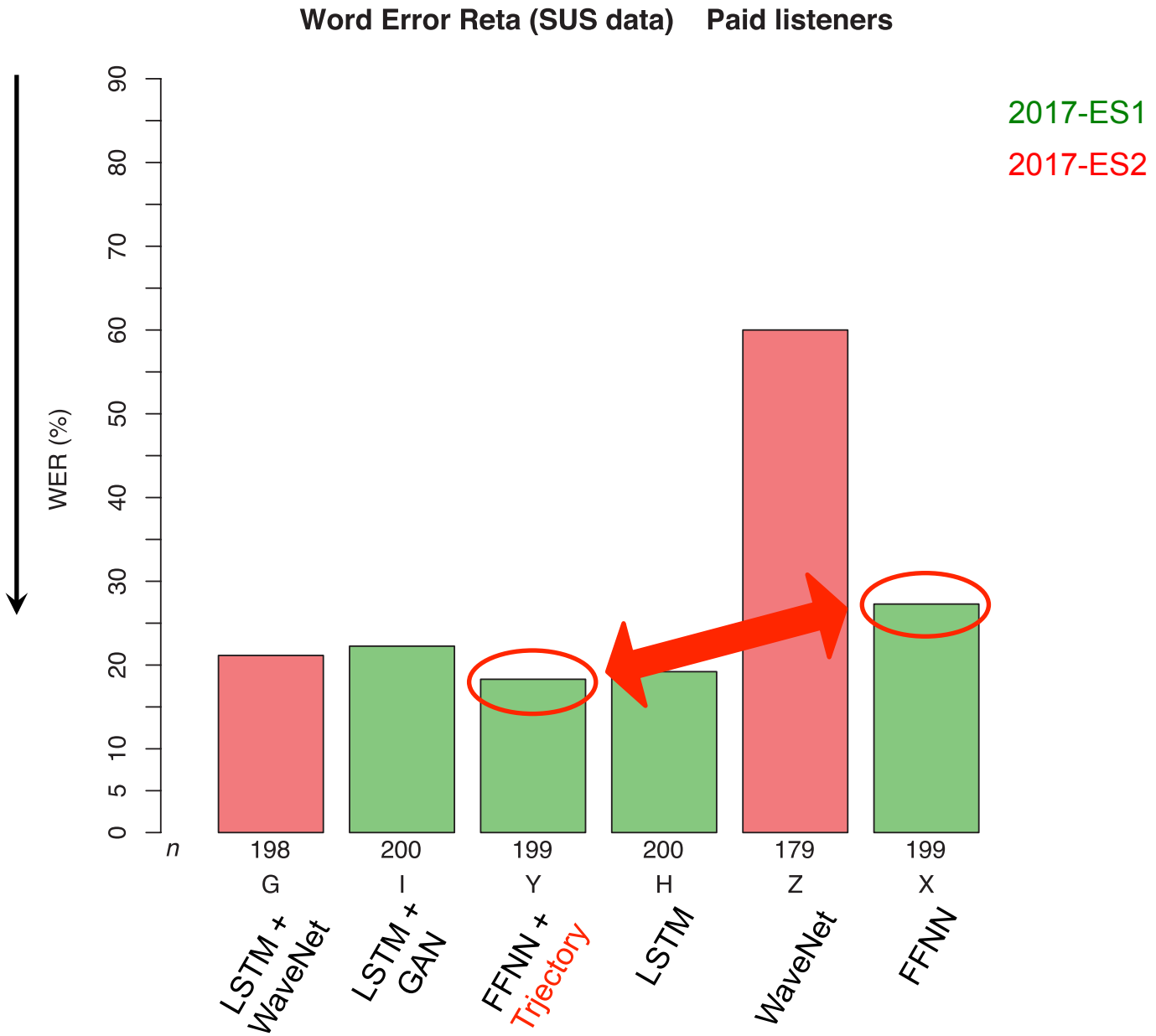
明瞭性に関する評価結果



明瞭性に関する評価結果



明瞭性に関する評価結果



むすび

- **Blizzard Machine Learning Challenge 2017**
  - 2017-ES1
    - *言語特徴量系列から音響特徴量系列の予測*
  - 2017-ES2
    - *言語特徴量系列から音声波形系列の予測*
  - 受聴試験
    - *自然性・話者類似性・明瞭性に関する主観評価実験*
  - 評価結果
    - *最先端の機械学習手法に基づくシステムは高いスコアを達成*

- **Blizzard Challenge 2018**
  - 2017と同様の児童書のオーディオブックを用いたタスク
  - システム提出締め切り: 4月26日
  - ベンチマークの学習スクリプトと合成音声を公開予定

HMM	DNN	DNN Trj