Blizzard Machine Learning Challenge 2017の概要*

◎沢田慶, 徳田恵一 (名工大), Simon King (エジンバラ大), Alan W Black (CMU)

1 はじめに

テキスト音声合成 (text-to-speech; TTS) システムのための，機械学習に基づく音響モデル構築に焦点を当てた評価会である Blizzard Machine Learning Challenge (BMLC) 2017 を主催した [1]．本稿では，BMLC 2017 の目的・タスク・データセット・システム・評価実験について述べる．

2 BMLC 2017

TTS システムとは，任意のテキストに対応する音声を自然かつ明瞭な合成音声として生成する技術のことである．TTS システムの構築手法は，計算機性能の向上に伴い，ヒューリスティックに基づく手法から統計的アプローチに基づく手法へと進歩してきた．近年では機械学習 (特にディープラーニング) 分野の知見を活かした様々な手法が導入され，合成音声の品質は劇的に向上している．今後も，機械学習分野の技術により合成音声の品質は改善していくと考えられる．しかし，数多く提案されている TTS システムのための手法を比較評価することは容易ではない．これは，TTS システムの比較評価に広く用いられている受聴試験による主観評価の結果は，データセット・タスク・受聴試験が異なると議論することが難しいからである．この問題への取り組みのひとつとして，共通のデータセットを用いて TTS システムを構築して受聴試験を行う Blizzard Challenge (BC) と題する評価会が 2005 年から毎年開催されている [2]．これまでに世界中の研究機関が BC に参加し，TTS システムに関する知見を共有してきた．しかし，十分な品質の合成音声を提出するためには，音声合成に関する知見を利用した数多くの作業を必要とした．これは，機械学習分野の研究者が BC へ参加する妨げになっていると考えられる．また，音声合成に関する知見が TTS システムの性能に影響するため，機械学習に基づく手法どうしを比較評価することはできない．

そこで，音声合成に関する知見を必要とする作業を取り除き，機械学習に基づく音響モデル構築のみに集中する評価会である BMLC を提案する．BMLC の開催により，機械学習研究者の新規参入と音声合成における機械学習手法の比較評価の実現を目指す．BMLC 2017 では，BC 2016 で配布された児童書のオーディオブックに対して機械学習の問題に集中することを可能とする事前処理を施した 2 種類のデータセットを参加者に配布した．参加者は，音声合成研究者と機械学習研究者が公平となるルールの下で音響モデルの構築を行う．そして，主催者による受聴試験により，機械学習に基づく音響モデルの性能を比較評価する．

3 タスク

BMLC 2017 では，音声合成分野と機械学習分野の研究者が公平な条件の下で音響モデルの構築を可能とする，2 つのタスク (2017-ES1, 2) を設計した．

3.1 2017-ES1

2017-ES1 は言語特徴量系列から音響特徴量系列を予測するタスクである．主催者は，フレーム単位の音響特徴量系列と言語特徴量系列の組を配布する．参加者は，言語特徴量系列から音響特徴量系列の予測を行う音響モデルを構築する．

3.2 2017-ES2

2017-ES2 は言語特徴量系列から音声波形系列を予測するタスクである．主催者は，音声波形系列と言語特徴量系列の組を配布する．参加者は，言語特徴量系列から音声波形系列の予測を行う音響モデルを構築する．

4 データセット

BMLC 2017 では，機械学習に基づく音響モデル構築に適するデータセットを，BC 2016 で用いられたデータから作成した．BC 2016 では，Usborne Publishing で市販されている児童書のオーディオブック計 50 冊の音声データとテキストが配布された．音声データは約 5 時間であり，イギリス英語の女性話者 1 名により収録されている．

4.1 音声波形

音声波形を直接モデル化する 2017-ES2 のために，文単位に分割された 44.1 kHz，16 bit，モノラルの WAVE ファイルを配布した．また，この WAVE ファイルから 2017-ES1 のための音響特徴量を抽出した．

4.2 音響特徴量

音響特徴量は，フレームシフトを 4.989 ms として抽出した無声部を有声部の値で線形補間した対数基本周波数，有声無声情報，スペクトル包絡から求めた 50 次元のメルケプストラム係数，非周期性指標から求めた 25 次元のメルケプストラム係数の計 77 次元の特徴量とした．また，音響特徴量系列から音声波形系列を生成する，Linux と macOS 用のバイナリファイルも配布した．

4.3 言語特徴量

言語特徴量は，HTS-2.3.1 デモスクリプト形式の計 687 次元の特徴量とした．各言語特徴量は最大値と最小値を用いて $0-1$ に正規化した．

音声波形，音響特徴量，言語特徴量はそれぞれサンプル単位，フレーム単位，音素単位と異なる時間単位から成る．FFNN や LSTM などの DNN を用いた (seq2seq モデルではない) 音響モデルの学習には，

---

*Overview of the Blizzard Machine Learning Challenge 2017.
by Kei Sawada, Keiichi Tokuda (Nagoya Institute of Technology),
Simon King (University of Edinburgh), Alan W Black (Carnegie Mellon University)

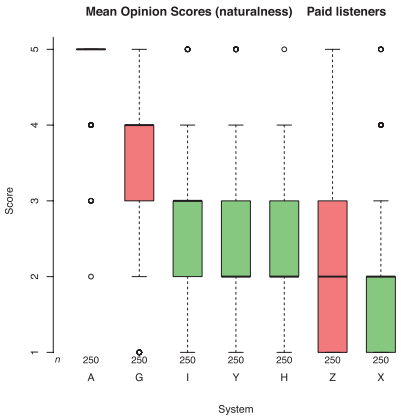

Fig. 1 MOS for naturalness

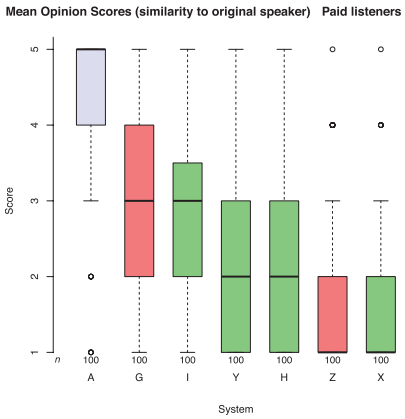

Fig. 2 MOS for speaker similarity

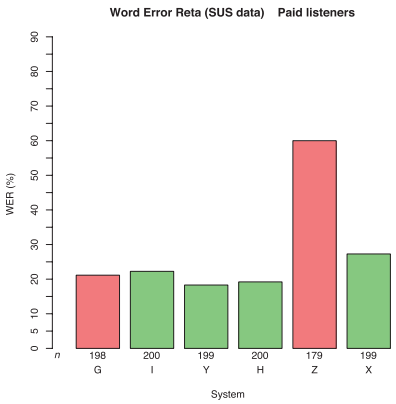

Fig. 3 WER of SUS

Table 1 Systems

ID	Category	Task	Model
A	Natural	–	–
X	Benchmark	2017-ES1	FFNN
Y	Benchmark	2017-ES1	FFNN + Trajectory training
H	Submitted	2017-ES1	LSTM
I	Submitted	2017-ES1	LSTM + GAN postfilter
Z	Benchmark	2017-ES2	WaveNet
G	Submitted	2017-ES2	LSTM + WaveNet

異なる時間単位の特徴量系列の時間対応 (アライメント) が必要となる．そこで，BMLC 2017 では，HMM を用いて音響特徴量系列と言語特徴量系列のアライメントを推定し，フレーム単位の言語特徴量系列を提供した．

4.4 学習データの取捨選択

BC 2016 で用いられた音声データの発話内容とテキストには不一致が存在する．さらに，児童書のオーディオブックであることから，叫び声や歌声などの過度な感情表現をした音声データも含まれている．このようなデータのアライメントを推定すると，アライメントエラーが多く発生する．アライメントエラーを含む音響特徴量系列と言語特徴量系列の組は音響モデル学習に悪影響を与えるため，アライメントエラーが多く含まれるデータはデータセットから取り除いた．学習データの取捨選択をした約 4 時間の音声波形，音響特徴量，言語特徴量をデータセットとして配布した．

5 システム

BMLC 2017 には，7 チームが登録し，3 チームがシステムを提出した．また，3 つのベンチマークシステムを用意した．Table 1 にシステムの一覧を示す．

6 評価実験

システムの性能を評価するために，主催者による主観評価実験を行った．

6.1 受聴試験

BMLC 2017 では自然性を評価する 5 段階 MOS 試験，話者類似性を評価する 5 段階 MOS 試験，明瞭性を評価する無意味文 (SUS) の書き取り試験の 3 つの評価基準による受聴試験を実施した．被験者は，50 人の英語母語話者である．

6.2 実験結果

自然性の MOS，話者類似性の MOS，SUS の単語誤り率 (WER) を Figs. 1, 2, 3 に示す．図中のシステムは，自然性の MOS を基準に並んでいる．

Fig. 1 より，WaveNet を用いたシステム G が合成音声の中で最も高い MOS を獲得した．一方，WaveNet を用いたシステム Z は高い MOS を得られなかった．これらのシステムの主な違いは，中間表現として音響情報を利用しているかどうかである．つまりこの結果は，音響情報を利用せずに言語特徴量のみを入力として高品質な音声を合成する WaveNet を学習するとは難しいことを示唆している．

Fig. 2 より，自然音声 A と合成音声の MOS には大きな差がある．しかし，2017-ES1 のシステムの中では，GAN に基づくポストフィルタを用いたシステム I が高い MOS を示したことから，GAN に基づく手法は話者類似性を改善する可能性がある．

Fig. 3 より，システム Z が他のシステムと比べ高い WER を示している．これは，言語特徴量のみを入力とした WaveNet の学習が不十分であり，発音が曖昧になったためだと考えられる．

7 むすび

本稿では，BMLC 2017 の概要を述べた．TTS システムのための，機械学習に基づく音響モデル学習に焦点を当てた 2 つのタスクを設計し，データセットを公開した．受聴試験の結果，最先端の機械学習に基づく手法は高いスコアを示した．今後も，音声合成分野と機械学習分野の研究者が参加しやすい評価会を設計し主催することで，音声合成研究の発展を促進する．また，テキスト系列から音響特徴量系列を予測する手法やテキスト系列から音声波形系列を予測する手法も提案され始めていることから，これらのタスクの導入についても検討する．

参考文献

[1] K. Sawada *et al.*, “The Blizzard Machine Learning Challenge 2017,” ASRU 2017, pp. 331–337, 2017.

[2] A. W. Black *et al.*, “The Blizzard Challenge — 2005: Evaluating corpus-based speech synthesis on common datasets,” Interspeech 2005, pp. 77–80, 2005.