テキスト音声合成におけるデータサブセット選択のための指標検討*

◎関 健太郎，高道 慎之介，佐伯 高明，猿渡 洋（東大院・情報理工）

1 はじめに

テキスト音声合成 (text-to-speech: TTS) コーパスの構築方法として，オーディオブックやインターネット上のデータ（ダークデータ）等の大量のデータから収集する方法が提案されている [1,2]．この方法はテキストを設計して読み上げる従来の方法よりも多様なデータを大量に収集することが可能であり，TTS モデルの表現力向上に向けて大きな注目を集めている．

TTS コーパスの大規模化は今後も継続すると予想されるが，記憶媒体の容量や学習時間といった制約から際限なく大きなコーパスを構築することは現実的でない．さらに実際のデータには分布の偏りが内在し，学習効率が低下すると考えられる．この問題に対処する枠組みとして，コアセットと呼ばれる高い学習効果を持つ小規模サブセットを用いた機械学習手法が研究されている [3,4]．テキスト音声合成分野においても，適切なサブセットを選択することでサイズ制約下で効率的なデータセットを構築することが望まれる．

そこで本研究では TTS コーパスに対してサブセット選択を実施するために，データサブセットの評価指標を検討する．評価の枠組みとしてエントロピー，多様性という 2 つの枠組みを検討し，複数の特徴量を用いた評価指標を提案する．各評価指標に基づく選択が TTS モデルの性能に及ぼす効果について，実験的に評価を行う．

2 関連研究

2.1 多話者 TTS コーパス

多話者 TTS システムは，テキストと話者表現の組を入力として受け取り対応する音声波形を出力するシステムである [5]．音声収録による多話者 TTS コーパス構築は多大なコストがかかるため，オーディオブック音声に対する音声強調 [1] やダークデータに対するデータを選択 [2] を用いた大規模多話者 TTS コーパス構築方法が提案されている．

2.2 基盤モデルによる特徴量抽出

大量のデータを用いて事前学習されたモデルは基盤モデルと呼ばれ，様々なタスクへの転移学習において高い性能を発揮している [6]．自然言語処理における BERT [7] や音声認識における wav2vec 2.0 [8] は基盤モデルの一種であり，これらのモデルは特徴量抽出器としても利用されている．

2.3 データ選択

2.3.1 音素バランス文セットの設計

バランスの取れた音声データセットを整備する古典的方法として，音素バランス文セットを構築する方法がある [9]．この方法は各音素の出現確率 p_i に基づいて計算されるエントロピー $H(\boldsymbol{p}) = -\sum_i p_i \log p_i$ を最大化することで，なるべく多くの音素が出現しかつその出現確率が均等になることを目指す手法である．音素バランスに基づいて，ATR 音素バランス文セット [10] や ITA コーパス [11] といった複数の音素バランス文セットが作成されている．

2.3.2 多様性評価に基づくデータ選択

大規模データから多様性に基づいて一部のデータを選択する手法は，機械学習におけるデータサブセット選択 [12] や推薦システム [13] などの場面において重要性が報告されている．多様性を定式化する方法としては類似度の最大値を用いる方法 [14] や類似度の総和を取る方法 [15]，行列式点過程を用いる方法 [16] など複数の方法が提案されている．

2.3.3 音声合成におけるデータ選択

データ収集による多話者 TTS コーパス構築手法 [2, 17] において提案されるデータ選択法は低品質データの除去を目的としたものであり，データセットとしてのバランスや分布は考慮されていない．学習の効率化方法を目的としたサブセット選択を提案する先行研究 [18] は多話者 TTS コーパスを話者単位で分割し小規模モデルを学習することで全データを用いた学習よりも品質が向上することを示し，話者間のデータの偏りが多話者 TTS モデルの品質に悪影響を与えることを示した．しかしこの手法は話者をクラスタに分割することを目的としており，元のデータセットに含まれる全話者を単一のモデルで学習することは想定されていない．すなわち，大規模モデルの効率的な学習を目的としてサブセットを選択する手法は検討されていない．

3 手法

本研究ではバランス・多様性という 2 つの枠組みにおいて，それぞれ入力特徴量・出力特徴量を用いたデータサブセット評価指標を検討する．

3.1 データサブセット評価指標

離散特徴量のバランス：音素バランスと同様に離散特徴量の出現確率 $\boldsymbol{p}$ から定まるエントロピー $H(\boldsymbol{p})$ を最大化することでサブセットのバランスを評価する．複数の離散特徴量を同時に考慮する場合，本研究ではそれぞれの離散特徴量に対するエントロピーの和を考慮することとする．

連続特徴量の多様性：連続特徴量に基づいてサブセットの多様性を評価する方法として，類似度の総和を取る方法 [15] が提案されている．特徴量が正規化されており $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$ であるとき，$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = 2 - 2\mathrm{cossim}(\boldsymbol{x}, \boldsymbol{y})$ より cos 類似度の最小化と二乗距離の最大化は同じである．このため，本研究では

$$V(S) := \frac{1}{|S|^2} \sum_{\boldsymbol{x},\boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

---

*Study on data subset selection metrics for text-to-speech, by Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari (The University of Tokyo).

の最大化に基づいてデータ選択を実施する．これにより，特徴量空間において広く分布したサブセットを選択する効果が期待される．

3.2 特徴量

入力離散特徴量：言語情報，話者情報を表現する最も単純な方法として，音素，話者 ID を用いる．

出力離散特徴量：音声波形を離散シンボル列として扱う方法として，Generative Spoken Language Model (GSLM) [19] が提案されている．この手法では音声波形を自己教師あり学習モデルによってエンコードし，得られたフレーム単位特徴量を量子化することによって離散シンボル列を得る．GSLM は音響的特徴量をエンコードしているため，音素の接続や話者間の差異を考慮できることが期待される．

入力連続特徴量：入力テキストを表現する方法として，事前学習済された言語モデルを用いて文ごとの特徴量を抽出する手法が考えられる．この特徴量に基づく多様性最大化は複数のドメインからバランス良く学習データを取得することに寄与すると考えられる．文を固定次元特徴量に変換する方法の一つに BERT [7] の出力層を平均する方法があり [20]，本研究ではこの平均として得られるベクトルを正規化することで言語特徴量ベクトルを抽出する．

本研究で扱う TTS システムは話者を x-vector [21] という固定次元ベクトルによって表現するため，話者特徴量としてこのベクトル表現を用いる．

出力連続特徴量：音声波形の特徴量として，自己教師あり学習モデルによる特徴量が音声認識 [8] および音声品質評価 [22] など複数のタスクにおいて有効性を示している．本研究では自己教師あり学習モデルの出力ベクトル列に対し，入力テキストと同様に平均化及び正規化の処理を行って音響特徴量を取得する．

3.3 データサブセット選択アルゴリズム

データサブセットの選択は組み合わせ最適化問題であり計算量が容易に爆発しうる．本研究ではデータセット規模及びデータサブセットのサイズが非常に大きな場合への応用を目的としており，空間計算量・時間計算量の大きなアルゴリズムを実施することは望ましくない．そこで本研究では，貪欲法を用いてデータサブセットの元を 1 つずつ増やすことによりサブセット選択を実施する．

3.4 データ選択手法

上述の枠組みと特徴量の組み合わせについて，それぞれにどのような効果が期待されるかを説明する．

音素バランス：従来手法である音素エントロピー最大化 [9] を実施し，効果を検証する．

入力バランス：音素エントロピーと話者 ID エントロピーとの和を最大化する．音素・話者の偏りを整えることで入力シンボルのバランスが取れたデータセットの構築を行う．

出力バランス：GSLM のエントロピーを最大化する．音響特徴量に基づいて音素・話者情報を考慮できるため，データとしてバランスが向上することが期待される．

入力多様性：言語特徴量と話者特徴量の結合ベクトル（入力特徴量）の多様性を最大化する．類似した入力を持つデータを避けて学習データセットを構築することで学習データの範囲が広がり，モデルの汎化性能向上に寄与すると予想される．

出力多様性：音響特徴量（出力特徴量）の多様性を最大化する．類似した音声データを避けることでモデルの汎化性能が向上すると考えられる．

入出力多様性：入力特徴量・出力特徴量をベクトルとして結合し，結合特徴量に基づいて多様性を最大化する．本手法では入力多様性に基づく選択おいて除外される，入力は似ているが出力の異なるようなデータが学習データに含まれると予想される．モデルは入力と出力の組み合わせを学習するため，組み合わせとしての多様性を最大化することで学習に効果的なデータセットが構築できると期待される．

4 実験的評価

4.1 実験条件

4.1.1 データセット

多話者 TTS コーパスを用いて実験を行う．実験は日本語と英語の 2 言語についてそれぞれ**単一言語の多話者 TTS モデル**を学習することで実施する．日本語多話者 TTS コーパスとして JVS [23] コーパスの parallel 100 及び nonparallel 30 のサブセットを利用し，英語多話者 TTS コーパスとして LibriTTS-R [1] の学習用 clean サブセット（train_clean_100 及び train_clean_360）を用いた．データセットサイズはそれぞれ 25 時間，243 時間である．サブセット間の重複を回避するためにサブセットサイズをコーパス全体の約 1 割に設定し，それぞれ 3 時間，25 時間のサブセットを選択した．また，3.4 節で説明したサブセットに加えて全データを用いた学習を行い，サブセット手法のデータ不足から生じる品質低下の評価を行った．

4.1.2 多話者 TTS モデルと学習

多話者 TTS モデルとして FastSpeech 2 [24] 及び HiFi-GAN ボコーダ [25] を用いた．モデルのサイズやハイパーパラメータは日本語版の公開実装[1]及び英語版の公開実装[2]のものを用いた．ただし話者表現は one-hot ベクトルの代わりに x-vector を利用し，線形層を通じて FastSpeech 2 のエンコーダ出力に加算した．x-vector の計算には公開モデル[3]を利用し，各話者について全発話の平均を正規化して話者単位の x-vector 表現を取得した．学習ステップ数は各データサブセットの規模に合わせて設定し，JVS, LibriTTS-R でそれぞれ $50k$, $300k$ ステップとした．また，ボコーダは学習済みモデルである UNIVERSAL_V1[4]を用いた．

4.1.3 特徴量抽出器

データ選択に用いる特徴量は学習済みモデルを用いて抽出した．日本語の言語特徴量抽出には DistilBERT [26]，英語の言語特徴量抽出には BERT[5]，話

[1] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[2] https://github.com/ming024/FastSpeech2
[3] https://github.com/sarulab-speech/xvector_jtubespeech
[4] https://github.com/jik876/hifi-gan
[5] https://huggingface.co/distilbert-base-uncased

者特徴量には学習時に用いる x-vector，音響特徴量抽出には wav2vec 2.0[6] を用いた．

4.2 評価基準

4.2.1 疑似 MOS 評価

合成音声の主観評価スコア (mean opinion score: MOS) を自動予測する手法が提案されている．既存の MOS 予測モデルは相対順位の比較において予測精度が良いことが報告されており [27]，さらに学習データ（英語・中国語）に含まれない日本語音声のドメインにおいても相対評価の有効性が確認されている [2]．そのため，本研究では事前学習された UTMOS [22] モデルの強学習器[7]による予測値による疑似 MOS 評価を行った．

読み上げ文として，日本語では ITA コーパス[8]の recitation サブセットから 324 文を，英語 TTS モデルでは LibriTTS-R のテスト用の clean サブセット（test_clean_100 及び test_clean_360）からランダムに抽出した 16 文を事前に用意した．

データセットに含まれる全ての話者について，それぞれの話者の全読み上げ文の疑似 MOS を評価する．ここで，学習時に含まれなかった話者についても x-vector を用いて評価を実施した．

モデル単位疑似 MOS 評価：多話者 TTS モデルの平均的な性能を定量的に比較するために，モデルごとに全発話の平均値を計算して評価を行った．

話者単位疑似 MOS 評価：多話者 TTS モデルでは話者ごとに合成品質が異なることが予想される．これを評価するために，それぞれのモデルについて各話者の疑似 MOS 平均を算出し，話者単位疑似 MOS 分布を比較した．

4.2.2 主観評価実験

疑似 MOS 評価で有効とされたサブセット選択手法の有効性について，主観評価実験によって検証した．JVS コーパスのサブセットで学習されたモデルの合成音声を用いてプリファレンステストを実施し，合成音声の品質を評価した．評価者数は 100 人とし，各評価者はランダムに選定された話者・文の組み合わせ 10 組について合成音声の自然性を評価した．比較は疑似 MOS 評価で有効とされたサブセット選択手法とその他の手法の全ての組み合わせについて行った．

4.3 評価結果

4.3.1 モデル単位疑似 MOS 評価

表 1 に各モデルのモデル単位疑似 MOS を示す．本研究ではサブセットサイズを全体の 1 割と低い値に設定したために全データを用いる手法がサブセットより高い品質を達成しており，サブセット選択手法に対して JVS, LibriTTS-R でそれぞれ 0.024，0.020 上回る値を達成している．全データと同等の品質を達成するのに必要なデータ量に対する調査は今後の課題である．

サブセット選択手法内での比較において入出力多様性サブセットがいずれのコーパスにおいても高い品質を達成しており，他の手法に対して各コーパスで 0.025，0.030 上回る値を達成している．この結果は入出力多様性に基づくサブセット選択手法の有効性を示していると言える．

JVS における結果と LibriTTS-R における結果を比較すると，音素バランスの相対順位が一致していないことが分かる．音素バランスサブセットでは話者間バランスが考慮されておらず，母分布に依存していることが予想される．話者間バランスの取れた JVS コーパスにおいて音素バランスサブセットが高い値を達成し，話者間バランスが比較的悪い LibriTTS-R コーパスにおいて音素バランスサブセットが低い値を達成していることは，話者間バランスの重要性を示唆している．

Table 1: 各手法・データセットで学習された TTS モデルにおける疑似 MOS の全話者平均．各データセット内での相対評価に関して，比較が可能である．

手法	JVS	LibriTTS-R
全データ	3.020	3.940
音素バランス	2.971	3.879
入力バランス	2.942	3.889
出力バランス	2.924	3.888
入力多様性	2.943	3.890
出力多様性	2.931	3.899
入出力多様性	2.996	3.920

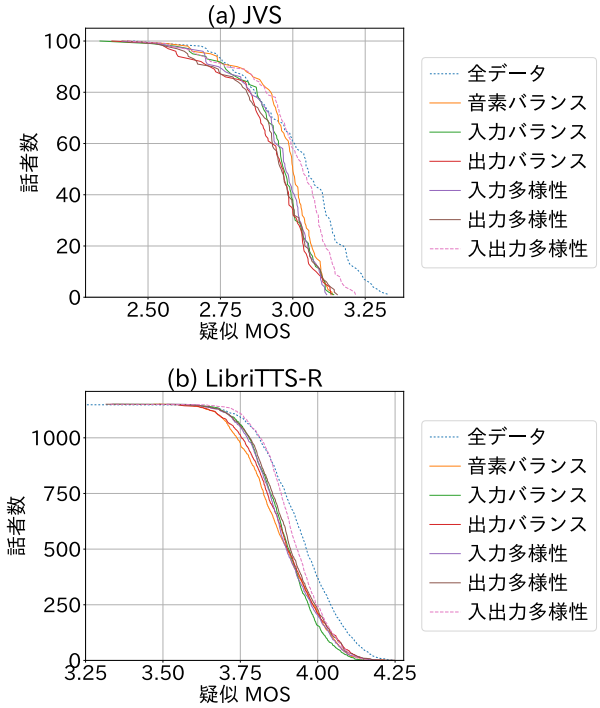

Fig. 1: 各話者の疑似 MOS の累積ヒストグラム．y 軸の値は，x 軸の値より高いスコアを持つ話者の数を表す．

4.3.2 話者単位疑似 MOS 評価

図 1 に話者単位疑似 MOS の累積ヒストグラムを示す．ほとんどの区間において入出力多様性サブセットの曲線が他のサブセットの上側にあり，入出力多様性に基づくサブセット選択の有効性を示している．

特に，JVS における疑似 MOS 2.9 付近や

[6] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[7] https://github.com/sarulab-speech/UTMOS22

[8] https://github.com/mmorise/ita-corpus

Table 2: 合成音声の自然性に関する主観評価結果（JVS コーパスから取得した入出力多様性サブセットと他のサブセットとの比較）

比較手法	スコア 比較	vs	入出力多様性	p 値
全データ	**0.601**	vs.	0.399	1.37×10^{-10}
音素バランス	0.438	vs.	**0.562**	6.02×10^{-5}
入力バランス	0.448	vs.	**0.552**	1.00×10^{-3}
出力バランス	0.457	vs.	**0.543**	7.29×10^{-3}
入力多様性	0.441	vs.	**0.559**	9.82×10^{-5}
出力多様性	0.475	vs.	0.525	1.03×10^{-1}

LibriTTS-R における疑似 MOS 3.7 付近など一部の区間において，入出力多様性サブセットは全データよりも高い話者数を達成している．これは全データにおいて話者特徴量空間上のデータの偏りが発生していることが原因と考えられる．

4.3.3 主観評価実験

疑似 MOS 評価において有効性が示された入出力多様性サブセットについて，主観評価実験によって他の手法との比較を行った．プリファレンステストの結果を表 2 に示す．いずれのサブセットに対しても入出力多様性サブセットが優れた値を取っており，入出力多様性に基づくサブセット選択の有効性が示された．各サブセットの結果を疑似 MOS の大小関係を比較する．疑似 MOS の高いサブセットほど品質が高く，したがって表に記された割合が低い値を取るはずである．しかし，実際には疑似疑似 MOS の高い音素バランスサブセットにおいて割合は高い値を取っている一方で疑似 MOS の低い出力多様性サブセットで低い値を取っており，必ずしも整合性があるとは言えない．プリファレンステストのスコアは平均値の差分のみを反映するものではなく分散によって変化するスコアであるため，音素バランスサブセットで学習された音声は品質の分散が大きいという可能性が示唆される．

5 まとめ

サブセット選択手法を検討し，従来的な音素シンボルの出現回数バランスに基づく方法に加え，基盤モデルを用いて抽出した特徴量を用いた GSLM シンボルバランス及び特徴量多様性最大化に基づく方法を検討した．検討した手法のうち，入力特徴量・出力特徴量をペアにした結合特徴量の多様性に基づいてデータを選択することが有効であることを示した．

今後の課題として，実際のダークデータを用いてさらに大規模なデータセットにおける実験を予定している．

謝辞: 本研究の一部は，科研費 22H03639 及び JST ムーンショット型研究開発事業 JPMJMS2011 の助成を受け実施した．

参考文献

[1] Y. Koizumi et al., “Libritts-r: A restored multi-speaker text-to-speech corpus,” *arXiv preprint arXiv:2305.18802*, 2023.

[2] K. Seki et al., “Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,” *arXiv preprint arXiv:2210.14850*, 2023.

[3] B. Mirzasoleiman et al., “Coresets for data-efficient training of machine learning models,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 6950–6960.

[4] K. Killamsetty et al., “Glister: Generalization based data subset selection for efficient and robust learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8110–8118.

[5] W. Ping et al., “Deep Voice 3: 2000-speaker neural text-to-speech.” in *ICLR*, vol. 79, 2018, pp. 1094–1099.

[6] R. Bommasani et al., “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021.

[7] J. D. M.-W. C. Kenton, L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[8] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[9] A. Kurematsu et al., “Atr japanese speech database as a tool of speech recognition and synthesis,” *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[10] 磯健一, “音声データベース用文セットの設計,” **昭** *63* **年春音講論**, *03*, 1998.

[11] 小口純矢 et al., “Ita コーパス: パブリックドメインの音素バランス文からなる日本語テキストコーパスの構築と基礎評価,” **研究報告音声言語情報処理** *(SLP)*, vol. 2021, no. 31, pp. 1–4, 2021.

[12] S. Agarwal et al., “Contextual diversity for active learning,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 137–153.

[13] A. Anderson et al., “Algorithmic effects on the diversity of consumption on spotify,” in *Proceedings of the web conference 2020*, 2020, pp. 2155–2165.

[14] J. Goldstein, J. Carbonell, “The use of mmr and diversity-based reranking in document reranking and summarization,” in *Proceedings of the 14th Twente Workshop on Language Technology in Multimedia Information Retrieval*, 1998, pp. 152–166.

[15] A. Borodin et al., “Max-sum diversification, monotone submodular functions and dynamic updates,” in *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, 2012, pp. 155–166.

[16] A. Kulesza et al., “Determinantal point processes for machine learning,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[17] K.-Z. Lee, E. Cooper, “A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis,” *Interspeech 2018*, vol. 12873, 2018.

[18] P. O. Gallegos et al., “An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets.” in *INTERSPEECH*, 2020, pp. 1758–1762.

[19] K. Lakhotia et al., “On generative spoken language modeling from raw audio,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[20] N. Reimers, I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[21] D. Snyder et al., “X-vectors: Robust DNN embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[22] T. Saeki et al., “UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022,” in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[23] S. Takamichi et al., “JVS corpus: free Japanese multi-speaker voice corpus,” *arXiv:1908.06248*, 2019.

[24] Y. Ren et al., “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *Proc. ICLR*, 2021.

[25] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[26] K. ”Koga et al., “Line distilbert japanese,” 2023.

[27] W. C. Huang et al., “The VoiceMOS Challenge 2022,” in *Interspeech*, 2022, pp. 4536–4540.