学習・評価ループを用いたデータ選択によるダークデータからの音声合成*

☆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

大規模音声コーパスと系列変換モデルの発展により，近年のテキスト音声合成 (text-to-speech: TTS) モデルは人間の発話に匹敵する高品質な音声合成を実現している [1]．TTS の拡張として所望の話者の特徴を持った音声を合成する，多話者 TTS の研究も盛んである [2, 3]．これらの学習には TTS コーパスと呼ばれるテキストと音声の対が必要だが，音声の収録にはコストがかかり，多話者 TTS コーパスの構築ではより一層コストがかかるため，TTS コーパスで網羅できる話者類はかなり限定される．この具体例として，日本語多話者 TTS コーパス (JVS [4]) と本論文で扱うコーパスの話者分布の比較を図 1 に示す．これは 2 つのコーパスの音声について x-vector [5] を計算し t-SNE [6] で可視化したものである．多話者 TTS コーパスでカバーできる話者は非常に限定されることがわかる．

これに対し，インターネット上には**ダークデータ**と呼ばれる利用可能性が未知の音声（YouTube 動画など）が大量に存在し [7]，これらの活用によって話者数を大きく増やすことができると期待される．これに関連して，TTS コーパス以外のコーパスから品質の良い発話のみを選択して TTS コーパスを構築する方法が提案されている [8]．一方で近年，学習データのノイズ（音響的雑音に限定されない）に頑健な TTS モデルの学習手法が提案されており [9]，これらの手法において品質の悪い発話が TTS モデルの学習に悪影響を与えるとは限らない．すなわち，TTS コーパスの構築は音声の音響品質ではなく，使用する TTS モデルに対する学習データとしての質に基づいて実施されるべきである．

本稿ではダークデータから TTS モデルを学習するためのデータ選択手法を提案する．提案手法では TTS モデルの学習と合成音声の知覚品質評価のループを通じて学習データの重要度をスコアリングし，音響品質ではなく学習データとしての質の観点からデータをフィルタリングする．加えて本稿ではダークデータの事前スクリーニングを提案し，ダークデータ収集から TTS モデル学習までの完全自動化を達成する．YouTube から取得した実際のダークデータを用いて実験的評価によって本手法の有効性を確認する．

2 関連研究

TTS モデル：多話者 TTS は話者表現によって合成音声の話者性を制御する [2]．種々の手法のうち，本稿では話者表現に x-vector [5] を用いる．また，学習データの雑音・残響を考慮して学習する TTS モデルが提案されている [9]．

TTS コーパス以外のデータの活用：学習データの話者数を増やす方法として音声認識コーパスからデータ選択する手法が提案されている．データ選択手法として音響品質やラベルノイズ（テキストと音声の不一致）を評価する方法が提案されている [3, 8]．これらは TTS モデルと独立にデータ選択を実施しており，近年のノイズに頑健な TTS モデルの学習において必ずしも適切とは言えない．さらにダークデータは巨大なコーパスを全自動で構成できる潜在能力が期待されており，音声認識の分野で数々の成功例が報告されている [10]．

合成音声の主観品質の自動評価：TTS モデルの評価にかかるコストを削減するために，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ある．自動予測における課題は汎化性能，すなわち学習データと評価データの乖離に対して頑健な予測の実現である．近年の深層学習ベースのモデルは，予測値自体は正確でなくても，音声サンプル間の相対順位の予測精度が比較的良いことが報告されている [11]．

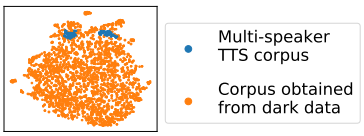


Fig. 1: 多話者 TTS コーパス [4] とダークデータ (YouTube [12]) の話者分布の違い．膨大なダークデータに比べ TTS コーパスの話者は局所的である．

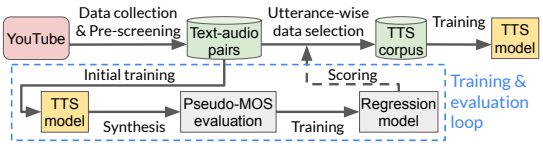


Fig. 2: 提案手法のフロー図．YouTube から取得したダークデータをモデル学習と合成音声評価を含むループによって評価し，データ選択を経てコーパスを構築する．

3 提案手法

3.1 データ取得と前処理

図 2 に示すように，まずデータ収集と前処理を実施する．この前処理により，品質が低すぎるデータを棄却する．

YouTube からデータを取得し，前処理は先行研究 [12] における音声認識用と話者識別用の両方の手法を適用する．以下に簡潔に概要を述べる．

テキストと音声の整合度に基づく処理：TTS の学習データの音声は，書き起こしテキストとよく整合している必要がある．そこで音声とテキスト（本稿では YouTube 字幕）の整合度を定量化するために，connectionist temporal classification (CTC) スコア [13] を算出する．音声は CTC セグメンテーションによっ

*Data selection for text-to-speech synthesis with training and evaluation in the loop, by Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari (The University of Tokyo).

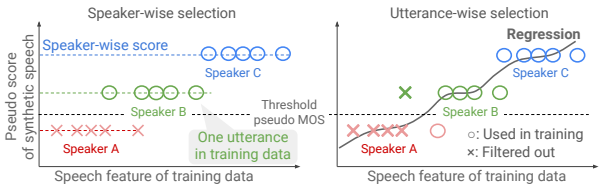


Fig. 3: 話者ごとの選択と発話ごとの選択の比較．回帰モデルを用いることで，擬似 MOS の高い話者の発話であっても低品質な発話を除去できる．

て発話ごとに分割され，スコアの低い発話は除外される．

話者分散に基づく処理：TTS の学習データは各話者について複数の発話があることが望ましい．そこで，各発話グループ（本稿では YouTube の 1 つの動画に属する発話セット）の各発話の x-vector を計算し，複数の発話において話者表現がどの程度安定しているかを話者コンパクト性スコアとして定量的に評価する．このスコアが大きい発話グループを除去し，一人の発話からなる発話グループのみを抽出する．

3.2　学習と評価のループを用いたデータ選択

本手法は学習データを**使用する TTS モデルに対する学習データとしての質**という観点で評価する．各発話に対して，そのデータを TTS モデルを学習させたときの合成音声の知覚品質を予測し，そのスコアを各データの評価スコアとして用いる．このスコアに基づいたデータ選択によって TTS コーパスを構成する．

3.2.1　初期学習

前処理で得たデータを全て用いて TTS モデルの初期学習を行う．

3.2.2　合成音声の品質評価

初期学習で得られた TTS モデルの合成音声の品質を評価する．簡単な評価方法は学習時の損失を用いることだが，これは必ずしも合成音声の知覚的品質と一致しないことが知られている [14]．知覚品質を直接的に反映するためには主観評価実験を行えば良いが，この手法では膨大な量のデータには対応できない．そこで本稿では，自動 MOS 予測モデルを用いて予測された擬似 MOS を活用する．2 節で述べたように，現在の MOS 予測モデルは高い汎化性能を達成している．そこで，学習済の MOS 予測モデルを用いて自然さの MOS を予測し，擬似 MOS として利用する．

この評価は学習データの評価に用いるため，それぞれのデータが合成音声の品質に与える影響の差を推定する必要がある．この違いをを評価する最も単純な方法は，学習データの文をそのまま合成して擬似 MOS を評価することであるが，擬似 MOS スコアは合成する文によって変化する [15] ため適切ではない．以上を踏まえて，各話者について共通の文セットの擬似 MOS の平均値を計算し，話者ごとの擬似 MOS 評価を行う．

3.2.3　各発話データのスコアリング

話者ごとの擬似 MOS に基づいて学習データをフィルタリングする．単純な方法は擬似 MOS の高い話者のデータを選択することであるが，同じ話者の音声であっても発話ごとにデータの質が異なると考えられるため，発話単位でのフィルタリングを行うことが望ましい．すなわち図 3 に示すように，擬似 MOS の高い話者であっても学習データとしての質が低い発話は除外し，逆も同様とするべきである．

このため，訓練データの各発話から話者ごとの擬似 MOS を予測する回帰モデルを学習する．ここで音響的に類似した学習データは合成される音声も類似した品質になると仮定し，さらに，回帰モデルは音響的に類似したデータに対して近い値を予測すると仮定する．これらの仮定から，我々は回帰モデルを用いることとする．前処理後の全てのデータを用いて回帰モデルを学習する．

3.3　データ選択と学習

回帰モデルによって学習データの各発話を評価し，スコアの低い発話を除外して TTS コーパスを構築する．最後にこのコーパスを用いて TTS モデルを学習する[1]

4　実験的評価

4.1　実験条件

4.1.1　データセット

先行研究 [12] の公開実装[2]を用いて YouTube から約 3,500 時間のダークデータを取得した．前処理は CTC スコアの閾値を -0.3，話者コンパクト性スコアの許容域を $[1, 7]$ として行った[3]．前処理後のデータは 2719 人の約 60,000 発話（合計 66 時間）であった．提案手法内における擬似 MOS 値評価には JVS コーパス [4] の 100 文を用いた．また最終的な合成音声の評価には ITA コーパス[4]の 324 文を用いた．

4.1.2　モデルと学習

多話者 TTS モデルとして FastSpeech 2 [16] と HiFi-GAN ボコーダ [17] の UNIVERSAL_V1[5]を用いた．モデルのサイズやハイパーパラメータは公開実装[6]を用いたが，話者表現は one-hot ベクトルの代わりに x-vector 抽出器[7]の出力を利用し，線形層を通じて FastSpeech 2 のエンコーダ出力に加算した．

TTS は JVS コーパス [4] の 10,000 発話によって事前学習した．事前学習はバッチサイズを 16 として 300k ステップ実施した．TTS モデルの学習は，この事前学習済みモデルを初期値とし，バッチサイズを 16 として 100k ステップ実施した．

合成音声の擬似 MOS の評価には事前学習された UTMOS [15] モデルの強学習器を用いた[8]．学習データ評価の回帰モデルは 1 層 256 ユニットの双方向 LSTM [18]，線形層，ReLU 活性化関数，線形層からなるモデルを用いた．この入力特徴量として自己

[1] 本稿では初期学習済みモデルからの finetuning ではなく，scratch 学習を行う．
[2] https://github.com/sarulab-speech/jtubespeech
[3] これらは先行研究 [12] と同じ値である．
[4] https://github.com/mmorise/ita-corpus
[5] https://github.com/jik876/hifi-gan
[6] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[7] https://github.com/sarulab-speech/xvector_jtubespeech
[8] https://github.com/sarulab-speech/UTMOS22

教師あり学習モデル wav2vec 2.0[9]によって抽出したフレームレベル特徴量を用いた．モデルによる各フレームの出力を平均したものを擬似 MOS とした．このモデルの学習はステップ数，ミニバッチサイズ，最適化手法，損失関数をそれぞれ 10k，12.，Adam [19]（学習率 0.0001），二乗誤差とした．

4.1.3 比較手法

本論文では以下のデータ選択手法を比較した．

Unselected: 前処理後の全てのデータ，60,000 発話を使用した．

Acoustic-quality: 各学習データの音響品質を発話単位で評価し，データ選択を行った．評価には深層学習モデルである NISQA [20] によって naturalness, noisiness, coloration, discontinuity, loudness の 5 指標を評価した．それぞれのスコアは $[1, 5]$ の値をとり，全ての指標が 3.5 以上の発話のみを選択した．データ量は約 12,000 発話となった．

Ours-Utt（提案手法）: 学習・評価ループを用いたデータ選択により発話単位の選択を行った．閾値は選択されたデータ量が “Acoustic-quality” と等しくなる値に設定した．

Ours-Spk: 学習・評価ループを用いたデータ選択だが，3.2 節で述べた話者単位の発話選択を行った．閾値は選択されたデータ量が “Acoustic-quality” とほぼ同一になるように設定した．

4.1.4 評価基準

実験結果を以下の基準で評価した．

高品質話者の増加量：提案手法によって，TTS モデルがより多くの話者について高品質な音声を合成できるようになるかを調査した．各手法について擬似 MOS の分布を調べ，閾値を超える話者（高品質話者）の人数を調査した．閾値には JVS コーパス [4] で別途学習した高品質な TTS モデルにおいて擬似 MOS の最も低い話者の値を用いた．

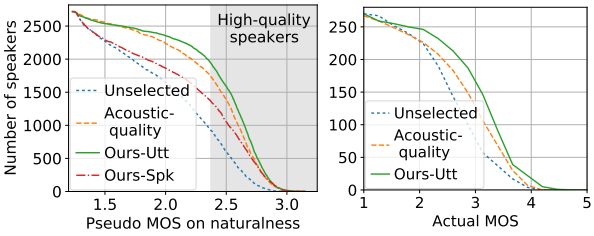
unseen 話者に対する有効性：多話者 TTS モデルの性能は学習データに含まれる話者（seen 話者）だけでなく，学習データに含まれない話者（unseen 話者）の合成音声の品質にも影響する．このため，unseen 話者の x-vector を用いて unseen 話者の高品質話者の人数を数えた．

高品質話者の多様性：高品質話者の x-vector の分布を調査し，提案手法が話者の多様性に寄与したかを検証した．定量評価として高品質話者の x-vector について Euclid 最小全域木を計算し，これを話者多様性スコアとした．

擬似 MOS と実際の MOS の比較：MOS 評価を実際に行い，提案手法の有効性を検証した．また合成音声の知覚的品質と擬似 MOS との関係を評価し，擬似 MOS の有効性を調査した．

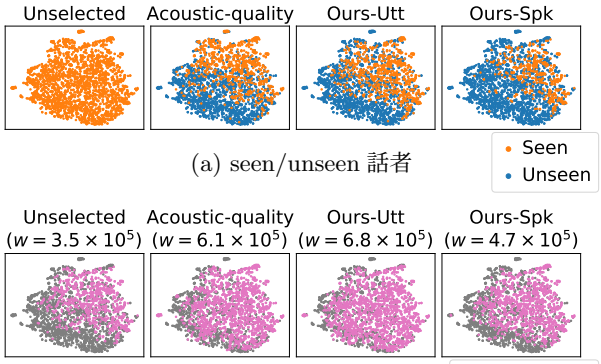
ここで，データ選択手法ごとに seen 話者と unseen 話者は異なることに注意する．“Unselected” では全ての話者が seen 話者である一方，他の手法ではそれぞれ一部の話者が unseen 話者である．各手法において学習データに選択されなかった話者は unseen 話者として扱った．特に断りのない限り，seen 話者と unseen 話者の結果を合算して記述する．

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



(a) 擬似 MOS　(b) 実際の MOS

Fig. 4: 擬似 MOS と実際の 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．



(a) seen/unseen 話者

(b) 高品質話者の分布．
w は話者多様性を表すスコアであり，
w が高いほど話者が多様であることを表す．

Fig. 5: 各手法における話者の分布

4.2 評価結果

4.2.1 高品質話者の増加量

図 4a に擬似 MOS の累積ヒストグラムを，表 1 に各手法における高品質話者の人数を示す．全手法の中で提案手法が最も高い値を示しており，提案手法を用いて学習した TTS モデルがより多くの話者について高品質な音声を合成できることが示された．

4.2.2 unseen 話者に対する有効性

図 5a は seen 話者と unseen 話者の分布を表す．“Acoustic-quality” と “Ours-Utt” は似た分布を持っていた．また，“Ours-Utt” は “Ours-Spk” に比べて広い範囲の話者をカバーしていた．表 1 に seen 話者と unseen 話者における高品質話者の人数及び割合を示す．“Ours-Utt” は “Acoustic-quality” に比べ seen 話者における高品質話者の割合が優れており，unseen 話者についても同様であった．また “Ours-Spk” は unseen 話者に占める高品質話者の割合は低かった．“Ours-Spk” の高品質話者人数が少ない原因は，unseen 話者の品質に起因すると考えられる．

4.2.3 高品質話者の多様性

図 5b に高品質話者の分布と話者多様性スコア w を示す．定性的にも定量的にも，“Ours-Utt” は他の手法と比較して話者多様性を増加させており，本手法が話者の多様性に寄与していることが分かる．

Table 1: seen/unseen 話者の人数．各セルの数字はそれぞれ高品質話者の人数，全話者の人数，２つの値の比率を表す．

手法	seen	unseen
Unselected	924/2719(34.0%)	-
Acoustic-quality	731/912(80.2%)	1006/1807(55.7%)
Ours-Utt	811/882(92.0%)	1131/1837(61.6%)
Ours-Spk	468/505(92.7%)	899/2214(40.6%)

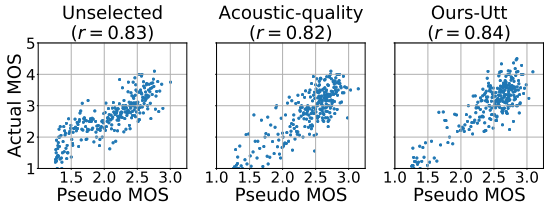


Fig. 6: 擬似 MOS と実際の MOS の比較．各点がそれぞれ話者を表す．

4.2.4　擬似 MOS と実際の MOS の比較

“Ours-Spk” を除いた TTS モデルについて，合成音声の自然さについて 5 段階の MOS 評価を行った．500 人の聴取者が参加し，各聴取者は 24 個の音声サンプルを聴いた．評価コストを削減するために 2,719 人の話者からサンプリングして評価を行った．各データ選択法において擬似 MOS を 272 区間に分割して各区間からランダムに 1 人の話者を選択し，各手法で 272 人（2,719 人のうち 10%）を用意した．実際の MOS を各話者について集計した結果を図 4b に示す．提案手法は全手法の中で最も高い値を示しており，提案手法が知覚的な音声品質を向上させることができることを示している．

これらの結果をさらに分析するため，擬似 MOS と実際の MOS との関係を調べた．図 6 は散布図と相関係数 r を示している．擬似 MOS 予測モデルの学習データは英語・中国語であり日本語を含んでいないにもかかわらず，どの手法に対しても高い相関性 ($r > 0.8$) があることが分かった．これは擬似 MOS の学習データに含まれていない言語に対しても擬似 MOS が有効であることを示しており，擬似 MOS 予測モデルの高い汎化性を示している．

5　まとめ

本稿では学習・評価ループを用いたデータ選択による TTS コーパス構成手法を提案した．YouTube から取得した実際のダークデータを用いた実験の結果，本手法は従来の音響品質に基づく手法を上回る性能を示した．

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] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, “Deep Voice 3: 2000-speaker neural text-to-speech.” in *ICLR*, vol. 79, 2018, pp. 1094–1099.

[3] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in neural information processing systems*, vol. 31, 2018.

[4] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari, “JVS corpus: free Japanese multi-speaker voice corpus,” *arXiv:1908.06248*, 2019.

[5] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “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.

[6] Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[7] Dimitar Trajanov, Vladimir Zdraveski, Riste Stojanov, and Ljupco Kocarev, “Dark data in Internet of things (IoT): challenges and opportunities,” in *7th Small Systems Simulation Symposium*, 2018, pp. 1–8.

[8] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[9] Takaaki Saeki, Kentaro Tachibana, and Ryuichi Yamamoto, “DRSpeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning,” in *Proc. Interspeech 2022*, 2022, pp. 793–797.

[10] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang *et al.*, “GigaSpeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio,” *arXiv:2106.06909*, 2021.

[11] Wen Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, “The VoiceMOS Challenge 2022,” in *Interspeech*, 2022, pp. 4536–4540.

[12] Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe, “JTube-Speech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification,” *arXiv:2112.09323*, 2021.

[13] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll, “CTC-segmentation of large corpora for German end-to-end speech recognition,” in *Proc. SPECOM*. Springer, 2020, pp. 267–278.

[14] Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe, “ESPnet-TTS: Extending the edge of TTS research,” *arXiv:2110.07840*, 2022.

[15] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, “UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022,” in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *Proc. ICLR*, 2021.

[17] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[18] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *Proc. ICLR*, 2015.

[20] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. Interspeech 2021*, 2021, pp. 2127–2131.