データ単位前処理自動選択による音声合成コーパスのデータクレンジング*

◎関 健太郎（東大），高道 慎之介（慶大／東大），
佐伯 高明，猿渡 洋（東大）

1 はじめに

近年，テキスト音声合成 (text-to-speech: TTS) モデルの学習において，より大規模な学習データを用いて性能の向上や新機能の実現を達成する手法が提案されている．このため学習データ規模拡大のために音声認識コーパスやインターネットデータといった大規模なデータセットから収集する方法が注目されている．これらのデータセットは音声データの品質が保証されていないため，品質の低いデータを除去するデータ選択の手法や，データ品質を向上させるデータクレンジングの手法といった前処理手順の重要性が高まっている．本研究では特に，データクレンジングの手法を扱う．

従来行われてきたデータクレンジングの方法は，単一の学習済み音声強調・復元モデルを用意し，これを全データに一律に適用する方法である [1,2]．しかし，事前学習済みのデータクレンジングモデルは特定の用途に合わせて学習されているおりドメイン外データにおいて性能が低下することが予想されるため，この手法はインターネットデータのような多様なデータに対して十分な性能を発揮しない．

本研究では，複数の音声強調・復元手法を自動的に選択し組み合わせるデータクレンジング手法を提案する．この方法はデータごとに最適なクレンジングモデルを選択的に適用するため，インターネットデータに対しても有効なデータクレンジングが実現可能となる．YouTube から収集した実際のダークデータを用いた実験により，提案手法の有効性を検証する．

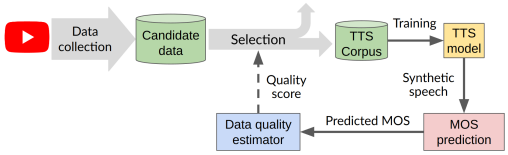
2 先行研究

先行研究ではインターネット上からのデータ収集からデータ選択，TTS モデルの学習までを自動化されたプロセスとして実施する方法を提案している [3]．この手法の全体像を図 1a に示す．従来手法では学習データの選択には音響品質などの事前決定された規範に基づいてデータ選択が実施されてきたが，学習データのノイズに頑健な TTS モデル学習手法の提案 [4] を背景にデータそのものとしての品質は必ずしも学習データとしての品質に一致しないと考えられる．そのため，この手法では，学習・評価ループに基づいてデータ品質を評価することによって，「学習データとしての品質」に基づくデータ選択を実現する．

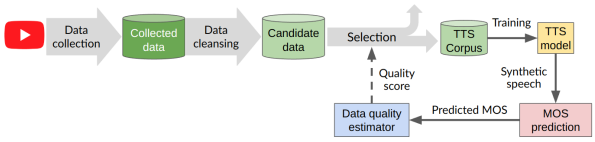
3 提案手法

本研究では，図 1b に示すように，データ収集とデータ選択の間にデータクレンジングを導入する方法を検討する．Sec. 1 で述べた通り，本研究ではインターネットデータの多様さに鑑み，複数の前処理手法を選択的に切り替えて適用する手法を提案し，データごとに最適なデータクレンジングの実現を狙う．提案するの前処理切り替えの基本的な考え方を，図 2 に示す．従来手法ではここでの前処理として画一的に音声強調・復元モデルが適用されたが，本研究では各データに対して複数の前処理手法を試行し，品質評価を行った上で最も高い品質スコアを得た手法を採用する．この前処理手法の選択をデータ単位で切り替えることにより，インターネットデータのようなドメインの広いデータに対しても効果的なデータクレンジングを実現する．

本研究では品質評価基準として音響品質を用いる基づく手法と，先行研究 [3] の提案する「学習データとしての品質」を用いる方法を検討する．後者では，学習・評価ループを用いた品質評価によってデータクレンジングを実施するため，提案手法はデータクレンジング・データ選択・学習が一つのループを形成する．



(a) 先行研究のコーパス構築手法 [3]．YouTube から取得したダークデータに対しモデル学習と合成音声評価を含むループによって評価される「学習データとしての品質」の尺度に基づいてデータ選択を実施し，コーパスを構築する．



(b) 本研究の提案手法．先行研究の手法 [3] にデータクレンジングを組み込む方法を検討する．

Fig. 1: 先行研究のコーパス構築方法と，本研究のコーパス構築方法を比較して示す．

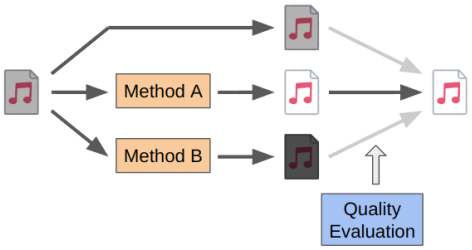


Fig. 2: 提案するデータ単位前処理選択．複数の前処理手法での結果を比較し，品質が最も高くなる前処理手法を選択する．この選択をデータごとに独立に行うことで，データ単位前処理切り替えを実現し，より効果的なデータクレンジングを実現する．

*Data Cleansing for Speech Synthesis Corpora through Automatic Data-Level Preprocessing Selection, by Kentaro Seki (The University of Tokyo), Shinnosuke Takamichi (The University of Tokyo ／ Keio University), Takaaki Saeki, and Hiroshi Saruwatari (The University of Tokyo).

3.1　音響品質に基づく手法

3.1.1　データ収集

本手法では先行研究 [3] の手法に基づいて，YouTube から手動字幕の付けられた動画を取得することで書き起こし文の付与された音声データを取得する．詳細な方法は先行研究を参照されたい．この手法によって，発話単位の音声データと，対応する書き起こしと，話者 ID を取得する．

3.1.2　データクレンジング

取得したデータに対して候補となる複数の前処理手法から最適な手法を選択してデータクレンジングを実施し，データ品質を改善する．ここでは図 2 に示すように，各データに対し候補となる前処理手法を適用し，それぞれの手法を適用した後の音響品質を評価する．前処理後の音響品質が最も高くなる手法を実際の前処理として選択し，採用する．つまり，各データに対し音響品質を最大化するような前処理を選択することとする．これによってデータセット全体の音響品質が最大限向上するため，効果的なデータクレンジングが実現することが期待される．

ここで，候補となる前処理手法には，音声強調・復元モデルに加え，取得した音声データをそのまま用いるという方法を採用し，候補の一つとして検討する．これは，十分に音響品質の高いデータに対しては前処理を適用することで手法の誤差によってむしろ音響品質が低下する恐れがあるためである．また，インターネットデータの多様性に鑑みて，どの音声強調・復元モデルも推論に失敗し音響品質の改善が行えないようなデータがあることが想定される．そのような場合に前処理の手続きを実施しないことで，データセット全体の品質が向上することが期待される．

3.1.3　データ選択

インターネットデータは品質に大きな幅を持つため，データクレンジングによって必ずしも十分な改善が実施されるとは限らない．そこで，データクレンジング後のデータを候補データとしてデータ選択を実施してコーパスを構築する．ここでは，音響品質の高いデータから順番に取得する方法を利用する．これによって，さらにコーパスの平均的な音響品質が向上することが期待される．

3.1.4　学習

候補データのうちデータ選択によって選択されたデータセットを TTS コーパスとし，TTS モデルの学習を実施する．

3.2　学習データ品質に基づく手法

Sec. 3.1 では，提案したデータ単位前処理選択によるデータクレンジング手法を先行研究 [3] のコーパス構築プロセスに導入した．この手法は音響品質を最大化するように前処理を選択しているが，先行研究では音響品質ではなく「学習データ品質」に基づいてデータ選択を実施することが TTS モデルの性能向上に有効であることが示されており [3]，前処理選択においても学習データ品質を最大化することによって効果的な前処理選択の実現が期待できる．したがって，

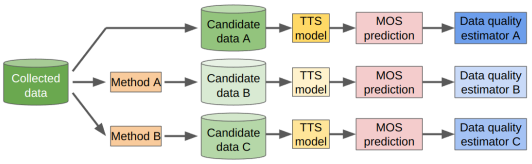


Fig. 3: 前処理手法ごとに学習データ品質を評価する方法．各前処理を画一的に適用して得られるデータセットに対して学習・評価ループを適用し，学習データ品質を評価する．

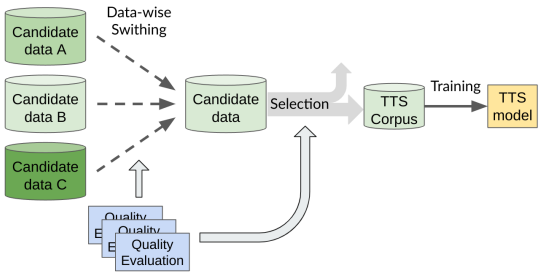


Fig. 4: 学習データ品質に基づくデータクレンジング・データ選択手法．学習データ品質に基づいて前処理を選択し，さらに学習データ品質に基づいてデータ選択を実施する．

本セクションでは学習データ品質に基づく前処理切り替えの手法を提案する．

3.2.1　データ収集・データ品質評価

Sec. 3.1.1 の手法に基づいて収集したデータに対し，データクレンジングを実施する．Sec. 3.1.2 と同様に，複数の前処理手法候補を用意し，各前処理手法を適用したときの学習データ品質を評価する．

各前処理手法ごとに学習データ品質を評価する方法を，Fig 1b に示す．提案手法では，各前処理手法を全データに画一的に適用し，先行研究 [3] 学習・評価ループによって学習データ品質を評価する．この手法は候補データを用いて TTS モデルを学習し，学習された TTS モデルの合成音声の品質を話者ごとに評価し，学習データから合成音声の品質を予測する回帰モデルを学習する方法である．これによって，各データに対しそのデータを用いて学習された TTS モデルの合成品質を予測することで，学習データとしての品質を評価する枠組みである．この手法は単一の多話者 TTS モデルであっても話者ごとに合成音声の品質に分散が生じることを利用して一回のモデル学習で各データの学習データ品質を評価しているが，本研究のように複数の前処理手法を組み合わせた場合，一回のモデル学習では前処理ごとに切り分けた評価は実現できない．そこで本研究では，前処理の手法ごとに別々に学習データ品質評価を実施する．

3.2.2　データクレンジング・データ選択・学習

学習データ品質評価に基づいて，図 4 に示す方法でデータクレンジング・データ選択を実施する．データクレンジングにおいては，学習データ品質が最も高くなる前処理をデータごとに選択し，採用する．さら

に，Sec. 3.2.1 の手法によって学習データ品質が評価されているため，それに基づいてデータ選択を実施する．具体的には，採用された前処理における学習データ品質をそのまま各データの学習データ品質とみなし，これが高いデータから順番に取得し，TTS コーパスを構築する．構築されたコーパスを用いて TTS モデルの学習を行う．

4　実験評価

4.1　実験条件

4.1.1　データセット

データ取得には先行研究 [5] の公開実装[1]を利用し，YouTube から約 3,500 時間のダークデータを取得した．ここでは 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] としてデータ取得時に利用可能なデータの選択を実施した[2]．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 Mean opinion score (MOS) 値評価には JVS コーパス [6] の 100 文を利用し，最終的な合成音声の評価には ITA コーパス[3]の 324 文を用いた．

4.1.2　モデルと学習

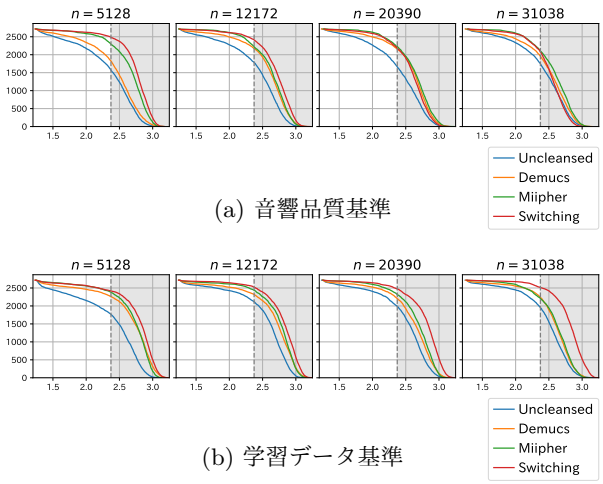
多話者 TTS モデルの音響モデルには FastSpeech 2 [7] を採用し，公開実装[4]のハイパーパラメータを用いた．話者表現として x-vector 抽出器[5]の出力を利用し，線形層を通じて FastSpeech 2 のエンコーダ出力に加算した．ボコーダには学習済みの HiFi-GAN ボコーダ [8] の UNIVERSAL_V1[6]をファインチューニングせずに利用した．

FastSpeech 2 は JVS コーパス [6] の 10,000 発話によって，バッチサイズを 16 として 300k ステップの事前学習を実施した．FastSpeech 2 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 ステップの学習を実施した．

合成音声の擬似 MOS の評価には事前学習された UTMOS [9] モデルの強学習器を用いた[7]．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 [10]，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教師あり学習モデル wav2vec 2.0[8]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学習データ品質とした．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 [11]（学習率 0.0001），二乗誤差とした．

4.1.3　データクレンジング・データ選択

音響品質に基づくデータクレンジング・データ選択では，深層学習モデル NISQA [12] によって評価された 5 つの指標の最小値を音響品質評価値とした．



(a) 音響品質基準

(b) 学習データ基準

Fig. 5: 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．グレーの領域は高品質な話者に対応する．

データクレンジング手法として，以下を検討した．

- **Unclenased**: データをそのまま用いた．スタジオ収録音声のような十分に品質の高い音声には，この手法が適していると予想される．
- **Demucs**: 前処理手法として，全データに対し事前学習済みの音声強調モデル Demucs [13] を適用する[9]．Demucs は音源分離モデルの一種であり，バックグラウンドミュージックを除去するように訓練されている．そのため，加法性ノイズを低減し，データ品質を向上させる効果が期待される．
- **Miipher**: 前処理手法として，全データに対し事前学習済みの音声復元モデル Miipher [14] を適用する[10]．このモデルは劣化した音声を復元するために訓練されている．したがって，録音機器や残響による環境歪みを低減し，データ品質を向上させる効果が期待される．
- **Switching**: 各データに対し，“Unclenased”, “Demucs”, “Miipher” のうちのいずれか一つを選択し，適用した．選択は音響品質もしくは学習データ品質に基づいて実施され，品質が最も高くなる手法を選択した．

データ選択におけるデータセットサイズは $n = 5128, 12172, 20390, 31038$ とした．

4.1.4　評価

以下の点を明確にするため，評価を実施した．

- **擬似 MOS が向上するか**：“Switching” 手法は他の方法に比べ全体的な合成品質を向上させることが期待される．そこで収集したデータに含まれる 2719 人の話者について合成音声の擬似 MOS を評価し，その分布を調べた．特に，閾値を超える話者を高品質話者と定め，その領域における擬似 MOS 分布に着

[1] https://github.com/sarulab-speech/jtubespeech
[2] これらは先行研究 [5] と同じ値である．
[3] https://github.com/mmorise/ita-corpus
[4] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[5] https://github.com/sarulab-speech/xvector_jtubespeech
[6] https://github.com/jik876/hifi-gan
[7] https://github.com/sarulab-speech/UTMOS22
[8] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[9] https://huggingface.co/spaces/Wataru/Miipher/tree/main
[10] https://github.com/facebookresearch/demucs

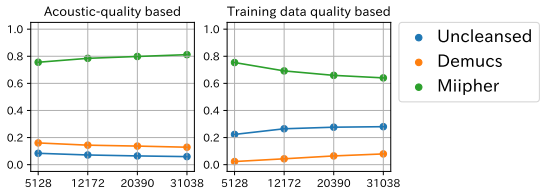


Fig. 6: “Switching” において各前処理手法が選択された割合．

目した．ただし閾値には JVS コーパス [6] 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．

- **どのデータクリーニング手法が実際に選択されたのか**：予備実験において，データクリーニング手法の中で “Miipher” が際立っており，“Miipher” が頻繁に選択される可能性が予想された．これを検証するため，どのデータが選択されたかを調査した．さらに，“Miipher” を単独で使用した場合との比較を行い，この選択が与える影響について議論した．

4.2 評価結果

4.2.1 擬似 MOS に基づく評価

図 5 に擬似 MOS の累積ヒストグラムを示す．学習データ基準の手法において，“Switching” はどの手法においても他の手法を上回っており，提案手法の有効性が確認される．一方で音響品質基準の手法においては，“Switching” は他の手法を下回る場合があり，特に $n = 31038$ において高品質話者の合成品質が “Miipher” よりも低下していることが分かる．この結果は，音響品質が学習データとしての品質評価に適切でないためと考えられる．すなわち，前処理選択のための学習データの品質評価においても，先行研究 [3] の学習データ品質評価指標が効果的であることが示唆される．

4.2.2 切り替え手法の調査

図 6 に，“Switching” 手法での各前処理手法の選択率を示す．全ての条件で，Miipher が半数以上のケースで選択されている．YouTube から得られたデータはスタジオ録音よりも低品質の録音が多いと推定されるため，Miipher による音声修復の効果が大きいと考えられる．実際，取得したデータのうち，音響品質が 4.0 を超える高品質な音声サンプルは 5128 件で，全体の 10% 未満である．

一方，学習品質に基づく選択では，“Uncleansed” がデータの 20% 以上で選ばれている．これは，YouTube から得られたノイズの多いデータの中にも，データクリーニングモデルを適用しない方が望ましい高品質なデータが一定割合含まれていることを示している．これらの結果は，ダークデータの品質分布が広範であるため，画一的な前処理手法には限界があることを示唆している．

5 まとめ

本研究では，先行研究のインターネットデータからの TTS コーパス構築手法にデータクレンジングを組み込む方法を検討した．提案手法は複数の前処理候補からデータごとに最適な手法を切り替えることで各データに適した前処理手法を選択し，効果的なデータクレンジングを実現する．さらに，本研究の前処理選択において，学習データとしての品質に基づいて前処理を選択する手法を提案した．

実験的評価によって，提案手法の学習データ品質に基づく前処理切り替え手法の有効性を確認した．さらに，音響品質を用いる前処理選択における実験結果との比較によって，前処理選択において学習データ品質を評価基準に用いることの有効性が確認された．

謝辞: 本研究の一部は，科研費 22H03639，24KJ0860 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] Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, and Hiroshi Saruwatari, “Coco-nut: Corpus of japanese utterance and voice characteristics description for prompt-based control,” *arXiv preprint arXiv:2309.13509*, 2023.

[2] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna, “LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus,” in *Proc. INTERSPEECH 2023*, 2023, pp. 5496–5500.

[3] Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari, “Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[4] Takaaki Saeki, Kentaro Tachibana, and Ryuichi Yamamoto, “DRSpeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning,” *Proc. INTERSPEECH*, 2022.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] Simon Rouard, Francisco Massa, and Alexandre Défossez, “Hybrid transformers for music source separation,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[14] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani, “Miipher: A robust speech restoration model integrating self-supervised speech and text representations,” *arXiv preprint arXiv:2303.01664*, 2023.