TTSOps 2.0: テキスト音声合成におけるデータ収集・前処理・学習プロセスの統合的最適化*

◎関 健太郎 (東大/慶應大)，齋藤 佑樹 (東大/産総研)，
高道 慎之介 (慶應大/東大)，佐伯 高明 (東大)，猿渡 洋 (東大)

1 はじめに

テキスト音声合成（text-to-speech; TTS）技術は機械学習モデルの進展に伴い急速に発展しており，この発展を支えている要因の一つが大規模データの活用である．特に YouTube やポッドキャストなどの Web スケールデータをいかに活用するかが重要な課題として注目されている．

Web スケールデータは無尽蔵に存在するため，その中からどのデータを選択して学習に用いるかという**データ選択手法**が極めて重要となる．代表的なデータ選択手法の分類を Fig. 1a に示す．

一つ目の軸は，model-agnostic と model-aware の区別である．従来手法 [1] はあらかじめデータを選択した後にモデルを学習するフィードフォワード型の枠組みに基づいており，人手で設計されたデータ選択基準の妥当性が最終的なモデル性能を大きく左右してきた．これに対し，我々は学習と評価を反復的に行う feedback-loop 構造を導入した TTSOps を提案した [2]．TTSOps では，学習済みモデルの合成結果を評価するループによって，そのデータがモデルの学習にどの程度寄与するかを直接的な基準として用い，model-aware なデータ選択を実現している．

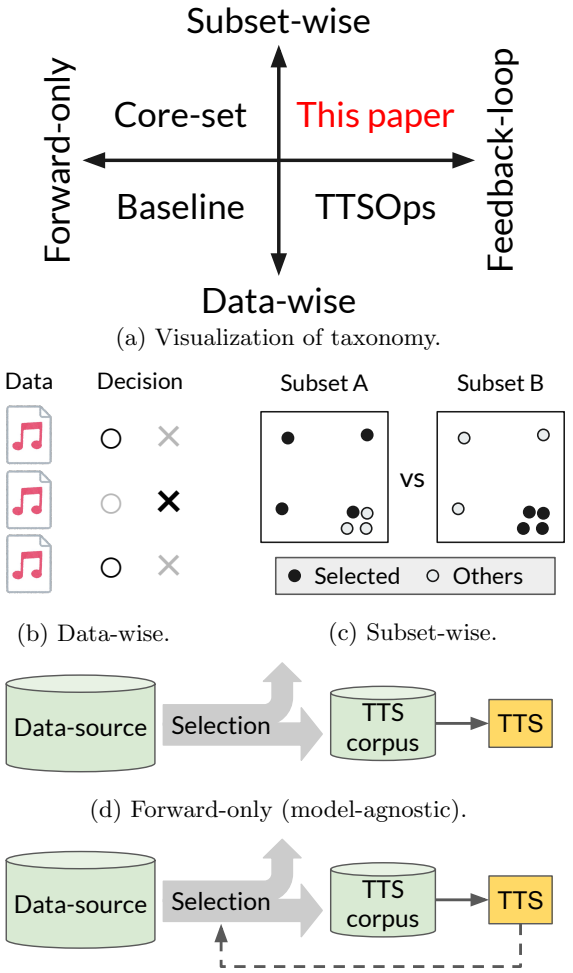
もう一つの軸は，data-wise と subset-wise の区別である．data-wise な手法では，各データを独立に品質評価してデータ選択を行うため，データ間の重複や分布の偏りといった集合全体の性質は考慮されないことが多い．しかし，同質なデータを過剰に収集しても，学習効率や話者バリエーションの観点からは必ずしも有益とはいえない．特に Web データは分布の偏りを含みやすく，集合全体の多様性を考慮したデータ選択が重要となる．この課題に対して，我々はコアセット選択に基づく subset-wise なデータ選択手法を提案した [3]．コアセット選択はデータ集合全体を俯瞰し，カバレッジを最大化するように部分集合を選択することで多様性の高いコーパス構築を目指す．

本研究では，feedback-loop に基づく model-aware なデータ品質評価と，subset-wise に不足領域を補完するデータ取得を統合したデータ選択手法を提案し，効率的なコーパス構築を実現する．

2 提案手法

本研究では，多話者 TTS における話者バリエーションの増強を目的とする．具体的には，一定以上の合成品質しきい値を満たす話者を「合成可能」と定義し，その話者数を最大化することを目標とする．

提案手法の概要を Fig. 2 に示す．Step 0 では Web から初期データ候補 C_0 を収集し，Step 1 において品質フィルタリングを行い初期データセット C_1 を構成する．Step 1 の品質フィルタリングには，TTSOps [2] で提案された feedback-loop 型の手法を用い，事前定義されたヒューリスティックではなく合成音声の自動品質評価に基づいてデータ選択を行う．続く Step 2 では，能動学習に基づきデータ追加を行う．具体的には，Step 1 で構築したフィルタ済みコーパスを用いて学習したモデルの合成性能に基づき，学習データが不足している話者の発話を Web から選択的に取得し，データセットに追加する．この際，すでに十分な高品質データを有する話者の発話は除外されるため，データセット全体としての冗長性が低減される．さらに，冗長性の評価に model-aware な基準を用いることで，事前定義された指標への依存を回避する．



(a) Visualization of taxonomy.

(b) Data-wise.

(c) Subset-wise.

(d) Forward-only (model-agnostic).

(e) Feedback-loop (model-aware).

Fig. 1: Taxonomy of data selection methods in TTS.

*Integrated Optimization of Data Collection, Data Selection, and Training Process for Text-to-Speech Synthetis, by Kentaro Seki (UTokyo/Keio), Yuki Saito (UTokyo/AIST), Shinnosuke Takamichi (Keio/UTokyo), Takaaki Saeki, and Hiroshi Saruwatari (UTokyo).

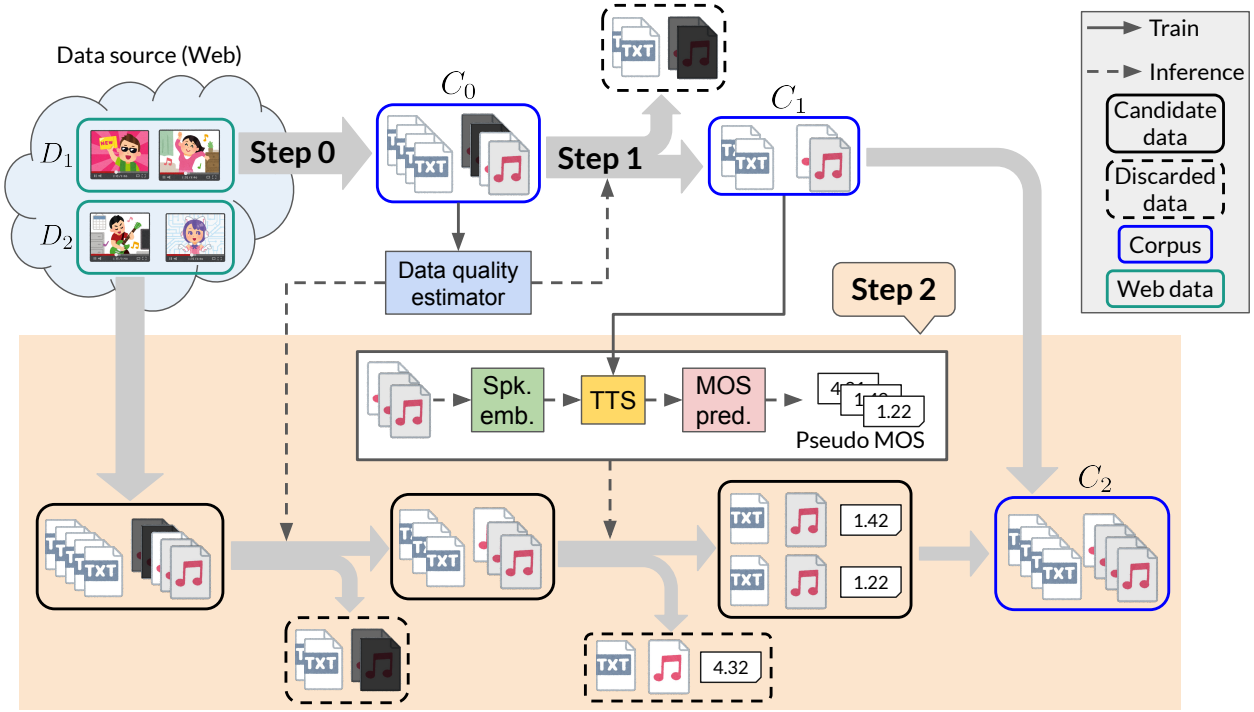


Fig. 2: Overall procedure of proposed method.

2.1 準備

先行研究 [4] に基づいて YouTube 上で動画を検索し，動画 ID の集合 D_{all} を作成する．動画 ID は，音声とテキストからなる完全なデータペアと比べて軽量であるため，大規模な事前収集が可能である．

合成品質のしきい値を定めるため，スタジオ品質データを用いて多話者 TTS モデルを学習する．このモデルの合成品質はスタジオ品質データと同等であるとみなせることから，話者ごとの pseudo-MOS の最小値を合成品質のしきい値 τ_{hq} と定義する．

2.2 Step 0: 初期データ候補の取得

動画 ID プール D_{all} から部分集合 D_1 をランダムにサンプリングし，各動画から音声とテキストのペアを収集することで初期コーパス C_0 を作成する．基本的な構造条件を満たさないデータを除外するため，先行研究 [2] と同様に，音声認識整合率および話者表現分散スコアに基づくフィルタリングを行う．こうして取得された各データ $x_i \in C_0$ は，音声 a_i，書き起こし t_i，および話者表現埋め込み s_i からなる組 (a_i, t_i, s_i) として表現される．

2.3 Step 1: 低品質データの除去

初期コーパス C_0 に含まれる低品質データを除外するため，先行研究 [2] に基づき，model-aware な品質フィルタリングを行う．本手法では，C_0 を用いて多話者 TTS モデルを学習し，すべての話者に対して pseudo-MOS を評価する．得られた話者ごとの合成品質に基づいて，学習データ品質推定モデル $\text{TQ}(\cdot;\theta_{\text{TQ}})$ を学習する．ここで θ_{TQ} はモデルパラメータである．TQ は，音声 a_i を入力として，対応する話者の pseudo-MOS を推定するように学習される．この推定値が合成品質のしきい値 τ_{hq} を上回る発話のみを選択することで，高品質な合成音声の生成に寄与する発話を残すことを目的とする．したがって，品質フィルタリング後のコーパス C_1 は，以下のように定義される．

$$C_1 = \{(a_i, t_i, s_i) \mid \text{TQ}(a_i;\theta_{\text{TQ}}) > \tau_{\text{hq}}\}. \quad (1)$$

2.4 Step 2: 追加データ取得

合成可能な話者数を増やすため，コーパス C_1 の学習データが不十分な話者に対し，追加の学習データを取得する．C_1 を用いて学習した TTS モデルを $\text{TTS}(\cdot,\cdot;\theta_{C_1})$ とし，固定されたテストテキスト集合 T を用いて，各話者 s_i の平均合成品質（SQ）を以下のように定義する．

$$\text{SQ}(s_i;\theta_{C_1}) = \frac{1}{|T|}\sum_{t\in T}\text{MOS}(\text{TTS}(t, s_i;\theta_{C_1})), \quad (2)$$

ここで MOS$(\cdot)$ は，合成音声の品質を推定する pseudo-MOS 予測モデルである．

追加データ候補 $D_2 = D_{\text{all}} \setminus D_1$ のうち，以下の二つの条件をともに満たすデータ $x_i = (a_i, t_i, s_i)$ のみを追加対象とする．

$$\text{TQ}(a_i;\theta_{\text{TQ}}) > \tau_{\text{hq}}, \quad (3)$$

$$\text{SQ}(s_i;\theta_{C_1}) < \tau_{\text{hq}}. \quad (4)$$

これにより，発話単体としての品質は十分である一方，合成性能が依然として不十分な話者のデータのみが選択的に追加される．その結果，データセット全体の冗長性を抑えつつ，データ容量に対して効率的な学習コーパスを構築できる．

以上の手続きによって得られた追加データを C_{add} とし，最終的な学習コーパスを $C_2 = C_1 \cup C_{\text{add}}$ と定義する．

Table 1: Quantitative comparison across different corpus construction strategies. Our proposed active learning method outperforms both the random baseline and core-set selection across all metrics for both FastSpeech 2 and Matcha-TTS.

	FastSpeech 2			Matcha-TTS		
metric	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)
Baseline	2.4192	1467	16.9047	2.1457	784	26.6982
Core-set	2.5055	1780	16.5465	2.2899	1111	25.6426
Active Learning	**2.6019**	**2185**	**16.0223**	**2.3846**	**1577**	**22.9432**

3 実験評価

3.1 実験設定

3.1.1 データセット

JTubeSpeech [5] の公開実装[1] を用いて取得した 2,719 本の動画 ID を D_{all} とした．このうち 10%（271 本）を D_1，残りの 2,448 本を $D_2 = D_{\text{all}} \setminus D_1$ とした．各動画からは平均して約 20 発話が得られた．

3.1.2 モデルおよび学習

本研究では，音響モデルと事前学習済み HiFi-GAN ボコーダ [6][2] を組み合わせることで，多話者 TTS モデルを構成した．音響モデルとして，FastSpeech 2 [7][3] および Matcha-TTS [8][4] を用いた．話者条件付けには，事前学習済みモデルから抽出した 512 次元の x-vector[5] を線形層により 256 次元に射影し，音響モデルへの条件入力として用いた．

学習対象は音響モデルのみとし，初期化には JVS コーパス [9] を用いた 300k ステップの事前学習済みモデルを使用した．その後，各ステップにおいて，当該モデルを 100k ステップファインチューニングすることで学習を行った．

pseudo-MOS の推定には UTMOS モデル [10] [6] を使用した．JVS コーパス [9] を用いて得られた合成品質のしきい値 τ_{hq} は，FastSpeech 2 では 2.4156，Matcha-TTS では 2.3693 であった．

3.1.3 比較手法

Baseline: TTSOps [2] に対応する手法として，学習データ品質の推定値が合成品質のしきい値 τ_{hq} を超えることをデータ選択の条件とする．すべてのサンプルは品質条件を満たす一方で，Web データに内在する分布の偏りにより，必ずしも十分な性能が得られない可能性がある．データ量を統一して比較するため，条件を満たすデータからランダムサンプリングすることで学習データを構成する．

Core-set: Baseline に model-agnostic な subset-wise selection を組み合わせた手法として，学習データ品質の推定値がしきい値 τ_{hq} を超えるデータを候補とし，コアセット選択 [3] により学習データを選択する．集合全体のデータ分布を考慮することで，合成可能な話者バリエーションの増加が期待される．

[1] https://github.com/sarulab-speech/jtubespeech
[2] https://github.com/jik876/hifi-gan
[3] FastSpeech 2: https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[4] Matcha-TTS: https://github.com/shivammehta25/Matcha-TTS
[5] https://github.com/sarulab-speech/xvector_jtubespeech
[6] https://github.com/sarulab-speech/UTMOS22

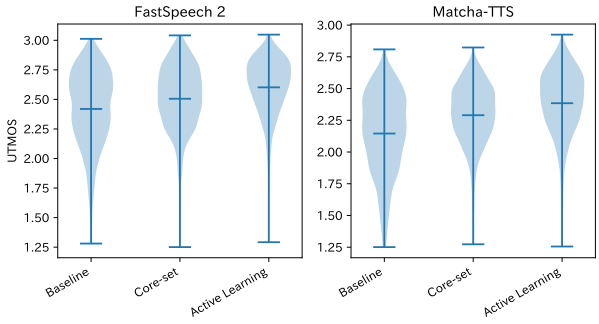


Fig. 3: Distribution of UTMOS scores for each method. Violin plots show the speaker-wise UTMOS distributions under three corpus construction strategies, with three horizontal lines indicating the maximum, mean, and minimum values respectively.

Active Learning: 提案手法では，model-aware な基準に基づき，集合単位でのデータ選択および追加取得を行う．

3.1.4 評価指標

2,719 話者すべてに対して主観評価実験を実施することは困難であるため，本研究では以下の自動評価指標を用いて評価を行う．

UTMOS: (↑) 自然性の擬似的な評価尺度として，UTMOS モデル [10] による pseudo-MOS 評価を行う．2,719 話者それぞれについて話者単位の pseudo-MOS を算出し，その平均値および分布を比較する．

 # HQ-Speaker: (↑) 本研究におけるコーパス最適化の目的関数として，話者平均の合成品質が合成品質のしきい値 τ_{hq} を超える話者数を評価する．

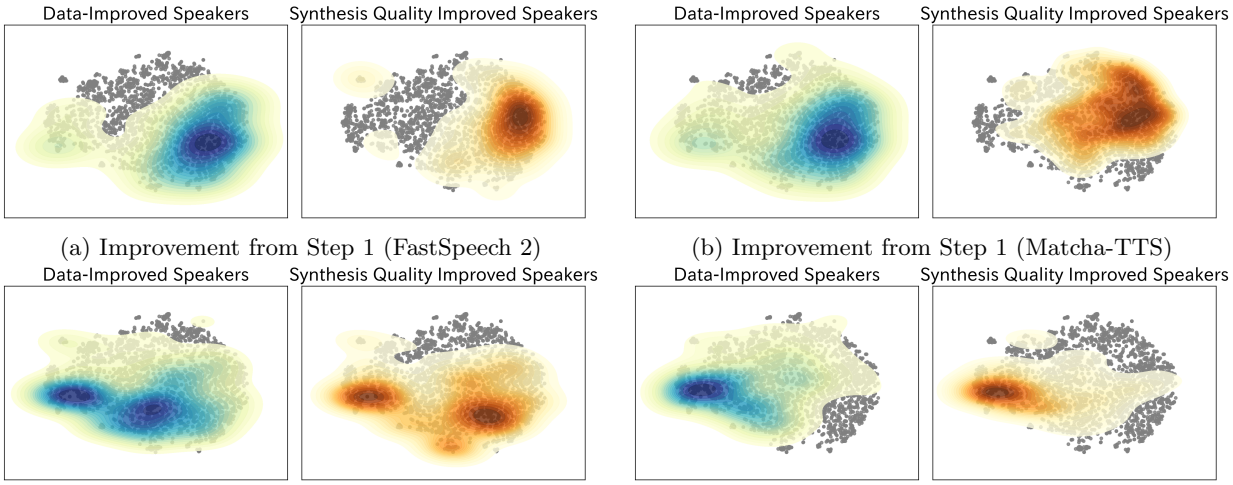
CER: (↓) 明瞭性を評価する指標として，合成音声の文字誤り率（CER）を算出する．音声認識には Whisper large-v3 モデル[7]を用いる．

3.2 結果 1. ベースラインとの比較

定量評価の結果を Table 1 に示す．UTMOS, #HQ-Speaker, CER のすべての指標において，“Active Learning” は “Baseline” および “Core-set” を上回る性能を示した．UTMOS の分布を Fig. 3 に示す．“Baseline” では低品質側まで広い分布を示しているのに対し，“Core-set” および “Active Learning” では分布全体が高品質側へとシフトしていることが確認できる．

この結果は，Web データに内在する分布の偏りにより，一部の話者に対して学習データが不足し，合成品質が低下していたことを示唆している．subset-wise

[7] https://github.com/openai/whisper



(a) Improvement from Step 1 (FastSpeech 2)

(b) Improvement from Step 1 (Matcha-TTS)

(c) Improvement from Step 2 (FastSpeech 2)

(d) Improvement from Step 2 (Matcha-TTS)

Fig. 4: Kernel density estimation (KDE) visualization of speaker distributions in the x-vector space. Colored regions indicate the density of speakers whose data or synthesis quality is improved at each step, while gray points represent the distribution of all speakers. These visualizations illustrate how data refinement (Step 1) and additional data acquisition (Step 2) contribute to synthesis quality improvement.

なデータ選択を導入することで，こうした偏りが緩和され，話者間の品質ばらつきが改善されたと考えられる．特に “Active Learning” は “Core-set” と比較しても優れた性能を示しており，model-aware な基準を組み合わせることで，subset-wise データ選択の有効性がさらに向上していることが分かる．

以上の結果から，同一データ量における多話者 TTS モデルの話者カバレッジを改善するという観点において，提案手法による model-aware かつ subset-wise なデータ選択が有効であるといえる．

3.3 結果 2. ステップごとの分析

提案手法の各段階におけるコーパス改善の効果を分析するため，Step 1 および Step 2 におけるデータセットの改善と，それに伴う TTS モデル性能の改善を比較した．

まずデータセットの改善を分析するため，コーパス更新によって学習データが改善された話者の分布を調べた．具体的には，Step 1 において低品質発話が除去された話者と，Step 2 において新たな発話が追加された話者について，それぞれ話者空間における分布を可視化した．

次に TTS モデルの改善を分析するため，各 Step において，擬似 MOS が合成品質のしきい値 τ_{hq} を新たに上回った話者，すなわち Step 実行前には高品質話者でなく，Step 実行後に高品質話者となった話者の分布を調べた．

これらの話者分布を可視化するため，t-SNE [11] により射影した x-vector 空間上において，カーネル密度推定を用いた可視化結果を Fig. 4 に示す．両モデルおよび両ステップにおいて，データ品質の改善が観測された話者分布と，合成品質の改善が観測された話者分布は，空間上で類似した領域に集中していることが確認できる．これは，各ステップで行われたデータ操作が，直接的に TTS モデルの品質改善に寄与していることを示している．さらに Step 1 と Step 2 を比較すると，それぞれ異なる領域の話者に対して改善が生じており，両ステップが補完的な役割を果たしていることが分かる．

4 おわりに

本論文では，テキスト音声合成におけるデータセット構築において，subset-wise data selection と feedback-loop data selection の二つのパラダイムを統合することで，データ効率の高いコーパスを構築する手法を提案した．提案手法は，データ品質フィルタリングと，不足している話者データを重点的に補完する二つのステップから構成され，段階的にデータセットを改善する．実験結果から，提案手法が model-aware なデータ選択として有効であることを示した．さらに，ステップごとの分析により，両ステップが互いに補完的な役割を果たしていることを明らかにした．

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] Heiga Zen *et al.*, in *Proc. Interspeech*, 2019, pp. 1526–1530.

[2] K. Seki *et al.*, *IEEE TASLP*, vol. 33, pp. 4956–4970, 2025.

[3] K. Seki *et al.*, in *Proc. ICASSP*, 2024, pp. 12 351–12 355.

[4] K. Seki *et al.*, in *Proc. ICASSP*, 2023, pp. 1–5.

[5] S. Takamichi *et al.*, *arXiv:2112.09323*, 2021.

[6] J. Kong *et al.*, *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[7] Y. Ren *et al.*, in *Proc. ICLR*, 2021.

[8] Shivam Mehta *et al.*, in *Proc. ICASSP*, 2024, pp. 11 341–11 345.

[9] S. Takamichi *et al.*, *arXiv:1908.06248*, 2019.

[10] T. Saeki *et al.*, *arXiv:2204.02152*, 2022.

[11] Laurens Van der Maaten *et al.*, *Journal of machine learning research*, vol. 9, no. 11, 2008.