音声・テキストペアが存在しない状況における TTS および STT による合成データ混合を用いた ASR 学習

野田 陽 [1] 酒井 眞 [1] 杉野 かおり [1] 田森 秀明 [1] 岡崎 直観 [2] 乾 健太郎 [3,4,5]
[1] 朝日新聞社 [2] 東京科学大学 [3]MBZUAI[4] 東北大学 [5] 理化学研究所
{noda-h3,sakai-m16,sugino-k,tamori-h}@asahi.com
okazaki@comp.isct.ac.jp kentaro.inui@mbzuai.ac.ae

概要

Automatic Speech Recognition（ASR）モデルの性能向上のための，Text-To-Speech（TTS）や Speech-To-Text（STT）を用いたデータ拡張手法が多数提案されている．従来は，既存の音声コーパスの音声部分から STT によりテキストを作成したペアデータや，付随するテキスト部分から TTS により合成音声を作成したペアデータを学習に用いることが多い．TTS データ単体では性能が悪化することが報告されており，STT で作成したデータや，TTS データの作成に使用した元データとの混合が ASR 性能向上に有効であると示されている．

本研究では，音声・テキストのペアデータが存在しない状況を想定し，TTS と STT に入力するデータの由来が異なる場合でも，混合合成データで性能向上が可能であることを確認した．さらに混合比率の影響や，TTS に用いる記事テキストの特徴が推論結果に与える影響についても検討する．

1 はじめに

近年，Whisper を代表とする大規模音声認識モデルのファインチューニングは広く用いられている．高精度化には，大量の音声とテキストのペアが必要であるが，収録や文字起こしを人手で行うにはコストがかかる．この課題に対し，TTS を用いたデータ拡張の研究が盛んである．近年の TTS モデルの性能向上によって，高品質な合成音声・テキストペアデータが作成可能になっている．一方で TTS データのみを用いた学習は，ASR 性能低下を招きやすいことが報告されており，自然音声データとの組み合わせが必要となっている．

日本語 ASR を対象とした研究では，TTS の入力テキストに，既存の音声コーパスに付随の書き起こし文を用いる例が多数である．しかし実運用では，自然音声と書き起こし文のペアを大量に収集することは難しく，日本語の大規模音声コーパスも限られている．一方で，テキストデータは比較的容易に収集可能である．自然音声と書き起こし文が存在しない場合であっても，テキストから作成した TTS データを有効に活用できることが望ましい．

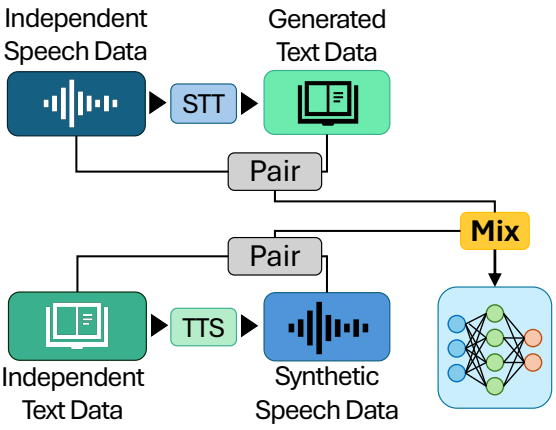


図 1 提案する合成データ作成方法の概要

そこで本研究では，音声・テキストのペアデータが存在しない場合を想定し，(i) 新聞記事テキストを用いて合成音声を作成した TTS ペアデータと，(ii) 自然音声データから STT によってテキストを作成した STT ペアデータを混合したデータセットを構築する．このデータを用いた学習を行い，ASR 性能向上が可能であるかを検証する．さらに，TTS データと STT データの混合比率を変化させることで，自然音声データがどの程度必要かを調査する．

加えて，新聞記事テキストを TTS データに用いた際の影響を把握するため，推論結果の誤り傾向（置換誤り）についても分析し，無制限に作成可能な TTS データを有効に活用するための適切な設計が必要であることの示唆を得る．

2 関連研究

2.1 TTS を用いたデータ拡張

Li ら [1] は，自然音声と合成音声の混合に着目し，合成音声のみでは性能が悪化するため，適切な比率で自然音声と混合する必要があると示している．また，Rosenberg ら [2] は，音響的多様性や語彙の多様性を考慮した合成音声の作成に注目し，合成音声を用いた学習の有効性が向上することを示す一方で，合成音声単独では実音声による学習との差は依然として存在し，実音声と合成音声を併用することの必要性を報告している．Yang ら [3] は，低リソース ASR タスクにおける TTS によるデータ拡張の有効性を示し，特に音声合成に入力するテキストの多様性が ASR の性能向上に寄与すると主張している．

2.2 合成データの影響

Mizumoto ら [4] は，大規模音声言語モデルを対象に，ASR 性能における合成データの影響について調査している．TTS データのみでは性能が低下する一方で，自然音声データや STT データと組み合わせることで性能向上が可能であることを示している．同研究では ReazonSpeech コーパスを基に合成データを作成しているが，由来が異なる合成データ同士での検討はされていない．

以上を踏まえ，本研究では，合成音声データと自然音声データの由来が異なる場合に着目する．音声・テキストのペアデータが存在しない状況を想定し，独立した由来の合成データの混合データセットによる Whisper のファインチューニングについて検証する．

3 新聞記事テキスト

本節では，TTS 入力テキストとして，朝日新聞社が保有する新聞記事テキストを使用する．テキストの特徴について調査するため，新聞記事テキスト 2 年分について分析した．STT の入力音声として使用した ReazonSpeech についても同様に分析し，比較する．

ReazonSpeech はテレビ放送から音声と字幕を取得し，音声と字幕のアラインメントを取ることで作成された大規模音声コーパスである．

語彙数　ユニークなトークン数をカウントし，それを語彙数としたところ，新聞記事テキストの語彙数は 297,613 語だった．ReazonSpeech の持つ語彙数は 255,661 語であり，新聞記事テキストと共通している語彙数は 148,089 語だった．

表 1　全トークンにおける主な品詞の出現率

品詞	新聞記事テキスト	ReazonSpeech
名詞	40.39%	29.84%
助詞	30.86%	32.29%
動詞	13.21%	14.06%
助動詞	8.10%	12.23%
形容詞	1.12%	1.62%
副詞	0.87%	2.46%
代名詞	0.51%	1.71%
接続詞	0.17%	0.44%

トークン中の品詞出現率　表 1 は，全トークン中における主な品詞の出現率である．ReazonSpeech のテキストと比較して，新聞記事テキストは記号類が多いため，全トークンから記号・補助記号を除外した．新聞記事テキストは ReazonSpeech に比べて，名詞の出現率が高い．これは小磯ら [5] で検証されている，フォーマルな文体ほど名詞率が高いということに当てはまる．名詞は，文章の中で語彙的な意味を持つ自立語に分類される．反対に文法的な機能を持つ機能語（助詞・助動詞・代名詞・接続詞）の出現率はいずれも ReazonSpeech の方が高くなっている．新聞記事テキストは特に名詞の出現率が高いため，総合して自立語全体の出現率が高くなっている．以上のことから，新聞記事テキストが ReazonSpeech と比較して，特に名詞が多く含まれる情報量の多いテキストだと考えることができる．情報量が多いテキストを音声合成に用いることで，ASR モデルがより広い語彙に対応できるようになることが期待できる．

4 実験設定

4.1 音声認識モデル

本研究ではファインチューニングするベースラインモデルとして，Whisper-medium[1)] を使用した．Whisper は，OpenAI が開発した音声認識モデルであり，約 68 万時間に及ぶデータで事前学習されている．ファインチューニングの設定は付録 A の表 3 に記載している．

1）　https://huggingface.co/openai/whisper-medium

4.2 TTS

音声合成には VOICEVOX[2] を使用した．合成音声に音響的多様性を持たせるため，話者数は許諾上問題のない範囲で最大話者数である 26 話者を使用した．

TTS データを作成するにあたって，発話の区切りの良い箇所で分割するため，記事テキストを句点で分割し音声合成した．また Whisper では推論時に音声を 30 秒に分割するため，それに合わせて，30 秒を超えてしまうものは分割し，ファインチューニングに使用した．

VOICEVOX による音声合成では，句読点やかぎ括弧部分で自然な間をとった発話となるため，これらの記号（、。「」）は残し，その他の記号は削除して音声合成とファインチューニングデータに使用した．

1 サンプルにランダムに 1 話者を割り当て，テキストと音声のペアを作成した．

4.3 STT

STT データの作成には，大量の音声データを高速に推論する実運用の観点から，ベースラインモデルとは異なる nvidia/parakeet-tdt_ctc-0.6b-ja[3] を使用した．推論する音声には，ReazonSpeech v1 コーパス [6] の音声データを使用した．

4.4 データセット設計

ファインチューニングには，ReazonSpeech v1 コーパスの約 18,000 時間中，自然音声・書き起こしテキストペアの Orig データ 5,000 時間分と，記事テキストから作成した TTS データ 5,000 時間分，そして ReazonSpeech v1 コーパスの音声部分を用いて作成した STT データ 5,000 時間分を使用した．

データセット設計は以下のようにした．TTS データと STT データの混合比率を変化させることで，TTS データの性能悪化がどれくらいの STT データとの混合で抑えられるかも調査する．

- Orig 5,000 時間分
- TTS 5,000 時間分
- STT 5,000 時間分
- TTS 2,500 時間分 + STT 2,500 時間分
- TTS 3,500 時間分 + STT 1,500 時間分
- TTS 4,500 時間分 + STT 500 時間分

2） https://voicevox.hiroshiba.jp/
3） https://huggingface.co/nvidia/parakeet-tdt_ctc-0.6b-ja

実験では比較のため，ReazonSpeech のペアデータ（Orig）も用いるが，本研究の主眼は音声とテキストのペアがない場合に，由来の異なる音声のみ・テキストのみから TTS や STT から作成したデータの混合データが有効かを検証する点にある．

5 評価方法

5.1 評価データセット

本研究では，評価データセットとして，JSUT Basic5000[7]，Common Voice Corpus 6.1(CV6.1)[8]，Inhouse テストデータ [9] を使用する．JSUT は，単一女性話者によって無響室で収録された音声コーパスである．CV6.1 は，録音環境や話者はサンプルによって異なり，1 サンプル中の話者数は 1 人である．Inhouse テストデータは，自社で公開前提で収録されたインタビュー音声がメインとなっており，1 サンプル中の話者数が複数のものが含まれている．録音環境や話者はサンプルによって異なる．

各データセットのノイズ　各データセットのノイズを可視化するために，参照音声を必要としない音声品質推定モデル NISQA[4] [10] で評価データセットと ReazonSpeech の音声，TTS データの音声についてノイズスコアを算出した．これは数値が大きいほどノイズが少ないことを表す．分布図は付録 B に記載している．TTS データ・JSUT・CV6.1 は数値が大きい方に分布が寄っており，ノイズの少ない音声が中心になっている．一方で ReazonSpeech と Inhouse テストデータは数値が小さい方に分布がよっており，ノイズの多い音声が中心になっていることがわかる．

5.2 文字誤り率：CER

文字誤り率（Character Error Rate: CER) とは，正解テキストと ASR モデルによる推論テキストを比較し，正解テキストに対して文字単位での挿入誤り，削除誤り，および置換誤りの文字数を数え，総文字数に対するその誤り数を正規化した指標である．CER が小さいほど正確に文字起こしできていると言える．以降，CER は百分率（%）で表記する．記号（、。「」）の出力傾向が条件間で大きく異なるため，CER は正解・推論テキストから記号を除去した上で算出した．

4） https://github.com/gabrielmittag/NISQA

表 2　各条件での CER(↓)

	JSUT	CV6.1	Inhouse
Baseline	10.45%	22.46%	35.45%
Orig 5k	9.03%	10.05%	31.06%
TTS 5k	10.09%	25.08%	96.40%
STT 5k	8.73%	10.23%	24.11%
TTS 2.5k+STT 2.5k	7.60%	10.19%	23.63%
TTS 3.5k+STT 1.5k	8.38%	10.10%	25.44%
TTS 4.5k+STT 0.5k	9.52%	11.43%	28.40%

5.3　置換誤り分析

TTS に新聞記事テキストを使用した影響を調査するために，トークン単位での置換誤りを分析した．正解・推論テキストにおいて，アラインメントが正しく取れているサンプルを使用するため，推論テキストの CER が 10 %以下のサンプルにフィルタリングした．正解・推論テキスト，どちらにもトークンが存在しているが一致しない場合を置換誤りとし，全評価データセットでの品詞別割合を算出した．

記事テキストは名詞の割合が高いため，名詞の置換割合が少なくなっていることを期待する．

6　実験結果とその考察

6.1　CER 結果

表 2 は，ベースラインおよび各データセットでのファインチューニングにおける CER 結果である．

TTS データ　TTS データのみでのファインチューニングは JSUT 以外でベースラインより性能が低下している．5.1 節より，TTS データと JSUT のノイズ状況が似ていることから，音響的特徴や録音環境がテストデータと似ている TTS データではそれほど性能が悪化しないと考えられる．しかし，工夫のない合成音声だけでは，性能向上に繋がらないということが改めて確認できた．

STT データ　STT データのみでのファインチューニングにおいては，JSUT と Inhouse において，Orig データのみの場合より CER が小さくなっている．これは ReazonSpeech の元のラベルの品質に悪い部分があり，STT により適切なラベルになったのではないかと考えられる．

TTS データ+STT データ　1:1 で混合した時，いずれのデータセットにおいても，CER は Orig データと同等かそれ以下になった．この結果から，由来の異なる TTS データと STT データを組み合わせた場合でも性能向上可能であることが示唆される．また，TTS の比率を上げた場合においてもベースラインを上回る結果となった．総合的にみると，STT の比率が大きいほど CER が小さくなる傾向があり，STT の自然音声が TTS の音響的な弱みを補完している可能性が考えられる．一方で，これらの結果は，一定量の自然音声さえあれば，大量の TTS データとの組み合わせによって ASR 性能向上ができることを示唆している．

6.2　置換誤り分析

置換誤りの品詞別割合には条件間で大きな差は見られなかった．記事テキストの名詞の多さを生かすには，TTS に入力するテキストの適切な設計が必要だと考えられる．結果については，付録 C の表 4 に記載している．

6.3　その他の記事テキストの影響

合成音声の，記号位置での間をとった発話により，TTS データでのファインチューニングの推論結果に，かぎ括弧や句読点が多く現れた．特に句読点は多く出力された．このことから，ファインチューニングに用いるデータに句読点やかぎ括弧を残しておくことで，推論後の後処理をしなくても自然な位置に句読点等を挿入できることが期待できる．

7　まとめ

本研究では，音声・テキストのペアデータが存在しない場合を想定し，自然音声から作成した STT データと，新聞記事テキストから作成した TTS データの混合データのファインチューニングで，Orig データでのファインチューニングと同等以上の性能向上ができることを示した．

一方で，記事テキストの特徴が生かされたような推論結果とはならなかった．このことから，ドメイン適応などにおいてテキストの特徴を生かしたい場合には，TTS に入力するテキストの適切な設計が必要だと考えられる．今後は，低リソース言語や方言に焦点をおき，10 時間程度など，より少量の自然音声での実験を行う予定である．また，異なる TTS モデルや，大規模言語モデルとの組み合わせにについても，今後の検討課題である．

参考文献

[1] Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin. Training neural speech recognition systems with synthetic speech augmentation. **arXiv preprint arXiv:1811.00707**, 2018.

[2] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In **IEEE automatic speech recognition and understanding workshop (ASRU)**, 2019.

[3] Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. Enhancing low-resource asr through versatile tts: Bridging the data gap. In **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2025.

[4] Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, and Yui Sudo. Is synthetic data truly effective for training speech language models? In **Interspeech**, 2025.

[5] 小磯花絵, 小木曽智信, 小椋秀樹, 宮内佐夜香. コーパスに基づく多様なジャンルの文体比較-短単位情報に着目して-. 言語処理学会第 15 回年次大会 (NLP2009), 2009.

[6] Yue Yin, Daijiro Mori, and Seiji Fujimoto. Reazonspeech: A free and massive corpus for japanese asr. 言語処理学会第 29 回年次大会 (NLP2023), 2023.

[7] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. **arXiv preprint arXiv:1711.00354**, 2017.

[8] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In **Proceedings of the twelfth language resources and evaluation conference**, 2020.

[9] 山野陽祐, 田森秀明, 杉野かおり, 黒田由加. アクティブラーニングによる音声認識モデルのための効率的なデータアノテーション手法. 人工知能学会全国大会第 38 回 (JSAI2024), 2024.

[10] 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 **Interspeech**, 2021.

[11] Nick Rossenbach, Mohammad Zeineldeen, Benedikt Hilmes, Ralf Schlüter, and Hermann Ney. Comparing the benefit of synthetic training data for various automatic speech recognition architectures. In **IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, 2021.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

Appendix

A　ファインチューニング設定

表 3　ファインチューニング設定

Optimizer	AdamW[12]
Peak Learning rate	2e-5
GPU	8× A100 80GB
Scheduler	linear
Batch size	32
Accumulation steps	2
Epoch	1.5

B　NISQA

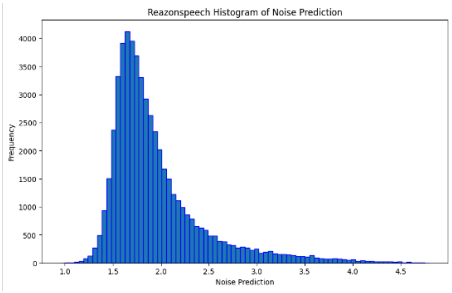
NISQA(Neural Intrusive and Non-Intrusive Speech Quality Assessment) とは，参照音声を必要とせず音声品質の人間の主観的評価を推定できるモデルである．総合 MOS に加えて，雑音，歪み，音色劣化，不連続性が推定される．これらの指標は 1 から 5 で評価され，1 だとノイズが多い，5 だとノイズが少ないと人間が感じることを指す．今回は各データセットのノイズについて視覚化するために，雑音スコアを算出した．ReazonSpeech と記事テキストから作成した TTS 音声データは全体の 0.5%をサンプリングしたものからスコアを算出した．図 2 が各データセットの分布図である．

C　置換誤り中の品詞別割合

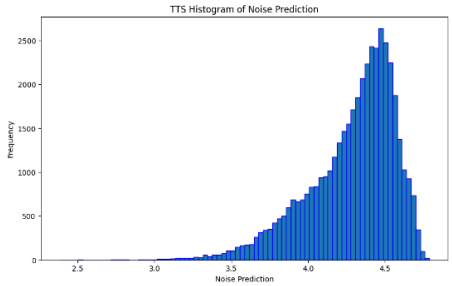
1:1 が TTS2.5k+STT2.5k，7:3 が TTS3.5k+STT1.5k，9:1 が TTS4.5k+STT0.5k の結果である．

表 4　置換誤り中の品詞別割合

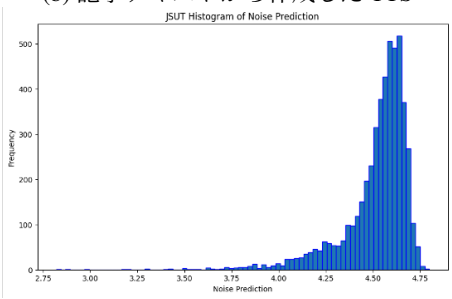
品詞	Orig	TTS	STT	1:1	7:3	9:1
名詞	59.90%	56.88%	58.19%	56.90%	57.44%	59.82%
助詞	3.31%	6.5%	5.01%	5.37%	5.66%	5.27%
動詞	24.21%	22.54%	24.05%	24.08%	23.36%	21.95%
助動詞	0.94%	1.31%	1.81%	1.55%	1.56%	1.25%
形容詞	3.21%	3.35%	2.89%	3.41%	3.45%	3.63%
副詞	1.96%	2.04%	2.00%	2.08%	2.26%	1.80%
代名詞	1.19%	1.69%	1.17%	1.52%	1.25%	1.00%
接続詞	0.24%	0.35%	0.32%	0.37%	0.34%	0.29%



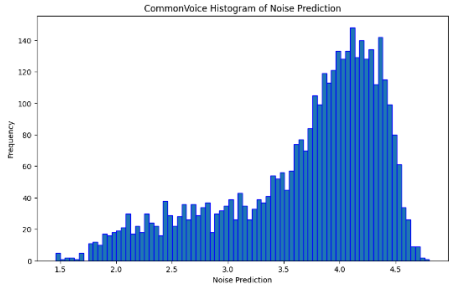
(a) ReazonSpeech



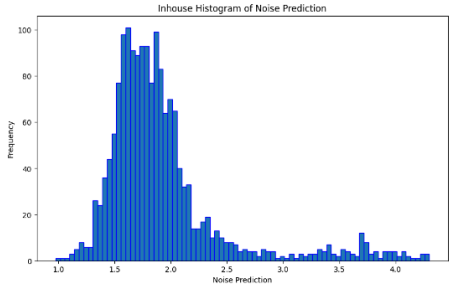
(b) 記事テキストから作成した TTS



(c) JSUT Basic5000



(d) Common Voice 6.1



(e) Inhouse

図 2　NISQA の雑音スコア