SS-JDSC：単一話者日本語構音障害音声コーパス *

小笠原 朝陽 (岩手大)*，○高道 慎之介 (慶大／東大)*，楊 家寧 (東大)*，
末永 剛 (クリエイター)，談 宜育 (岩手大)

1 はじめに

音声認識（ASR）は音声入力に基づく AI 利用 [1, 2] を促進するが，構音障害者はその恩恵を享受していない．この情報享受格差は，既存の音声認識が構音障害音声を精度よく認識できないこと [3] に起因する．

この発展には，構音障害音声認識に資するオープンコーパスの整備が急務である．Table 1 のコーパスが整備されているが，(1) 構音障害が音声に与える影響は言語に依存する [4] にも関わらず，英語・中国語以外のコーパスがほとんど存在せず，(2) 個人特化 ASR の性能は学習データ量に依存する [5] 一方で，話者あたりのデータ量が限定的であるという欠点がある．

そこで本研究では，日本語における初めての構音障害音声オープンコーパス SS-JDSC (single-speaker Japanese dysarthric speech corpus) を構築する．単一話者のみのコーパスだが，既存コーパスに比べ話者あたりのデータ量は非常に大きく，また，音声認識のために設計された多様な音声から成る．本論文では，本コーパスを用いた音声認識実験と，学習した音声認識がコミュニケーションに与える影響を報告する．

Table 1 構音障害音声コーパスの比較．

コーパス	言語	話者数	時間数 (全体/話者) [h]	公開
Whitaker [11]	En	6	unk / unk	
UA-Speech [12]	En	19	unk / unk	✓
TORGO [13]	En	8	23 / 2.88	✓
Euphonia [14]	En	1000+	1300+ / 1.3	
SAP [15]	En	400+	500+ / 0.8	✓
CU DYS [16]	Zh	11	7.5 / 0.68	✓
MSDM [17]	Zh	25	6.8 / 0.23	
CDSD [5]	Zh	44	44 / 1	✓
EasyCall [18]	It	31	unk / unk	✓
(Unnamed) [19]	Ja	16	unk / unk	
SS-JDSC	**Ja**	1	15.1 / **15.1**	✓

Table 2 Hard サブセットのテキストの例

カテゴリ	テキスト
CF	ピカピカに磨いたパイプをポケットにしまった．
PCS	タイタニックとダイナミックな展開に驚いた．
SUS	カメラがペンを持ってタクシーを呼んだ．

2 コーパスデザイン

本コーパスは 3 つのサブセットから成る．

2.1 Basic サブセット

Basic サブセットは，通常音声コーパスである JSUT BASIC5000 [6] と同じ文を発話した音声から成る．JSUT BASIC5000 の文は，日本語の常用漢字を全て含んでおり，音声合成 (TTS) [7] の学習や，ASR の評価 [8] に適している．

2.2 Hard サブセット

Hard サブセットは，構音障害の影響が強く表れる音声から成る．例を Table 2 に示す．日本語の構音障害音声は，破裂音 /p, t, k/や破擦音 /ts/ などで聞き取り困難が生じる場合が多い．本サブセットの phonemic confusion (CF) カテゴリは，そのような音素を多く含む．また，音声的に混同しやすい文 (PCS) と合文法無意味文 (SUS) のカテゴリも含む．

2.3 Daily サブセット

Daily サブセットは，日常的に使う文から構成される．構音障害話者にインタビューを行い，当該話者が ASR を必要とする日常シーンを想定し everyday, research, work, emotion, others のカテゴリとその発話音声を用意した．

2.4 収録とアノテーション

収録． 先天性鼻咽腔閉鎖不全症を持つ，20 代の日本語母語話者である男性 1 名の音声を収録した．収録には Shure MV5 マイクロフォンを用いた．サンプリング周波数，ビット深度，ファイルフォーマットは，それぞれ 44.1 kHz，16-bit，RIFF WAV 形式とした．全ての収録を日本の一般的な家屋環境で行い，話者自身が音声収録を監督した．

アノテーション． 収録後に発話を手動でチェックし，障害に起因しないと思われる収録誤りをコーパスから除外した．最終的な内容量を Table 3 に示す．

構音障害の深刻度． FDA-2 [9] のような，構音障害の深刻度に関する情報はコーパスに含まれない．代わりに，他の構音障害音声コーパスと比較してその深刻度を間接的に評価する．次節に示す通り本コーパスの単語誤り率 (WER) は 93%であり，この数値は TORGO コーパス [10] における最も深刻なクラス “severe” と同程度である．なお，両コーパスで言語が異なるためこの比較は参考程度とされたい．

3 音声認識実験

3.1 実験条件

データ． サブセットごとに約 450 発話からなる評価セットを用意した．残るデータは学習セットおよび検証セットして利用した．学習セットのサイズは 13.0 時間であり，30 発話（サブセットごとに 10 発話）を，大規模言語モデル (LLM) のための検証セットとして利用した．サンプリング周波数は 16 kHz とした．

音声認識モデル． `whisper-large-v3` (1.54B

*Equal contribution.

Table 3　コーパスのスペック．

	カテゴリ	説明	発話数	時間数 [h]
Basic	-	通常音声とパラレル	4701	6.89
Hard	CF	障害の影響を受けやすい文	2094	1.84
	PCS	音声的に混同しやすい文	75	0.07
	SUS	合文法無意味文	73	0.09
Daily	Everyday	家族や友達との会話文	2098	2.62
	Research	研究に関する議論	927	1.39
	Work	公共の場における会話文	721	1.06
	Emotion	感情を表現した文	414	0.54
	Others	緊急，季節，趣味	399	0.63
Total	-	-	11502	15.12

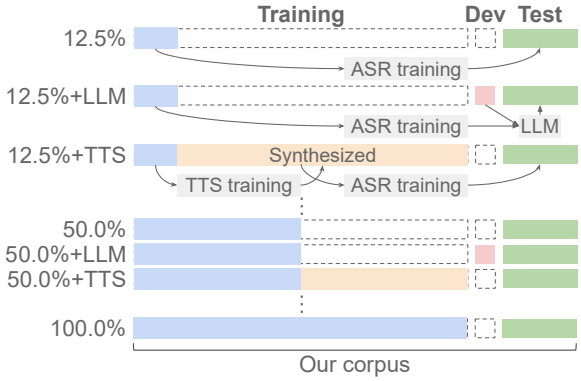

Fig. 1　コーパスの分割．全設定で共通の評価セットを用いる．100% は 13.0 時間の学習データに相当する．

parameters)[1]をファインチューニングした．ファインチューニング時は言語 ID を指定した．実装は Hugging Face `transformers`[2]，特徴量抽出器 `WhisperProcessor` と，学習管理 `Seq2SeqTrainer` を用いた．最適化には beta = (0.9, 0.999)，epsilon = 1×10^{-8}，学習率 1×10^{-5} の Adam を用いた．スケジューラには，500 warm-up step を設けた線形減衰を用いた．ミニバッチサイズは 32 とし，学習には単一の NVIDIA GH200 を用いた．

学習データ拡張と推論データ補正． 3.3 節では，TTS に基づく学習データ拡張 [10] と，LLM に基づく推論データ補正 [20] を実施した．TTS モデルとして，VITS [21] に基づく `Style-Bert-VITS2`[3]を使用した．学習時のハイパーパラメータは公式実装の既定値とし，単一の NVIDIA V100 GPU を用いて学習した．データ補正時には `gpt-4o` API を使用した．

評価指標． WER と BERTScore [22][4]を用いた．

3.2　実験 1：学習データサイズの影響

Fig. 1 に示すように，学習データサイズを 13.0 時間 (100%), 6.5 時間 (50%)，... と対数的に減衰させて，音声認識精度を評価した．減衰時には，各サブセットの含まれる割合が変わらないように発話を乱択した．参考として，Basic 評価セットに対応する通常音声を JSUT BASIC5000 から選択し，通常音声に対する `whisper-large-v3` の音声認識精度も評価した．

[1] `https://huggingface.co/openai/whisper-large-v3`
[2] `https://huggingface.co/docs/transformers`
[3] `https://github.com/litagin02/Style-Bert-VITS2`
[4] `https://pypi.org/project/bert-score/`

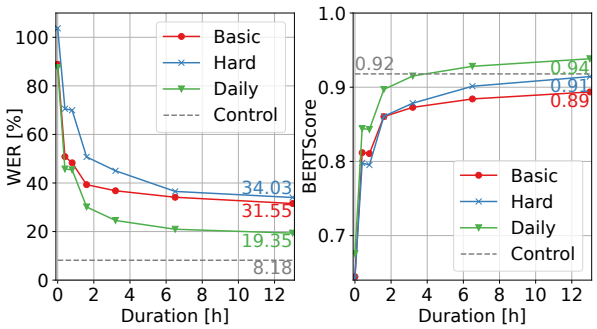

Fig. 2　学習データサイズを変化させたときの WER (左) と BERTScores (右)．

Fig. 2 に示すように，学習データサイズを増加させると WER と BERTScore は単調に改善する．WER については 13.0 時間の学習データにおいてほぼ収束傾向にあることから，本コーパスは WER 改善にあたり十分な量の発話を含んでいるといえる．一方，BERTScore においては 13.0 時間を超えても改善が期待される．

サブセット同士を比較すると，混同しやすい音素を多く含む Hard サブセットの WER が最も悪い．Daily サブセットは，他のサブセットよりも比較的単純な語彙や文脈であるため，最も良い WER および BERTScore を獲得したと思われる．しかしながら，学習データサイズによる改善はみられるものの，あらゆる実験条件の WER は，通常音声の WER を下回っていない．

3.3　実験 2：学習データ拡張と推論データ補正

学習データ拡張と推論データ補正の効果を検証した．下記の 2 手法は Fig. 1 の “*+TTS” と “*+LLM” にそれぞれ対応する．

- **TTS に基づく学習データ拡張**：TTS モデルを学習データでファインチューニングしたのち，ファインチューニングに使用していないテキストから音声を合成し，ASR の学習データに追加した．例えば，12.5% の学習データをファインチューニングに用いるとき，残る 87.5% のテキストから音声を合成した．
- **LLM に基づく推論データ補正**：推論時に音声認識結果を LLM で補正する．補正時には正解文–推論文の対を含むテキストプロンプトを LLM に与え，few-shot learning による補正を試みる．プロンプトの具体例を Fig. 3 に示す．

Fig. 4 に結果を示す．“Finetune” は，Fig. 2 と同様に，学習データ拡張と推論データ補正を使用しない設定である．まず，“Finetune” と “+TTS” を比較すると，学習データ補正は，学習データサイズが小さいときに，特に大きなスコア改善をもたらすことがわかる．しかしながら，“+TTS” の最良設定でも，13.0 時間の学習データを用いた “Finetune” の性能には至っていない．すなわち，TTS 音声は実収録音声

You are a professional at correcting speech recognition errors. Below are pairs of speech recognition results and their correct transcriptions. The errors include the insertion, substitution, or deletion of specific phonemes.
—
Speech recognition: [text]
Reference: [text]
...
—
Correct the following speech recognition result; explanations are not required.

Fig. 3　推論データ補正に用いるテキストプロンプト．

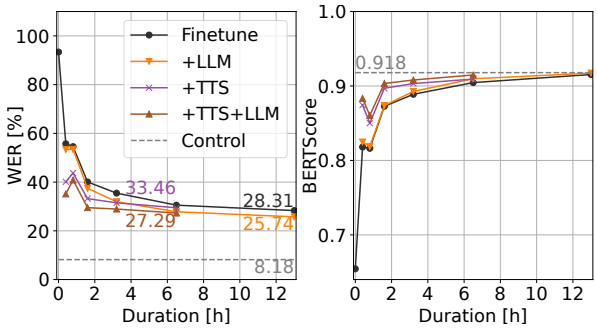

Fig. 4　TTS に基づく学習データ拡張と，LLM に基づく推論データ補正を実施した際の WER (左) と BERTScores (右)．全ての評価セットの結果を平均した値を表示している．

を代替できるほどの品質にないこと（例えば jitter の再現 [23]）がわかる．

推論データ補正の影響は，学習データ補正と逆の傾向にある．学習データ量が少ないときに補正の効果は限定的だが，量が増加すると補正の効果も増大する．さらに，学習データ拡張と組み合わせることで，13.0 時間の学習データを用いた “Finetune” の WER を超える．一方で，通常音声に対する音声認識精度には至っていない．

3.4　実験 3：人間の認識能力との比較

ASR の音声認識精度を，人間の認識能力と比較した．話者との社会的関係の異なる 2 種類の評価者に対し，構音障害音声の書き起こし試験を実施した．

- **Human (close)**: 話者自身が縁故法で依頼した，家族や友人などの，話者と頻繁に会話する成人 4 名．
- **Human (far)**: クラウドソーシング[5]で依頼した，構音障害者との会話経験のない成人 50 名．

全ての評価者は，聴覚に障害のない日本語母語話者であった．共通のウェブサイトを用い，再生音声の発話内容を書き起こさせた．書き起こしにあたる制限時間は設けず，聴きなおしは何度でも認めた．評価セットは，各サブセットの評価セットから 10 発話ずつ選択した，合計 30 発話である．評価者の負担を軽減するために，長い発話は選択対象から除外した．

Fig. 5 に結果を示す．0.4 時間の学習データを用いる時点で，ASR の WER と BERTScore は “Human (far)” を超えることがわかる．さらに，データ拡張／

[5] https://www.lancers.jp

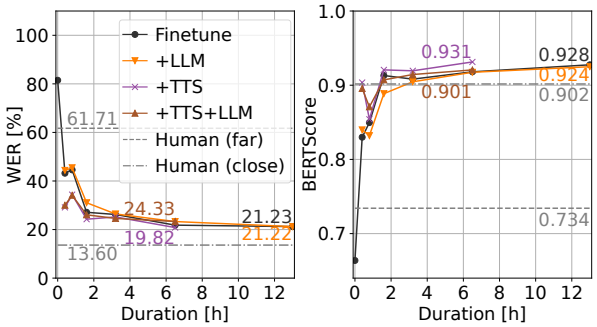

Fig. 5　人間と音声認識モデルの WER (左) と BERTScore (右)．全ての評価セットの結果を平均した値を表示している．

補正を用いることで更に少ないデータでそれが達成されることがわかる．“Human (close)” と比較すると，十分な学習データがあるとき（具体的には，ファインチューニングのみでは 3.2 時間以上，データ拡張／補正があるときは 0.8 時間以上）に ASR は BERTScore において “Human (close)” を上回る．以上の結果より，現在の音声認識性能は話者の家族や友人の文字起こし精度 (WER) には及ばないものの，意味的な精度 (BERTScore) では同程度以上に至るといえる．

3.5　実験 4：意図再現度に関する主観評価

最後に，ASR がコミュニケーションに与える影響を調査した．既存研究 [24] に示されるように，WER が低いことと，音声認識結果が発話の意図を正確に再現していることは必ずしも一致しない．そこで本既存研究に倣い，意図再現度に関する主観評価を実施した．音声認識結果と正解文の対を評価者に提示し，意図再現度を 5 段階で評価させた．評価者への指示文は表 4 のとおりである．日常的なコミュニケーションにおける影響を調査するため，Daily サブセットの評価セットからランダムに 100 発話を選択し評価させた．400 人の日本語母語話者が評価に参加し，各評価者は 50 対を評価した．

Fig. 6 左は，各設定における MOS (mean opinion score) 値である．学習データ量が増加するにつれ MOS 値は単調に増加し，学習データ量が 13.0 時間を超えてもさらなる改善が期待される．また，LLM に基づく推論データ補正は顕著な改善を示し，13.0 時間を超える学習データ量でも同様に改善が期待される．対して，TTS に基づく推論データ補正は，MOS 値の改善をもたらすものの，その改善は飽和傾向にある．

最も優れた設定 (MOS = 4.20) において発話ごとに 5 段階スコアを平均した．そのヒストグラムを Fig. 6 に示す．スコアが 4 を上回る (“ほぼ一致” あるいは “完全一致”) ときにコミュニケーション品質が十分に高いと仮定すると，約 70%の発話がその品質に至っている．一方で残る 30%はその品質が十分でない．本件については，今後の改善が必要である．

Table 4　主観評価時の指示文．音声認識文と正解文のペアに対しいずれか 1 つを回答する．

	指示
1	**完全不一致**：認識文は正解文と大きく異なり、意図を全く反映していない
2	**意図不一致**：単語は似ていても、意図が異なる内容になってしまっている
3	**部分一致**：内容の一部が誤っていて、正しい意図を理解するには推測が必要
4	**ほぼ一致**：一部に単語の誤りや欠落があるが、全体として意図は損なわれていない
5	**完全一致**：認識文は正解文とほぼ同一、あるいはわずかな表現の違いがあっても意味・意図に差はない

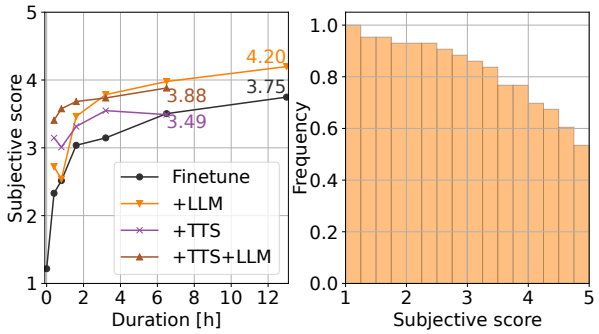

Fig. 6　意図再現度に関する主観評価結果．左は各実験条件における MOS 値．右は，左図において最も優れる 13.0 h+LLM における，発話ごとの MOS 値の逆順累積ヒストグラムである．

4　おわりに

本研究では，日本語単一話者音声コーパスを構築しその性能を音声認識において評価した．音声認識実験では，学習データサイズの影響，人間の聴取能力との比較，意図再現度に関する主観評価結果を報告した．

本コーパスは，プロジェクトページ[6]にて入手可能である．

参考文献

[1] Y. Shen et al., “Taskbench: Benchmarking large language models for task automation,” in *Proc. NeurIPS*, 2025.

[2] G. Comanici et al., “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities,” *arXiv preprint*, 2025.

[3] I.-T. Hsieh, C.-H. Wu, “Dysarthric Speech Recognition Using Curriculum Learning and Multi-stream Architecture,” in *Proc. Interspeech*, 2025, pp. 2210–2214.

[4] S. Pinto et al., “A cross-linguistic perspective to the study of dysarthria in Parkinson ’ s disease,” *Journal of Phonetics*, vol. 64, 2017.

[5] Y. Wan et al., “CDSD: Chinese dysarthria speech database,” in *Proc. Interspeech*, 2024.

[6] R. Sonobe et al., “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” *arXiv preprint*, 2017.

[7] N. Hojo et al., “Revisiting WFST-based Hybrid Japanese Speech Recognition System for Individuals with Organic Speech Disorders,” in *Proc. Interspeech*, 2025.

[8] Y. Nakagome, M. Hentschel, “WCTC-Biasing: Retraining-free Contextual Biasing ASR with Wildcard CTC-based Keyword Spotting and Inter-layer Biasing,” in *Proc. Interspeech*, 2025, pp. 5178–5182.

[9] P. M. Enderby, R. J. Palmer, *Frenchay Dysarthria Assessment: Second Edition (FDA-2)*, Austin, TX, 2008.

[10] W.-Z. Leung et al., “Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis,” in *Proc. Interspeech*, 2024, pp. 2494–2498.

[11] J. R. Deller et al., “The whitaker database of dysarthric (cerebral palsy) speech,” *J. Acoust. Soc. Am.*, vol. 93, 1993.

[12] H. Kim et al., “Dysarthric speech database for universal access research,” in *Proc. Interspeech*, 2008.

[13] F. Rudzicz et al., “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation*, vol. 46, 2012.

[14] R. L. MacDonald et al., “Disordered speech data collection: Lessons learned at 1 million utterances from project euphonia,” in *Proc. Interspeech*, 2021.

[15] C. Zwilling et al., “The speech accessibility project: Best practices for collection and curation of disordered speech,” in *Proc. Interspeech*, 2025.

[16] K. H. Wong et al., “Development of a Cantonese dysarthric speech corpus,” in *Proc. Interspeech*, 2015.

[17] J. Liu et al., “Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis,” *Biomedical Signal Processing and Control*, vol. 79, 2023.

[18] R. Turrisi et al., “EasyCall corpus: A dysarthric speech dataset,” in *Proc. Interspeech*, 2021.

[19] T. Takiguchi, “Speech disorder diversity and its speech communication support technology,” *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 81, 2025.

[20] Y. Hu et al., “Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models,” in *Findings of ACL*, Aug. 2024.

[21] J. Kim et al., “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. ICML*, 2021.

[22] T. Zhang et al., “BERTScore: Evaluating text generation with BERT,” in *Proc. ICLR*, 2020.

[23] J. Li et al., “Synthetic Dysarthric Speech: A Supplement, Not a Substitute for Authentic Data in Dysarthric Speech Recognition,” in *Proc. Interspeech*, 2025, pp. 2755–2759.

[24] B. Phukon et al., “Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches,” in *Proc. Interspeech*, 2025.

[6] https://huggingface.co/datasets/Asahi-Ogasawara/SS-JDSC