J-CHAT: 音声対話言語モデルのための大規模日本語対話音声コーパス *

◎中田 亘 *，関 健太郎 *，谷中 瞳，齋藤 佑樹（東大），
高道 慎之介（慶大／東大），猿渡 洋（東大）

1 はじめに

音声対話システムは，人–AI のコミュニケーションの手段として大きな注目を集めている [1]．音声には，トーンやイントネーションに加え，笑い声や叫び声などの非言語発話が含まれる．これらの表現を音声対話システムで実現することにより，豊かな人–AI コミュニケーションが実現可能である．これらの表現を実現するために，音声の言語的構造をモデル化する Spoken Language Model (SLM) は重要な役割を持つ．代表的な SLM の例として，HuBERT [2] 由来の特徴量を離散化することで音声を離散トークン列として扱い，書き言葉と同様の言語モデルを学習する Generative SLM [3] が挙げられる．Generative Pre-trained Transformer (GPT) [4] による自己回帰的なテキスト生成と同様に，GSLM は入力音声の続きを生成する能力を持つ．

SLM の学習には，大規模な音声コーパスが必要となる．これは，一般に SLM の性能はモデルサイズに対して単調増加であり [5]，モデルサイズの増加に伴いより大きな音声コーパスが必要となるからである．中でもオープンソースなコーパスは研究の参入障壁を下げ，再現性のある研究を可能とすることから重要である．しかしながら，オープンソースの対話音声コーパスの構築は限定的であり，そもそも大規模なオープンソースコーパスの構築法は確立されていない．対話音声コーパス構築の代表的な方法としては電話の録音 [6] やスタジオ収録 [7] がある．これらの方法は人的資源や金銭的資源が大量に必要となるためスケール性に乏しく，大規模な対話音声コーパス構築方法として適さない．加えて，既存の大規模な対話音声コーパスは主に英語で構築されており，他の言語での研究が遅れる要因となっている．

本研究では，インターネットから自動的に対話音声を収集することで，大規模な対話音声コーパスを構築する方法を提案する．また，実際に提案手法によって日本語対話音声コーパス J-CHAT を構築し，公開する．これにより，in-the-wild な対話音声データが収集できるため，自発性に富み多様な話題を網羅した音声対話を収集することが可能となる．また，提案手法では音源分離モデルを用いて BGM 等を除去することで，ノイズの含まれないクリーンな音声コーパスを提供する．本研究の貢献は以下のとおりである．

- 大規模かつオープンソースな音声コーパスの構築，公開．
- 自発性が高く，クリーンな音声コーパスを提供することによる対話音声合成研究の促進．
- 言語に依存しないコーパス構築方法の開発．

2 関連研究

2.1 音声言語モデル（SLM）

大規模な SLM は，様々な音声情報処理タスクに対して有効であると報告されている．しかしながら，SLM の学習には大規模な音声コーパスが必要となる．例えば Google の SoundStorm [8] では，対話音声合成のために 10 万時間の対話音声を使用している．さらに，音声認識では 68 万時間の音声データが使用されており [9]，音声翻訳では 450 万時間 [10] の音声データが使用されている．これらの研究により，学習データ量の増加がモデル性能の向上に寄与することが報告されており，大規模な音声コーパスの構築の重要性を示唆している．

しかしながら，これらの研究ではモデル構造の改善が主眼であり，実験に使用されたデータセットは公開されておらず，その構築方法も大部分が不明である．このようなデータセットに関する情報の不足は，SLM において大きな問題となっている [11]．本研究は特に対話ドメインの SLM に焦点を当て大規模な対話音声コーパスを構築する手法を提案し SLM 構築に必要となる大規模音声コーパス構築に関する議論の促進を目指す．

2.2 既存の音声コーパス

対話音声モデリングのための SLM を構築するためには，自発性が高くノイズの少ない大規模な対話データが必要となる．Table 1 に本研究に関連する音声コーパスを示す．STUDIES コーパス [7] や DailyTalk コーパス [12] は，オープンソースなコーパスであり，複数ターンからなる対話で構成されている．しかしながら，これらのコーパスはスタジオ収録により構築されており，金銭的制約からスケールアップは困難である．実際，その時間数は少ない．Fisher コーパス [6] は時間数において前述のコーパスに勝るがオープンソースでない．GigaSpeech コーパス [13] は，大規模かつオープンソースなコーパスであるが，音声認識のために設計されており発話レベルの音声しか使用できないため，対話音声モデリングに使用できない．またデータの収集先にオーディオブックを含むため，自発性という観点において好ましくない．

---

* Japanese Large Scale Dialogue Corpus for Human-AI Talks, by Wataru Nakata, Kentaro Seki, Hitomi Yanaka, Yuki Saito (The University of Tokyo), Shinnosuke Takamichi (Keio University / The University of Tokyo), and Hiroshi Saruwatari (The University of Tokyo). * indicates equal contribution. * で示された著者の貢献度は同じ

Table 1: Comparison of speech corpora related to this study, sorted by size.

Corpus Name	Size (hours)	Open-source	dialogue	Spontaneous	Clean speech
STUDIES [7]	8	✅	✅		✅
DailyTalk [12]	20	✅	✅	✅	✅
Fisher [6]	2k		✅	✅	✅
GigaSpeech [13]	33k	✅			
J-CHAT (This work)	69k	✅	✅	✅	✅

3 コーパス構築方法

Fig. 1 にコーパス構築の流れを示す．所望の言語の対話音声コーパス（本研究では日本語）を構築するためにインターネットから音声を収集し，言語，対話，音響的な観点からフィルタリングを実施した．

3.1 データ収集

先行研究 [14] では多様な音声が YouTube から収集できることが示されているため，YouTube をデータ収集先として採用した．YouTube からの収集では，Wikipedia 上のページ題目から作成したキーワードを用いてランダムな検索を行った [15]．これにより，ファイル数にして 60 万，時間数として 18 万時間となる音データを収集した．

YouTube から収集した音データには，対話音声に加えて独話や音楽データが多く含まれていたため，十分な量の対話データを確保することが難しかった．そのため本研究では，YouTube に加えて Podcast からのデータ収集を行なった．Podcast は，音声が主たるメディアのため効率の良いデータ収集が可能である．加えて，Podcast の URL リストを提供するサイトである PodcastIndex[1]では，Podcast の内容の言語などのメタデータが配布されているため，これを利用した．まず，PodcastIndex から，日本語とされている Podcast の RSS フィードを収集した．その後，収集した RSS フィードに含まれている音ファイルの URL を用いてダウンロードを行った．これによりファイル数にして 88 万ファイル，時間数にして 14 万時間の音データを収集した．

3.2 データ選択

3.2.1 日本語音声データの抽出

日本語でないデータを除去するために，Whisper [9] の言語識別機能を利用した．Whisper から得られる日本語である確度 p が 0.8 より大きい場合を日本語とし，そうでないデータを除去した．これにより，YouTube から集めた音データの 55.7%，Podcast から集めた音データの 84.7%を抽出し日本語音声データとした．

3.2.2 対話音声データの抽出

抽出した日本語音声データから，対話データを抽出した．これには，音データから対話区間を検出することが必要となる．本研究では話者ダイアライゼーション (Speaker Diarization; SD) を用いてこれを行った．

[1] https://podcastindex.org

Table 2: Corpus statistics by its subsets, YouTube and Podcast. # means “number of”.

feature	YouTube	Podcast	Total
total duration[hr]	11,001	57,891	68,892
# dialogue	1,013,488	3,924,009	4,937,497
mean duration [s]	39.07	53.11	50.23
mean # turns	7.58	10.68	10.10
mean # speakers	3.23	3.12	3.14

SD では，誰が，いつ喋っているかの情報が得られる．本研究では SD モデルとして PyAnnote [16] を使用した．得られた音声区間（ターン）を元に，5 秒以上発話がない区間を区切りとして音声を切り分け対話とした．その後独話を排除するために，対話の中で80%以上を同じ話者が発話しているものを除外した．これにより，2 人以上の話者が対話しているデータのみを抽出した．抽出されたデータ量は YouTube において 41.9%，Podcast において 45.0%である．

3.3 データクレンジング

Podcast には多くの場合 BGM が含まれているが，これは音声生成タスクではノイズとなるため除去する必要がある．本研究では Section 3.2 で抽出した対話データ全てに対して音源分離モデル Demucs [17] を使用することで，ノイズ除去を行なった．

最終的に得られるコーパスは，Demucs を適用した対話音声と，SD によって付与された音声区間情報とのペアデータである．

4 コーパス分析

4.1 データセットの規模

Table 2 に構築したコーパスである J-CHAT の統計量を示す．J-CHAT は 6.9 万時間の日本語音声から成る．その内 YouTube から収集した音声が 1.1 万時間，Podcast が 5.8 万時間となっている．

Podcast と YouTube の違いとしては，Podcast において対話の平均時間が YouTube のおよそ 1.4 倍であることである．これにより，Podcast では一対話あたりのターン数が多くなっている．一方で，対話あたりの話者の数は Podcast，YouTube の間で同様な値となった．

4.2 音韻的多様性

J-CHAT の音韻的多様性を確認するために，音声埋め込みモデルであり，音韻的な特徴を表す Hu-

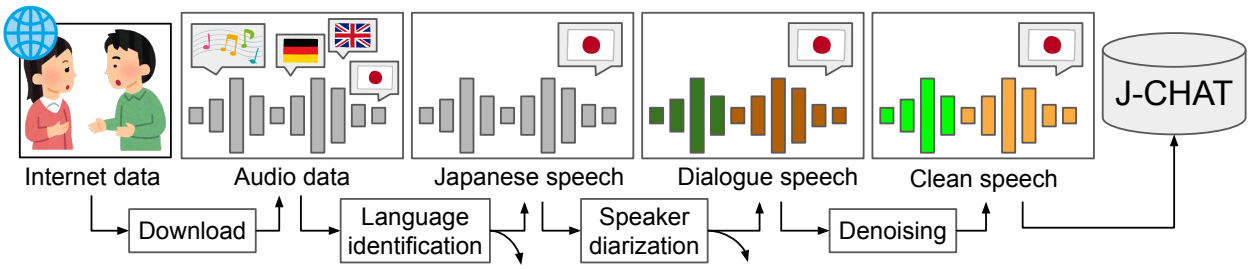

Fig. 1: Corpus construction methodology proposed in this work.

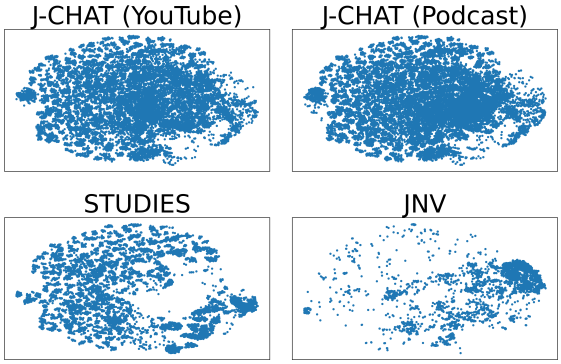

Fig. 2: Distribution of HuBERT [2] features extracted from J-CHAT (ours), STUDIES (simulated dialogue), and JNV (non-verbal expression).

BERT [2] から得られる特徴量の分布を Fig. 2 に示す．HuBERT 特徴量の抽出に際しては，比較したコーパスである STUDIES [7]，JNV [18]，および J-CHAT の YouTube から収集したサンプル，Podcast から収集したサンプルからそれぞれ 1000 サンプルをランダムに選択し，各発話のランダムな区間 5 秒から HuBERT 特徴量を抽出した．その後，t-SNE [19] を用いて次元圧縮を行い 2 次元のプロットとした．Figure 2 から，読み上げ音声から成る STUDIES や非言語発話から成る JNV では，分布が偏っている一方で，J-CHAT ではより広い分布を持つことがわかる．このことから，J-CHAT の音韻的多様性が示された．

5 実験

J-CHAT を用いて対話用の SLM が学習可能であることを確認するために，先行研究で提案されている対話 SLM である dGSLM [11] を学習・評価した．dGSLM は GSLM [3] で提案された枠組みを用いており，音声を離散化しその離散化音声 (unit) で Transformer [20] 言語モデル（Dialogue Language Model; DLM）を学習する．最後に，ボコーダを用いて離散化音声から音声波形を生成する．これにより，対話音声の生成が可能となる．

音声の離散化には，音声由来の HuBERT 特徴量を k 平均法を用いてクラスタリングしたものを使用する．以後，これを speech-to-unit とする．

実験では 4 つのモデルを比較した．

- resynth: speech-to-unit から得られた unit からボコーダを用いて再合成した音声
- dGSLM-YouTube: J-CHAT のうち YouTube から収集したサンプルで学習した dGSLM
- dGSLM-Podcast: J-CHAT のうち Podcast から収集したサンプルで学習した dGSLM
- dGSLM-J-CHAT: J-CHAT 全体で学習した dGSLM

5.1 実験条件

まず，J-CHAT 全体を対話のターン切り替わりにおいて出力チャネル変えることにより 2 チャネルへと分割し，dGSLM の学習に適用できるようにした．その後，J-CHAT を train/dev/test へとランダムに分割した．YouTube から収集したサンプルに関しては，train/dev/test が，それぞれ 1,003,397/9,991/100 対話であり，Podcast から収集したサンプルに関してはそれぞれ 3,885,007/38,902/100 サンプルである．

speech-to-unit には，HuBERT モデルには，日本語事前学習済みの HuBERT [21] を使用した．クラスタ数は 1000 とした．

DLM のモデル設定は先行研究 [11] に倣った．パラメータ数は，23,638,529 である．DLM の学習には，32 枚の NVIDIA V100 GPU を用いて 48 時間行なった．学習ステップ数は 10 万ステップであり，学習率は 2×10^{-4} である．

離散化された音声から音声を生成するボコーダには，XVector [22] から得られる話者情報で条件付けした HiFi-GAN を使用した [23]．なお，ボコーダの学習は，JVS [24] および JVNV [25] コーパスを用いて行なった．それぞれ，多話者日本語読み上げ音声，多話者日本語非言語発話コーパスである．ボコーダの学習には，4 枚の NVIDIA V100 GPU を使用して 48 時間行なった．

5.2 評価

評価には先行研究同様 [11]，自然性，有意味性の観点における Mean Opinion Score (MOS) による主観評価を使用した．それぞれの主観評価において評価者数は 60 人であり，各評価者は 8 対話を評価した．

評価用の音声の生成にあたっては最初の 5 秒に test セットを入力し，その後に続く 25 秒の対話を予測した．サンプリング手法にはビームサイズ 5 のビーム

Table 3: Result for MOS tests with their 95% confidence intervals.

Model	Naturalness	Meaningfulness
resynth	2.55 ± 0.18	2.48 ± 0.18
dGSLM-YouTube	1.44 ± 0.13	1.56 ± 0.14
dGSLM-Podcast	1.44 ± 0.13	1.52 ± 0.13
dGSLM-J-CHAT	2.28 ± 0.19	2.18 ± 0.19

サーチを利用した．また，ボコーダの推論時に条件付けする話者には，JVS コーパス [24] の JVS001（男性）話者を第一チャネルに使用し JVS002（女性）話者を第二チャネルに使用した．

5.3 結果

結果を Table 3 に示す．これらの結果から，dGSLM で生成したサンプルのうち，dGSLM-J-CHAT が有意味性，自然性において最も良いことが確認された．これは，J-CHAT の対話 SLM 構築における有用性を示している．また，dGSLM-YouTube と dGSLM-Podcast の間で学習データ量が大きく異なるにも関わらず有意な差は確認されなかった．これは，データ量を単純に大規模化するだけでは dGSLM の性能は上がらないということを示唆している．また，dGSLM-J-CHAT と resynth の間には有意な差が有意味性，自然性の双方において確認された．事実生成されたサンプルは時折理解可能な単語を生成するものの，一貫性にはかけている．

6 まとめ

本研究では，大規模な対話音声コーパスである J-CHAT の構築方法を示した．実験結果から，YouTube や Podcast の複数ドメインからデータを収集することが対話 SLM の学習に有益であることを示している．

今後の課題として，他の言語におけるコーパスの構築や，対話の一貫性が担保された対話モデリングの方法が挙げられる．

謝辞本研究の一部は，産総研覚醒プロジェクト及び 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] K. Mitsui et al., “Towards human-like spoken dialogue generation between ai agents from written dialogue,” in *arXiv*, 2023.

[2] W.-N. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] K. Lakhotia et al., “On generative spoken language modeling from raw audio,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[4] A. Radford et al., “Improving language understanding by generative pre-training,” 2018.

[5] S. wen Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 1194–1198.

[6] C. Cieri et al., “The Fisher Corpus: a resource for the next generations of speech-to-text,” in *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, M. T. Lino et al., Eds., May 2004.

[7] Y. Saito et al., “STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent,” in *Proc. Interspeech 2022*, 2022, pp. 5155–5159.

[8] Z. Borsos et al., “Soundstorm: Efficient parallel audio generation,” *arXiv preprint arXiv:2305.09636*, 2023.

[9] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[10] L. Barrault et al., “Seamless: Multilingual expressive and streaming speech translation,” *arXiv preprint arXiv:2312.05187*, 2023.

[11] T. A. Nguyen et al., “Generative spoken dialogue language modeling,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[12] K. Lee et al., “DailyTalk: Spoken dialogue dataset for conversational text-to-speech,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] G. Chen et al., “GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio,” in *Proc. Interspeech 2021*, 2021, pp. 3670–3674.

[14] K. Seki et al., “Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[15] S. Takamichi et al., “JTubeSpeech: corpus of Japanese speech collected from youtube for speech recognition and speaker verification,” in *arXiv*, 2021.

[16] A. Plaquet, H. Bredin, “Powerset multi-class cross entropy loss for neural speaker diarization,” in *Proc. Interspeech 2023*, 2023, pp. 3222–3226.

[17] S. Rouard et al., “Hybrid transformers for music source separation,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[18] D. Xin et al., “JNV corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions,” *Speech Communication*, vol. 156, p. 103004, 2024.

[19] L. Van der Maaten, G. Hinton, “Visualizing data using t-SNE,” *Journal of machine learning research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[20] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30. Curran Associates, Inc., 2017.

[21] K. Sawada et al., “Release of pre-trained models for the Japanese language,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari et al., Eds., May 2024, pp. 13 898–13 905.

[22] D. Snyder et al., “X-Vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[23] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[24] S. Takamichi et al., “JSUT and JVS: Free japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[25] D. Xin et al., “JVNV: A corpus of japanese emotional speech with verbal content and nonverbal expressions,” *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.