Moshi に基づく音声対話モデルの日本語ファインチューニングにおける対話データ特性の影響 *

☆阿部 雄斗 (早大/NII LLMC)　佐伯 真於 (早大/エキュメノポリス)
大橋 厚元 (名大)　高道 慎之介 (慶大)　藤江 真也 (千葉工大/早大)
小林 哲則 (早大)　小川 哲司 (早大)　東中 竜一郎 (名大/NII LLMC)

1　はじめに

大規模な日本語雑談対話音声を用いて全二重 (full-duplex) 音声対話モデルを構築し，ファインチューニングおよび推論に用いる対話データの特性がモデル性能に与える影響を調査した．

人間同士の会話では，話し手と聞き手が同時に発話したり，相槌で相互行為を調整したりする．これを実現する全二重音声対話の枠組みが注目されており，Moshi [1] や J-Moshi [2] に代表される end-to-end 型モデルは，シームレスな対話生成を可能とする一方，学習データに含まれる対話構造・発話様式の影響を強く受ける．特に日本語では，相槌頻度，オーバーラップ発話，沈黙長などのターンテイキング特性が英語とは異なることが指摘されている [3]．この背景から，LLM-jp 対話ワーキンググループでは，商用利用可能な日本語音声対話コーパスの整備と，日本語全二重音声対話モデルの構築を進めている．

本稿では，Moshi に基づくモデルを基盤として，学習に用いる対話データの特性がモデル性能に与える影響の予備的評価を行う．具体的には，ファインチューニングに用いるデータが異なる複数のモデルに対し，雑談対話音声を用いた対話継続タスク (人間による音声を短時間与えた後にモデル応答を生成) で比較し，音質客観評価（NISQA [5]），言語的客観評価（LLM-as-a-Judge；LLM-AJ [6]），および主観評価により分析する．雑談対話音声としては，LLM-jp 対話ワーキンググループが構築した大規模日本語雑談対話コーパス LLM-jp-Zoom1 (後述) と，日本語 CallHome [4] を用いる．

本稿の構成は以下の通りである．まず 2 章で用いた音声コーパスについて述べ，3 章で対話継続実験の設定と結果を示す．最後に 4 章でまとめと今後の課題を述べる．

2　音声対話コーパス

2.1　LLM-jp-Zoom1

LLM-jp-Zoom1 は，日本語音声言語モデルの学習を主目的として構築した，総計約 1000 時間の大規模二者対話コーパスである．日本語母語話者による二名ペアの対話を Zoom 上で実施し，静音環境での収録を前提に，話者ごとに分離した音声トラックとして録音した．各セッションの対話時間は 30 分で，NII 提供の話題リストに基づく自由度の高い日常会話を中心に設計し，全体の約半数はカジュアルな口調（友人同士の会話に相当）で構成した．コーパスは計 2000 対話から成り，同一話者ペアは最大 10 回まで参加可能とした（話題は各回で変更，口調は固定）．音声に加えて，参加者属性・性格特性アンケート，対話ごとの主観評価アンケート，終了後アンケートなどのメタデータを付与している．これら一式を，LLM-jp により日本語音声言語モデルの研究・開発に資する公開データセットとして提供する予定である．

2.2　LLM-jp-Zoom1 の分析

客観的音声品質評価として NISQA を用い，MOS 値として 2.96 を得た．この値は，タスク指向対話コーパス Tabidachi（旅行代理店対話，2.98） [7] と同程度であり，自由度の高い日常対話でありながら一定の音声品質が確保されていることを示す．

また，LLM-jp-Zoom1，Tabidachi，日本語 CallHome のターンテイキング特性 (IPU，Pause，Overlap，Gap) [8] を算出し，表 1 に示す．IPU は Silero VAD[1] を用いて推定した．Tabidachi と比較すると，LLM-jp-Zoom1 では Overlap の累積時間が長く，異なる話者間の IPU 間における無音時間（Gap）が短い傾向が観測された．これは，相槌や短尺応答が高頻度に出現する，高密度なターンテイキング構造を有していることを示している．さらに，話者間の発話量の対称性が日本語 CallHome と類似していることから，本データセットは，日本語日常会話に典型的な協調的かつ雑談的な対話特性を反映していると考えられる．

3　対話継続実験

本章では，構築した音声対話モデルの対話継続性能を評価した．各試行で，10 秒の実音声（人間話者同士）を入力とし，その直後にモデルに続きの 20 秒間

---

* Effects of Dialogue Data Characteristics on Fine-Tuning of Moshi-Based Japanese Spoken Dialogue Models. by ABE, Yuto (Waseda Univ./NII LLMC), SAEKI, Mao (Waseda Univ./Equmenopolis), OHASHI, Atsumoto (Nagoya Univ.), TAKAMICHI, Shinnosuke (Keio Univ.), FUJIE, Shinya (Chiba Tech.), KOBAYASHI, Tetsunori, OGAWA, Tetsuji (Waseda Univ.), and HIGASHINAKA, Ryuichiro (Nagoya Univ./NII LLMC)

[1] https://github.com/snakers4/silero-vad

Table 1　LLM-jp-Zoom1，Tabidachi，日本語 CallHome 各コーパスにおけるターンテイキング特性.

	Number of occurrences / 20s				Cumulative duration / 20s			
	IPU	Pause	Gap	Overlap	IPU	Pause	Gap	Overlap
LLM-jp-Zoom1	A:5.58 B:4.92	A:1.18 B:0.48	3.68	5.54	A:11.64 B:10.14	A:0.76 B:0.25	5.23	3.89
Tabidachi	A:5.06 B:2.12	A:2.88 B:0.48	3.18	1.16	A:10.73 B:2.91	A:3.11 B:0.32	7.74	0.56
日本語 CallHome	A:5.48 B:5.44	A:2.02 B:1.72	6.08	2.46	A:7.56 B:8.43	A:1.59 B:1.17	9.32	0.87

の対話音声を生成させた．生成音声について，音質・自然性・意味性の観点から客観評価および主観評価を行った.

3.1　実験条件

以下の 3 種類のモデルを用いて，同一手順で評価した．このうち 1 種類は既存の J-Moshi モデルであり，残る 2 種類を新たに構築した．いずれのモデルも，事前学習には共通して J-CHAT [9] を用いた.

- **J-Moshi:** 既存の日本語全二重型音声対話モデル `nu-dialogue/j-moshi`[2] をそのまま用いる．本研究における追加学習・改変は行わない．同モデルは Tabidachi (旅行代理店コーパス，タスク指向でオペレータ役が一方的に喋る傾向がある)，名古屋大学で構築された雑談・相談対話コーパス，日本語 CallHome，日本語自然発話コーパス (CSJ) でファインチューニングされている.
- **LLM-jp-Moshi:**　Moshi を初期値とし，J-CHAT により事前学習した後，LLM-jp-Zoom1 のみでファインチューニングしたモデル.
- **LLM-jp-Moshi+:**　Moshi を初期値とし，J-CHAT により事前学習した後，LLM-jp-Zoom1 に加えて，京都観光案内対話データベース (KTD，53 時間の旅行相談対話) および VisualBank[3](300 時間のコールセンター，インタビュー，対面接客・営業の対話) でファインチューニングしたモデル．両データはタスク指向対話が中心であり，LLM-jp-Zoom1 の雑談寄りの特性とは性質が異なる.

学習手順　Moshi の公開済み事前学習チェックポイント `kyutai/moshiko-pytorch-bf16`[4] を初期化に用い，名古屋大学から公開されているファインチューニングスクリプト[5]で学習した．最適化は AdamW，J-CHAT による事前学習は 1 エポック (バッチサイズ 512)，各ファインチューニングは 7 エポック (バッチサイズ 16) とした (他のハイパーパラメータは J-Moshi [2] に準拠).

推論設定　各モデルに対し，学習データに含まれていないテストデータとして用意した LLM-jp-Zoom1 もしくは日本語 CallHome から選択した 10 秒の実音声入力の直後から 20 秒間の応答音声を自動生成した.

評価プロトコル　音質客観評価として NISQA による推定 MOS を算出し，言語面の客観指標として LLM-AJ[6] を用いた．LLM-AJ では，ASR[7] により得た書き起こしに対して，一貫性・自然さ・関連性・指示遵守・ターンテイキング・総合評価の 6 つの観点を 1–10 点 で評価するプロンプトを用いた．主観評価はクラウドソーシング[8]で実施し，自然性 (人間のような自然な対話に聞こえるか) および意味性 (音声の内容を理解可能か) を 5 段階尺度で評定した．各データセットからランダムに抽出した 50 サンプルを用い，被験者 50 名がモデルの匿名化・提示順のランダム化のもとで評価した.

3.2　客観評価実験結果

表 2 および 表 3 に，LLM-jp-Zoom1 評価セットと日本語 CallHome 評価セットに対する客観評価 (NISQA による推定 MOS，および LLM-AJ 各指標) の結果を示す.

両表より，評価データセットにより最良モデルが異なる傾向が確認された．すなわち，

- **LLM-jp-Zoom1** に対しては，LLM-jp-Moshi が全指標で最良の性能を示した.
- **日本語 CallHome** に対しては，NISQA を除き LLM-jp-Moshi+ が最良となった.

この差異は，推論データとファインチューニングデータの対話形式の近さが LLM-AJ 指標 (一貫性・自然さ・関連性・指示遵守・ターンテイキング・総合評価) に寄与するという既報の知見 [10] と整合している．すなわち，

[2] https://huggingface.co/nu-dialogue/j-moshi
[3] https://qleandataset.visual-bank.co.jp/
[4] https://huggingface.co/kyutai/moshiko-pytorch-bf16
[5] https://github.com/nu-dialogue/moshi-finetune
[6] https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4
[7] https://huggingface.co/reazon-research/reazonspeech-espnet-v2
[8] https://crowdworks.jp/

- **LLM-jp-Moshi** は LLM-jp-Zoom1 のみでファインチューニングされており，同一スタイルの評価セット (LLM-jp-Zoom1) で優位となった．
- **LLM-jp-Moshi+** は LLM-jp-Zoom1 に加え，観光案内 (KTD) やコールセンタ (VisualBank) といった性質の異なる対話も併用してファインチューニングされているため，学習データに含まれない日本語 CallHome に対して LLM-AJ において頑健に高い性能が得られたと解釈できる．

3.3 主観評価実験結果

表 4 および表 5 に，LLM-jp-Zoom1 評価セットと日本語 CallHome 評価セットに対する主観評価 (自然性・意味性) の結果を示す．

両表より，評価セット・評価指標に依らず LLM-jp-Moshi+ が最良となり，LLM-jp-Moshi もベースライン (J-Moshi) を一貫して上回る傾向が確認された．すなわち，雑談寄りの対話 (LLM-jp-Zoom1) に加え，タスク指向 対話 (KTD，VisualBank) を併用してファインチューニングした多様性の高いモデルほど，知覚的な対話品質 (自然性・意味性) が向上することを示唆している．

一方で，主観評価と客観評価は異なる傾向を示した．たとえば，LLM-AJ の自然性指標 (NAT) でさえ，主観評価の自然性と逆転・乖離するケースが観測された．この結果は，日本語音声対話モデルに対する客観評価尺度の精緻化 (例えば，音声認識誤りの影響低減，韻律・相槌・オーバーラップ発話など日本語会話特性の反映) が今後の重要課題であることを示している．

4 まとめ

本研究では，Moshi アーキテクチャに基づく音声対話モデルにおいて，日本語音声対話データの特性がファインチューニングに与える影響を，客観評価と主観評価の両面から分析した．

構築した LLM-jp-Moshi および LLM-jp-Moshi+ は，全ての評価においてベースラインの J-Moshi を上回る性能を示した．また，実験結果から，評価データセットや評価手法の違いにより，同じモデルでも評価結果が大きく異なることが確認された．客観評価では評価データセットにより最良モデルが異なり，主観評価では多様な対話データを併用したモデルが優位となるなど，評価手法によって傾向が異なることが観測された．

これらの結果より，音声対話モデルの評価や比較には，複数の評価データセットと複数の評価手法を用いた包括的な評価が必要であり，目的に応じた学習データ設計が重要であることが示唆される．また，日本語音声対話モデルに対する客観評価尺度の精緻化が今後の課題である．

謝辞 産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した．

参考文献

[1] A. Défossez *et al.*, “Moshi: a speech-text foundation model for real-time dialogue,” arXiv preprint arXiv:2410.00037, 2024.
[2] A. Ohashi *et al.*, “Towards a Japanese full-duplex spoken dialogue system,” in *Proc. Interspeech*, 2025.
[3] N. Ward and W. Tsukahara, “Prosodic features which cue back-channel responses in English and Japanese,” *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.
[4] A. Canavan and G. Zipperlen, “CALLHOME Japanese Speech LDC96S37,” Philadelphia: Linguistic Data Consortium, 1996.
[5] G. Mittag *et al.*, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. Interspeech*, 2021.
[6] L. Zheng *et al.*, “Judging LLM-as-a-judge with MT-bench and CHATbot arena,” in *Proc. NeurIPS*, vol. 36, pp. 46595–46623, 2023.
[7] M. Inaba *et al.*, “Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 9, Article 130, 2024.
[8] T. A. Nguyen *et al.*, “Generative spoken dialogue language modeling,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.
[9] W. Nakata *et al.*, “J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling,” arXiv preprint arXiv:2407.15828, 2024.
[10] Y. Abe *et al.*, “Effects of dialogue corpora properties on fine-tuning a Moshi-based spoken dialogue model,” in *Proc. IWSDS*, 2026.

Table 2　学習データ条件および **LLM-jp-Zoom1** の継続対話音声に対する客観評価の結果. 音声品質は NISQA を用いて計測し，意味性は以下の指標で評価した：LLM-as-a-Judge による評価: COH = coherence(一貫性), NAT = naturalness(自然さ), REL = relevance(関連性)，INS = instruction following(指示遵守), TUR = turn taking(ターンテイキング), OVE = overall(総合評価).

		J-Moshi	**LLM-jp-Moshi**	**LLM-jp-Moshi+**	**llmjp-zoom1 (実音声)**
Pre-training		J-CHAT (69k hours)	J-CHAT (69k hours)	J-CHAT (69k hours)	-
Fine-tuning		Tabidachi 内製データ (200 hours) 日本語 CallHome, CSJ	llmjp-zoom1	llmjp-zoom1 京都観光案内対話データベース VisualBank	-
NISQA (MOS) (1–5)		3.35	**3.72**	3.69	2.96
LLM-AJ (1–10)	COH	4.12	**4.24**	3.59	6.92
	NAT	5.33	**5.53**	4.90	7.73
	REL	3.18	**3.47**	3.02	5.82
	INS	1.84	**2.02**	1.73	4.16
	TUR	4.31	**4.67**	4.12	6.80
	OVE	4.07	**4.37**	3.69	6.09

Table 3　学習データ条件および **日本語 CallHome** の継続対話音声に対する客観評価の結果.

		J-Moshi	**LLM-jp-Moshi**	**LLM-jp-Moshi+**	**日本語 CallHome (実音声)**
Pre-training		J-CHAT (69k hours)	J-CHAT (69k hours)	J-CHAT (69k hours)	-
Fine-tuning		Tabidachi 内製データ (200 hours) 日本語 CallHome, CSJ	LLM-jp-Zoom1	LLM-jp-Zoom1 京都観光案内対話データベース VisualBank	-
NISQA (MOS) (1–5)		2.41	**3.03**	2.91	2.47
LLM-AJ (1–10)	COH	3.14	3.10	**3.60**	5.56
	NAT	4.18	4.37	**4.98**	6.67
	REL	2.39	2.41	**2.88**	4.69
	INS	1.18	1.22	**1.76**	3.25
	TUR	3.37	3.55	**4.08**	5.65
	OVE	3.08	3.16	**3.62**	5.45

Table 4　**LLM-jp-Zoom1** の継続対話音声に対する 5 段階の主観評価

	J-Moshi	**LLM-jp-Moshi**	**LLM-jp-Moshi+**	**LLM-jp-Zoom1 (実音声)**
Pre-training	J-CHAT (69k hours)	J-CHAT (69k hours)	J-CHAT (69k hours)	-
Fine-tuning	Tabidachi 内製データ (200 hours) 日本語 CallHome, CSJ	LLM-jp-Zoom1	LLM-jp-Zoom1 京都観光案内対話データベース VisualBank	-
自然性 (1–5)	2.36	3.02	**3.19**	4.31
意味性 (1–5)	2.14	2.61	**2.80**	4.46

Table 5　**日本語 CallHome** の継続対話音声に対する 5 段階の主観評価

	J-Moshi	**LLM-jp-Moshi**	**LLM-jp-Moshi+**	**日本語 CallHome (実音声)**
Pre-training	J-CHAT (69k hours)	J-CHAT (69k hours)	J-CHAT (69k hours)	-
Fine-tuning	Tabidachi 内製データ (200 hours) 日本語 CallHome, CSJ	LLM-jp-Zoom1	LLM-jp-Zoom1 京都観光案内対話データベース VisualBank	-
自然性 (1–5)	2.35	2.82	**3.12**	4.03
意味性 (1–5)	1.54	2.34	**2.56**	4.10