Effects of Dialogue Corpora Properties on Fine-Tuning a Moshi-Based Spoken Dialogue Model

Yuto Abe[1,2], Mao Saeki[1,3], Atsumoto Ohashi[4], Shinnosuke Takamichi[5], Shiyna Fujie[6], Tetsunori Kobayashi[1], Tetsuji Ogawa[1], Ryuichiro Higashinaka[4,2]

[1]Waseda University, [2]NII LLMC, [3]Equmenopolis, Inc., [4]Nagoya University, [5]Keio University, [6]Chiba Institute of Technology

Correspondence: abe@pcl.cs.waseda.ac.jp

Abstract

We study how the turn-taking properties of spoken dialogue corpora shape the learning and behavior of full-duplex speech dialogue models. Beyond acoustic and linguistic quality, effective systems must reproduce task-dependent dynamics such as conversational tempo and turn-taking. We analyze multiple Japanese dialogue corpora using *i)* NISQA for speech quality, *ii)* LLM-as-a-Judge for linguistic/semantic appropriateness, and *iii)* four timing indicators, inter-pausal units, pause, gap, and overlap, to quantify interactional style. A curriculum strategy then fine-tunes a Moshi-based full-duplex model by incrementally combining corpora with distinct turn-taking profiles. On a dialogue-continuation task, corpus-specific turn-taking patterns reliably shaped model behavior: chat-style corpora yielded more natural rhythms with moderate overlaps and gaps, whereas consultation-style corpora promoted slower, deliberate timing. Fine-tuning on high-quality audio improved perceptual naturalness, while mixing task-mismatched data reduced linguistic coherence.

1 Introduction

Full-duplex spoken dialogue models, such as Moshi (Défossez et al., 2024), J-Moshi (Ohashi et al., 2025), FreezeOmni (Wang et al., 2024), and SyncLLM (Veluri et al., 2024), have recently shown that temporal transformer architectures can listen and speak simultaneously, enabling low-latency turn exchange and more natural interaction (Ma et al., 2025). While scaling speech data improves acoustic and linguistic quality, an effective system must also reproduce task-dependent interactional dynamics—notably conversational tempo and turn-taking patterns.

A long line of work analyzes human turn-taking via temporal cues such as pauses, gaps, and overlaps, connecting these statistics to conversational rhythm and speaker coordination (Nguyen et al., 2023). Informal chats between close friends tend to be fast and overlap-rich; consultation dialogues are typically slower and more one-sided (Yamaguchi et al., 2016). However, most prior analyses rely on small or text-derived datasets, and systematic, corpus-level comparisons across large spoken resources are limited.

In parallel, curriculum learning has been used to stabilize and improve large speech models by moving from noisy, large-scale data to smaller, higher-quality corpora (Wen et al., 2025; Rouditchenko et al., 2025). Yet, curricula are rarely designed with explicit attention to corpus turn-taking profiles (e.g., overlap frequency, gap duration). Because such properties plausibly shape perceived naturalness and interactivity, understanding their role is essential for controllable dialogue behavior.

We study how turn-taking characteristics of multiple Japanese spoken dialogue corpora affect a Moshi-based full-duplex dialogue model. We quantify four timing features, inter-pausal units (IPU), pause, gap, and overlap, and relate them to the naturalness, coherence, and turn-taking behavior of generated dialogues under a curriculum that combines corpora with distinct interaction styles.

Our contributions are twofold:

1. **Corpus-level analysis**: We provide quantitative, interpretable profiles of Japanese dialogue corpora using IPU, pause, gap, and overlap, revealing clear differences in conversational tempo and speaker asymmetry.
2. **Curriculum linked to turn-taking**: We show that curricula which respect corpus turn-taking properties yield models with more controllable dialogue rhythm: chat-style corpora encourage moderate overlaps and responsive timing, whereas formal/consultation corpora promote deliberate, stable turns.

The remainder of this paper is organized as follows. Section 2 describes the corpora, measures,

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 104–108
February 26–March 1, 2026.

and statistical analysis. Section 3 reports dialogue-continuation experiments with the Moshi-based model and evaluates speech quality, linguistic appropriateness, and turn-taking characteristics. Section 4 concludes and outlines future work.

2 Analysis of Spoken Dialogue Corpora

This section presents the dialogue corpora used in the study, the metrics employed for analysis, and the main results.

2.1 Datasets

We analyze four Japanese spoken dialogue corpora: *J-Chat* (web-derived) (Nakata et al., 2024), *J-Chat-Clean* (a quality-filtered subset of J-Chat), the Corpus of Spontaneous Japanese (*CSJ*) (Maekawa, 2003), and a travel-agency task corpus (*Tabidachi*) (Inaba et al., 2024). These corpora differ in recording conditions, speaker relations, and domains, yielding a broad range of interactional styles: *J-Chat* is large-scale but acoustically variable; *J-Chat-Clean* removes low-quality segments to preserve more natural turn-taking; *CSJ* comprises well-structured formal speech; and *Tabidachi* consists of guided travel-consultation dialogues with characteristically asymmetric, system-led turns. Collectively, they span a continuum from large, noisy conversational data to smaller, high-quality task-specific recordings—an effective basis for curriculum-style fine-tuning.

For *J-Chat*, we applied monaural speech separation with Asteroid/Conv-TasNet[1], automatic transcription with ReazonSpeech-espnet[2], and word-level time alignment with WhisperX (Bain et al., 2023)[3].

2.2 Metrics

Turn-taking characteristics were analyzed using four timing-based indicators: IPU, pause, gap, and overlap. Speech segments were extracted with the Silero Voice Activity Detector (VAD)[4], and statistics were computed over 20-second windows. These measures capture conversational tempo and interactivity: for example, frequent overlaps signal natural responsiveness, whereas longer gaps reflect more formal or deliberate rhythms (Ward and Tsukahara, 2000).

Linguistic and semantic quality was evaluated using the LLM-as-a-Judge framework (Zheng et al., 2023), which scores coherence (COH), naturalness (NAT), relevance (REL), instruction-following (INS), turn-taking (TUR), and overall quality (OVE) on a ten-point scale. Acoustic quality was assessed using the NISQA model (Mittag et al., 2021), which predicts MOS scores for perceptual naturalness. Together, these metrics provide a comprehensive view of both linguistic and paralinguistic properties of each corpus.

2.3 Results

Table 1 summarizes turn-taking statistics, and Table 2 reports acoustic and linguistic quality scores.

Turn-taking characteristics. *J-Chat* exhibits frequent overlaps (e.g., 5.06 per 20 s in the Podcast subset) and balanced IPU counts between speakers (about 5–6 each), reflecting spontaneous, chat-style interactions. The curated subset *J-Chat-Clean* exhibits fewer overlaps (3.86 per 20 s) and longer gaps (10.82 s), reflecting calmer, more separated turns. *CSJ* shows fewer overlaps (2.30 occurrences) and long gaps (10.25 s), indicating structured, deliberate speech. *Tabidachi* has the fewest overlaps (1.16 occurrences) and the greatest asymmetry between speakers (IPUs: A = 5.06, B = 2.12), consistent with task-oriented dialogues characterized by stable, one-sided turns. Together, these patterns confirm distinct conversational rhythms and coordination styles across corpora.

Acoustic and semantic quality. *CSJ* attains the highest acoustic naturalness (NISQA = 3.14) and semantic naturalness (LLMAJ-NAT = 6.96), reflecting clean recordings and consistent structure. *Tabidachi* leads in coherence (6.77), relevance (5.60), and instruction-following (4.53), indicating strong task alignment. By contrast, web-derived *J-Chat*, especially the YouTube subset, shows lower acoustic quality (NISQA = 1.94) and weaker semantic scores (e.g., LLMAJ-COH = 4.63), highlighting the trade-off between scale/diversity and consistent conversational and recording quality.

Summary. Corpus properties, particularly turn-taking structure and recording quality, clearly differentiate datasets. Consequently, corpus selection and ordering should be guided not only by data volume but also by the target dialogue style (spon-

[1] https://huggingface.co/JorisCos/ConvTasNet_Libri2Mix_sepclean_16k

[2] https://huggingface.co/reazon-research/reazonspeech-espnet-v2

[3] https://github.com/m-bain/whisperX

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

	Number of occurrences / 20s				Cumulative duration / 20s			
	IPU	Pause	Gap	Overlap	IPU	Pause	Gap	Overlap
J-Chat (Podcast)	A:5.26 B:6.00	A:1.10 B:2.04	4.34	5.06	A:8.69 B:10.87	A:0.66 B:1.29	7.98	3.01
J-Chat (YouTube)	A:5.00 B:4.98	A:1.46 B:1.86	3.78	3.92	A:7.58 B:9.51	A:1.30 B:1.95	8.41	3.03
J-Chat-Clean	A:5.38 B:6.18	A:1.42 B:1.78	5.90	3.86	A:8.20 B:8.90	A:1.18 B:1.45	10.82	2.19
CSJ	A:3.92 B:5.12	A:1.10 B:2.44	4.18	2.30	A:5.57 B:11.22	A:0.98 B:1.63	10.25	1.20
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

Table 1: Turn-taking statistics across corpora. "Number of occurrences / 20s" shows average number of events in 20-second segment, and "Cumulative duration / 20s" indicates their total duration (in seconds) within same segment.

		J-Chat (Podcast)	J-Chat (YouTube)	J-Chat-Clean	CSJ	Tabidachi
NISQA (MOS) (1–5)		2.27	1.94	2.51	**3.14**	2.98
LLMAJ (1–10)	COH	5.69	4.63	5.54	6.00	**6.77**
	NAT	6.67	5.52	6.62	**6.96**	6.87
	REL	4.69	3.67	4.77	5.06	**5.60**
	INS	3.27	2.33	3.08	3.42	**4.53**
	TUR	5.76	4.59	5.67	5.94	**6.36**
	OVE	5.52	4.40	5.54	5.68	**5.99**

Table 2: Scores of speech quality (NISQA) and semantic appropriateness (LLM-as-a-Judge). COH = coherence, NAT = naturalness, REL = relevance, INS = instruction following, TUR = turn taking, OVE = overall.

taneous, formal, or task-oriented), a consideration that is central to effective curriculum design for full-duplex spoken dialogue models.

3 Dialogue Continuation Experiment

This section examines how fine-tuning on corpora with different turn-taking characteristics affects full-duplex speech dialogue models. In a **dialogue continuation** setup, each model received a 10-second audio prompt from a held-out *Tabidachi* split and then generated the next 20 seconds of dialogue, allowing assessment of adaptation to conversational rhythm and turn-taking under realistic conditions.

3.1 Model and Training Setup

All experiments used the **Moshi** full-duplex architecture, which encodes stereo inputs and generates time-synchronized multi-channel speech in real time.[5] We adopted a three-stage **curriculum**.

1. **Pre-training** on large, noisy *J-Chat* to acquire general conversational structure.
2. **Intermediate fine-tuning** on cleaner *J-Chat-Clean* or formal *CSJ* to improve robustness and temporal alignment.
3. **Final fine-tuning** on task-specific, high-quality *Tabidachi* to refine interaction patterns for travel consultation.

This progression gradually shifts the model from broad/noisy behaviors to domain-specific, high-quality interaction.

3.2 Speech Quality and Semantic Validity

Table 3 summarizes the objective results.

Speech quality (NISQA). Model 1 (*Tabidachi only*) achieves the highest NISQA (3.12), followed by Models 4 (3.07), 2 (3.02), and 3 (2.90). Thus, fine-tuning solely on clean, domain-matched data yields the most natural acoustics; adding larger but more variable corpora can slightly degrade perceptual quality.

Linguistic/semantic quality (LLM-as-a-Judge). Model 2 (*CSJ* + *Tabidachi*) scores best on **Coherence** (4.50), **Relevance** (3.73), **Instruction-Following** (2.86), and **Overall** (4.20), indicating that formal, well-structured speech (*CSJ*) strengthens discourse organization and task alignment. Model 3 (*J-Chat-Clean* + *Tabidachi*) attains comparable **Naturalness** and **Relevance**,

[5] Our models were further trained based on the pre-trained checkpoint `kyutai/moshiko-pytorch-bf16` available on HuggingFace (https://huggingface.co/kyutai/moshiko-pytorch-bf16), using the fine-tuning scripts provided in https://github.com/nu-dialogue/moshi-finetune.

		Model 1	Model 2	Model 3	Model 4
Pre-training		J-Chat (69k hours)	J-Chat (69k hours)	J-Chat (69k hours)	J-Chat (69k hours)
Fine-tuning		–	–	J-Chat-Clean (300 hours)	J-Chat-Clean (300 hours)
		–	CSJ (12 hours)	–	CSJ (12 hours)
		Tabidachi (115 hours)	Tabidachi (115 hours)	Tabidachi (115 hours)	Tabidachi (115 hours)
NISQA (MOS) (1–5)		**3.12**	3.02	2.90	3.07
LLMAJ (1–10)	COH	4.32	**4.50**	4.39	4.20
	NAT	5.36	**5.48**	**5.48**	5.25
	REL	3.57	**3.73**	**3.73**	3.45
	INS	2.66	**2.86**	2.84	2.75
	TUR	4.52	**4.61**	4.52	4.39
	OVE	3.88	**4.20**	4.11	4.09

Table 3: Training configurations and objective evaluation results. Speech quality is reported with NISQA; semantic appropriateness is evaluated with LLM-as-a-Judge: COH = coherence, NAT = naturalness, REL = relevance, INS = instruction following, TUR = turn taking, OVE = overall.

	Number of occurrences / 20s				**Cumulative duration / 20s**			
	IPU	Pause	Gap	Overlap	IPU	Pause	Gap	Overlap
Model 1	A:4.23 B:1.57	A:2.50 B:0.11	1.84	0.98	A:13.86 B:1.43	A:3.05 B:0.08	4.94	0.37
Model 2	A:4.59 B:1.52	A:2.68 B:0.18	2.30	0.84	A:12.38 B:1.78	A:4.04 B:0.14	6.26	0.37
Model 3	A:4.64 B:1.45	A:2.86 B:0.34	2.32	0.66	A:12.61 B:1.26	A:3.69 B:0.30	6.32	0.27
Model 4	A:4.55 B:1.39	A:2.75 B:0.11	2.14	0.82	A:12.13 B:1.72	A:3.82 B:0.09	4.86	0.40

Table 4: Turn-taking statistics of generated dialogues. "Number of occurrences / 20 s" is average count of events per 20-second segment, and "Cumulative duration / 20 s" is their total duration (in seconds) within same segment.

suggesting large-scale clean conversational data improves fluency but not necessarily task coherence. Model 4 (*J-Chat-Clean* + *CSJ* + *Tabidachi*) does not surpass simpler curricula (slight drops in **Overall**=4.09, **Coherence**=4.20), implying redundancy/over-regularization when mixing multiple high-quality yet stylistically mismatched datasets.

3.3 Turn-taking Behavior

Table 4 summarizes turn-taking statistics for the generated dialogues. Model 3 (*J-Chat-Clean* + *Tabidachi*) yields the longest and most frequent gaps (2.32 occurrences; 6.32 s total per 20 s) and the fewest overlaps (0.66; 0.27 s), producing a slower, highly orderly rhythm with clear separation between turns. Notably, relative to Model 1 (*Tabidachi* only), gaps increase while overlaps decrease, indicating stronger turn separation.

These tendencies suggest that *J-Chat-Clean* promotes structured, non-overlapping exchanges, whereas *Tabidachi* reinforces natural pausing and alternation. As a result, Model 3 produces smooth and polite turn-taking but exhibits reduced spontaneity—mutual overlaps, a hallmark of lively conversation, are suppressed.

Overall, the turn-taking profile of the fine-tuning corpus, e.g., its balance of overlaps and gaps, has a greater impact on achieving the desired dialogue style than dataset size per se, suggesting that style-aware corpus curation should take precedence over scale.

4 Conclusion

We analyzed how corpus-specific turn-taking features (IPU, pause, gap, overlap) influence Moshi-based models. *J-Chat-Clean* promotes rhythmically stable, well-separated turns; *CSJ* yields cautious, formally structured timing; *Tabidachi* supports natural pausing and cooperative, task-oriented exchanges. Rather than naively concatenating corpora, **strategic selection and ordering** should reflect the target dialogue style—spontaneous, formal, or collaborative. Turn-taking statistics thus provide a practical basis for **corpus-aware fine-tuning** and controllable interactional style in full-duplex spoken dialogue systems.

Acknowledgments

We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use".

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Preprint*, arXiv:2303.00747.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2024. Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(9).

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025. Language model can listen while speaking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24831–24839.

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: its design and evaluation. In *ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. INTERSPEECH 2021*.

Wataru Nakata, Kentaro Seki, Hitomi Yanaka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling. *Preprint*, arXiv:2407.15828.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, and Ryuichiro Higashinaka. 2025. Towards a Japanese Full-duplex Spoken Dialogue System. In *Proc. INTERSPEECH 2025*, pages 1783–1787.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-R1: Do you really need audio to fine-tune your audio LLM? *Preprint*, arXiv:2505.09439.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. *Preprint*, arXiv:2409.15594.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. *Preprint*, arXiv:2411.00774.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. 2025. SARI: Structured audio reasoning via curriculum-guided reinforcement learning. *Preprint*, arXiv:2504.15900.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni technical report. *Preprint*, arXiv:2503.20215.

Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi, Nigel G Ward, and Tatsuya Kawahara. 2016. Analysis and prediction of morphological patterns of backchannels for attentive listening agents. In *Proc. 7th International Workshop on Spoken Dialogue Systems*, pages 1–12.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-Judge with MT-bench and chatbot arena. *Proc. NeurIPS*, 36:46595–46623.