Reproducing Proficiency-Conditioned Dialogue Features with Full-duplex Spoken Dialogue Models

Takao Obi[1,2], Sadahiro Yoshikawa[1], Mao Saeki[1], Masaki Eguchi[1,3], Yoichi Matsuyama[1]

[1]Equmenopolis, Inc., [2]Institute of Science Tokyo, [3]Waseda University,

Correspondence: t.obi@equ.ai

Abstract

Real-time, human-centered conversational AI requires systems that handle spoken dialogue with overlap and rapid turn-taking. Although full-duplex models promise these capabilities, empirical work applying them to conversational AI is still nascent. To fill this gap, this study investigates whether the full-duplex model can reproduce the human dialogue features. We adapt a full-duplex spoken dialogue model to a large corpus of second-language (L2) learner interviews and train proficiency-conditioned models. We then conduct real-time interview sessions between these models and a spoken dialogue system designed to elicit spontaneous learner speech, and analyze reaction time, response frequency, and fluency metrics across aggregated CEFR levels (A/B/C). Our results show that proficiency-conditioned models partially reproduce levelwise trends and distributions observed in human interviews across multiple metrics. These findings suggest that full-duplex models can reproduce dialogue features of human dialogues and offer a promising foundation for conversational AI systems.

1 Introduction

Real-time, robust, and human-centered conversational AI demands systems that interact with users at human-like granularity, including overlapping speech, barge-ins, backchannels, and rapid adjustments.

Full-duplex spoken dialogue models have attracted attention as a foundation for such systems because they enable simultaneous, bidirectional interaction without explicit turn segmentation. Moshi, a representative full-duplex model, achieves real-time speech-to-speech generation by modeling user and system audio in parallel, a capability that conventional turn-based architectures struggle to provide (Défossez et al., 2024). Although full-duplex models promise these capabilities, the work of applying them to conversational AI is still nascent.

In this work, we investigate whether the full-duplex model can reproduce the human dialogue features. We adapt the full-duplex model Moshi to a large corpus of second-language (L2) learner interview dialogues and train proficiency-conditioned models that generate responses at CEFR (Common European Framework of Reference for Languages) levels A/B/C (North and Piccardo, 2020). We then conduct real-time interview sessions between these models and InteLLA, a spoken dialogue system designed to elicit spontaneous learner speech (Saeki et al., 2024), and analyze the sessions in terms of reaction time, response frequency, and fluency metrics. The results show that CEFR-conditioned full-duplex models reproduce levelwise trends observed in human interviews and exhibit partial distributional alignment with human dialogues across multiple metrics. These findings indicate that full-duplex models can reproduce human dialogue features and provide a promising foundation for conversational AI systems.

2 Related Work

2.1 Speech Foundation Models

In human conversation, finely timed turn-taking is central (Sacks et al., 1974; Heldner and Edlund, 2010; Skantze, 2021). Speech foundation models have progressed from textless spoken language modeling to high-fidelity audio generation. dGSLM demonstrates dialogue generation with laughter and fluid turn-taking directly from audio units, without text supervision (Nguyen et al., 2023). AudioLM models long-range structure and style from discrete tokens (Borsos et al., 2023a). Neural codec language models enable zero-shot speaker and style transfer, and long-range prosody (Wang et al., 2023). SoundStorm extends this line of work with bidirectional, parallel decoding of codec tokens, achieving faster generation

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 43–51
February 26–March 1, 2026.

while preserving voice consistency (Borsos et al., 2023b). Collectively, these works highlight speech-level interactional phenomena that are central to human-centered dialogue and set the stage for dialogue modeling beyond text-only or turn-based pipelines.

2.2 Full-duplex Spoken Dialogue Models

Full-duplex spoken dialogue models have attracted growing attention because they support simultaneous, bidirectional interaction akin to human conversation. A representative model, Moshi, models user and system audio in parallel and performs multi-stream speech-to-speech generation, thereby enabling real-time interaction without explicit turn segmentation (Défossez et al., 2024). Other full-duplex model designs include time-synchronous LLMs (SyncLLM) (Veluri et al., 2024) and schemes that combine control tokens with explicit state mechanisms (Wang et al., 2024). On the application and evaluation fronts, researchers have proposed benchmarks for overlap handling (Lin et al., 2025), developed Japanese full-duplex systems (Ohashi et al., 2025), aligned models from interaction logs (Wu et al., 2025), and optimized conversational behaviors via reinforcement learning (Chen et al., 2025). While full-duplex research is expanding with these works, applying them to conversational AI is still nascent.

2.3 Automated Assessment of Language Proficiency

As an application area for real-time spoken dialogue, automated L2 proficiency assessment has seen growing interest. InteLLA was recently proposed to elicit spontaneous learner speech for proficiency assessment, accelerating research on automated L2 evaluation (Saeki et al., 2024). Related efforts include multimodal proficiency assessment frameworks (Takatsu et al., 2025), fluency estimation directly from speech (Matsuura et al., 2025), and the use of speech LLMs for oral proficiency scoring (Ma et al., 2025). Despite this progress, large-scale quality assurance and evaluation still require substantial human effort, motivating systems that can reproduce learner-like interactional behavior. If full-duplex models can reproduce proficiency-conditioned dialogue features, these models offer a promising foundation for building proficiency-conditioned user emulators.

Figure 1: Real-time spoken interaction over WebRTC between a dialogue system InteLLA (left) and a full-duplex model (right).

3 Proficiency-Conditioned Full-duplex Spoken Dialogue Models

We adapt the full-duplex spoken dialogue model, Moshi, to a large corpus of L2-learner interview dialogues to examine whether it can reproduce proficiency-conditioned dialogue features.

3.1 Moshi

Moshi consists of a neural audio codec named Mimi and a large spoken language model named RQ-Transformer.

Mimi : Mimi consists of a SEANet autoencoder (Tagliasacchi et al., 2020) and a residual vector quantizer (Zeghidour et al., 2021). The encoder discretizes 24 kHz audio into 8 RVQ codebooks at 12.5 Hz (80 ms frames), and the decoder reconstructs waveforms from RQ-Transformer outputs. Mimi encodes the user stream and decodes the model's speech tokens in real time.

RQ-Transformer : RQ-Transformer consists of a Temporal Transformer and a Depth Transformer. The Temporal Transformer models token sequences at 12.5 Hz and produces a hidden state z_s at time s from tokens up to $s-1$. A linear head then samples time-aligned text tokens t_s. The Depth Transformer models audio tokens along the depth dimension. To stabilize audio quality, a one-step delay is used for acoustic tokens, and PAD tokens are inserted where no text token is emitted.

3.2 Fine-tuning

We fine-tuned Moshi on a large corpus of L2 learner interview dialogues collected with InteLLA over three academic years, as shown in Table 1. Each dialogue is accompanied by an automatically assigned CEFR label from InteLLA. We aggregated labels into three levels (A/B/C) and fine-tuned a separate model for each level.

Table 1: Training data by aggregated CEFR level.

CEFR	Dialogues	Total (h)	Extracted (h)
A	8,171	722	367
B	13,031	1,804	1,132
C	3,118	645	437
Total	24,320	2,451	1,936

We trained for three epochs per level using the publicly released moshi-finetune[1] recipe. We conducted training on 8 NVIDIA H200 GPUs, with OneCycleLR (max_lr=2×10^{-6}), batch size 8, input length 100 seconds, and weight decay 0.1. The number of training steps was 621 (A), 1911 (B), and 738 (C). Similar to the original Moshi, PAD token losses were reduced by 50%, and the loss ratio between semantic tokens and acoustic tokens was set to 100:1.

For training, we extracted segments spanning from the system utterance end to the user utterance end. Using full interview data often caused excessive generation silence at inference time, even after hyperparameter tuning. Segment extraction mitigated this issue and enabled robust evaluation.

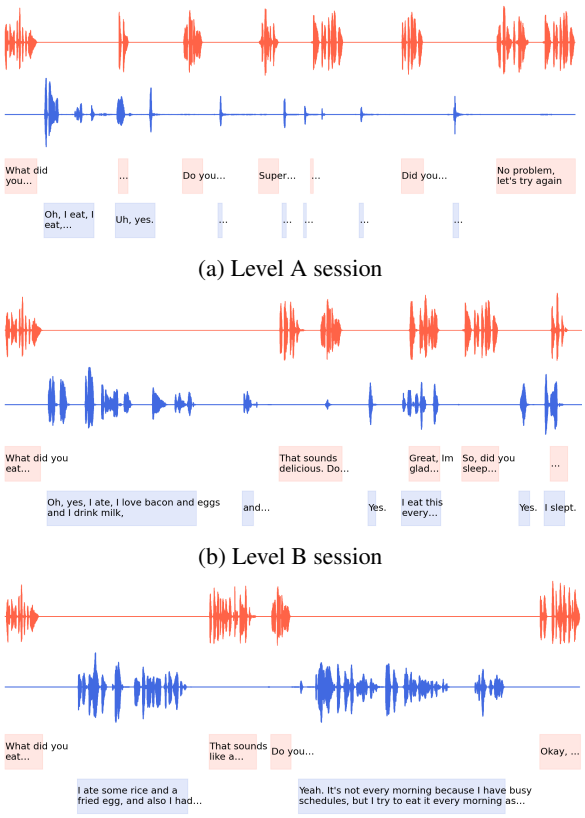
3.3 Evaluation Metrics

Because the full-duplex models were trained on L2 learner interview dialogues, we evaluate whether they can reproduce the dialogue features of a real interview in the same way a human learner does. We focused on the following indicators and analyzed how well the distributions obtained through multiple dialogues reproduce the distributions of actual human interviews:

Reaction time : We defined the model's reaction time as $\Delta t_m = t_m^{\text{start}} - t_s^{\text{yield}}$, where t_s^{yield} denotes the time when the system's question finishes, and t_m^{start} denotes the time when the model's response onset. Because full-duplex dialogue permits barge-ins, we allowed $\Delta t_m < 0$. If the system repeats the same question, we took the first t_s^{yield} as reference. When no response occurred before a new question, the case was treated as a no-response and excluded from reaction time analysis but included in response frequency analysis.

Response frequency : We defined the dialogue-level response frequency as $f_m = c_m^{\text{res}} / c_{s\to m}^{\text{pass}}$, where $c_{s\to m}^{\text{pass}}$ denotes the number of questions that require a response in a dialogue, and c_m^{res} denotes the number of responses produced by the model.

[1]https://github.com/kyutai-labs/moshi-finetune



(a) Level A session

(b) Level B session

(c) Level C session

Figure 2: Audio waveform examples with transcripts for proficiency-conditioned models using the same scenario (0–40 s). Orange: InteLLA; blue: model.

Fluency : We computed fluency metrics using a machine learning-based fluency annotation model (Matsuura et al., 2025), as shown in Table 2[2].

4 Dialogue Experiments

4.1 System Setup

To evaluate how full-duplex models reproduce human dialogue features, we built a real-time spoken interview system connecting these models with the interviewer agent InteLLA over WebRTC, as shown in Figure 1.

Full-duplex models: We used three full-duplex spoken dialogue models, each fine-tuned for one aggregated CEFR level. At inference time, we set the sampling temperature to 0.95 to increase

[2]In addition to the listed metrics, the annotation model can output mid-clause pause duration, end-clause pause duration, and mean pause durations. However, preliminary experiments indicated insufficient reliability for these metrics, so we omitted them from the fluency metrics.

Table 2: Fluency metrics used for the proficiency-conditioned full-duplex model.

Metric	Description
Articulation rate (AR)	Number of syllables per speech duration excluding pauses.
Mid-clause pause ratio (MCPR)	Number of mid-clause silent pauses per syllable.
End-clause pause ratio (ECPR)	Number of end-clause silent pauses per syllable.
Pause ratio (PR)	Number of silent pauses irrespective of pause location.
Disfluency ratio (DR)	Number of disfluency words (repetitions, self-corrections, and false starts) per syllable.
Speech rate (SR)	Number of syllables per speech duration.
Mean length of run (MLR)	Mean number of syllables of speech separated by pauses.

speaking propensity. Sampling temperature was determined based on speaking propensity and the appropriateness of the speech content.

Interviewer agent: We used InteLLA, a spoken dialogue system designed to elicit spontaneous learner speech, as the interviewer agent. Since longer interviews occasionally trigger prolonged silences in full-duplex models, we prepared interview scenarios focused on a single topic. To reduce the effects of scenarios, we prepared four distinct scenarios. The scenario topics were selected from InteLLA's A-level topic pool to control topic complexity across models.

Audio streaming: We used Agora RTC[3] for bidirectional audio streaming, with typical end-to-end latency ≤ 200 ms.

4.2 Procedure

We conducted 400 sessions (100 sessions per scenario) for each level model. After each session, both audio streams and InteLLA's system logs were stored. Response timing was automatically annotated in milliseconds by InteLLA's Turn Management Module. Timestamps reference InteLLA's server clock. Sessions were independent; the model was restarted before every session. After the sessions, we computed the evaluation metrics defined in Section 3.3.

Table 3 shows session examples for proficiency-conditioned full-duplex models using the same scenario. To complement these examples, Figure 2 shows the audio waveforms with transcripts for the same sessions shown in Table 3 (0–40 s).

4.3 Trend Direction Across CEFR Levels

We evaluated whether full-duplex models reproduce levelwise trends observed in human interviews on each metric. For human data, we randomly selected 400 samples from interviews with fluency metrics (A: 3117, B: 5579, C: 778 dialogues) to compare statistical power with the same sample size as in the model sessions.

We applied the Kruskal-Wallis (KW) test to each metric to detect overall differences in levels. Significant differences were found across levels in both human interviews and model sessions ($p < 0.01$), except for the model session's response frequency. We then performed pairwise Mann–Whitney U tests with Holm–Bonferroni correction and recorded the direction of the difference (A–B, B–C, A–C). We also confirmed the absolute Cliff's delta as an effect size.

Table 4 shows pairwise trend directions and absolute Cliff's delta. Although no significant difference was found in the model's response frequency in the KW test, we also included the results of a pairwise comparison to compare with other metrics. Bold cells indicate that the direction matches that of human data. The rightmost column counts agreements only for pairs in which the human comparison was significant.

4.4 Distributional Comparison with Human Interviews

We compared the distributions of each metric between human interviews and model sessions for each CEFR level. For each metric, let Δ denote the model–human mean difference, and let W_1 be the one-dimensional Wasserstein distance. We used the median difference in only the reaction time analysis because some interviews and sessions exhibited substantial response delays.

Using the human dialogues for the level, we bootstrapped to the corresponding model sample size for that level and estimated 95% percentile confidence intervals (CIs) of Δ and W_1 over 2000

[3]https://github.com/AgoraIO-Community/Agora-Python-SDK

Table 3: Session examples for proficiency-conditioned full-duplex models using the same scenario (excerpted).

Level A session

Time (s)	Speaker	Utterance
0.00 - 2.35	InteLLA	What did you eat for breakfast this morning?
2.78 - 6.28	Model	Oh, I eat, I eat, yeah.
7.78 - 8.44	Model	Uh,
7.99 - 8.70	InteLLA	Awesome!
10.19 - 10.36	Model	Yes.
12.53 - 13.92	InteLLA	Do you eat that every morning?
14.98 - 15.28	Model	Yes.
17.84 - 19.24	InteLLA	Super interesting!

Level B session

Time (s)	Speaker	Utterance
0.00 - 2.35	InteLLA	What did you eat for breakfast this morning?
2.66 - 15.66	Model	Oh, yes, I ate, I love bacon and eggs and I drink milk, and yogurts.
17.28 - 18.88	InteLLA	That sounds delicious.
19.86 - 21.26	InteLLA	Do you eat that every morning?
22.86 - 23.36	Model	Yes.
24.89 - 27.39	Model	I eat this every morning.
25.44 - 27.40	InteLLA	Great, I'm glad to hear that.

Level C session

Time (s)	Speaker	Utterance
0.00 - 2.35	InteLLA	What did you eat for breakfast this morning?
4.96 - 12.66	Model	I ate some rice and a fried egg, and also I had a cup of milk with me.
14.13 - 17.45	InteLLA	That sounds like a delicious and filling breakfast.
18.44 - 19.84	InteLLA	Do you eat that every morning?
20.36 - 34.76	Model	Yeah. It's not every morning because I have busy schedules, but I try to eat it every morning as much as I can. It's a good way to extend my life.
37.16 - 40.00	InteLLA	Okay, that's a great way to start the day!

resamples, yielding the human variability band. We then bootstrapped the model sessions (2000 resamples) to obtain CIs of Δ and W_1 for the models. We judged that the model reproduces the human distribution strictly if the model CI lies within the corresponding human band; partial overlap indicates partial reproduction (Vasishth and Gelman, 2021). As a baseline (not CEFR-conditioned), we also ran the same real-time interview sessions with the original Moshi model (without fine-tuning) under the same procedure and decoding settings.

Figure 3 and 4 summarize the CIs of Δ and W_1 (band (red): human; bars (blue): fine-tuned models; bars (gray): original Moshi). For the fine-tuned models, in comparisons of Δ, partial overlaps were observed for response frequency at all levels, AR at level A, MCPR at levels B and C, ECPR at levels A and B, PR at all levels, DR at all levels, SR at level A, and MLR at levels A and B. In comparisons of W_1, overlap was not observed across metrics. Compared with the original Moshi baseline, fine-tuning tends to reduce $|\Delta|$ and W_1 across several metrics and levels.

5 Discussion

5.1 Reproducing Levelwise Trends Across CEFR

Table 4 shows that proficiency-conditioned full-duplex models reproduce several levelwise trends observed in human interviews. In particular, AR, PR, and SR matched the human direction across all three level pairs, suggesting that the models capture proficiency effects on these fluency dimensions. For the other metrics, MCPR, ECPR, MLR, and

Table 4: Pairwise Mann-Whitney tests with Holm correction and absolute Cliff's delta (parentheses) in CEFR levelwise for human and full-duplex dialogues.

	Human			Full-duplex model*			Trend-agreement
Metric	A–B	B–C	A–C	A–B	B–C	A–C	count†
Reaction time	A $\gg$ B	B $\gg$ C	A $\gg$ C	A $\ll$ B	**B $\gg$ C**	**A $\gg$ C**	2/3
	(0.291)	(0.119)	(0.405)	(0.077)	(0.161)	(0.075)	
Response frequency	A $\ll$ B	B $\sim$ C	A $\ll$ C	A $\sim$ B	B $\sim$ C	A $\sim$ C	0/2
	(0.271)	(0.005)	(0.264)	(0.014)	(0.026)	(0.040)	
AR	A $\ll$ B	B $\ll$ C	A $\ll$ C	**A $\ll$ B**	**B $\ll$ C**	**A $\ll$ C**	3/3
	(0.679)	(0.533)	(0.843)	(0.220)	(0.392)	(0.537)	
MCPR	A $>$ B	B $\gg$ C	A $\gg$ C	A $\sim$ B	**B $\gg$ C**	**A $\gg$ C**	2/3
	(0.110)	(0.498)	(0.308)	(0.030)	(0.305)	(0.229)	
ECPR	A $\gg$ B	B $\gg$ C	A $\gg$ C	A $\sim$ B	**B $\gg$ C**	**A $\gg$ C**	2/3
	(0.267)	(0.740)	(0.790)	(0.048)	(0.193)	(0.193)	
PR	A $>$ B	B $\gg$ C	A $\gg$ C	**A $>$ B**	**B $\gg$ C**	**A $\gg$ C**	3/3
	(0.170)	(0.278)	(0.339)	(0.079)	(0.185)	(0.218)	
DR	A $<$ B	B $\gg$ C	A $\sim$ C	A $\sim$ B	**B $\gg$ C**	A $>$ C	1/2
	(0.112)	(0.184)	(0.049)	(0.020)	(0.164)	(0.112)	
SR	A $\ll$ B	B $\ll$ C	A $\ll$ C	**A $\ll$ B**	**B $\ll$ C**	**A $\ll$ C**	3/3
	(0.934)	(0.843)	(0.983)	(0.424)	(0.637)	(0.826)	
MLR	A $<$ B	B $\ll$ C	A $\ll$ C	A $\sim$ B	**B $\ll$ C**	**A $\ll$ C**	2/3
	(0.097)	(0.706)	(0.696)	(0.114)	(0.437)	(0.478)	

$\gg$: $p < 0.01$, $>$: $p < 0.05$, $\sim$: $p \geq 0.05$. * Bold indicates the same direction as human.
† Counted only where the human comparison was significant.

DR, agreement held only for a subset of pairs, and none adequately captured the differences between levels A and B. This suggests that the model has difficulty capturing differences between levels A and B on these dimensions.

For reaction time, the model's trend between levels A and B was the opposite of that in humans. Response frequency also did not differ by level in the model sessions. Given Moshi's tendency toward silence, residual silences may not have been suppressed sufficiently to reproduce these levelwise trends; stronger suppression of this tendency could reveal human-like trends.

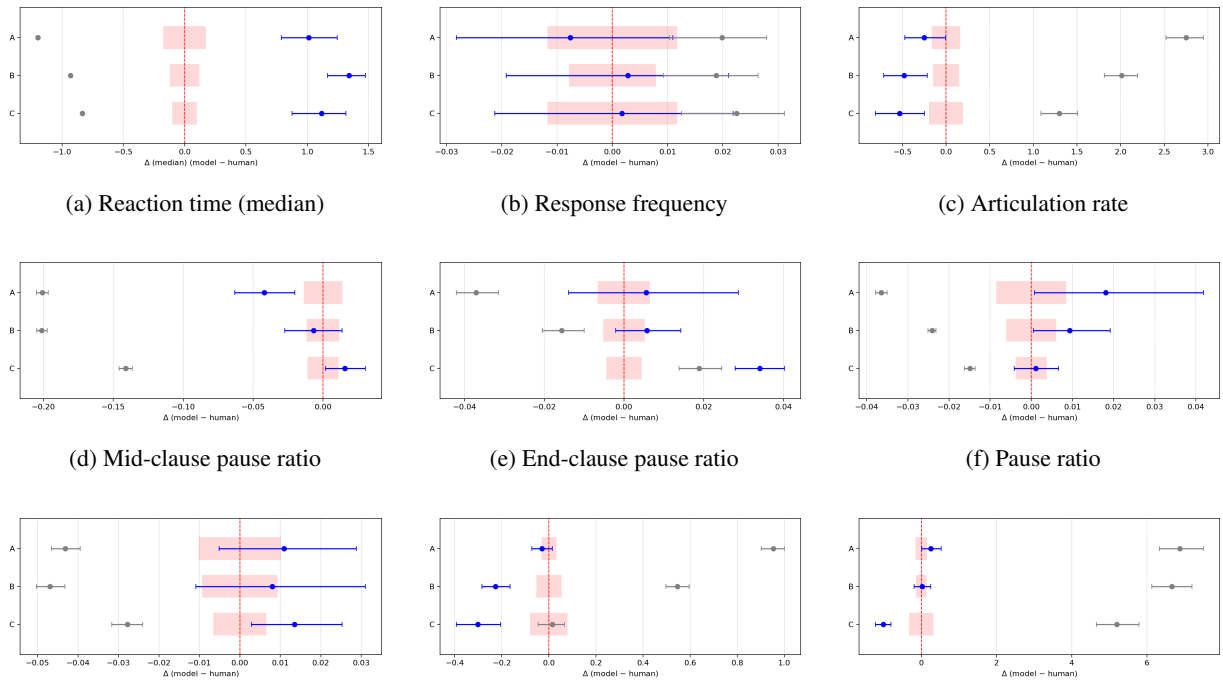
5.2 Distributional Similarity and Practical Reproducibility

In the distributional analysis, we observed overlap between the model CIs and the human variability bands for several pairs of Δ, whereas overlaps in W_1 are generally absent. Because the models began to align with human central tendencies, we expect that further improvements will bring the metrics closer into agreement with human distributions. The remaining gap may be narrowed by more selective data extraction, for example, by filtering out silence-driven outliers based on the distribution of silent time.

5.3 Propensity to Silence in Full-duplex Generation

Our full-duplex models exhibited longer reaction latencies (see Figure 3). Moreover, response frequency did not differ across levels in model sessions (see Table 4). To curb excessive silence, we (i) trained on dialogue segments spanning from the end of the system's utterance to the end of the user's utterance, (ii) constrained each session to a single topic per scenario to stabilize the dialogue, and (iii) used a higher sampling temperature (0.95) to increase speaking propensity. These choices suppressed residual silence, but a tendency to delay onset or remain silent persisted.

Beyond level effects, we examined response frequency at the scenario level by aggregating all model sessions and found topic-specific dips. Using a KW test followed by pairwise Mann–Whitney U tests with Holm–Bonferroni correction ($p < 0.01$), we found that one scenario elicited markedly fewer responses than the others. Because all four scenarios were selected from InteLLA's A-level topic pool and no significant differences in response frequency were observed across our A/B/C



(a) Reaction time (median) (b) Response frequency (c) Articulation rate

(d) Mid-clause pause ratio (e) End-clause pause ratio (f) Pause ratio

(g) Disfluency ratio (h) Speech rate (i) Mean length of run

Figure 3: CIs of the model–human difference for each metric (median in reaction time; mean in others). Red bands indicate human variability; blue and gray bars indicate the fine-tuned models and the original Moshi, respectively. All CIs are 95% bootstrap percentile intervals (2000 resamples).

models (see Table 4), this pattern appears scenario-specific rather than a consequence of model level or topic difficulty. This suggests the effectiveness of complementary methods for mitigating silence, such as model scenario design.

6 Limitations and Future Work

6.1 Heuristic Operating Points

During fine-tuning, we trained on segments spanning from the end of the system utterance to the end of the user utterance to mitigate prolonged silences at inference. At decoding, we fixed the sampling temperature at 0.95 to increase speaking propensity. These heuristic choices were based on preliminary experiments and adapted to ensure stable real-time operation. While these methods would be effective, they may bias the learned and measured timing distributions. We will replace these heuristics with principled procedures by systematically comparing segmentation windows versus full-dialogue training and operating temperatures.

6.2 Evaluation Metrics

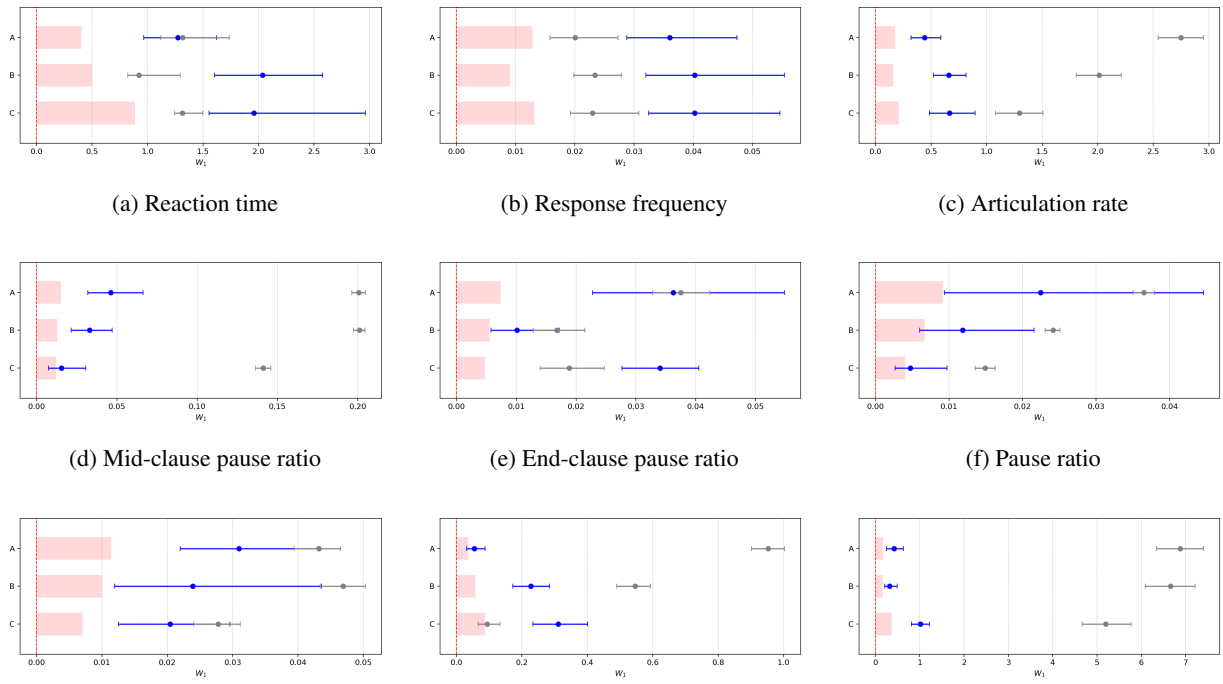
We evaluated the reproducibility of dialogue features using reaction time, response frequency, and a set of fluency metrics. While these indicators effectively capture temporal and linguistic aspects of dialogue, they do not encompass all interactional phenomena in full-duplex conversation. In particular, we did not explicitly quantify overlap and backchannel behavior in this study, as our primary focus was on timing and fluency in the question–answer interview dialogues. Future work will extend the metric coverage to include these aspects, as well as acoustic-prosodic, lexical, and grammatical features, to more comprehensively assess proficiency-conditioned behavior.

6.3 Data Imbalance Across Levels

Our training data exhibited substantial imbalance across aggregated CEFR levels (see Table 1). The weaker separability between A and B in some metrics may be partly due to this imbalance, as the B-level model had roughly three times as much training data as the A-level model. Future work will expand A/C-level data and explore balancing strategies to isolate the effect of data volume from proficiency-conditioned reproducing.

6.4 Context Specificity

We demonstrated that Moshi can reproduce dialogue features through the experiments. However, Moshi's behavior may change depending on the



(a) Reaction time (b) Response frequency (c) Articulation rate

(d) Mid-clause pause ratio (e) End-clause pause ratio (f) Pause ratio

(g) Disfluency ratio (h) Speech rate (i) Mean length of run

Figure 4: CIs of the model–human one-dimensional Wasserstein distance for each metric. Red bands indicate human variability; blue and gray bars indicate the fine-tuned models and the original Moshi, respectively. All CIs are 95% bootstrap percentile intervals (2000 resamples).

context, as we confirmed in Section 5.3. We plan to broaden both training and evaluation to diverse dialogues and to verify whether the full-duplex models reproduce human-like dialogue features beyond interviews.

7 Conclusion

We adapted the full-duplex model Moshi to a large corpus of L2 learner interview dialogues and trained proficiency-conditioned models at CEFR levels A/B/C. We then conducted real-time interviews between these models and InteLLA and analyzed the sessions in terms of reaction time, response frequency, and fluency metrics. The results showed that CEFR-conditioned full-duplex models reproduce levelwise trends observed in human interviews and exhibit partial distributional alignment with human interviews across multiple metrics. These findings indicate that full-duplex models can reproduce human dialogue features and provide a promising foundation for conversational AI systems.

In future work, we will improve model reproduction, including data extraction, and expand the metrics to more comprehensively assess full-duplex dialogue features. We further plan to extend the full-duplex model to multimodal dialogue by incorporating nonverbal cues such as facial expressions and gaze, aiming to achieve more human-like dialogues.

Acknowledgments

This research was supported by the project "Innovative Information and Communication Technology (Beyond 5G / 6G) Fund Program: Research on an Automatic Evaluation Infrastructure for Highly Reliable Multimodal Conversational AI Agents in the Beyond 5G Era (JPJ012368C-10301)" funded by the National Institute of Information and Communications Technology (NICT), and "Adaptable and Seamless Technology transfer Program through Target-driven R&D (A-STEP) / Development of a Conversational AI Agent Platform for Diagnostic Assessment and Learning Assistance (JPMJTT24J3)" by Japan Science and Technology Agency (JST). In addition, model training was conducted using the ABCI 3.0 system provided by the National Institute of Advanced Industrial Science and Technology (AIST) and AIST Solutions.

References

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. Audiolm: a language modeling approach to audio generation.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *Preprint*, arXiv:2305.09636.

Chen Chen, Ke Hu, Chao-Han Huck Yang, Ankita Pasad, Edresson Casanova, Weiqing Wang, Szu-Wei Fu, Jason Li, Zhehuai Chen, Jagadeesh Balam, and Boris Ginsburg. 2025. Reinforcement learning enhanced full-duplex spoken dialogue language models for conversational interactions. In *Second Conference on Language Modeling*.

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.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38:555–568.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung yi Lee. 2025. Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models. *Preprint*, arXiv:2507.23159.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J.F. Gales. 2025. Assessment of L2 Oral Proficiency using Speech Large Language Models. In *Proceedings of INTERSPEECH 2025*, pages 5078–5082.

Ryuki Matsuura, Shungo Suzuki, Kotaro Takizawa, Mao Saeki, and Yoichi Matsuyama. 2025. Gauging the validity of machine learning-based temporal feature annotation to measure fluency in speech automatically. *Research Methods in Applied Linguistics*, 4(1):100177.

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

Brian North and Enrica Piccardo. 2020. *Companion volume COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT Companion volume Language Policy Programme Education Policy Division Education Department Council of Europe*.

Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, and Ryuichiro Higashinaka. 2025. Towards a japanese full-duplex spoken dialogue system. In *Proceedings of INTERSPEECH 2025*, pages 1783–1787.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Mao Saeki, Hiroaki Takatsu, Fuma Kurata, Shungo Suzuki, Masaki Eguchi, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama. 2024. InteLLA: Intelligent language learning assistant for assessing language proficiency through interviews and roleplays. In *Proceedings of SIGDIAL 2024*, pages 385–399.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. Seanet: A multi-modal speech enhancement network. *Preprint*, arXiv:2009.02095.

Hiroaki Takatsu, Shungo Suzuki, Masaki Eguchi, Ryuki Matsuura, Mao Saeki, and Yoichi Matsuyama. 2025. Gnowsis: Multimodal multitask learning for oral proficiency assessments. *Computer Speech & Language*, page 101860.

Shravan Vasishth and Andrew Gelman. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, 59(5):1311–1342.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In *Proceedings of EMNLP 2024*, pages 21390–21402.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024. A full-duplex speech dialogue scheme based on large language model. In *Proceedings of NeurIPS 2024*.

Anne Wu, Laurent Mazaré, Neil Zeghidour, and Alexandre Défossez. 2025. Aligning spoken dialogue models from user interactions. In *Forty-second International Conference on Machine Learning*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *Preprint*, arXiv:2107.03312.