DiaFill: 短い発話でフィラーが豊富な音声対話台本の自動生成ツールキット

水本智也　小島淳嗣　藤田雄介　周藤唯
SB Intuitions 株式会社
{tomoya.mizumoto}@sbintuitions.co.jp

概要

音声対話は，テキスト対話とは異なり，フィラーや相槌，短い発話によって会話の流れが維持される．しかし，既存の対話データセットの多くは書き言葉のスタイルに偏っており，音声対話モデルの構築には適していない．この問題の解決のため，音声対話特有の特徴を豊富に含む日本語対話台本を生成するためのツールキットを提案する．フィラーや短い発話を含む実際の音声対話の書き起こしを用いて大規模言語モデルを訓練し，音声対話特有の特徴を持つ対話台本の自動生成を実現した．評価実験を通じ，DiaFill は汎用モデルと比較して，多くのターン数，短い発話，そしてフィラーを多く含む対話台本を生成できることを示し，音声対話研究の加速に貢献する実用的な基盤ツールであることを実証した．

1　はじめに

音声対話は，会話の流れを維持するために話し言葉特有の発話行為を不可欠としている点で，テキストチャットとは本質的に異なる．具体的には，フィラー（例.「えーっと」「あのー」），相槌（例.「うん」「はい」），そして短い発話（例.「わかった」「なるほど」）などが挙げられ，これらは円滑で協調的なコミュニケーションを支える役割を果たす．フィラーは単なるノイズではなく，発話の遅延を予告するシグナルとして機能し，話者が発話権を保持するのを助ける [1]．相槌は，話し手に対しリアルタイムなフィードバックを提供し，ターンの調整を促進する [2, 3]．さらに，短い発話は，会話の漸進的かつ協調的な性質に寄与している [4]．これらの特徴は，音声対話のリズムや自然さを形成するものであり，テキストチャットに見られる長く整ったターンとは明確に区別される．

図 1 に，同一の問い合わせタスクにおけるテキスト対話と音声対話の違いを示す．音声対話の例にはフィラー，相槌，短い発話が含まれており，より漸進的なやり取りとなっている．テキスト対話では情報が一度にまとめて要求されることが多いが，音声対話においては，情報はターンを細かく重ねながら段階的に引き出されるのが一般的である．

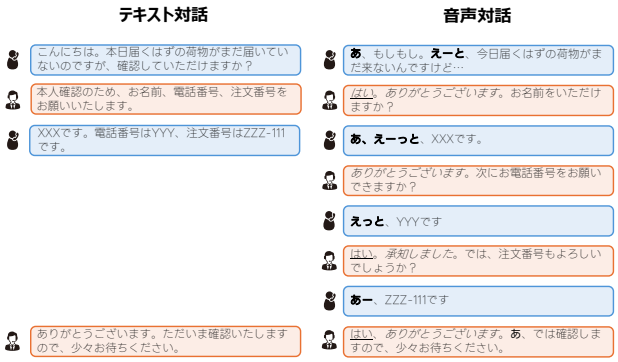

図 1: テキスト対話と音声対話の比較．太字はフィラー，下線は相槌，斜体は短い発話をそれぞれ表す．

しかし，このような特徴があるにもかかわらず，既存の対話データセットの多くは書き言葉のスタイルに偏重しており [5, 6, 7, 8]，フィラーや相槌，短い発話といった音声対話特有の要素が欠如している．その結果，これらのコーパスで学習されたモデルは，文脈的な一貫性はあるものの，実際の音声対話としては不自然な，書き言葉の応答を生成する傾向にある [9, 10]．自然な話し言葉による対話モデルを構築するためには，大規模な話し言葉スタイルの対話データが必要不可欠である．

音声対話データの構築に関する先行研究では，主に手作業によるコーパス作成が行われてきた [11, 12, 13]．これらのコーパスは自然な音声対話の現象を捉えているものの，音声収録や書き起こしに多大なコストを要するため，大規模でかつ多様性のあるデータの構築は困難である．この制約を克服するため，近年では大規模言語モデル（LLM）を用

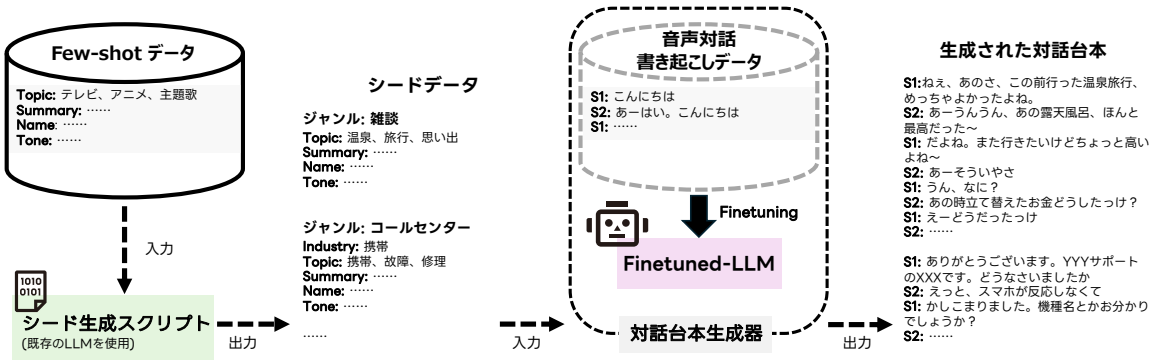

図 2: DiaFill の流れ．シード生成と対話生成の 2 段階で，音声対話特有の特徴を持つ対話台本を生成する．

いた対話データ生成も試みられている [14, 15, 16]．しかし，これらのアプローチはあくまで特定のデータセットを作成することに焦点を当てており，データ生成ツールの提供には至っておらず規模やドメインの多様性に制約が残る．また，DiaSynth [17] のような対話データ生成ツールキットも提案されているが，これらはフィラーや短い発話といった音声対話特有の特徴生成をターゲットとしていない．

音声対話データの不足という課題に対し，本研究では，音声対話特有のスタイルの日本語対話台本を生成するためのツールキット “**DiaFill**” を提案する．生成テキストをそのまま使うだけなく，音声合成や話者による読み上げをすることも想定しているため，これを「台本」と呼ぶ．DiaFill は雑談およびタスク指向対話の双方に対応しており，ユーザは任意のトピックや規模で音声対話データを構築できる．本ツールキットを用いることで，フィラーや短い発話のやり取りといった自然な音声対話の特徴を保持した大規模な対話台本を生成可能である．生成したデータを使うことで，LLM に音声対話特有の特徴を学習させることができ，より自然な発話ができるようになることが期待される．さらに音声合成ツールと組み合わせることで，生成された対話は音声対話モデルの構築に直接利用できる．

2 DiaFill ツールキット

DiaFill は，フィラーや短い発話といった音声対話の特徴を含む日本語対話台本を生成するためのツールキットである．図 2 に示すように，本ツールキットは「シード生成」と「対話台本生成」の 2 段階のパイプラインで構成される．

メインである対話台本生成器は，トピックや口調などの基本条件を指定したシードデータを入力とし，複数ターンの対話台本を出力する．このモデルは，フィラーを豊富に含む実際の音声対話データを用いてファインチューニングされており，音声対話独自のスタイルを再現可能である．

ユーザが自分でシードデータを用意する手間を省くため，シードデータを自動生成する補助モジュールも提供する．また，特定の意図を持った対話を生成したい場合は，それに合わせたシードデータを用意することで，ユーザに合わせた対話台本の生成が可能である．音声対話台本生成モデルは Hugging Face [1] で公開している．シード生成用スクリプト，および対話台本生成用のサンプルスクリプトは GitHub [2] で公開している．

2.1 対話台本生成

本節では，短い発話かつフィラーを豊富に含む対話を生成するモデルについて述べる．本生成器は，フィラーや短い発話を含む対話データを用いて LLM を教師ありファインチューニング（SFT）することで構築した．対話データセットとしては，NEDO の「生成 AI 開発加速に向けた新たなデータセットの構築に関する調査」 [3] プロジェクトにおいて作られたデータセットを使用した．学習時には，対話内容と話者情報などをシードデータとして与え，対応する音声対話台本を生成するようにモデルを訓練することで，音声対話特有のリズムや漸進的な特徴の再現を可能にしている．

対話生成の条件となるシードデータは，主に以下の 2 つの要素から構成される．一つ目は，対話内容であり，ジャンル，業界，トピック，および概要が含まれる．ジャンルは「雑談」または「コールセンター」を指定し，コールセンターの場合はさらに

1) https://huggingface.co/collections/sbintuitions/diafill
2) https://github.com/sbintuitions/diafill-toolkit
3) https://www.nedo.go.jp/koubo/IT3_100322.html

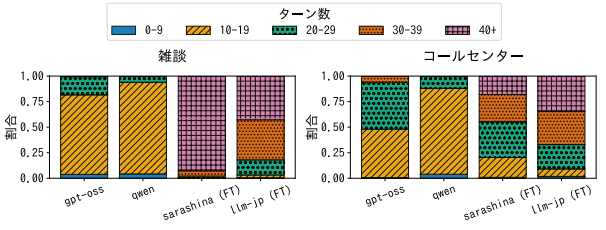

図 3: モデルおよびジャンルごとのターン数の分布．各棒グラフは，異なるターン数に含まれる対話の割合を表す．

「金融」「物流」などの具体的な業界カテゴリを指定する．トピックは 3 つのキーワードからなり，概要にはそれらに関連する簡潔な説明を記述する．二つ目は，話者情報であり各話者の名前と口調が含まれる．多種多様なシードを手動で大量に作成することは大変なため，次節で述べるシードを自動で生成するためのスクリプトも提供する．

音声対話台本生成モデルは，日本語の指示チューニング済みモデルをベースに，音声対話データの書き起こしを用い，標準的な教師ありファインチューニング設定で訓練した．パラメータサイズが小さく取り回しが良い点，多様な用途での利用が可能であるライセンスにするため，ベースモデルには “sarashina2.2-3b-instruct-v0.1” および “llm-jp-3-13b-instruct3” を採用した．訓練データには，NEDO のプロジェクトで構築された音声対話データを使用した．このデータは，Zoom を通じて収録され，自動で書き起こしし，人手修正がされており，訓練には約 20,000 件を使用した．10 種類の業界カテゴリを模したコールセンター対話が約 8,000 件，雑談対話が約 12,000 件からなる．元のデータには，話者名などの生成のシードとなる情報は付与されていないため，Qwen2.5-32B-Instruct を用いて対話テキストから自動で抽出した．

2.2 シードデータの自動生成

効率的な対話台本生成を支援するため，対話台本生成器の入力となるシードデータを自動生成する補助モジュールも提供する．本モジュールは，前節で説明した対話台本生成モデルの入力となるメタデータを自動的に作成する．各シードは，指定されたジャンルに基づいて生成される．具体的な処理としては，汎用的な指示チューニング済み LLM に対し，属性情報と few-shot の事例を与えることで，トピック，概要，話者情報を生成させる．

生成プロセスは以下の手順で行われる．まず，モデルは与えられたジャンルや業界，few-shot 事例を入力として，3 つのキーワードからなるトピックセットを生成し，続いて 2 つの話者名を生成する．次に，生成されたトピックセット，そのトピックセットを生成した時と同じ情報を入力として，対話の概要と各話者の口調を生成する．生成されるデータの多様性を担保するため，重複したトピックセットは再生成し，また過去生成された直近 100 件の履歴を参照し重複を避けるようプロンプトに加えた．

3 対話台本生成モデルの評価

本節では，2.1 で述べた対話台本生成モデルの評価を行う．評価用シードデータの生成は，2.2 節で述べた方法を用いた．雑談，コールセンタージャンルの評価それぞれ行った．コールセンタージャンルでは，実際の利用場面に即して訓練データにない業界カテゴリ 10 個を用意した（詳細は付録 A 参照）．[4] シードデータ生成には，LLM として Qwen2.5-32B-Instruct を用い，雑談・コールセンターそれぞれ 10,000 件のシードを生成した．各シードは 3 つのトピックを含むため，雑談・コールセンターごとに 30,000 件のトピックが得られ，ユニークトピック数は雑談で 3,421 件，コールセンターで 4,699 件であった．制約の緩い雑談ドメインではトピックの重複が多く見られた．

対話台本生成の実験では，計算コストを考慮し各設定からランダムにサンプリングした 1,000 件のシードを使用した．比較対象として，2.1 節で述べたファインチューニング済みモデル（以下 sarashina (FT), llm-jp (FT)）に加え，オープンなベースラインモデルとして gpt-oss-20b および Qwen2.5-32B-Instruct（以下 gpt-oss, qwen）を用いた．ベースラインモデルには，提案モデルと同一のシード情報をプロンプトとして与えた．なお，生成時の最大長は 1,000 トークンとし，同じ文字列の繰り返しで終了した出力の場合は再生成を行った．

3.1 対話台本生成の構造的特徴の分析

生成された対話の構造的特徴を明らかにするため，(1) ターン数，および (2) 平均発話長（文字数）の 2 点を分析した．ターン数は対話の双方向性を反映し，平均発話長は発話の簡潔さとターン交代のスタイルを示す指標となる．

4) 学習データに含まれる業界での評価結果は付録 B に示す．

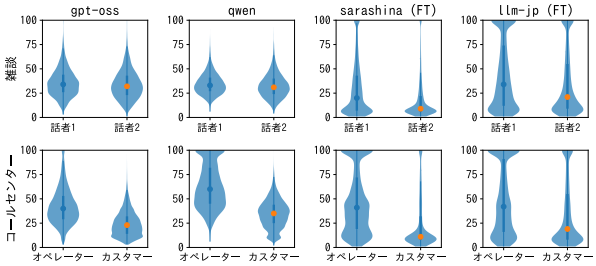

図 4: モデル，ジャンル，および話者役割ごとの発話長の分布．バイオリンプロットは発話長（文字数）のばらつきを示しており，極端な外れ値を抑制するため 100 文字で winsorization 処理を行った．

表 1: ドメインごとの自動評価結果．

モデル	Top.	Sum.	Tone	Fil.	Nat.
雑談					
gpt-oss	1.00	.937	.963	.080	.888
qwen	1.00	.968	.983	.447	.855
sarashina (FT)	.967	.279	.655	.991	.843
llm-jp (FT)	.923	.269	.420	.998	.791
コールセンター					
gpt-oss	1.00	.983	.959	.135	.949
qwen	1.00	.986	.997	.438	.987
sarashina (FT)	.995	.500	.922	.982	.698
llm-jp (FT)	.960	.453	.704	.980	.344

図 3 に示すように，ファインチューニングしたモデル（sarashina (FT), llm-jp (FT)）は 30 ターン以上の対話をする傾向があった．一方，ベースラインモデル（gpt-oss, qwen）は 20 ターン未満で終了するものがほとんどであった．この傾向はジャンルを問わず一貫しており，構築したモデルがよりインタラクティブなやり取りを生成できることを示している．

図 4 に発話長の分布を示す．ファインチューニングしたモデルは，20 文字以下の短い発話を生成する傾向がある．一方で，ベースラインモデルは長い発話を生成しており，書き言葉的な傾向が強い．

3.2 対話台本生成の品質の評価

次に，LLM-as-a-judge フレームワーク [18] を用いて対話品質を評価する．評価モデルには gpt-oss-120b を使用し，指定されたトピックが使えているか (Top.)，指定された概要と整合しているか (Sum.)，指定された口調か (Tone)，音声対話としてフィラーが適切か (Fil.)，全体的な対話の自然さ (Nat.) の 5 つの観点で評価した評価は各観点について適切か否かの二値判定（0/1）で行い，適切とされた割合を精度とする．トピックは，指定された 3 つのうち 1 つでも使えていれば適切と判断した．

sarashina (FT)
S1: えーとー最近あった楽しいことは何ですか？
S2: はい。えーっと、あのー手作りキャンディを作って、そのーま子供とかに配ったりしたんですけど
S1: あーそうなんですね。

qwen
S1: 手作りのキャンディを作ったって言ってたよね？
S2: はい、そうです。手作りキャンディを作りました。
S1: 美味しかったよ！特にチョコレート味のやつ

図 5: トピック「飴作り」における sarashina (FT) と qwen の出力例．下線はフィラーを示す．

表 1 に評価結果を示す．ベースラインモデルはトピックや概要において高いスコアを示し，プロンプトの指示に忠実であることが確認できた．一方で，フィラーのスコアは著しく低く，音声対話特有のフィラーをほとんど生成できていない．対照的に，ファインチューニング済みモデルは全てのジャンルにおいて 0.98 を超える極めて高いフィラー品質を達成した．しかし，概要に関するスコアはベースラインよりも低い．この主な要因は，学習データの概要が自動抽出されたものであり，必ずしも対話内容と完全に一致していないことが影響したと考えられる．ただし，トピックスコアは十分な精度を示しており，実用する上では大きな問題はないと考える．

図 5 に生成例を示す．ベースライン (qwen) がフィラーを一切用いていないのに対し，sarashina (FT) は「えーっと」「あー」などのフィラーを自然なタイミングで挿入している．また，「あーそうなんですね」のように，関心を示しつつ発話権を返すような，音声対話特有の表現も観察された．

4 おわりに

本稿では，音声対話特有の特徴を再現できる大規模な日本語対話台本生成ツールキット “DiaFill” について詳述した．実際の音声対話データを用いて大規模言語モデルをファインチューニングすることで，豊富なフィラー，短い発話，および漸進的なターン交代のパターンを捉えた台本生成を実現した．評価実験の結果，DiaFill によって生成された対話は，汎用の LLM が生成するものと比較して，音声対話としての性質を強く有していることが示された．今後は，生成された台本に音声合成で音声を付与し，音声対話モデリングや対話エージェントの学習といった下流タスクでの有効性を検証する予定である．

参考文献

[1] Herbert H. Clark and Jean E. Fox Tree. Using uh and um in spontaneous speaking. **Cognition**, Vol. 84, No. 1, pp. 73–111, 2002.

[2] Victor H Yngve. On getting a word in edgewise. In **Papers from the sixth regional meeting Chicago Linguistic Society**, pp. 567–578, 1970.

[3] Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. **Journal of Pragmatics**, Vol. 32, No. 8, pp. 1177–1207, 2000.

[4] Herbert H Clark and Edward F Schaefer. Contributing to Discourse. **Cognitive science**, Vol. 13, No. 2, pp. 259–294, 1989.

[5] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical Analysis of Training Strategies of Transformer-based Japanese Chit-chat Systems, 2021.

[6] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing**, pp. 986–995, 2017.

[7] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 2204–2213, 2018.

[8] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会 第 29 回年次大会発表論文集, pp. 108–113, 2023.

[9] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 15757–15773, December 2023.

[10] Yinghao Aaron Li, Xilin Jiang, Jordan Darefsky, Ge Zhu, and Nima Mesgarani. Styletalker: Finetuning audio language model and style-based text-to-speech model for fast spoken dialogue generation. In **Proceedings of First Conference on Language Modeling**, 2024.

[11] Ge Zhu, Juan-Pablo Caceres, and Justin Salamon. Filler Word Detection and Classification: A Dataset and Benchmark. In **Proceedings of Interspeech 2023**, 2022.

[12] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 39088–39118, 2023.

[13] Kikuo Maekawa. Corpus of spontaneous japanese: its design and evaluation. In **Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition**, pp. 7–12, 2003.

[14] Sehun Lee, Kang-wook Kim, and Gunhee Kim. Behavior-SD: Behaviorally Aware Spoken Dialogue Generation with Large Language Models. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 9574–9593, Albuquerque, New Mexico, 2025.

[15] Alkis Koudounas, Moreno La Quatra, and Elena Baralis. DeepDialogue: A Multi-Turn Emotionally-Rich Spoken Dialogue Dataset, 2025.

[16] Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 6626–6642, 2024.

[17] Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications. In **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 673–690, 2025.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023.

訓練時の業界カテゴリ
アパレルEC　自治体　生命保険　証券　テレビショッピング　化粧品　健康食品　旅行代理店　家電　通信事業者

評価用業界カテゴリ
銀行　電力　ガス　クレジットカード　宅配便　不動産　家電量販店　ドラッグストア　自動車保険　スーパーマーケット

図 6: コールセンターの業界カテゴリ．訓練に用いた業界と評価用に用いた業界

表 2: コールセンタージャンルで訓練に用いた業界を評価で使用した場合の性能.

モデル	Top.	Sum.	Tone	Fil.	Nat.
gpt-oss	1.00	.981	.956	.135	.936
qwen	1.00	.979	.997	.471	.977
sarashina (FT)	.997	.561	.937	.988	.741
llm-jp (FT)	.970	.443	.832	.984	.503

A　コールセンタージャンルの業界カテゴリ

図 6 に本文中で訓練と評価に用いた業界カテゴリを示す．一部「生命保険」と「自動車保険」など訓練と評価で似た業界もあるが，基本的にはそれぞれ独立した業界になっている．

B　訓練データに含まれる業界カテゴリを用いた場合の評価結果

表 2 に訓練データに含まれる業界を用いた場合の評価結果を示す．表 1 と比較し，ファインチューニングしたモデルの自然さのスコアは，訓練データに含まれる業界の方が高くなっており，訓練時の業界の方がより自然な対話を生成できていることがわかる．一方で，他のスコアに関しては訓練時にない業界を使った場合も，訓練時に含まれる業界の場合のスコアと遜色がない．