応答内容・順序に着目した音声対話ベンチマークの構築

渡邉一功 [1]　水本智也 [2]　周藤唯 [2]　河原大輔 [1]
[1] 早稲田大学　[2]SB Intuitions 株式会社
{ittsu120219, dkw}@waseda.jp
{yui.sudo, tomoya.mizumoto}@sbintuitions.co.jp

概要

本研究では，音声対話モデルが生成する応答内容の適切性を評価するため，対話行為に基づく音声対話ベンチマークを構築する．既存の音声対話ベンチマークは音声認識精度や音響的ロバスト性といった部分的能力に焦点を当てており，応答内容・順序を評価する枠組みは十分に整備されていなかった．本研究では，音声対話モデルの生成応答をテキストとして抽出し，対話文脈に基づく最適応答を対話行為系列として定義し評価する．その結果，音声対話モデルの性能を定量化し，最適な返答パターンとの乖離度や乖離の原因（冗長性，内容の不足など）を捉えることができた．

1　はじめに

近年，音声対話エージェントの高度化に伴い，その性能を適切に測定するための多様な音声対話ベンチマークが提案されてきた [1, 2, 3]．これらのベンチマークは，音声認識精度，環境音理解，話者変動へのロバスト性，あるいは応答タイミングといった観点から，音声対話モデルの能力を定量的に比較する上で重要な役割を果たしている．

一方で，音声対話は単なる音声認識や音響的ロバスト性の問題ではなく，ユーザー意図の推定，状況依存的な判断など，複数の要素が統合されて成立する複合的なタスクである．そのため，実世界で有用な音声対話エージェントを評価するには，生成された応答が文脈に照らして適切な順序で適切な内容の返答をできているかという内容的妥当性を評価することが必要不可欠である．しかし，既存の音声対話ベンチマークでは，このような内容的妥当性を体系的に評価する枠組みは十分に整備されていない．

本研究では，音声対話モデルの評価で一般に扱われる音響的品質（発話の明瞭性，プロソディ，雑音耐性など）には着目せず，対話として生成される応答内容の適切性を評価するための新たなベンチマークを構築する．具体的には，音声対話モデルが生成した応答をテキストとして抽出し，その内容がユーザーの要求や対話文脈に対して意味的に適切であるかを対話行為（Dialogue Act）系列の順序・要素の一致度から評価する．

2　関連研究

2.1　音声対話モデルの評価指標

近年，音声対話モデルの性能評価のベンチマークが数多く提案されている．AudioBench [1] は Question-Answering の形式を中心に環境音や対話の理解を評価し，VoiceBench [2] は 5 段階評価の LLM-as-a-Judge を中心に全般的な対話性能を評価している．また Full-DuplexBench [3] は，応答開始タイミングや割り込み処理など，リアルタイム対話におけるインタラクション能力に焦点を当てている．さらに，指示追従能力（Instruction Following）[4]，マルチターンの対話整合性 [5] といった観点からの評価も提案されている．

これらは音声対話モデルの要素的能力を評価する上で有用である一方で，Question-Answering の形式では実用上の対話の能力は測定できず，LLM-as-a-Judge での 5 段階評価では評価基準が曖昧になってしまう．最適な応答は応答の内容とその順序で決まると抽象化した時，そのような応答の内容・順序を直接評価する枠組みは限定的である．

2.2　対話行為

対話行為（Dialogue Act）は，発話が対話中で果たす機能を「Question」「Answer」などのカテゴリとして表現する枠組みであり，対話の意図や構造を記述・分析するために広く用いられている [6]．本研究では，生成応答を対話行為系列として表現することで，応答内容に過不足がないかを解釈可能な単位

で診断する．すなわち，従来のベンチマークでは捉えにくい「内容的妥当性」を，応答の機能構造に基づいて直接比較することが可能となる．

一方で，対話行為体系には粒度の異なる設計が存在し，注釈の再現性の観点から一概に細かい分類が良いとは言い難い．本研究では，この点を踏まえた上で評価ベンチマーク向けの対話行為体系を設計する（詳細は 3.2 節）．

2.3 本研究の位置付け

本研究は，生成応答の**内容的妥当性**を直接評価する枠組みを設計する．具体的には，応答を対話行為系列として構造化し，対話文脈に対する適切性を定量的に評価するベンチマークを構築する．これにより，音声対話モデルの応答の最適性を一貫した基準で比較可能とする点に本研究の新規性がある．

3 ベンチマークの設計

3.1 評価対象

本研究では，音声対話モデル／音声対話システム [7, 8, 9] が生成する応答内容の評価対象を，音声信号そのものではなく，**応答テキスト**に限定する．一方で，本ベンチマークにおける**モデルへの入力は音声**であり，評価対象モデルは，音声入力に基づいて応答を生成する．

音声対話モデルや音声対話システムは，音声出力を生成する前段階として，内部的に応答をテキストとして生成（あるいはテキスト表現を保持）している場合も少なくない．そのため，本研究では，各モデルからこの応答テキストを取得できることを前提とし，取得したテキストに対して対話行為系列により応答の適切性を評価する．

3.2 対話行為の定義

対話行為は，発話が対話の中で果たす機能（質問，回答，依頼など）を明示化する枠組みであり，対話の構造や意図を分析・評価する基盤として広く用いられている．代表的な体系として，国際標準である ISO 24617-2 [6] が知られている．ISO 24617-2 は，対話の多層的性質を精緻に記述できる一方で，分類体系が 15 個近くあり非常に細かく，注釈設計・アノテーションのコストが高くなりやすい．

本研究では，応答を複数の対話行為の系列（sequence）として表現する．具体的には，応答を対話行為の機能的要素に分解し，各要素を (i) 出現順序 と (ii) 要素の重要度（importance） を伴って記述する．これにより，「応答の中で何が，どれくらい重要な要素として，どの順番で現れているか」を明示的に表現でき，複数モデル間での比較や誤り分析を行いやすくなる．

対話行為の要素としてはメールコーパスにおける対話行為注釈で用いられている大分類 [10] を採用する．具体的には，**Question**（質問），**Check-Question**（確認質問），**Answer**（回答），**Inform**（情報提供），**Request**（依頼），**Suggestion**（提案），**Commissive**（約束・承諾）の 7 種類を基本ラベルとして定義する．この大分類は，対話の機能を十分に表現しつつも粒度が適度であり，注釈の再現性と運用可能性を両立できる点で，本研究の目的に適している．

3.3 データセット構築

本研究では，スマートスピーカーおよびスマートフォン上の音声アシスタント利用を想定し，単一ターンの対話データ（ユーザー発話とアシスタント応答）を合成的に構築する．まず，各シナリオの状況を仮定した上で，ユーザーの発話（input）とアシスタントのサンプル応答（output）を大規模言語モデル（LLM）に自然言語で生成させる．このとき，生成結果には重要度付き対話行為系列（output_dialogue_acts）も付与し，以降の評価で利用できる形式に統一する．対話行為系列とは，生成された応答文を先頭から順に構成要素へ分解し各要素に重要度を付与した列である．LLM に与えた詳細なプロンプト設計（システム指示，分類体系，few-shot 例，出力スキーマの制約）については，付録 A に記載する．

次に，生成したユーザー発話およびアシスタント応答のテキストを，音声入出力を伴う対話環境を模擬するために Tsukasa-Speech [11] を用いて音声化する．音声データは評価実験で一貫した条件となるよう，サンプリング周波数 24 kHz，量子化ビット数 16 bit，モノラルチャネルとし，音声開始前後にそれぞれ 120 ms の無音区間を付与した．以上により，テキストベースの対話内容と，音声入出力を模擬した音声データの両方を備えたデータセットを構築した．

4 評価

4.1 評価指標

本研究では，音声対話モデル / システムの応答に対してもデータセット生成に用いたものと同じ LLM を用いて対話行為系列を生成し，正解応答の対話行為系列と生成応答の対話行為系列の類似度を定量的に比較する．評価指標としては **Weighted LCS** と **重み付き対話行為編集距離（Weighted Dialogue Act Edit Distance; WED）** を用いる．Weighted LCS は, 共通する部分系列（Longest Common Subsequence; LCS）がどれだけ重要であるかを，対話行為の重要度に基づいて定量化する指標である．重み付き対話行為編集距離は正解の対話行為系列との一致 / 差異のパターンを定量化する指標である．

4.1.1 Weighted LCS

正解（Ground Truth; GT）の対話行為系列と，モデル生成応答の対話行為系列の間で，両者に共通して出現する部分系列（subsequence）のうち，**モデル生成側の重要度の合計**が最大となるものを抽出し，その合計値をスコアとして用いる．これは生成側が“重要だと思った要素”をどれだけ含めたかを測るためである．例えば，

$$\begin{aligned}\text{GT} &= [(\text{Question}, 0.7), (\text{Suggestion}, 0.1), (\text{Inform}, 0.2)],\\ \text{Pred} &= [(\text{Question}, 0.7), (\text{Suggestion}, 0.15),\\ &\qquad (\text{Answer}, 0.05)]\end{aligned}$$

のとき，共通部分系列の候補の中で重要度合計が最大となる $[(\text{Question}, 0.7), (\text{Suggestion}, 0.15)]$ が選択され，Weighted LCS は $0.7 + 0.15 = 0.85$ となる．

4.1.2 重み付き対話行為編集距離 (WED)

正解の対話行為系列と予測された対話行為系列を，それぞれ次のように定義する。

$$R = \left[(r_1, w_1^r), (r_2, w_2^r), \ldots, (r_n, w_n^r)\right], \tag{1}$$

$$P = \left[(p_1, w_1^p), (p_2, w_2^p), \ldots, (p_m, w_m^p)\right], \tag{2}$$

ここで，r_i および p_j は対話行為ラベルを表し，$w_i^r, w_j^p \in [0, 1]$ はそれぞれの対話行為に対応する重要度を表す。重要度は，以下の条件を満たすように正規化されているものとする。

$$\sum_{i=1}^{n} w_i^r = 1, \qquad \sum_{j=1}^{m} w_j^p = 1. \tag{3}$$

編集操作とコスト　本研究では，以下の 3 種類の編集操作に基づいて重み付き編集距離を定義する。ただし，予測系列 P を正解系列 R に変換する編集操作を考える．また，$p_i = r_j$ の場合の置換コストに関しては重要度があくまで各系列内で正規化された指標であることから，w_i^p, w_j^r の値に関わらず 0 と定義している．

- **削除（Deletion）**：予測系列中の対話行為 p_i を削除する操作

$$\text{cost}_{\text{del}}(p_i) = w_i^p. \tag{4}$$

- **挿入（Insertion）**：正解系列中の対話行為 r_j を挿入する操作

$$\text{cost}_{\text{ins}}(r_j) = w_j^r. \tag{5}$$

- **置換（Substitution）**：予測の対話行為 p_i を正解の対話行為 r_j に置き換える操作

$$\text{cost}_{\text{sub}}(p_i, r_j) = \begin{cases} 0, & p_i = r_j \text{ の場合}, \\ \dfrac{w_i^p + w_j^r}{2}, & p_i \neq r_j \text{ の場合}. \end{cases} \tag{6}$$

最終スコア　予測系列 P を正解系列 R に変換するために必要な編集操作（削除・挿入・置換）のうち，**総コストが最小となる編集系列**を考える．このとき，その最小総コストを正解系列 R と予測系列 P の間の重み付き対話行為編集距離として定義する．

$$\text{WED}(P, R) = \min_{\mathscr{E} \in \mathscr{A}(P,R)} \sum_{e \in \mathscr{E}} \text{cost}(e), \tag{7}$$

ここで $\mathscr{A}(P, R)$ は，P を R に変換可能なすべての編集操作系列の集合を表し，$\text{cost}(e)$ は各編集操作 e に対応するコストである．

4.2 実験設定

本研究では，データセット生成および対話行為アノテーションを，(1) gemini-2.5-flash により作成したデータセット，(2) gpt-5-nano により作成したデータセット，の 2 種類で行った．各データセットに対して，応答生成モデルとして gemini-2.5-flash と gpt-4o-audio-preview の 2 モデルを評価し，Weighted LCS および重み付き対話行為編集距離を算出した．

モデルへの入力は 付録 B に記載した．評価対象のモデルには入力**音声**を与えるとともに，推論時には**システムプロンプト（テキスト）**を併用し，音声入力とテキスト指示に基づいて生成された**応答テキスト**を取得して評価した．

表 1　データセット生成モデルおよび応答生成モデル別の評価結果

生成データセット	応答生成モデル	Weighted LCS (↑)		WED (↓)		WED 内訳 (Mean)		
		平均	標準偏差	平均	標準偏差	削除	挿入	置換
gemini-2.5-flash	gemini-2.5-flash	**0.662**	0.360	**0.549**	0.353	0.198	0.083	**0.268**
	gpt-4o-audio-preview	0.611	0.382	0.575	0.362	**0.168**	**0.043**	0.364
gpt-5-nano	gemini-2.5-flash	**0.654**	0.370	**0.561**	0.394	**0.183**	**0.055**	0.324
	gpt-4o-audio-preview	0.597	0.363	0.661	0.377	0.222	0.133	**0.306**

表 2　編集距離特性別の代表例

ID	概要	入力テキスト	正解応答テキスト	正解 DA	生成応答テキスト	生成 DA	応答生成モデル
1	挿入コスト最大 (=0.85)	10 分のタイマーをセットして	はい、10 分のタイマーをセットしました。名前を付けますか？（例：「パスタ」）時間になったらお知らせします。	[(Commissive, 0.60), (Question, 0.25), (Inform, 0.15)]	はい、10 分のタイマーをセットしました。	[(Inform, 1.00)]	Gemini
2	削除コスト最大 (=0.7)	留守番電話に伝言を残して。	伝言の宛先と内容を教えてください。録音を開始して伝言を残します。	[(Question, 0.65), (Inform, 0.15), (Commissive, 0.20)]	それでは、どんな内容の伝言を残しますか？お相手の名前や伝言したい内容を教えてください。	[(Request, 0.70), (Question, 0.30)]	GPT
3	置換コスト最大 (=1.0)	12 月 25 日にクリスマスパーティーの予定を追加して。	はい、12 月 25 日に「クリスマスパーティー」の予定を追加しました。何時からにしますか？場所も登録しますか？	[(Commissive, 0.60), (Question, 0.20), (Question, 0.20)]	はい、12 月 25 日にクリスマスパーティーの予定ですね。カレンダーに登録しておきました。時間は後からでも変更できますよ。	[(Check-Question, 0.30), (Inform, 0.50), (Inform, 0.20)]	GPT
4	編集距離最小 (=0.0)	設定済みのリマインダーを教えて。	現在、設定されているリマインダーは 3 件です。明日の午前 9 時には「ゴミ出し」、来週月曜日の午後 3 時には「歯医者の予約」があります。残りのリマインダーも確認しますか？	[(Inform, 0.65), (Question, 0.35)]	「明日の朝 8 時に会議」「午後 3 時に買い物」「夜 7 時に友達との電話」が設定されています。何か変更したいことはありますか？	[(Inform, 0.75), (Question, 0.25)]	Gemini

4.3　結果と考察

結果を表 1 に示す．また，編集距離特性ごとの代表例を表 2 に示す．

表 1 より，gemini-2.5-flash は gpt-4o-audio-preview と比較して，Weighted LCS が高く，WED が低い傾向を示した．

表 2 に示す具体例から各コストの内訳の特徴が把握できる，挿入コストは正解応答に含まれる重要な対話行為の欠落を意味しており，表 2 の ID 1 の場合を例に取ると正解 DA において重要な (Question, 0.60) の欠落を意味している．削除コストは生成応答における対話行為の冗長さを意味しており，ID 2 の場合を例に取ると生成 DA における (Request, 0.70) を指している．置換コストは似たような重要度の対話行為の相違を表しており，ID 3 の例に表れている．一方，編集距離が最小となる例では，ID 4 のように生成応答と正解応答の対話行為構成が一致している．

上記の事実を踏まえると概して gemini-2.5-flash の方が応答生成として優れているが gpt-5-nano のデータにおいては gpt-4o-audio-preview の性能は他と比べて不安定で特に挿入コストが大きいことから他のモデルに比べて重要な対話行為の欠落が多いことがわかる．

これらの結果より，提案指標は単なる文字列類似度では捉えにくい対話構造上の妥当性や冗長性を定量的に捉えており，本ベンチマークが内容的妥当性の評価に有効であることが示唆される．

5　結論と今後の課題

本研究では，音声対話モデルが生成する応答テキストを対象に，対話行為系列へ変換して評価するベンチマークを構築した．具体的には，単一ターンの音声アシスタント対話データを合成的に作成し，生成応答に付与した対話行為系列の類似度を Weighted LCS および重み付き対話行為編集距離により定量化した．その結果，複数モデル間の性能差や傾向を一貫した枠組みで捉えられることを示し，対話としての内容的妥当性を評価するための基盤を提供した．

今後は，(i) 返答内容のパターンの評価には表れないより詳細な内容の評価手法を確立する．(ii) 本研究で対象外とした音響的品質，応答タイミング，対話の自然さ，安全性なども含めた**総合的な音声対話評価ベンチマーク**として拡張し，実運用に近い条件での比較可能性を高める．(iii) 最後に，自動評価の妥当性を担保するため，**人手評価（主観評価・専門家評価）との比較**を行い，指標が捉えている品質特性や限界を明確化する．これらを通じて，音声対話モデルの実用的な改善に資する，より信頼性の高い評価基盤の確立を目指す．

謝辞

本研究は SB Intuitions 株式会社と早稲田大学の共同研究により実施した．実験には東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用した．

参考文献

[1] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models, 2025.

[2] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024.

[3] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities, 2025.

[4] Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. Ifeval-audio: Benchmarking instruction-following capability in audio-based large language models, 2025.

[5] Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction, 2025.

[6] Harry Bunt, Jan Alexandersson, Jean Carletta, Alex Chengyu Fang, Kees van Kuppevelt, Volha Petukhova, Andrei Popescu-Belis, and David Traum. Iso 24617-2: A semantically-based standard for dialogue act annotation. Technical report, International Organization for Standardization, 2012.

[7] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind, 2025. Accessed: 2026-01.

[8] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2025.

[9] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025.

[10] 上田良寛, 谷口友紀, 三浦康秀, 大熊智子. Dialogue act 情報を付与したメールコーパス. 言語処理学会 第 25 回年次大会 発表論文集, pp. 542–545. 言語処理学会, March 2019.

[11] Respaired. Tsukasa speech: A frontier japanese speech generation network. https://github.com/Respaired/Tsukasa-Speech, 2025. Accessed: 2025-12-27; Open-source text-to-speech system supporting expressive Japanese speech synthesis.

A　データセット生成に用いたプロンプト

単一ターンの音声アシスタント対話データを生成する際にモデルへ与えたプロンプト（日本語）を示す．本プロンプトは，(i) システム指示，(ii) 対話行為分類，(iii) few-shot 例，(iv) タスク指示から構成される．出力は JSON のみとし，指定されたキーを必ず含める．

A.1　システム指示（System Instruction）

```
あなたはスマートフォン／スマートスピーカー向けの音声アシスタント対話データ
（単一ターン）を生成します。


ルール:
1) 「ユーザーの発話（input）」は 1 つ，
   「アシスタントの発話（output）」も 1 つの単一ターンとする。
2) アシスタントの発話は単一メッセージ内で最適行動をとる。
3) 出力は JSON のみとし，マークダウンやコメントは禁止する。
4) 言語は日本語（ja-JP）とし，簡潔で自然な表現を用いる。
5) 実運用上，文脈として自明な情報は適切に仮定する。

  例:
  ・「今日の天気は？」と聞かれた場合，
    端末の位置情報などからユーザーが東京にいると仮定してよい。
  ・「迂回ルートを検索して」と言われた場合，
    すでにナビゲーションが開始されていると仮定する。
  ・「3 分タイマーかけて」と言われた場合，
    料理中などの自然な状況を仮定してよい。


6) JSON は以下の 4 キーを必須とする:
  ・scenario_context
  ・input
  ・output
  ・output_dialogue_acts

7) scenario_context.background には，
    対話が行われている状況（ユーザーの状態，端末の状態，
    暗黙の前提）を 1〜3 文程度で記述する。

8) output_dialogue_acts.steps の各要素は，
    act，importance（[0,1)），rationale（任意）を必ず含める。
```

A.2　対話行為分類

```
対話行為の分類:
・Question（質問）: 情報を求める発話
・Check-Question（確認質問）: 想定した答えの確認を求める発話
・Answer（回答）: 質問への直接的な返答
・Inform（情報提供）: 回答以外の補足情報や通知
・Request（依頼）: 相手の行動を求める発話
・Suggestion（提案）: 利益につながる提案
・Commissive（約束・承諾）: 宣言・受諾・拒絶


importance（重要度）:
・各 step に importance を付与する。
・importance は [0,1) の小数とし，steps 全体の合計は 1 とする。
・返答の中心となる行動ほど高く，補助的な行動ほど低く設定する。
```

A.3　Few-shot 例

```
{
  "scenario_context": {
    "background": ユーザーは自宅のリビングにいて、スマホの音声アシスタントに『今日の東京の天気は？』と尋ねている。位置情報からユーザーが東京都内にいることは端末が把握している。""
  },
  "input": {
    "role": "USER",
    "text": 今日の東京の天気は？"",
    "locale": "ja-JP"
  },
  "output": {
    "role": "ASSISTANT",
    "text": 東京は本日、晴れのち曇り、最高気温は度です。週末の天気もお調べしますか？"25"
  },
  "output_dialogue_acts": {
    "steps": [
      {
        "act": "Answer",
        "importance": 0.78,
        "rationale": 本日の天気を簡潔に伝える。""
      },
      {
        "act": "Suggestion",
        "importance": 0.22,
        "rationale": 週末予報を提案し付加価値を与える。""
      }
    ]
  }
}
```

A.4　タスク指示（Task）

```
タスク:
・上記スキーマに厳密に従い，単一ターンのサンプルを 1 件生成する。
・ユーザー発話はスマートフォン／スマートスピーカーの
  音声アシスタント利用として自然であること。
・アシスタント応答は，Dialogue Act 分類に基づく
  最適行動を 1 メッセージで表現すること。


指示:
{USER_INSTRUCTION}


出力は JSON のみとし，
必ず scenario_context，input，output，
output_dialogue_acts の各キーを含めること。
```

B　応答生成に用いたプロンプト

評価実験において音声対話モデルに与えたシステムプロンプトおよびユーザープロンプト（テキスト）を示す．モデルには入力音声を与えるとともに，以下のテキスト指示を併用して応答テキストを生成させた．

B.1　システムプロンプト

```
あなたは親切な日本語音声対話アシスタントです。
```

B.2　ユーザープロンプト

```
この音声の内容を理解し、音声対話として適切な返答をしてください。
出力には返答文以外の一切のテキストを含めず、返答文のみを返してください。
```