LLM を用いた音声対話による 接客訓練・半構造化インタビュー

Yifan Ma[1]　花一傑[1]　高橋空大[1]　長谷川遼[1]　土田陸斗[1]　園田哲也[2]　宇津呂武仁[1]
[1]筑波大学 システム情報工学研究群
[2]山梨大学 工学部 メカトロニクス工学科

概要

本研究では，状態遷移モデルと大規模言語モデル（LLM）を用いたテキストベースの接客訓練および半構造化インタビューシステムを拡張し，より現実的かつ実践的な訓練環境の提供を目的とする。従来のシステムは，構造化された段階的指導と，AI による柔軟な応答生成を可能にし，効率的なスキル習得を支援してきた。一方で，対話がテキストに限定されているため，実際の接客場面で生じる心理的緊張感やストレスを十分に再現できないという課題があった。この課題を解決するため，本研究では自動音声認識（ASR）および音声合成（TTS）を訓練フレームワークに統合し，ユーザとシステムの間で音声ベースの対話を行う構成とした。これにより，実環境に近いストレス下での訓練を可能とし，接客スキル習得におけるリアリティおよび訓練効果の向上を実現する。

1　はじめに

カスタマーサービス業務では，顧客対応スキルの効率的な習得が求められている．従来は OJT やロールプレイによる研修が行われてきたが，時間的・人的コストが高く，訓練内容の標準化が難しいという課題があった．近年では，シミュレーション型訓練，特に対話型 AI を用いた模擬練習が注目されており，従来手法より効率的なスキル習得が可能であることが報告されている [3]．教育・医療・サービス分野においても，AI を用いた模擬コミュニケーション訓練は，判断力や応答力，心理的耐性の向上に有効であるとされている．一方，多くの既存システムはテキストチャットによる対話を中心としており，実際の顧客対応で生じる心理的緊張やストレスを十分に再現できないという指摘がある [8]．実環境では，音声による即時応答や語調，感情的ニュアンス，沈黙などの非言語的要素を適切に処理する能力が求められるが，テキストベースの訓練ではこうした要素を体験することが難しい．特に音声対話には言い淀みや間投詞，感情的発話などテキストに現れにくい情報が含まれており，これに慣れることが実践的訓練の重要な要素である [7].

一方，職務面接の分野では，半構造化インタビューが柔軟かつ効果的な評価手法として広く用いられている [8]．半構造化インタビューは，事前に定めた質問項目を基盤としつつ，応募者の回答に応じて質問内容や順序を動的に調整することで，公平性と柔軟性を両立する手法である．面接官は対話を通じて応募者の能力や人柄を深く把握でき，心理的負荷を軽減しながら自然な応答を引き出すことができる [7]．近年では，こうした半構造化面接を AI で模倣し，回答内容に基づいてフォローアップ質問を自動生成する研究も進んでおり [8]，客観性と一貫性を高める新たなアプローチとして注目されている．

以上を踏まえ，本研究では，大規模言語モデルを用いた従来のテキストベース顧客対応訓練と半構造化インタビューシステムに音声入出力インターフェースを統合し，実環境に近いストレス条件下での訓練を可能とすることを目的とする

2　関連研究

2.1　接客訓練

近年，音声対話と VR を組み合わせたマルチモーダル接客訓練システムが提案されている．たとえば高橋ら [4] は，空港のクレーム対応を想定した VR ベースの接客訓練システムを開発し，訓練者が音声を通して顧客アクターと対話し，適切な言葉遣いや姿勢・お辞儀などの動作を統合的に練習できる環境を実現した．同システムでは，顧客アクターの発話・表情・動作が事前に定義されたシナリオに基づ

いて制御され，リアルな顧客応対体験を再現している．また，高橋らの研究ではそのシステムの対話部分に着目し，状態遷移モデルと大規模言語モデルを統合することで，訓練シナリオ中の柔軟な対話遷移を実現している点が特徴である．このアプローチにより，従来の線形的な訓練進行では困難であった顧客状態や反応に応じたリアルタイムな応答制御が可能となり，より効果的な接客訓練対話が展開できることが示された．

2.2 半構造化インタビュー

大規模言語モデルの発展に伴い，半構造化インタビューの対話進行を自律的に制御する試みが増えている．長谷川ら [2] は，状態遷移モデルと大規模言語モデルを統合した半構造化インタビュー対話制御フレームワークを提案した．長谷川さんの手法では，LLM をインタビュア役として用い，発話生成・スロット生成・スロットフィリングを担わせることで，回答内容に応じた柔軟な質問生成と条件分岐を実現している．また，複数のインタビュー対象者との対話を通じて生成されたスロットを蓄積し，後続の面談で参照可能とする「スロット蓄積型半構造化インタビュー」を実装した．

2.3 音声対話システム

音声対話システムの分野では，音声入出力を備えた対話エージェントの研究が進展している．Nicmanis らは，コールセンター業務を対象に，自動音声認識（ASR），言語理解（NLU），対話管理，および表現力豊かな音声合成（TTS）を統合した音声対話ボットを試作した [1]．同システムでは，ユーザ発話を音声で入力し，有限状態機械に類似したグラフ構造に基づいて，あらかじめ定義されたトピックや質問分岐に従った応答生成を行っている．また，採用面接や教育分野においても音声対話システムの活用が進んでいる．大学では，AI 面接官による模擬面接システムが導入され，学生は 3D アバターとの音声対話を通じて自己紹介や質疑応答の練習を行うことが可能となっている．中央大学が導入した「Chu 活ボット」は，生成 AI を用いた傾聴対話システムであり，音声対話形式の模擬面接を繰り返し実施し，即時に AI からフィードバックを得られる点が特徴である．こうした音声対話システムは，対面面接に不慣れな学習者にとって実践的な訓練環境を提供している．さらに教育領域では，語学学習者向けに音声対話を用いた会話練習システムも注目されている．音声認識および言語処理技術の進歩により，第二言語での対話練習や口頭試験のシミュレーションが可能となり，音声対話技術が学習者の会話能力向上に寄与することが報告されている [6]．このように，近年の音声対話技術の発展により，高速かつ音声で応答可能な対話システムが実用段階に入りつつある．

3 接客訓練・音声対話システム

本研究で提案する接客訓練向け音声対話システムは，高橋ら [4] による接客訓練モデルを基盤としたタスク指向型の設計である．図 1 に示すとおり，システムは (1) 音声認識 (ASR)，(2) 対話管理 (DM)，(3) 対話状態追跡 (DST)，(4) 行動選択 (Policy)，(5) 応答文生成 (NLG)，(6) 音声合成 (TTS) の 6 要素から構成される．

ASR にはローカル動作の日本語版 Whisper[1] 学習済みモデルを用いてユーザ発話を逐次テキスト化する．DST は ASR 出力およびシステム (顧客役) 発話の生成テキスト履歴を保持し，対話状態の参照に供する．DM・Policy・NLG は ChatGPT-4o[2] API を中核とし，LangGraph[3] により状態遷移とプロンプト連携をオーケストレーションする．まず顧客役 AI がシナリオ定義に基づくクレーム発話を提示し，ユーザ応答は ASR で認識される．続いて LLM は最新の対話履歴を事前定義の模範シナリオと照合し，(i) 合格の場合は LangGraph に従って次段階へ遷移し，該当段階の応答文を生成する．(ii) 不整合 (要改善) の場合は，クレーム対応の判定結果・根拠・改善点をテキストで即時提示し，同一クレームを再合成して再試行を促す．

TTS には軽量な Tsukuyomi-chan[4] モデルを採用し，即時性を損なわず自然性の高い音声出力を実現する．本構成により，状態遷移に裏打ちされた指導可能性と LLM による柔軟な言語生成を両立し，実務場面に近い音声インタラクション下の接客訓練を提供する．

4 半構造化インタビュー・音声対話システム

図 2 に示すとおり，本システムの全体構成 (ASR/DST/DM/Policy/NLG/TTS) は接客訓練向け音声

1) https://github.com/openai/whisper
2) https://openai.com/index/hello-gpt-4o
3) https://www.langchain.com/langgraph
4) https://huggingface.co/offtoung/tsukuyomi-chan-vits

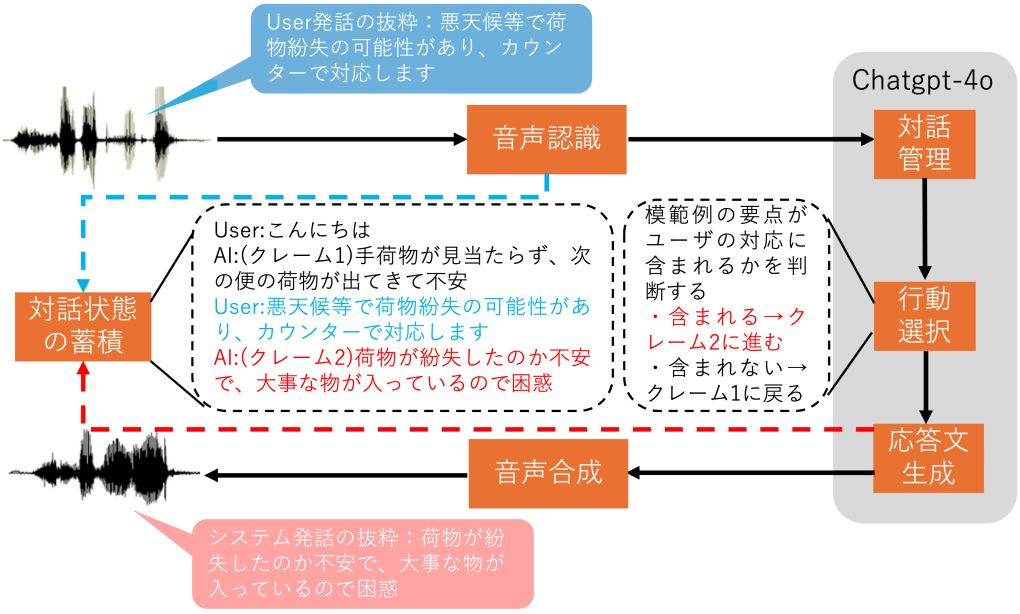


図 1: 接客訓練における音声対話システム

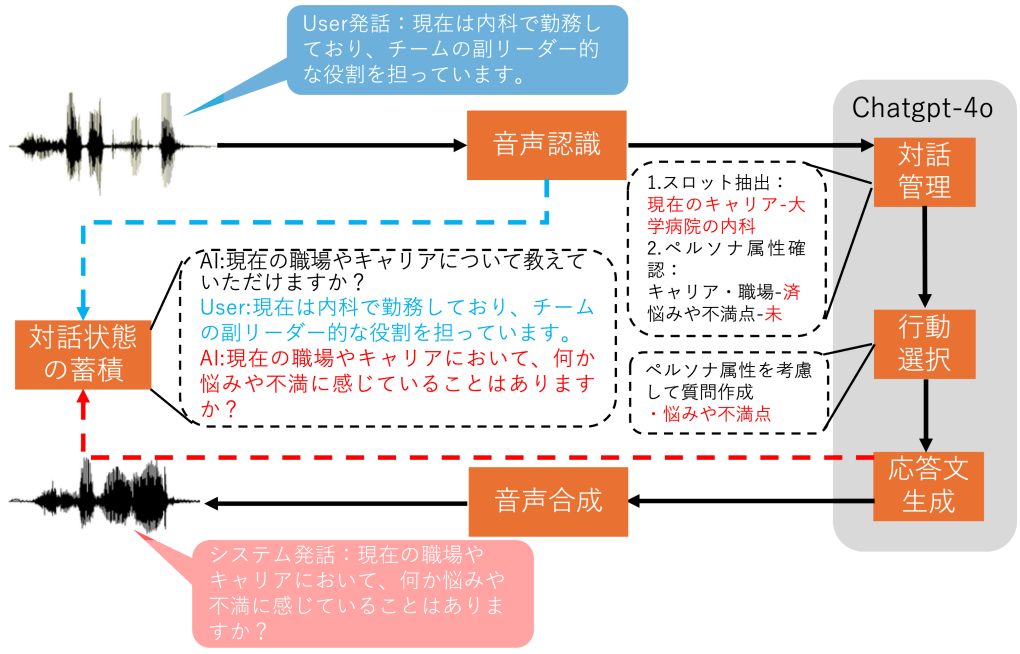


図 2: 半構造化インタビューにおける音声対話システム

対話システムと同一である．一方で，対話管理 (DM) および行動選択 (Policy) は，半構造化インタビューには接客訓練のような規定の模範シナリオが存在しないことを踏まえ，スロット抽出を中核とする設計へ再構成した．

処理フローは次の通りである．まず，ASR により得たユーザ発話からスロット (候補属性) を抽出し，該当するペルソナ属性を同定した上で DST(対話履歴) へ永続化する．抽出スロットが既存の推定ペルソナ集合に含まれない場合は，新たなペルソナ属性を生成し，当該スロットを紐づけて拡張する．DM/Policy は，更新後の属性集合に対して重要度 (優先順位) を推定し，NLG に対し次ターンの質問意図を指示する．NLG はこの指示と対話履歴を条件として，抽象度の高いフォローアップ質問を生成し，TTS により音声提示する．以上により，本システムはスロット中心の知識獲得ループを通じて，事前定義シナリオに依存せず，対象者の回答に応じた動的な質問選択と属性カバレッジの拡充を実現する．

5　評価

本節では，提案システムの処理遅延を，接客訓練シナリオおよび半構造化インタビューの実行ログに基づき評価する．音声対話パイプラインは **ASR**：Whisper Medium(日本語)，**DM/Policy/NLG**：ChatGPT-4o(LangGraph によるオーケストレーション)，**TTS**：Tsukuyomichan で構成する．

5.1　評価設定

本研究では，提案する音声対話システムの実運用可能性を検証するため，各処理モジュールに要する応答時間を計測し，対話の即時性の観点から評価を行った．

接客訓練システムでは，顧客役生成用 LLM と応答判定用 LLM を別々に用いており，ユーザ発話の ASR 処理時間，顧客発話生成に要する LLM 時間，および応答判定に要する LLM 時間を個別に計測した．

一方，半構造化インタビューシステムでは，対話制御のために複数の LLM 処理が直列に実行される構成となっているため，(1) 質問生成，(2) スロット抽出，(3) ペルソナ属性生成，(4) 分岐先決定，(5) 新規スロット生成の各処理について個別に処理時間を計測した．

5.2　結果

5.2.1　接客訓練（図 3a）

接客訓練における処理時間の一例を示す．ユーザ発話「こんにちは」に対する ASR に約 0.7 s，顧客役 AI による第 1 クレーム生成に約 1.7 s を要した．次に，ユーザ応答に対する ASR に約 2.1 s，応答妥当性判定に約 2.3 s，第 2 クレーム生成に約 3.5 s を要した．各ターンの遅延は数秒程度であり，音声対話として実用的な応答速度が得られた．

5.2.2　半構造化インタビュー（図 3b）

半構造化インタビューでは，質問生成に約 3.9 s，ユーザ発話 ASR に約 1.9 s を要した後，スロット抽出 4.1 s，ペルソナ属性生成 7.3 s，分岐先決定 0.8 s，新規スロット生成 3.8 s と続き，1 ターンあたり合計で 15 s 以上の遅延が発生した．その結果，音声対話としての即時性は十分とは言えない水準であった．

5.3　限界と今後の改善

評価結果より，接客訓練システムでは状態遷移モデルと LLM による判定・生成処理により，音声対話として実用的な応答時間が確保できることが確認された．これは，事前定義されたシナリオ構造により，LLM の役割が発話生成と簡易判定に限定されているためと考えられる．

一方，半構造化インタビューシステムでは，スロット抽出，属性推定，分岐決定，新規スロット生成といった複数の LLM 処理が逐次実行されるため，1 ターンあたりの遅延が大きく，音声対話としての即時性が確保できなかった．

以上より，接客訓練システムは音声対話型訓練として実運用が可能である一方，半構造化インタビューの実用化には，LLM 呼び出し回数の削減やルールベース手法との併用といった高速化が不可欠である．

実際に実験してみた結果を表 1 (付録の A 節) に示すように，スロット抽出やペルソナ推定などの LLM 呼び出しをスキップすることで，各ターンの応答時間が大幅に短縮されることが確認された．

6　おわりに

本研究では，状態遷移モデルと LLM を用いた接客訓練および半構造化インタビューシステムに ASR・TTS を統合し，音声対話による訓練環境を構築した．評価の結果，接客訓練では実用的な応答速度が得られた一方，半構造化インタビューでは複数の LLM 処理に起因する遅延が大きく，現状ではリアルタイム音声対話としての実用化は困難であることが明らかとなった．

今後の課題として，(1) Whisper のストリーミング認識を導入したリアルタイム化，(2) LLM 応答の並列生成およびプロンプト圧縮による遅延短縮，(3) TTS の逐次再生による体感即応性の向上，(4) WER・主観評価・訓練効果の多面的評価，(5) VAD(Voice Activity Detection) [5] により無音区間を検出することにより高速化を実現する，が挙げられる．これらを通じて，より自然で応答性の高い音声ベース AI トレーニング環境の構築を目指す．

謝辞

本論文は，一部，科研費 25K03416 の支援を受けたものである．本研究の実施にあたり，山梨大学 阪口直紀氏，レオチーシャン助教，西崎博光教授に協力頂いた．

参考文献

[1] Davis Nicmanis and Askars Salimbajevs. Spoken Dialogue System for Call Centers with Expressive Speech Synthesis. In *Proceedings of Interspeech 2022*, pp. 5215–5216, 2022.

[2] 長谷川遼, 花一傑, 宇津呂武仁, 橋本慧海, 中野幹生, 白松俊. 状態遷移モデルおよび大規模言語モデルを用いた半構造化インタビューのモデル化の評価. 第 39 回人工知能学会全国大会論文集, 2025.

[3] Irene Suarez-Garcia, Rocio Perez-Moreno, Ramon Rojas-Burke, and Juan M. Garcia-Gomez. Dialogue: A generative ai–based pre–post simulation study to enhance diagnostic communication in medical students. *JMIR Medical Education*, Vol. 11, No. 1, p. e63082, 2025.

[4] 高橋空大, 花一傑, 長谷川遼, 宇津呂武仁, 星野准一, 西崎博光. 状態遷移モデルおよび大規模言語モデルを用いた複数顧客接客訓練対話のモデル化. 第 39 回人工知能学会全国大会論文集, 2025.

[5] Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

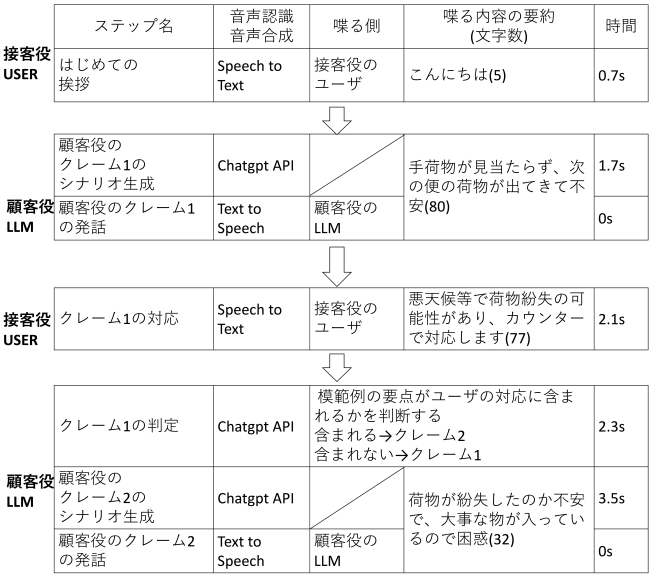
[6] Veronika Timpe-Laughlin, Tetyana Sydorenko, and Phoebe Daurio. Using spoken dialogue technology for l2 speaking practice: what do teachers think? *Computer Assisted Language Learning*, Vol. 35, No. 5-6, pp. 1194–1217, 2022.

[7] Charles Welch, Allison Lahnala, Vasudha Varadarajan, Lucie Flek, Rada Mihalcea, J. Lomax Boyd, and Joao Sedoc. Isca: A framework for interview-style conversational agents. *arXiv preprint arXiv:2508.14344*, 2025.

[8] He Zhang, Yueyan Liu, Xin Guan, Jie Cai, and John M. Carroll. Harnessing the power of ai in qualitative research: Role assignment, engagement, and user perceptions of ai-generated follow-up questions in semi-structured interviews. *arXiv preprint arXiv:2509.12709*, 2025.

表 1: ターンごとの処理時間

LLM 数	動作する LLM	ターン 1	ターン 2	ターン 3
5	全 LLM	13.09s	25.77s	24.55s
1	質問生成 LLM のみ	13.30s	19.55s	18.70s
0	無	11.13s	15.43s	15.52s



(a) 接客訓練における音声対話シナリオの実例

(b) 半構造化インタビューにおける音声対話シナリオの実例

図 3: 半構造化インタビューと接客訓練の実例

A　LLM 数別ターン別応答時間の比較および音声対話シナリオ実例

表 1 に，各 LLM 数条件におけるターン別応答時間を示す．LLM 数が減れば応答時間は短縮されるが，それでもリアルタイム応答には至っておらず今後の改善を要する．図 3 に，両システムの音声対話シナリオの実例を示す．