音環境に適応する音声合成能力を搭載した音声対話システムの構築と実証実験に基づく検討

武 伯寒[1]　高道 慎之介[2,1]　関 健太郎[1]　猿渡 洋[1]

概要：本稿では環境雑音や対話相手の発話といった周囲の音環境に応じて発話する音声対話システムを構築し，実証実験によってシステムのユーザーにどのような体験の変化を齎すかについて調査した．構築した音声対話システムには，一人称視点での聴覚情報を利用してその場の音環境を考慮して話し方を変化させる音声合成技術を搭載した．このシステムは人間の音環境に応じて適した発話を生成することを模擬し，様々な実環境でユーザーにとって自然で円滑な対話体験を提供することを目指す．雑音環境下でシステムと対話を行う実証実験では，定量評価により構築したシステムによるユーザーの対話体験に改善が見られなかったことがわかった．この結果に基づき．システム全体の連携を見据えて音環境への適応をモデリングする必要性について考察した．また実証実験で行われた対話の収録物を定性的に解析し，音環境によってシステムやユーザーにどのような対話の変化が見られたかについて調査した．

1. はじめに

音声対話システムは対話ロボットなどで応用されるにあたって，様々な実環境の中で自然で円滑な音声コミュニケーションを提供することが求められる．対話ロボットに搭載される音声対話システムは，ショッピングモールや観光地といった，環境雑音が存在する様々な場所での運用が想定される．環境雑音の他にも，実環境での音声対話システムは対話相手であるユーザーとのコミュニケーションにおいてユーザーの話し方や，ユーザーとの位置関係といった環境要素が現れる．本稿ではこれら周囲の環境要素を「音環境」と総称する．人間同士の音声コミュニケーションにおいては，それぞれが見聞きして得た視聴覚の知覚情報をもとに，音環境の要素に応じてその場に適した話し方で発話を生成する [1], [2], [3]．これにより，例えば騒がしい環境では声を張り上げる [4]，対話相手の声の明るさに自身の発話音声も同調する [5]，といった行動が現れる．このようにして，人間は実環境での音声コミュニケーションを互いにとって自然で円滑なものにする．同様にして，音声対話システムも一人称視点での知覚情報に基づき音環境を考慮して適応し，ユーザーにとって自然だと感じられる，淀みがないコミュニケーションを実現する話し方での発話を計画する必要がある．

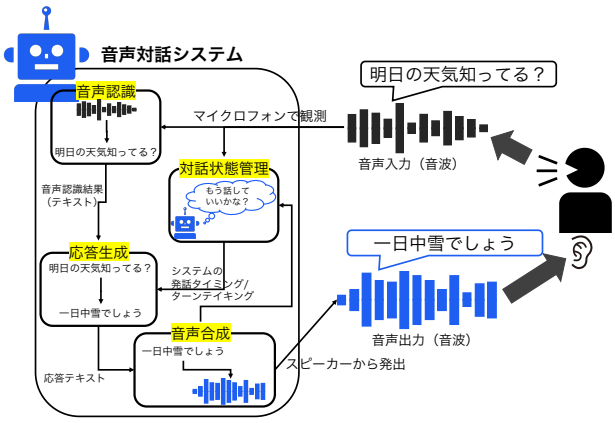


図 1　各種音声や対話情報を処理するモジュールを結合することで実現する音声対話システムのブロック図.

音声対話システムは音声や対話情報を処理するモジュールを結合する方式によって実現され [6]，このうちシステムが発出する発話音声の合成にはテキスト音声合成 (text-to-speech; TTS) [7] が用いられてきた．モジュール結合型の音声対話システムでは**図 1** に示すように，ユーザーからの発話を音声認識モジュールに書き起こした後，得られた発話テキストに対する応答を応答生成モジュールで生成する．生成した応答テキストを入力として，TTS は人間の音声生成を模擬してそのテキストを読み上げる発話音声を合成する．対話状態管理モジュールによって応答生成や音声合成のタイミングを制御することにより，図 1 に示した音声対話システムはリアルタイムにユーザーと音声コミュニ

[1] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[2] 慶應義塾大学
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522, Japan.

ケーションをとることが可能である．

音声対話システムが様々な音環境に適応して自然で円滑な音声コミュニケーションを提供できるようになるためには，音声合成モジュールに音環境に適応した話し方の音声を合成する能力を搭載し，評価する必要がある．前述した TTS による発話音声合成を用いた音声対話システムは，様々な実環境においてはその情報を利用せず，応答テキストのみからそれを読み上げる音声を合成して発出する．そのため，例えば騒がしい環境でも静音環境での読み上げ音声を模擬して合成することで，ユーザーにとって聞き取りづらい発話音声を伝達してしまい，円滑な意思疎通を妨げてしまう．この問題に対処するため，deep neural network (DNN) を用いた，一人称視点の聴覚情報をテキストと同時に入力し，音環境に適応した対話音声を合成する機構が提案されている [8]．その一方で，音声対話システムの一要素としての音声合成モジュールの，音環境に適応する発話音声合成能力については検討が為されていない．実際，これまでの研究開発では，ある一定の実環境において，音声合成モジュールに TTS を用いたシステムの応答生成を主に評価しており [9]，音声合成部の音環境適応能力に着目した研究は行われていない．

以上の議論を踏まえ，本稿では一人称視点の知覚情報を用いて音環境に適応する対話音声合成 (dialogue speech synthesis exploiting egocentric information to be environment-adaptive; EgoEA-DSS) を用いた音声対話システムについて検討する．まず，先行研究 [8] で開発された EgoEA-DSS モデルを用いて構築した音声対話システムについて報告した．その後，2 種類の実環境において，音声対話システムを搭載したアバターと被験者が対話する実験を行った．この実験における被験者の対話体験に基づき，EgoEA-DSS モデルを搭載した音声対話システムの自然で円滑な音声コミュニケーションを提供する能力を比較評価した．また，実環境での音声対話における，被験者とアバターの対話内容や音環境への適応を，収録したユーザーやアバターの対話音声の特性に基づいて分析した．

2. 関連研究

2.1 音環境に適応した発話音声の合成

音環境の各要素に適応することを試みた音声合成・加工の手法はこれまでに様々な研究で提案されてきた．周囲の環境雑音への適応を実現するために，信号処理による出力音声の加工 [10] や DNN を用いた TTS [11], [12] といった音声合成手法が提案されている．しかしこれらの手法は朗読音声の合成を主眼としており，自発的な対話における環境雑音への適応を検討していない．その他に，対話相手の発話音声に適応する自発的な対話音声合成の研究 [13] も為されている．しかしこの先行研究では静音環境での対話を想定しており，様々な実環境で運用される音声対話システムには不向きである．

人間は対話の際に周囲の状況を絶えず見聞きして捉え，その情報を以後の発話計画にフィードバックして speech chain [1] を形成することが知られている．音声合成の機構も同様に speech chain を形成することで，その場に応じた自然で円滑なコミュニケーションを実現することが期待できる．この speech chain を形成する音声合成を実現する，一人称視点の視聴覚情報と発話テキストを入力として，その場に応じた適した話し方の発話音声を合成するタスクを EgoEA-DSS と定義する．我々は以前この EgoEA-DSS に必要な, 話者の発話音声と一人称視点の受聴音，そして視点映像を時刻同期して収録した自発対話コーパス SaSLaW を整備し，実際に受聴音を入力する，DNN に基づく EgoEA-DSS システムを構築した [8]．本稿は，この EgoEA-DSS を音声対話システムの一要素と見立てて，音声対話システムとの対話体験に基づく EgoEA-DSS システムの評価を行った初めての報告である．

2.2 音声コミュニケーションにおける発話音声の音環境への適応のモデリング

音声コミュニケーションにおける音環境への適応の補助や自動化はその場に応じた自然なコミュニケーションと円滑な情報伝達に必要である．そのため，音声対話システムに限らない様々なツールにおいて，この音環境への適応をモデリングする技術が検討されてきた．例えばロボットが伝達した音声を，対象者までの距離 [14] や，それに加えて周囲の環境雑音 [15] に応じてその場に適した聞き取りやすいものになるように調整する手法が提案されている．他にも，テレプレゼンスシステムの操作者の発話音声の音量を，遠隔地の環境雑音や特定のユーザーとの距離に応じて調整する手法が提案されている [16].

音声対話を行うシステムについては，実際に周囲の環境の様子や雑音によって，自然で聞き取りやすいと感じられるシステムの話し方が変化する [17] ことが報告されている．本研究ではこの適した話し方をシステムに自動で獲得させることを目指し，本稿ではそのために構築した音声対話システムをユーザーとの対話体験により評価した．

3. 音環境に適応した音声合成能力を有する音声対話システムの構築

本節では音声対話システムとの対話体験に基づく EgoEA-DSS システムの評価を行うために構築した音声対話システムについて述べる．

3.1 受聴音を入力する EgoEA-DSS モデル

本節で構築した音声対話システムには，**図 2** に示す，先行研究 [8] で開発した一人称視点の受聴音を入力する EgoEA-DSS モデルを搭載できるようにした．このモデル

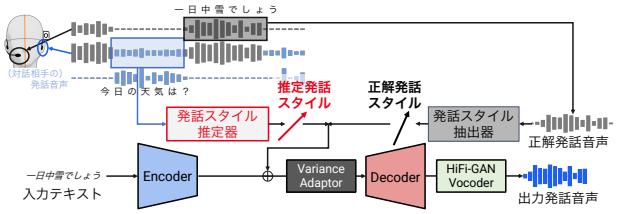


図 2 本稿で構築した音声対話システムで用いた，受聴音を入力する EgoEA-DSS モデルのパイプライン．

は FastSpeech 2 [18] をベースとした DNN モデルにより入力テキストと一人称視点の受聴音から発話音声のメルスペクトログラムを推定し，HiFi-GAN [19] によりメルスペクトログラムに対応する音声波形を推定する．受聴音としては，モデル自身が発話を合成する直前に対話相手が発話していた区間のものを用いる．入力した受聴音から，Global Style Token [20] により抽出された発話スタイル特徴量を発話スタイル推定器という DNN モデルにより推定し，得られた発話スタイル特徴量を FastSpeech 2 に条件付ける．これにより，音環境に応じて適した発話スタイルの発話音声を合成することを期待した．

本節で構築した音声対話システムの音声合成モジュールとしては，自発対話コーパス SaSLaW [8] に収録されている話者 spk06 のデータを用いて学習した，ベースラインとなる TTS モデル (“TTS”) と，図 2 のモデル (“EgoEA-DSS”) の 2 種類を採用した．TTS は，テキストのみを入力しメルスペクトログラムを出力する従来の FastSpeech 2 と HiFi-GAN を結合したモデルである．これらのモデルの詳細なアーキテクチャや学習条件は先行研究での報告 [8] を参照されたい．

3.2 構築した音声対話システムのソフトウェア構成

実際に構築した，音環境に適応した話し方を変化させた発話音声を合成する音声対話システムと，それを搭載したアバターの処理フローを描いたブロック図を**図 3** に示す．図 1 で示したモジュール結合型のシステムをベースとしつつ，音声合成部に EgoEA-DSS システムを採用することで，音環境に適応した発話音声を合成する能力を搭載した．

このアバターの構築にあたっては，realtime multimodal dialogue system toolkit (Remdis) [21] というツールキットで構築したリアルタイムに視覚・音情報のやり取りを行えるモジュール結合型の音声対話システムをベースとした．具体的には，Remdis が提供する，図 3 の各モジュールを並列して逐次動作させる音声対話システムと，MMDAgent-EX [22] というアバター管理のインターフェースを組み合わせたシステムをベースとしてアバターを開発した．

音声合成部に採用した EgoEA-DSS モデルに入力する一人称視点の受聴音として，アバター自身が発話する直前のユーザーが発話しているターンの間に計測した受聴音を用いた．これは図 3 の中では，対話状態管理モジュールによって発話直前のユーザーのターンと判定された区間において，ユーザーが “明日の天気知ってる？” と発話している際のアバターの一人称視点の受聴音に相当する．

なお，音声入力のチャネルとして，3.1 項で説明した EgoEA-DSS モデルに入力する一人称視点の受聴音のチャネルとは別に，音声認識と対話状態管理モジュールに入力する発話音声のみの外部雑音が抑制された音データのチャネルを用意した．これは音声入力を受聴音の 1 チャネルのみにした場合，音声認識と対話状態管理モジュールがユーザーとシステムの発話を区別できずに混線してしまうのを防ぐためである．

3.3 構築した音声対話システムのハードウェア構成

構築した音声対話システムを搭載したアバターとユーザーが対話している際の外観のイメージ図と，実際にユーザーと対話している様子を撮影した写真を，それぞれ**図 4** に示す．ユーザーはノート PC のディスプレイに投影された MMDAgent-EX のアバターに話しかけ，ノート PC 上部に配置したスピーカーからシステムが合成した音声が発出されることにより，ユーザーとアバターの音声コミュニケーションが行われる．

音声対話システムが音データを計測する装置として，接話マイク (SHURE PGA31-TQG ワイヤレス用ヘッドセットマイク[*1]) と耳掛けバイノーラルマイク (Sound Professional MS-EHB-2[*2]) を採用した．このうち接話マイクは，ユーザーの発話音声を録るためにユーザーが自身の頭部に装着した．これにより，接話マイクでは発話音声以外の雑音を抑制したデータを，耳掛けマイクからは周囲の環境雑音やアバター自身の発話音声も含む一人称視点のステレオの受聴音を計測する．

3.4 音声対話システムの各種モジュールの詳細設定

本項では構築した音声対話システムの図 3 における各種モジュールの詳細設定について述べる．

音声入力．接話マイクと耳掛けマイクではいずれもサンプリング周波数 16 kHz で音データを計測した．接話マイクではモノラル信号，耳掛けマイクではステレオ信号を計測した．

音声認識．音声認識モジュールには Remdis における設定と同様に，ストリーミング音声認識サービス Google Cloud Speech-to-Text[*3]の API を用いた．

対話状態管理．音声対話におけるシステムとユーザーの発話権が交替したかどうか，すなわちターンテイキングを管理する対話状態管理モジュールには，DNN で表現した

[*1] https://shure.com/ja-JP/products/microphones/pga31
[*2] https://soundprofessionals.com/product/MS-EHB-2/
[*3] https://cloud.google.com/speech-to-text

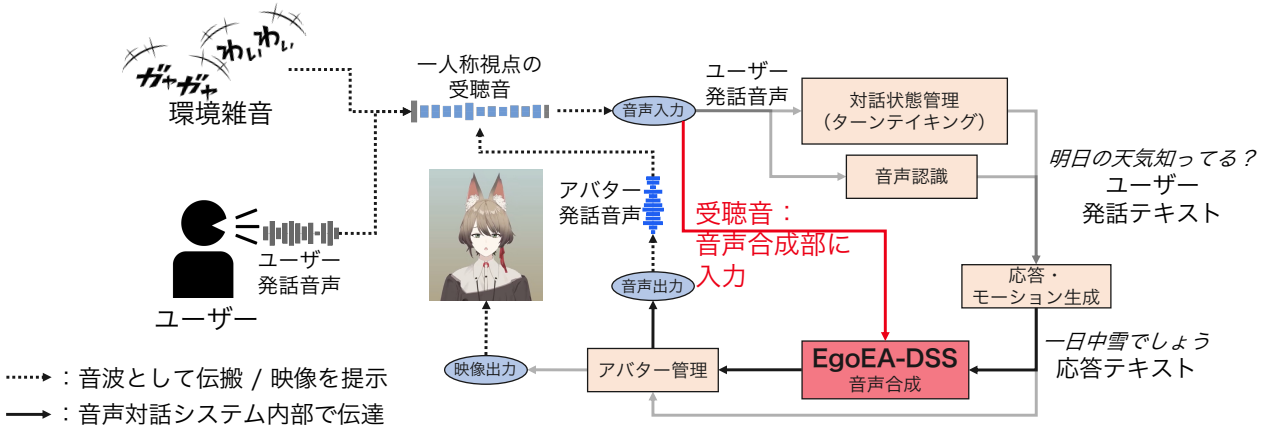


図 3 本稿で構築した，受聴音を入力する EgoEA-DSS モデルにより音環境に適応した発話音声合成を試みるアバターの処理フロー．

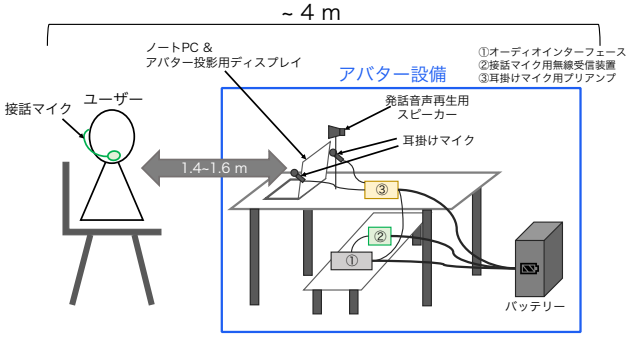


図 4 構築した音声対話システムを搭載したアバターを設営したものと，実際に対話するユーザーとアバターの位置関係を示したイメージ図．

音声活動予測 (voice activity projection; VAP) モデル [23] を，複数の日本語の音声対話データセットで学習したもの [24] を用いた．このモデルを用いて，推論では 25 ミリ秒前までのシステムの発話音声とユーザーの発話音声を入力とし，推論した瞬間からその 600 ミリ秒後までの間にシステムが発話権を有する確率 p_{now} と，推論時点から数えて 600 ミリ秒から 2000 ミリ秒までの間にシステムが発話権を有する確率 p_{future} を推定した．

得られた p_{now} と p_{future} に基づき，発話権の決定を次のようにヒステリシスが現れる閾値処理によって行った．ある時点において，その直前に VAP モデルがユーザーに発話権があると判断した場合，$p_{now} \geq 0.65, p_{future} \geq 0.5$ であればシステムに発話権があると判断する．対して直前に VAP モデルによってシステムが発話権を有すると判断されていた場合は，$p_{now} < 0.375, p_{future} < 0.5$ であればユーザーに発話権があると判断する．

応答・モーション生成．応答・モーション生成モジュールには，音声認識モジュールから得られたユーザーの発話の書き起こしを大規模言語モデルである GPT-4o mini[*4] に入力し，応答テキストとそれに最も整合する MMDAgent-EX 出力用の感情・仕草の種類を出力させた．GPT-4o mini に

[*4] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

```
==
あなたはユーザと雑談するアシスタントで，名前は「ウカ」(Uka) です．次のユーザ発話に対する気の効いたリアクションや返答を作成し，句読点(、,。,！,？) で分割して出力してください．最後にアシスタントの感情(0_平静,1_喜び,2_感動,3_納得,4_考え中,5_眠い,6_ジト目,7_同情,8_恥ずかしい,9_怒り) と動き (0_待機,1_ユーザの声に気づく,2_うなずく,3_首をかしげる,4_考え中,5_会釈,6_お辞儀,7_片手を振る,8_両手を振る,9_見渡す) を出力してください．返答の文面は極力 1 文，最大でも 2 文で作成してください．出力は以下のフォーマットに従ってください
==
こんにちは。よろしくお願いします。/0_平静,2_うなずく
==
```

図 5 応答生成部の GPT-4o mini に読み込ませたシステムプロンプト．

はシステムの初回起動時に，応答生成のためのシステムプロンプトとして**図 5** のものを読み込ませた．得られた出力から応答テキストと，表情・仕草の種類を抽出し，それぞれ音声合成とアバター管理モジュールに送信した．応答テキストの送信の際には，文章を句読点で分割してから音声合成モジュールに送信した．

音声合成．音声合成モジュールには，比較評価のため 3.1 項で紹介したベースラインの TTS と，図 2 のアーキテクチャを持つ EgoEA-DSS の 2 種類を採用した．受信した応答テキストは，句読点に分割された各要素ごとに音声合成モジュールに入力して，発話音声が合成され次第逐次オーディオ出力に送信した．

音声出力．音声合成モジュールから受け取った発話音声は，サンプリング周波数 16 kHz のモノラル信号としてスピーカーから発出した．

アバター管理．MMDAgent-EX のアバターの外見としては，公式で整備されている Uka（うか）[*5] というサイバネティックアバターのキャラクターモデルを用いた．

4. アバターとの対話体験に基づく EgoEA-DSS システムの評価実験

本節では 3 節で構築した音声対話システムを搭載したアバターと被験者が実際に対話を行う実験と，その結果に

[*5] https://github.com/mmdagent-ex/uka

基づいた EgoEA-DSS システムの比較評価について報告する．

この実験には 20 代の男性 4 名，女性 1 名の合計 5 名が被験者として参加した．実験において，被験者は 2 箇所の実環境に赴き，実際にアバターと対面して自由に対話を行った．被験者はこの対話体験において，アバターは音環境に適応して自然で聞き取りやすい発話を行い，自然で円滑なコミュニケーションを提供したと感じたか，について質問紙への回答を通して評価した．この評価に基づき，図 2 のアーキテクチャを持つ EgoEA-DSS モデルは，ベースラインモデルである受聴音を無視した TTS モデルに比べて良好な対話体験を提供したかについて調べた．さらに，被験者とアバターの対話音声を収録し，その収録物をもとに実環境における対話音声の特性について分析した．

4.1 実験環境・設備の準備

本実験では屋内環境と屋外環境の 2 種類の実環境で，被験者とアバターの対話が行われた．屋内環境として東京大学本郷キャンパス工学部 2 号館展示室，屋内環境として東京大学本郷キャンパス工学部前広場を選定した．それぞれの環境における周囲の騒がしさを測るため，被験者が赴いて実験を行うごとに，対話が行われていない時に騒音計でその場の騒音値を騒音計（サンワサプライ デジタル騒音計 CHE-SD1[*6]）により計測した．計測した結果，屋内環境の騒音値は 33 dB 以下，屋外環境の騒音値は 46 から 51 dB 程度であった．

各実環境での実験を行うにあたって，図 4 に示したイメージ図に従い，各実環境において被験者と音声対話を行うアバターが対面するための実験設備を設営した．アバターは 3 項で述べたものをそれぞれの場所で構築した．被験者とアバターを表示したディスプレイの間には 1.4 m から 1.6 m の距離を設け，ディスプレイに表示したアバターと被験者が向かい合うようにした．被験者は音声認識モジュールと対話状態管理モジュールの入力のために用いる接話マイクを装着した．

対話実験の間に，アバターの耳掛けマイクと被験者が装着した接話マイクが計測した音データは，アバターと被験者の対話音声の解析のために時刻同期して収録した．

4.2 対話実験の内容とその手順

被験者は実験を通して，5 回アバターと対話タスクをこなし，その 5 回の対話体験について質問紙への回答により評価する，という対話・評価タスクを繰り返した．このタスクを 2 箇所の環境でそれぞれ 4 回ずつ，合計 8 回繰り返した．屋内と屋外のどちらで先に 4 回の対話・評価タスクをこなしたかは被験者によって異なる．

EgoEA-DSS システムの比較評価のため，各対話・評価タスクごとに TTS と EgoEA-DSS のどちらを音声合成モジュールに組み込むかを次のように設定した．まず各被験者による 8 回の対話・評価タスクの実行を 2 回ずつ 4 セットの実行とみなした．このセットにおいて，TTS と EgoEA-DSS が 1 回目と 2 回目のタスクいずれかでそれぞれ音声合成モジュールに組み込まれた．各セットごとにどちらのモデルが先に対話・評価タスクで組み込まれるかはランダムに決定した．これにより，各被験者は各実環境ごとに，一方の音声合成モデルを搭載したアバターについて 2 回評価タスクをこなした．

1 回の対話タスクではまず，被験者はアバターとの対話の話題の指針となるトークテーマを指示された．被験者による自由で自発的な音声対話を観測するために，実験実施の前に雑談やディベートのテーマを事前に選定し，被験者への指示に用いた．テーマの指示後はアバターが起動し，起動完了後に被験者はアバターと 1 分から 2 分程度の台本のない自発的な音声対話を行った．時間が経過して対話が終了したところでアバターがシャットダウンされ，被験者は次の対話のトークテーマを受け取った．このテーマの指示，アバターの起動，アバターとの対話・アバターの終了で 1 回の対話タスクが構成された．

被験者はこの 5 回の対話タスクを終了後に評価タスクに移った．評価タスクではそれまでの 5 回の対話体験について，次の 3 つの観点に基づいて被験者に評価されることを狙った．

Q1． アバターは周囲の音環境に応じた自然な話し方であったか？

Q2． アバターの発話音声は聞き取りやすかったか？

Q3． 全体として，アバターは自然で円滑な対話体験をユーザーに提供できたか？

これらの観点で定量的な評価を行うために，書籍 [6] や関連研究 [17] における質問項目を参考にして，**表 1** に示す 6 つの質問項目を作成した．各質問項目に対応する質問文を被験者に提示し，被験者はそれぞれについて 1 を “全く同意しない”，4 を “どちらとも言えない”，7 を “非常に強く同意する” とした 7 段階の Likert 尺度で回答した．1 回の評価タスクにおいて，被験者は表 1 の質問項目に先立って周囲の環境の様子について記述式で回答し，その後 6 つの質問についての回答を行った．各被験者は各実環境ごとに一方の音声合成モデルを搭載したアバターについて 2 回評価タスクをこなし，結果各質問項目について 2 度回答した．5 人の被験者を合わせて，各実環境・各音声合成モデルごとに，表 1 の 6 つの質問項目それぞれに 10 回答得られた．

被験者は 8 回の対話・評価タスクをこなした後，自身がそれまで実験で得られたアバターとの対話の体験から感じたことについて，種類の如何を問わず自由記述式で質問紙に回答した．

[*6] https://www.sanwa.co.jp/product/syohin?code=CHE-SD1

表 1 対話実験の評価タスクにおいて被験者に提示した質問項目．質問の概要と質問文，そしてそれぞれの質問がどの観点に対応するかを明示した．

観点	質問の概要	質問文
Q1	総合的な音環境適応力	アバターの話し方は周囲の環境に適応し自然だと感じた
	対話相手への注意	アバターはあなたに注意を向けていると感じられた
	周囲の環境への注意	アバターは周囲の環境を気にしているように感じた
Q2	発話の聞き取りやすさ	アバターの発話は聞き取りやすかった
Q3	アバターの人間らしさ	アバターは人間らしいと感じた
	対話の楽しさ	アバターとの対話を楽しめた

表 2 静かな屋内環境でのアバター対話実験において、得られた Likert 尺度の回答の統計と Wilcoxon の符号付順位検定の結果。μ, σ は回答の Likert 尺度の平均と標準偏差を表す。

	TTS		EgoEA-DSS			
質問内容	μ	σ	μ	σ	p 値	Cohen's d
総合的な音環境適応力	4.8	0.98	4.7	1.10	0.65	-0.14
対話相手への注意	5.2	0.98	5.0	1.48	0.48	-0.22
周囲の環境への注意	3.4	0.66	3.0	1.10	0.34	-0.31
発話の聞き取りやすさ	5.3	0.45	4.9	0.94	0.10	-0.57
アバターの人間らしさ	3.8	1.25	4.0	1.10	0.41	0.25
対話の楽しさ	4.9	0.98	4.8	0.98	0.71	-0.11

表 3 屋外環境でのアバター対話実験において、得られた Likert 尺度の回答の統計と Wilcoxon の符号付順位検定の結果。μ, σ は回答の Likert 尺度の平均と標準偏差を表す。

	TTS		EgoEA-DSS			
質問内容	μ	σ	μ	σ	p 値	Cohen's d
総合的な音環境適応力	4.6	0.92	4.3	0.78	0.18	-0.44
対話相手への注意	5.1	1.22	4.6	1.56	0.34	-0.32
周囲の環境への注意	3.3	1.27	3.6	1.91	0.38	0.26
発話の聞き取りやすさ	3.9	1.30	3.5	1.57	0.55	-0.23
アバターの人間らしさ	4.3	1.35	4.0	1.10	0.48	-0.26
対話の楽しさ	4.7	1.1	4.5	1.63	0.59	-0.16

4.3 対話体験についての回答に基づく評価結果と考察

5 名の被験者から得られた表 1 の 6 つの質問項目に対するサンプルサイズ $N = 10$ の Likert 尺度の回答について，各実環境ごとに音声合成モデル間でその値に差があるかを統計的な検定により調べた．統計的な検定は有意水準 $\alpha = 0.05$ での Wilcoxon の符号順位検定により行った．

表 2 と**表 3** に屋内・屋外環境それぞれにおける Likert 尺度における回答とその統計的な検定の結果を示す． これらの結果を観察すると，屋内・屋外のいずれにおいても，全ての質問項目について音声合成モデル間に有意差は見られなかったことがわかった．すなわち，受聴音を入力しない従来の TTS としてのベースラインモデルと，図 2 にある受長音を入力して音環境を考慮する機構を持つ EgoEA-DSS モデルの間に，被験者の感じた対話体験の差は **Q1** から **Q3** のいずれの観点で見ても現れなかったことがわかった．

この結果が現れた原因として，本節の実験のデザインに問題があったことが考えられる．

まず実験が実施された実環境は全体的に静かであり，屋外でもその騒音値は 46 から 51 dB 程度であった．これに加え，質問紙に被験者が記入した周囲の環境の様子についての記述式回答を集計したところ，被験者 5 名から得られた 20 件の屋外環境についての回答のうち，6 件がその環境を “静かな屋外” と記したことがわかった．このような静かな雑音環境下では，ベースラインモデルの TTS よりも EgoEA-DSS の方が自然で聞き取りやすい音声を合成すると感じられ，対して騒がしい環境では EgoEA-DSS の方が自然で聞き取りやすいと評価されたことが報告されている [8]．本節における，ベースラインの TTS モデルと EgoEA-DSS モデルを搭載したアバター間に対話体験の差が見られなかったという結果は，この先行研究の結果とも整合する．本研究で構築した音声対話システムの音環境に適応した発話合成能力を検証するためには，人通りの多い公園やショッピングモールといった，より騒がしい実環境下での実験に基づく EgoEA-DSS システムの比較評価が必要である．

また本実験では，被験者が表 1 に掲げた質問項目に Likert 尺度で回答する際に，音声合成の受聴ではなく，ターンテイキングや生成された応答の理解といった部分を主軸に回答した可能性がある．例えば「対話相手への注意」の質問項目について，被験者は発話音声が物理的に自然で聞き取りやすいかではなく，生成した応答内容がそれまでの文脈と整合しているか，ターンテイキングが不自然ではないか，といった観点で評価した可能性が考えられる．実際，自由記述式の回答においては対話時の音環境の観点から見たアバターとの対話体験についての記述はなく，次のように応答生成やターンテイキングについての記述が見受けられた．

- 対話をする際のタイミングがかみ合わず，スムーズなやりとりができないことが何度かあった
- アバターからの返答は当たり障りのない内容が多く，通常の会話よりも話を広げづらいことがあった

この結果と考察を踏まえて，今後はより音声合成部に注目した評価が明確にできるように，実験手順のデザインや質問項目の構成を改善する必要がある．また，被験者から得られた自由記述式の回答を踏まえると，EgoEA-DSS システムの改善に基づく音声合成部の性能向上だけではなく，対話システム全体の連携を見据えた音声合成手法の考案も

必要であると言える．

4.4 被験者とアバターの対話音声の分析

本項では被験者 1 名の対話実験にて収録したアバターと被験者の発話音声を解析した結果とその考察について報告する．アバターと被験者の発話音声に，どのような音環境への適応が現れたかを調べた．

対話音声の解析にあたって，被験者とアバターの発話音声を収録した一連の音データをターンごとの発話単位に分割した．まず接話マイクの収録データに対して pyannote.audio [25] を用いて発話区間を自動で検出し，その区間で接話マイクの収録データを分割して被験者の発話音声を得た．次にアバターの発話音声を抽出するために耳掛けマイクで収録した受聴音データに Demucs v4 [26] を適用し，外部雑音を抑制した音声成分を抽出した．このもとで，被験者の発話区間でない部分は全てアバターの発話区間であるとして，抽出した耳掛けマイク収録データの音声成分をこの区間で分割した．その後自動で分割したデータに対して，アバターが発話していない，あるいは発話を為していない音声を手動で排除して，アバターの発話音声を得た．

対話音声の分析として，アバターのある発話音声と，その直前のターンの被験者の発話音声の音量を計算し，その組の分布を観察した．この観察により，被験者の発話音声の環境雑音への適応と，対話相手の発話音声の適応が現れたかについて調べた．発話音声の音量の計算のために，Harvest [27] を用いて得た有声フレームの信号の振幅の root mean square (RMS) を計算し，そのフレーム間平均値を発話音声の音量とした．**図 6** に得られたアバターと被験者の発話音声の音量の組の分布を示す．このプロットの際に，対話したアバターが TTS と EgoEA-DSS のどちらを音声合成部に搭載したものかを明示した．

図 6 の分布を観察したところ，アバターと被験者の発話音声の音量の間には相関が認められなかった．実際，有意水準 $\alpha = 0.05$ の Spearman の相関係数検定を実施したところ，いずれの実環境でどちらのアバターを用いても，発話音声の音量に有意な相関は認められなかった．この観察は，アバターとユーザーのいずれにおいても，発話音声に対話相手への発話音声への適応が現れなかったことを示唆している．この結果は，EgoEA-DSS システムの開発に用いた SaSLaW コーパスの収録物には，静かな雑音環境において話者間の発話音声の相関が現れなかったという報告 [8] と整合する．4.3 項で述べた考察と同様に，今後の EgoEA-DSS の研究においては，より騒がしい実環境での対話実験の実施による比較評価が必要である．

次に図 6(b) の青い丸で囲った部分の発話音声の音量に注目されたい．この部分の発話が発出された雑音環境を確認すると，屋外にていわゆる強い “風が吹く音” である非定常な低周波成分の強い雑音が現れていた．この雑音環境下で，ユーザーの発話音声の音量は屋内や屋外の他の発話音声のものに比べて大きくなっていた．その一方で，EgoEA-DSS を搭載したアバターの発話音声の音量には，屋外でのその他の発話音声のものからの変化が現れなかった．この観察は，強い風が吹く雑音が現れた環境での対話において，人間は環境雑音に適応して発話を生成した一方で，開発した EgoEA-DSS は環境雑音への適応をしないまま発話を合成したことを示唆している．

この結果が現れた原因として，EgoEA-DSS の開発に用いたコーパス SaSLaW [8] の収録において，環境雑音に風が吹く音のような非定常な低周波成分が強く現れる雑音が含まれなかったことが考えられる．この考察を鑑みると，今後の EgoEA-DSS の研究では，より多様な環境雑音が存在するもとでの対話データの収集・活用を行うべきである．

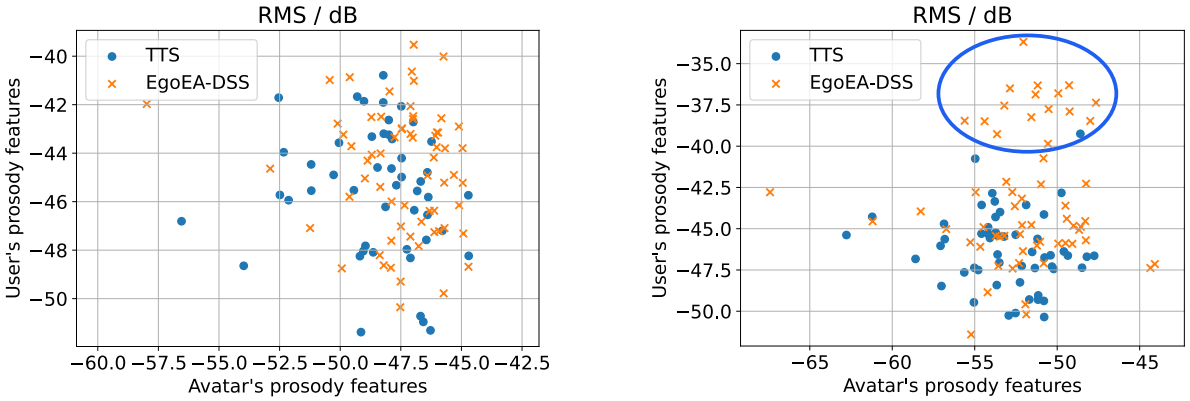
5. おわりに

本稿では，音環境に適応した発話音声を合成する能力を搭載した音声対話システムを構築し，このシステムが様々な実環境においてユーザーに自然で円滑なコミュニケーションを提供できるかを評価した．まず先行研究で開発した一人称視点の受聴音を活用した音環境に適応する対話音声合成 (EgoEA-DSS) システムを組み込んだ，モジュール結合型の音声対話システムを構築した．この音声対話システムの一要素としての EgoEA-DSS システムを，アバターと被験者が対話を行う実証実験により評価した．被験者がアバターとの対話体験に関する質問項目に回答した結果を集計したところ，EgoEA-DSS システムとベースラインの音声合成システムを搭載したアバターの間に，実環境にて自然で円滑なコミュニケーションを提供する能力に差が認められなかった．また，被験者とアバターの対話の中で，EgoEA-DSS を搭載したアバターの発話音声には周囲音環境への適応が現れなかった．この結果を鑑みて，今後はより多様な実環境での対話実験の実施や，質問項目などの実験デザインの再考を行いたい．また，EgoEA-DSS システムの性能改善を目し，より多様な実環境における自発対話のデータ収集・活用や，対話システム全体と情報をやりとりできる EgoEA-DSS システムの開発を行っていきたい．

謝辞 本研究の一部は，科研費 22H03639，23K18474，JST 創発的研究支援事業 JP23KJ0828，JST ムーンショット型研究開発事業 JPMJMS2011，及び Google Gemma 2 Academic Program GCP Credit Award の助成を受け実施しました．

参考文献

[1] Denes, P. B. and Pinson, E.: *The speech chain*, Macmillan (1993).
[2] Cooke, M., King, S., Garnier, M. and Aubanel, V.: The listening talker: A review of human and algorithmic context-induced modifications of speech, *Computer*



(a) 静かな屋内での対話音声　(b) 屋外での対話音声

図 6 アバターの発話音声と，その直前のターンの被験者の発話音声の音量 (RMS) の組の散布図．青い丸で囲った部分に含まれる発話音声は，その発出の際に風が強く吹いていた．

Speech & Language, Vol. 28, No. 2, pp. 543–571 (2014).

[3] Hazan, V. and Baker, R.: Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions, *The Journal of the Acoustical Society of America*, Vol. 130 4, pp. 2139–52 (2011).

[4] Lombard, E.: Le signe de l'élévation de la voix, *Annales des Maladies de L'Oreille et du larynx*, Vol. 37, pp. 101–119 (1911).

[5] Levitan, R., Gravano, A., Willson, L., Štefan, B., Hirschberg, J. and Nenkova, A.: Acoustic-Prosodic Entrainment and Social Behavior, *Proc. NAACL-HLT*, pp. 11–19 (2012).

[6] 井上昂治，河原達也：音声対話システム –基礎から実装まで–，オーム社 (2022).

[7] Xu, T.: *Neural Text-to-Speech Synthesis*, Springer Nature (2023).

[8] Take, O., Takamichi, S., Seki, K., Bando, Y. and Saruwatari, H.: SaSLaW: Dialogue Speech Corpus with Audio-visual Egocentric Information Toward Environment-adaptive Dialogue Speech Synthesis, *Proc. INTERSPEECH*, pp. 1860–1864 (2024).

[9] Minato, T., Higashinaka, R., Sakai, K., Funayama, T., Nishizaki, H. and Naga, T.: Overview of dialogue robot competition 2023, *arXiv preprint arXiv:2401.03547* (2024).

[10] Zorila, T.-C., Kandia, V. and Stylianou, Y.: Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression, *Proc. INTERSPEECH*, pp. 635–638 (2012).

[11] Bollepalli, B., Juvela, L. and Alku, P.: Lombard Speech Synthesis Using Transfer Learning in a Tacotron Text-to-Speech System, *Proc. INTERSPEECH*, pp. 2833–2837 (2019).

[12] Novitasari, S., Sakti, S. and Nakamura, S.: A Machine Speech Chain Approach for Dynamically Adaptive Lombard TTS in Static and Dynamic Noise Environments, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2673–2688 (2022).

[13] Guo, H., Zhang, S., Soong, F. K., He, L. and Xie, L.: Conversational End-to-End TTS for Voice Agents, *Proc. SLT*, pp. 403–409 (2021).

[14] Fischer, K., Naik, L., Langedijk, R. M., Baumann, T., Jelínek, M. and Palinko, O.: Initiating Human-Robot Interactions Using Incremental Speech Adaptation, *Proc. HRI*, pp. 421–425 (2021).

[15] Ren, Q., Hou, Y., Botteldooren, D. and Belpaeme, T.: No More Mumbles: Enhancing Robot Intelligibility Through Speech Adaptation, *IEEE Robotics and Automation Letters*, Vol. 9, No. 7, pp. 6162–6169 (2024).

[16] Hayamizu, A., Imai, M., Nakamura, K. and Nakadai, K.: Volume adaptation and visualization by modeling the volume level in noisy environments for telepresence system, *Proc. HAI*, pp. 67–74 (2014).

[17] Tuttosi, P., Hughson, E., Matsufuji, A., Zhang, C. and Lim, A.: Read the Room: Adapting a Robot's Voice to Ambient and Social Contexts, *Proc. IROS*, pp. 3998–4005 (2023).

[18] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *Proc. ICLR* (2021).

[19] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Proc. NeurIPS*, pp. 17022–17033 (2020).

[20] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F. and Saurous, R. A.: Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis, *Proc. ICML*, Vol. 80, pp. 5167–5176 (2018).

[21] Chiba, Y., Mitsuda, K., Lee, A. and Higashinaka, R.: The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models, *Proc. IWSDS*, pp. 1–6 (2024).

[22] Lee, A.: MMDAgent-EX (version 1.0.0) (online), DOI: 10.5281/zenodo.10427369 (2023).

[23] Ekstedt, E. and Skantze, G.: Voice Activity Projection: Self-supervised Learning of Turn-taking Events, *Proc. INTERPEECH*, pp. 5190–5194 (2022).

[24] 佐藤友紀，千葉祐弥，東中竜一郎：複数の日本語データセットによる音声活動予測モデルの学習とその評価，人工知能学会 言語・音声理解と対話処理研究会（第 100 回），pp. 192–197 (2024).

[25] Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, *Proc. INTERSPEECH*, pp. 1983–1987 (2023).

[26] Rouard, S., Massa, F. and Défossez, A.: Hybrid Transformers for Music Source Separation, *Proc. ICASSP*, pp. 1–5 (2023).

[27] Morise, M.: Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals, *Proc. INTERSPEECH*, pp. 2321–2325 (2017).