URL https://journal.ntt.co.jp/article/39223
DOI https://doi.org/10.60249/26055501

「ヒトのように会話するAI」をめざして —— Full-duplex型音声対話AIの研究開発

NTTでは，ヒトと同じように素早い応答や適切な相槌を行い，自然な抑揚で話すことができる音声対話AI（人工知能），「Full-duplex型音声対話AI」の研究開発を進めています．その実用化に向けた第一歩として，2025年のNTT R&Dフォーラムにおいて，コールセンタ自動応対AIを模した展示を行い，大きな反響をいただきました．本稿では，Full-duplex型音声対話AIの研究開発に関する一連の取り組みについて紹介します．

安藤 厚志（あんどう あつし）
高島 悠樹（たかしま ゆうき）
村田 俊樹（むらた としき）
NTT人間情報研究所

はじめに

音声による対話は，ヒトにおけるもっとも手軽なコミュニケーション手段の1つです．この音声対話の機能を持ったAI（人工知能），すなわち音声対話AIの研究開発が，世界中で盛んに行われています．音声対話AIの活用範囲は非常に幅広く，接客ロボットやコールセンタなどの対話が必要とされる業務の自動化，運転中の車内や高所作業などの手が塞がった状態での機械への指示，スマートグラスやスマートスピーカにおける手軽な入力インタフェースなどの実用化などが進められています．さらに，雑談を通じて私たちの生活に寄り添ったり，語学学習のための話し相手になったりする音声対話AIも社会に広がりつつあります．

音声対話AIにおける最新研究の1つとして，Full-duplex型音声対話と呼ばれる技術があります．従来の音声対話AIは，ユーザ（話し手）とAIが交互に話し，かつ相槌なども非常に少ない，トランシーバのような対話（Half-duplex：単方向通信）が主流でした．しかし最新の音声対話AIでは，話し声に対して自然な相槌を打つ，ユーザの話し終わりを待たずとも先回りしてAIが話し始めるなど，ヒトのように聞きながら話す対話（Full-duplex：双方向通信）ができるようになっています．加えて，抑揚や話し方なども多様かつ自然となり，まるで本当の人間と話しているかのような印象を与える音声対話AIが社会に現れようとしています．

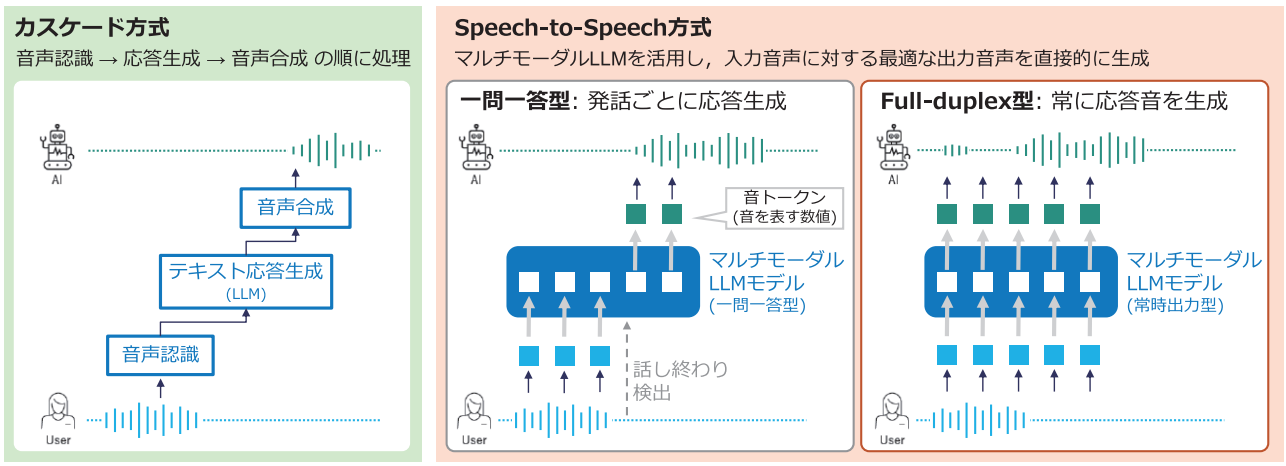
NTTにおいても，2025年度からFull-duplex型音声対話AIの研究開発を新たに立ち上げました．さらに新たなチャレンジとして，NTT R&Dフォーラム2025においてコールセンタ自動応対AIのProof-of-Concept（PoC）展示を実施し，非常に大きな反響をいただきました．本稿では，Full-duplex型音声対話AIの仕組みと，NTTにおける実用化に向けた取り組みを紹介します．

音声対話AIの分類：カスケード方式とSpeech-to-Speech方式

音声対話AIの実現方式は，大きく分けて，カスケード方式とSpeech-to-Speech方式の2つに分類できます（**図1**）．

■カスケード方式

カスケード方式は，特定の処理を行うブロックを直列に組み合わせて対話を実現する方式です．一般には，音声認識・テキスト応答生成・音声合成の3つのブロックで構成されることが多いです．ユーザから音声対話AIに向けた声は，まず音声認識に



Speech-to-Speech方式は，一問一答型とFull-duplex型の2種類にさらに区分できる．

図1　音声対話AIの分類

よって話している内容を表すテキストに変換されます．その後，テキストからユーザの話した内容を理解し，AIが話すべき応答をテキストとして生成します．最後に，この応答テキストを音声合成によりAIの声に変換して出力することで，AIによる声での応答を行う，という仕組みです．2026年現在でも，実用化されている多くの音声対話AIではカスケード方式が採用されています．

カスケード方式の最大の長所は，個々のブロックを差し替えることができる点にあります．例えば，音声合成ブロックを別の人物の声に差し替えることで，AIの声色を変えることができます．さらに近年では，テキスト応答生成に最新のLLM（Large Language Models）を用いることで，ユーザの意図を正確に理解し，複雑かつ知的な応答テキストを生成できるようになりました．

しかし，カスケード方式の短所の1つとして，ユーザの声に対する反応が遅れやすいという点があります．これは，3つの処理ブロックが直列につながっており，前のブロックの結果がそろってから次のブロックの処理を行うために遅延が積み重なってしまうことが原因です．ほかにも，AIの声の抑揚が乏しくなりやすい，音声認識によりユーザの声をテキスト化することで話し方や抑揚などの情報が失われ，これらを考慮した受け答えができない，などの課題もあります．

■Speech-to-Speech方式

Speech-to-Speech方式（End-to-End方式とも呼ばれます）は，音声認識や音声合成などのブロックを明確に分けず，ユーザからの入力音声に対してAIの声を直接的に出力するという方式です．この方式は，LLMを音声に拡張した，マルチモーダルLLMと呼ばれる技術を基礎としています．Speech-to-Speech方式では，大量の音声やテキストを与えて，入力に対してどのような声を出すべきかを直接的に学習させます．これにより，ユーザの声に含まれる話し方や年齢・性別などのテキスト化すると失われてしまう情報を扱うことができ，それらに適した抑揚を持つAIの声を出力することができるようになっています．

Speech-to-Speech方式の音声対話AIは，さらに2つの種類に分けることができます．1つは，一問一答を繰り返して対話を行うタイプです．この方式は，ユーザの話し始めや話し終わりを検知する技術（発話区間検出）と組み合わせて利用されます．一問一答タイプはいわば「話す内容」と「声色や抑揚」をAIが予測するというもので，ある程度多くの学習データ（入力に対してどのような声を出すべきかを教えるデータ）さえあれば，AIが習得可能な課題であるといえます．さらに，一問一答のやり取りは1つの会話に何度か表れることもあるため，AIの学習データを数多く集めやすいという利点もあります．しかし，「話すタイミング」は発話区間検出によって決定されるため，ユーザが相槌を打つだけでも過敏に反応してしまったり，ユーザの話し終わりに対する反応が若干遅れてしまったりすることがあるという課題もあります．

Speech-to-Speech方式のもう1つの種別が，常にAIが声を発し続けることができるタイプで，こちらがFull-duplex型音声対話AIと呼ばれています．AIはユーザの声をごく短い時間間隔（例えば0.08秒ごと）に解析し，その時刻において出すべき短い音の断片を出力します．AIが話すべきでないタイミングでは，「何もない」音を出すようにします．これを繰り返すことで，ユーザの話し声に対する素早い応答や，自然なタイミングでの相槌を実現できることが最大の利点です．しかしこの方式は，「話す内容」「声色や抑揚」に加えて「話すタイミング」を非常に短い間隔で推定し，かつ会話全体でみても意味が正しく自然な声となるように音の断片を作成しなければなりません．加えて，音声対話データは一問一答データに比べて集めることが難しいため，技術的な側面でも，学習データ収集の側面でも，非常に難易度が高いといえます．

コールセンタ自動応対AIの実現に向けた取り組み

私たちは，Full-duplex型音声対話AIの有望な応用先の1つとして，コールセンタにおける自動応対AIを考えています．現在でも，コールセンタ向け音声対話AIを実現する製品はありますが，話し方が人間のオペレータとは全く異なっており，顧客に受け入れられないことがあるという課題があります．Full-duplex型音声対話AIが持つ素早い応答や自然な相槌などの機能，さらに人間に近い抑揚などの特性は，対話AIに不慣れな方を含めたあらゆる顧客に対して，自然で親しみやすい自動応対サービスの提供につながるのではないかと考えました．

このような発想のもと，私たちの研究チームでは「NTT R&Dフォーラム2025において，Full-duplex型音声対話技術を活用したコールセンタ向け自動応対AIのPoCを展示する」という目標を立てました．これは，Full-duplex型音声対話技術の研究開発は世界中で盛んに進められていることから，スピーディな実用化によって他社優位性を高めたいと考えたためです．しかし，この計画を立てたのは2025年4月，R&Dフォーラム当日の11月までは8カ月弱しかありません．データやプログラムなどほぼ何もそろっていない状況からの，非常に厳しいチャレンジでした．

PoC実現に向けたさまざまな取り組みの中で，NTT独自の工夫を加えた点を2点紹介します．1点目は，音声対話AIに学習させる対話データの高品質化です．実は，Full-duplex型音声対話AIの学習においては，雑音などが含まれず明瞭な音声だけを含む人間どうしの対話データに加えて，どの時刻にどの単語を話したのかという精緻な書き起こし情報も必要です．これらをそろえるにあたり，NTT人間情報研究所やNTTコミュニケーション科学基礎研究所で長年研究されてきた音声処理技術，例えば雑音抑圧や音声強調，音声認識などを積極的に導入しました．これにより，高品質な対話AI学習用データの構築を短期間で実現できました．

2点目は，対話AIに学ばせたい振る舞いを含む音声対話データを人工的につくる技術（疑似音声対話生成技術）を新たに開発した点です．今回のPoCでは，AIが「用件や顧客名の復唱」を必ず行うことをめざしました．しかし，前述の方法で構築した学習用データには，復唱が含まれる対話は少数しかなく，復唱の学習は極めて困難でした．この課題を解決するため，復唱や決められた案内など，AIに獲得してほしい振る舞いを含んだ疑似音声対話データを生成

し，実際の音声対話データと合わせて対話AIに学習をさせています．疑似音声対話生成においては，LLMを活用して対話内容を生成することで言葉選びの違いに対する頑健性を向上させたり，NTT人間情報研究所が開発した高品質な音声合成技術を利用することで個々の話し声の自然性を上げたりといった，さまざまな工夫を加えました．それだけでなく，疑似音声対話と実際の音声対話データをうまく組み合わせることができる，NTT独自の対話AI学習手法も創出しました．これにより，話す内容は疑似音声対話をできるだけ踏襲しつつ，抑揚のつけ方や声色は実際の音声対話データから獲得できるようになり，高い音声品質を保ちながら特定内容の対話を行うことができる音声対話AIを構築できました．

向かって左側のモニタでは従来のカスケード方式音声対話AIの例を動画再生し，右側では研究員と音声対話AIがその場で対話を行った．

図2　NTT R&Dフォーラム2025におけるFull-duplex型音声対話AIの展示ブース

NTT R&Dフォーラムへの出展と反響

NTT R&Dフォーラムでは，コールセンタ通話の冒頭部分（用件確認・復唱，顧客氏名確認，顧客電話番号確認，保留の案内）を想定し，顧客役の人間と対話AIがその場で対話をする，ライブデモ方式を採用しました（**図2**）．来場者の声が入らないように部屋を仕切った以外は，実際の利用環境とほぼ同等といえます．また，多くの生成AIと同様に，人間が同じように話しても対話AIの応答は毎回少しずつ変化します．対話AIが想定どおりの応答をしない可能性もある，挑戦的な展示でした．

結果として，R&Dフォーラムにご来場いただいた方々からは，私たちの想像をはるかに超えた反響をいただきました．多くの方から驚きと称賛のお言葉をいただき，技術内容や実用化時期などたくさんの質問を頂戴しました．「AIではなく人間が応答しているかと思った」という意見もしばしばいただくほどでした．Full-duplex型音声対話AIは基調講演でも紹介され[(1)]，R&Dフォーラム後にはNTTグループの事業会社だけでなくグループ外企業・研究機関の皆様からも，非常に多くの問合せをいただきました．音声対話AIに向けた社会からの期待は非常に大きいものであると，私たちは強く認識しました．

今後の展開

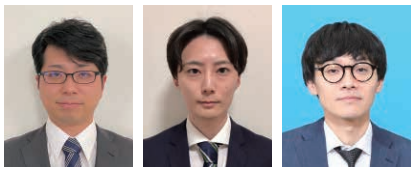
私たちは，Full-duplex型音声対話AIの早期実用化に向けた技術改良を続けています．例えば，コールセンタ向け自動応対AI領域では，R&Dフォーラムでのデモよりもさらに長く複雑な応対（例えば，特定サービスの予約や顧客情報の変更依頼の通話冒頭から終了まで）においても正確な受け答えを実現する技術や，顧客管理システムと連携することで自動応対AIだけで通話対応を可能とする技術の研究開発を進めています．また，上記の研究開発と並行し，2026年度中の実証実験に向けた準備を始めました．今後も自動応対AIの早期実用化をめざして，さまざまな取り組みを加速させる予定です．

さらに，Full-duplex型音声対話AIの活用先は，コールセンタ向け自動応対AIにとどまらないと私たちは考えています．冒頭で述べたように，音声対話AIはさまざまなデバイスにおけるユーザフレンドリーな入出力インタフェースにもなります．雑談ができる音声対話AIは，単に日常的な話し相手にもなり得るだけでなく，特定のキャラクタや人物を模倣させることで「いつでも，どこでも，話したい存在と話すことができる」体験を提供できるでしょう．このような音声対話AIの応用領域を広げる研究にも着手しており，現在はNTTコミュニケーション科学基礎研究所の研究員とも組織横断的な連携を進めています．

これからも，Full-duplex型音声対話AIの研究開発や社会実装を通じて，ヒトとAIとの新たなコミュニケーションを創出し，ヒトとAIが共生する社会の実現をめざします．

■参考文献

(1) https://youtu.be/7vwsIOtqJOU?t=1726s



（左から）　安藤　厚志／高島　悠樹／村田　俊樹

NTT人間情報研究所では，Full-duplex型音声対話AIの研究開発や社会実装を通じて，ヒトとAIとの新たなコミュニケーションを創出し，ヒトとAIが共生する社会の実現をめざします．

◆問い合わせ先

NTT人間情報研究所
デジタルツインコンピューティング研究プロジェクト