音声対話システムの開発プロセス

Fairy Devices Inc.

2021–10–07

1　概要

近年の音声言語処理技術の発展を背景として、個人用・家庭用・産業用など幅広い用途において、音声対話システムの重要性は増しています。しかしながら音声対話システムには、音声信号処理・自然言語処理・対話制御など技術的な側面、人間と機械のあいだのコミュニケーションに関する心理学的な側面、多数のハードウェア・ソフトウェアのコンポーネントから構成されるシステムとしての側面など、さまざまな要因が絡んでおり、誰でも容易に開発できるとは言いがたい状況です。本稿では、音声対話システムの典型的な開発方法について説明します。特にビジネス目的で製品化に向けて音声対話システムを開発する場合の、典型的な開発プロセスに焦点を当てます[*1]。

音声対話システムの典型的な開発プロセスの概要を図 1 に示します。典型的な開発プロセスは以下の 3 つのフェーズから構成されます。

- デザイン
- プロトタイピング
- 製品化

まず始めに、システムそのものの「人格」でありユーザーと直接インタラクションを行う主体でもある「エージェント」のデザインを行います。次に決定したデザインに基づいてプロトタイプ版を開発します。いずれのフェーズでもユーザー評価により問題点を明らかにし、要求する水準に達するまで当該フェーズを繰り返します。最後にプロトタイプ版の仕様に基づいて、製品版を開発します。以下では各フェーズの内容と進め方について説明します。

2　デザインフェーズ

デザインフェーズではシステムの果たすべき目的を定め、エージェントのデザインを決定し、その妥当性を検証します。このフェーズでは以下のことを実施します。

- 音声対話システムの目的を明確化・エージェントのデザインを決定
- 対話ルールのフレームの決定・対話シナリオの作成
- WOZ 方式によるユーザー評価試験
- 必要に応じて上記を繰り返す

デザインフェーズの終了時には、これから開発する音声対話システムがどのように振る舞うべきかの基準が明確に

[*1] 本稿はあくまで典型的な開発プロセスを紹介するものです。個々の事例においてはそれぞれ状況が異なり、必ずしも本稿の説明が当てはまるとは限りません。実際の開発においては、さまざまな相違が発生する可能性があります。また研究目的、個人での開発などビジネス目的以外の開発において、本稿で紹介する開発プロセスを推奨するものではありません。

図 1　開発プロセスの概要

なっているはずです。この基準が曖昧なままステップを進めてしまうと、開発プロセスが迷走して行き詰まったり、ちぐはぐで混乱したシステムや出来上がったりするリスクがあります。

2.1　音声対話システムの目的

始めに、音声対話システムの目的を明確にする必要があります[*2]。ここで定めた目的が以降の開発プロセスにおける指針となります。

音声対話システムの目的とは

音声対話システムが果たすべき役割、存在意義、実現するべき価値、ユーザーへもたらすべき影響などの総称

2.2　エージェントのデザイン

策定したシステムの目的に基づいてエージェントのデザインを行います。エージェントとは、自ら判断し行動する主体 [1] のことを指します。全ての音声対話システムは会話を通してユーザーと協調しながら課題を遂行するため、ユーザにとっては、音声対話システムそのものがエージェントであるとみなされます[*3]。エージェントはシステムに対してユーザーが持つ印象[*4] をうまく利用して、これを裏切らないように慎重にデザインすることが重要です。

エージェントのデザインは以下の 3 要素から構成されます (図 2 を参照)。

ペルソナ　エージェントの性格特性（パーソナリティ [6] 及び行動の特徴的なパターン [7])。ペルソナに一貫性を持たせることで、システムの目的が明確になり、またシステムの振る舞いが予測可能になります [8]。

[*2] 音声対話システムはしばしば目的志向・非目的志向に分類されます。前者は実行すべきタスクが明確なもの（例えばチケット予約システムなど)、後者はそうでないもの（例えば雑談対話システムなど）です。しかしこの分類は必ずしもバイナリなものではなく、多くの音声対話システムが両方の特性を兼ね備えています。例えば雑談対話システムでも作成者は何らかの目的を設定している事が多いでしょう。本稿では音声対話システムの目的を明らかにした上で開発を進めるという前提の下で議論を進めます。

[*3] 音声対話システムが自ら判断する能力（自律性）を十分に備えていなかったとしても、ユーザの発話意図を適切に理解し、ユーザーの課題達成を支援するために必要な機能を備えていればエージェントの要件を満たします。

[*4] ユーザーがシステムに対して抱く「こう操作すれば、こう動くであろう」という予測や期待をメンタルモデルと呼びます [2]。メンタルモデルは必ずしも機械の実際の動作を正確に反映しているわけではなく、擬人化 [3]、シグニファイア（アフォーダンス）[4] や身近なものからの類推など、さまざまなユーザーの解釈を通して形成されます。エージェントはユーザーの先入観やイメージをうまく利用して、機械の実際の振る舞いと、ユーザーのメンタルモデルの仲介を行う存在であるとも言えます。

アビリティ　エージェントの能力（何ができるか・できないか）、知識（何を知っているか・知らないか）、知覚（視覚・聴覚など）。システムが実際に提供するサービスをエージェントデザインに反映させてわかりやすくユーザーに提示する必要があります。

ボディ　エージェントの見かけ・外見、及びどのような手段で情報を表現するのか。スマートスピーカーであれば筐体の外観、サイネージ・アプリであればディスプレイ、ロボットであれば可動部を考慮した上で、ビジュアル表現、音声の声質[*5] 、その他のパラ言語・非言語的表現（イントネーションの同調 [9] や、うなずき・視線・ジェスチャーなど [10]）をデザインします。

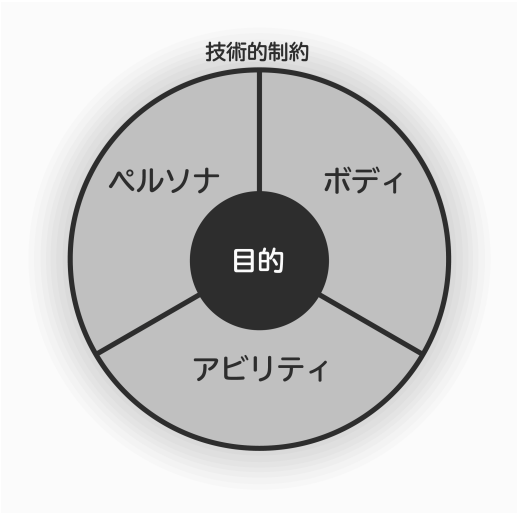

図 2　エージェントデザインの 3 要素

これらの 3 要素が互いに一貫性を保つようにデザインを行うことが重要です[*6]。さらに技術的妥当性の制約を考慮する必要があります。すなわち、現在の技術水準において実現可能なものであるか、専門的な知見に基づいて検証を行う必要があります。ただしデザインフェーズの段階では、まだシステム構成、ハードウェア・ソフトウェア設計、音声言語処理の技術選定など技術面での具体的な事項については検討する必要はありません。

ここでエージェントデザインはシステムの目的、および技術的妥当性の制約から論理的に決定すべきものであることに注意してください。言い換えると、エージェントデザインの各要素を恣意的に選んでよい訳ではありません。エージェントデザインの注意点を以下にまとめます。

エージェントデザインの注意点

- 音声対話システムの目的に基づいてデザインする。
- ペルソナ・アビリティ・ボディの 3 要素が一貫性を保つようにデザインする。
- 技術的妥当性の制約の下でデザインする。

これらの点に注意しながらシステムの目的に沿った一貫したエージェントデザインを決定します。エージェントデ

[*5] システム応答の声質に生の人間の声を使うのか、より機械的な音声を使うのかは、他の要素を考慮して決定します。人の生声を使ったシステムは人と同等の会話能力をユーザーに期待させるため、実際の機能との齟齬がユーザーの失望を招く場合があります [11]。例えば、単純なタスクでは流暢な人の声よりも機械的な合成音声のほうがタスク成功率が上がることがわかっています [12]。

[*6] 一般的な機械のデザインにおいて機能（アビリティ）と外観（ボディ）の 2 つの要素の一貫性が重要なのと同様に、音声対話システムのデザインにおいてはこれにペルソナを加えた 3 要素の一貫性を考慮する必要があります。機能や外観と一致していないペルソナを与えられた機械はユーザーにとって使いづらく、支離滅裂な印象を与えてしまいます [13]。

ザインを適切に行うことにより、音声対話システムがどのように振る舞うべきかの基準が明確になります。

2.2.1 擬人化について

エージェントデザインにおいて過度な擬人化は原則的に避けた方がよいでしょう。人間を基準として、あるいはできるだけ人間に似せてエージェントデザインを決めてしまうと、ユーザーは音声対話システムに対して、人間と同じような能力や行動パターンを期待しがちです [14]。しかしながら現在の技術水準では、「人間のような」システムを実現するのは困難であり、結果としてユーザーの違和感・失望を招く事態になってしまいます [15]。したがって音声対話システムのエージェントデザインにおいて過度な擬人化は避けた方がよいでしょう[*7]。(ただし、あまりにも人間から遠いエージェントデザインを選択してしまうと、ユーザーとのコミュニケーションが困難になってしまう可能性もあるため注意が必要です。)

2.2.2 エージェントの外見について

ボディがディスプレイを備えた筐体である場合、あるいはロボットを自由に設計できる場合、エージェントの外見は一貫性を考慮して決める必要があります。エージェントの外見は、必ずしも生物的なアバターである必要はありません。外見のビジュアル表現を使わずに、音声による対話のみを行う選択肢もあります。

2.3 WOZ によるユーザー評価試験

デザインフェーズの最後に、デザインの妥当性を検証するためのユーザーによる評価試験を行います。この時点での評価対象はデザインのみですから、システム設計・実装や音声言語処理の精度による影響はできるだけ除外して、デザインに従って理想的に動作した場合の評価を行う必要があります[*8]。このため、音声対話システムの役割を人間が担うこととします。このような方式を Wizard of Oz (WOZ) と呼びます。具体的には以下の手順で WOZ によるユーザー試験を実施します。

- 評価試験のための準備
  - 試験用の環境構築・システム開発
  - 対話シナリオ作成
- 評価指標の策定
- 評価試験の実施
- 評価結果の分析と課題の洗い出し

ユーザー評価試験終了後には、以下のことが実現されているはずです。

ユーザー評価試験により達成されること

- 評価基準が定まる
- エージェントデザインが適切であるか・次のフェーズで進んでもよいかの判断が得られる

この時点で評価基準が定まっていないまま開発プロセスを進めてしまうと、次フェーズ以降においてシステム設

[*7] 機械には(人間とは異なった)機械としての知性を想定すべきであるとも言えます。この事は、動物を過度に擬人化して捉えると、動物の知性の理解を妨げてしまうという事からも理解できるかもしれません。人間とは異なる知性を持ち、かつ、人間とコミュニケーション可能な存在として音声対話システムのエージェントデザインを行うべきでしょう。

[*8] 統計的機械学習モデルがシステムにとって本質的な役割を果たしている場合には、WOZ 方式によるユーザー評価試験が意味をなさない可能性もあります。例えば、特に制約のない(汎用の)雑談対話システムの場合には、WOZ 方式でユーザー評価試験を行なっても、エージェントデザインの設計の妥当性を評価することは難しいでしょう。このような場合には、本ステップは必然的にスキップすることになります。

計・開発の状況などの内的要因、あるいは開発チーム外からの意見など外的要因により基準自体が迷走してしまう可能性があります。またこの時点で、決定したエージェントデザインが妥当であったかの検証を行うことが重要です。ユーザー評価試験の結果、エージェントデザインが一貫していない、あるいは目的に沿っていないことが判明した場合にはエージェントデザインを再度行ってください。

2.3.1 環境構築とシステム開発

まず始めに、評価試験用の環境構築を行います。システムがどのような環境で（室内か屋外か、周囲に雑音はあるのか、どれくらいの人がいるのかなど）、またどのような状況で（ユーザーは一人か複数人か、動き回るのかどうかなど）使用されるのかは、システム構成を決定する上で重要になるので、この時点で具体化しておく必要があります。さらに、同じ音声対話システムでも環境が異なれば、ユーザーの印象や話し方が変わる可能性があるため [16][17]、最終的な製品版がリリース後に利用される環境を可能な範囲で再現することが好ましいと言えます。例えばリビングルーム・展示会場・駅構内・工場の作業場など、あるいはこのような環境を模擬した試験環境を利用することが考えられます。ここで準備した試験環境は、以降のフェーズにおいても繰り返し使用されます。

さらに、WOZ 方式によるユーザー評価試験を実施するための、最低限のシステム開発を行う必要があります。検討したエージェントのデザインに沿ってモックアップのシステム（例えば [18]）を開発します。複数のデザイン要素によって構成されたシステムは、一部の部品のデザインのわずかな変化がシステム全体のユーザー体験を大きく変えることが知られています [19]。特にエージェントのボディはユーザーの印象に大きな影響を与える可能性があるため、可能な範囲で製品版で想定されているボディと近いものを利用することが望ましいと言えます。エージェントのペルソナはタスクの目的や主要なユーザー層のメンタルモデルに合致したデザインが望ましく、WOZ によるエージェントの比較実験が有効です [20]。対話の体験の質を左右するエージェントの見た目（ビジュアル表現）[21] やシステム音声の声質 [22] に関しては、エージェントデザインに従ったものを準備してください。パラ言語表現も対話の印象に寄与しますので、システム役の人間が簡単な操作だけで実行を指示できる動き（うなずき [23]、LED の既定パターンの明滅 [24] など）はこの時点で実現しておいた方が好ましいでしょう。

WOZ のシステム開発においては、システム役の人間の存在をユーザーに意識させないことが重要です。そのため、主に以下の点に注意してください。

- ユーザーにはエージェントのボディは見えるが、システム役の人間は見えないこと
- ユーザーの声がシステム役の人間に聞こえること
- システム役の人間がシナリオに沿ってシステム応答を選択するための補助があること（シナリオ・応答文を掲載したドキュメント、応答文を画面上で選べる簡単なアプリケーションなど）

マイクやスピーカーは試験実施のために都合のよいものを使用してください。この時点ではエージェントのボディと接続されていなくても、ボディのデザインに沿っていなくても構いません。例えばボーカルマイクと小型スピーカー、あるいはヘッドセットを用いてもよいでしょう。

WOZ におけるシステム応答の音声の生成については、エージェントデザインの方針に応じていくつかの選択肢が考えられます。例えば予め録音しておいて音声を再生する、システム役の人間が入力したテキストから音声合成する、システム役の人間の声を声質変換して使う、システム役の人間の声をそのまま使う、などです。これに応じて、システム役の人間がシステムをどのように操作する必要があるか（応答文を画面上から選択するか、テキスト入力するか、声で発声するか）が決まります[*9]。

[*9] 音声対話システムでは応答のタイミングが極めて重要で、早過ぎたり遅すぎたりするとユーザーに不信感を与えたり不満を与えたりします。そのため WOZ の試験では、最終的な製品版で技術的に可能な応答速度を可能な限り再現することが望ましいでしょう。WOZ システムの応答のタイミングも、この遅延を考慮して決定する必要があります。

なお複数人のユーザが同時にシステムを利用する場合でも、WOZ は有効な評価手段です [25]。

2.3.2 対話の枠組み決定・シナリオ作成

続いて、エージェントのデザインに沿って対話シナリオを作成します。WOZ の実験においてシステム役の人間は、このシナリオに沿ってユーザーと対話を行います。対話制御にはさまざまな枠組みが考えられますが、ここではまずはユーザー発話を機械に理解させるための枠組みを考えます[*10]。ここでは詳しい説明は省略し、概念の説明のためにインテンション・スロットという枠組みを焦点を当てます。この枠組みでは以下のような入出力と応答決定ルールを考えます。

入力　システムが認識するユーザー発話のインテンション・スロットを定める
出力　システムの実行可能なアクション、および応答文を定める
対話ルール　ある状態である入力を受け立った時に、どの出力を実行し、どの状態に遷移するかのルールを定める

インテンションはユーザーの意図を意味し、例えば「照明を付けるコマンド」「スケジュールを確認するコマンド」などが挙げられます。スロットは変数に相当するもので、例えば「行き先駅」「日付」などが挙げられます。このようにユーザー発話を離散的な記号に帰着させることで、多種多様なユーザー発話のパターンを吸収し、曖昧さを回避します [27]。WOZ 方式では人間が既定のインテンション・スロットを判断し、プロトタイピング以降では機械学習モデルなどにより自動で推定することになります。

また、ここでシステムが実行可能なアクションを決定します。例えば「(ユーザー発話が理解できなかったため) ユーザーに再度発話を促す」「推定した意図が正しいか確認する」「照明を付ける（ために外部機器を操作する)」「カレンダーから特定の日付の予定をリストアップする」などが挙げられます。WOZ による評価試験を通して、対話システムに対してユーザーがどのような反応を返すか、取るべきアクションに漏れがないかを洗い出します [28]。さらにシステムの応答文を定める必要があります。応答文の準備の仕方にはいくつかの方法があります。例えば、固定文、スロットつきのテンプレート文、WOZ によるユーザー試験ではシステム役の人が応答文を考える・プロトタイピング以降では機械学習モデルにより応答文生成を行う、などの方法が考えられます。

最後に、対話ルールの大枠を決めておく必要があります。システムは対話の流れに応じた状態を持つことを想定します。対話ルールとは、ある状態において、ある入力を受け付けた時にどのような出力を選択し、次にどの状態に遷移するかを表します。入力と出力はエージェントデザインに従って網羅的に定義してください。エージェントのペルソナに合致しない発話、アビリティを損なう冗長すぎたり不明瞭すぎたりする発話、ボディからユーザが得る印象に著しくそぐわない発話などは慎重に検討し、除外する必要があります。一方でこの段階においては、WOZ 方式によるユーザー試験を実施するためのシナリオが準備できていれば十分であり、細部まで網羅的なルールを考える必要はありません。例えば、ある状態で受け付ける入力は 1 通りのみに制限してしまってもよいかもしれません。ただし、エージェントのアビリティ・ペルソナの妥当性を検証するために十分なバリエーションのシナリオを準備してください。さもないとエージェントデザインの検証が不十分なものになってしまう可能性があります。

プロトタイピング以降では機械学習によって応答を決定する場合にも、デザインフェーズではシナリオベースで WOZ 方式によるユーザー評価試験を行うことは可能です。このような場合には、機械学習モデルと準備したシナリオがかけ離れたものになる可能性があるため、エージェントデザインの検証を行うのは簡単ではありません。した

[*10] プロトタイピング以降では（POMDP[26] などの）統計的手法を用いる場合であっても、入力・出力の枠組みは有効です。このような場合には、WOZ によるユーザー評価試験では人手で設定した対話ルールを用いてください。プロトタイピング以降では、入力・出力の枠組みを維持したまま対話制御を統計的手法に置き換えてください。また、同様の手順でユーザー評価試験を実施することが可能です。雑談システムの場合には本ステップはスキップされます。例えば用例ベースによる雑談対話システムについては、ここでは考慮から除外されています。この他にも WOZ によるユーザー評価試験が適さない場合があります。

がってプロトタイピングフェーズにおいて、イテレーションを繰り返して機械学習モデルを改善する必要があります。

2.3.3 評価指標

デザインフェーズでは WOZ 方式によって、またプロトタイピングフェーズではプロトタイプ版システムを使用することによって、ユーザー評価試験を実施します。これらの試験の目的は各々のフェーズでエージェントデザイン・対話ルール設計、プロトタイプ版システムを目的に沿って評価することです。対話ログの収集・分析により問題を発見することも重要ですが、開発プロジェクト全体に渡って適切な評価を行うためには、どの程度の水準で音声対話システムの目的を達成できているかを測る一貫した基準が必要です。例えば以下のような評価軸が考えられます。

- 既定の条件下において、既定のタスクを円滑な対話により解決できたか（解決できたタスクの割合など）
- ユーザーがいつどこでタスク解決を諦めたか（解決できなかったタスクの割合、諦めたタイミングなど）
- アンケート調査法を用いたユーザーによる主観評価（Godspeed 尺度 [29] やロボット否定的態度尺度 [30] など、広く信頼性と妥当性が確認されている評価尺度を使用します）

評価基準は、実行可能な評価手順を伴ったものでなければなりません。例えば、抽象的な文章だけであったり、著しく測定の再現性を欠いていたり、長期間の観測が必要となる基準は適していません。一貫した評価基準がないと、開発計画における現在地を見失ってしまうリスクがあります。

プロトタイピングフェーズにおいては各技術要素の精度（音声認識の単語誤り率など）も定量的な指標として考えられます。これらの精度を高めることは重要ですが、それがシステム全体の評価の向上に有意に寄与するとは限らない点には注意が必要です。ある要素技術のわずかな精度向上は、ユーザーの印象を変えるほどの効果を及ぼさない可能性があります。一方で特定の要素のわずかな違いがユーザーの印象に決定的に大きな影響を与える場合もあります [19]。したがって、各要素技術の精度はシステムをどのように改善すればよいかを知るためには有用ですが、システム全体の評価の代替にはならない事に注意してください。

2.4 次フェーズへ進むべきかの判断

デザインフェーズ終了時には、以下のことが実現されているはずです。

デザインフェーズで達成されること

- 音声対話システムの目的に基づいたエージェントのデザインが定まっている
- 評価用の環境が整い、評価基準が定まっている
- ユーザー評価試験の結果が得られている

エージェントデザインが一貫していること、目的に沿っていること、および期待する水準に達していることが確認できるまで、デザインフェーズを繰り返し実施します。

また、以下のような判断に至った場合には、音声対話システムの目的自体を見直す必要があります。

- 技術的制約を考慮すると期待した水準で目的を達成できない
- 期待した水準で目的を達成するためには技術的制約を満たせない

場合によっては、現在の技術水準や想定するプロジェクト規模に対して開発プロジェクト自体の目標が過大である可能性もあります。このような場合には開発プロジェクト自体を見直す必要があります。デザインフェーズでユーザー試験評価を行わないと、このような重大な問題を見過ごしたままプロトタイピングや製品化フェーズに進んでしまうリスクがあります。

3　プロトタイピングフェーズ

プロトタイピングフェーズの目的は前フェーズでデザインしたエージェントを実際に動作するシステムとして実現し、その評価を行うことです。このフェーズでは以下のことを実施します。

- 対話ルール作成
- システム設計・実装
- ユーザー評価試験
- 必要に応じて上記を繰り返す

本フェーズ終了時には、エージェントデザインが対話ルールおよびシステム設計に落とし込まれ、実際に動作するプロトタイピング版が実現されているはずです。ユーザー評価試験の結果を受けて、プロトタイプ版を仕様として決定し製品化フェーズに進みます。

エージェントデザインを対話ルール・システム設計に落とし込むことは自明な作業ではありません。特にシステムに対するユーザーの評価は予測がつきにくい側面があり、個人差が大きく、エージェントデザインとユーザーの趣向には複雑な相互作用がある [31] ため、多かれ少なかれ試行錯誤が必要になります。したがって、本フェーズでは設計・実装とユーザー評価試験を繰り返すことによってシステムのクオリティを向上させていくことが要求されます。

3.1　対話ルール作成

デザインフェーズでは対話ルールの枠組みを決定し、WOZ によるユーザー試験のためのシナリオを作成しました。プロトタイピングフェーズでは、選択した枠組みに基づいて網羅的な対話ルールを作成します。対話ルールの作成が初回だけで完了することは、ほとんどありません。ユーザー評価によるフィードバックに基づく対話ルールの改善を繰り返し行う必要があります。例えば、想定していなかったユーザーの行動パターン[*11] にどのように対応すべきかを決定し、対話ルールに反映していくことが重要です。

ユーザーが不特定多数である場合には、ユーザーから不適切な発言・攻撃的な発言が投げかけられる可能性があります。このようなユーザーの振る舞いに、どのように対応するかを決定する必要があります [33]。

対話ルールの決定において、デザインフェーズで決定したエージェントデザインが判断の基準となります。システムがエージェントデザインから逸脱した振る舞いをしないように注意を払ってください。この段階で、デザインにない機能を追加するなど、場当たり的なエージェントデザインの変更は避けなければなりません。

3.2　システム設計

音声対話システムは、エージェントデザインに基づいて論理的に設計されなければなりません。言い換えるとシステム設計を恣意的に決めてはいけないという事です。したがって音声対話システムの適切な設計はエージェントデザインによって異なったものになります。以下では、まず典型的なシステム構成を紹介し、次に設計上のポイントを説明します。

[*11] ユーザーの想定外の行動として、話しかけるタイミングがわからなかった、タスク達成までの筋道がわからなくなった、認識できない語彙を頻繁に使用した、想定していない機能を使おうとしたなど、さまざまなパターンが考えられます [32]。

3.2.1　典型的なシステム構成

まずは典型的なシステム構成を紹介します。ここでは音声言語処理の各要素技術をコンポーネントとして扱い、これらのコンポーネント間の情報の流れに着目します。図 3 は典型的なシステム構成図を表します。ここでインターフェースは筐体、ディスプレイ、LED、マイク、スピーカーなどを表します。また外部 API とはシステムがアクションを実行するためにアクセスする外部サービスを表します。例えば駅案内システムの場合には、経路検索 API が該当します。その他の各要素技術について以下で簡単に説明します。

図 3　典型的なシステム構成

前処理　マイクから取得した音声から必要なデータ（ユーザーの発話）を取り出します。ユーザー発話の抽出のために、自己発話抑制、雑音抑制、残響抑制、音源分離、音声区間抽出などの技術をシステムの利用環境に応じて適宜選択し、使用します。さらに、ユーザーの対システム発話のみを判別するためにウェイクワード検知が用いられる場合もあります。

音声認識　ユーザー発話を自然言語に変換します。音声認識に加えて感情認識、パラ言語認識、話者認識などが用いられる場合もあります。

意図理解　自然言語として表されたユーザー発話の意図を推定します。

対話制御　ユーザー発話の意図からシステムのアクションを決定します。ルールベース、機械学習モデル、それらのハイブリッドなどの手法があります。決定すべきアクションには外部サービスの操作、システム応答発話、筐体・ディスプレイ上の動作、LED 明滅パターンなどが含まれます。

言語生成　システムアクションに応じたシステム発話文を作成します。固定文、用例ベース、機械学習モデルによる生成などの手法があります。

音声合成　システム発話文から音声データを生成します。

エージェントデザインに応じて、図 3 とは異なるシステム構成になる場合もあります。上記のコンポーネントがすべて使用されるとは限らず、また他のコンポーネントが必要となる場合もあります。

3.2.2　ソフトウェアの構成

システム設計の 1 つのポイントは、ソフトウェア構成の決定です。すなわち必要十分なコンポーネントをリストアップし、適切な連携方法を判断する必要があります。各コンポーネントについて、計算資源の割り当ても問題になります。音声対話システムはさまざまな要素技術を必要とするため、ユーザーインターフェースの役割を果たすハードウェア本体（以下ではクライアントサイドと呼びます）だけですべての処理を実行しきれない場合が多々あります。またインターネット経由で外部サービスにアクセスすることも多いでしょう。そのため、しばしばクライアントとは別にクラウド上の計算資源（以下ではサーバーサイドと呼びます）を利用します。各コンポーネントをクライアント・サーバーサイドどちらに配置するかは重要な選択となります。これはクライアント・サーバー間でどのようなタイミ

ングで、どのようなデータを送受信するかに大きく関わります。各コンポーネントが必要とする計算資源と実行速度、コンポーネント間を流れるデータのサイズと発生タイミング、自然で円滑なコミュニケーションを実現するために必要なシステムアクションと許容されるレイテンシなどの要因を勘案して、コンポーネントの構成を決定します。

3.2.3　モデル・アルゴリズムの選択

各要素技術のモデル・アルゴリズムの選択は、システム設計上のもう 1 つのポイントです。まずはエージェントデザイン、および対話ルールの枠組みに基づいて、各要素技術に対する要求事項を明らかにする必要があります。その上で利用可能なモデル・アルゴリズムをリストアップします。各々の選択肢の特性、期待される性能、準備するためのコスト、インターフェース、メンテナンス性などを考慮して、使用するモデル・アルゴリズムを決定します。既存のウェブサービスや訓練済みモデルが利用できる場合には、必ずしも自分たちで機械学習モデルを訓練する必要はありません。また、要求を満たすのであれば必ずしも機械学習モデルを用いる必要はありません。例えば前処理については信号処理ベースのもの、対話制御についてはルールベースのものが適切な選択肢である場合もあります。適切な選択には専門知識を必要とするため、専門家の助言の下で判断することが必須となります。

3.2.4　ハードウェアの構成

さらにクライアンドサイドのハードウェア構成の決定も、システム設計上のポイントです。クライアンドサイドのハードウェアはエージェントのボディに相当し、ユーザーに対するインターフェースの役割を担います。外見や動きについてはデザインフェーズでシステムの目的に基づいて決定されているはずです。一方でマイク、スピーカー、カメラ、ディスプレイなどのスペックはクライアントサイドで使用される要素技術の要求に満たすように決定する必要があります。例えばマイクの個数・配置、スピーカーの配置はほとんどの場合にクライアントサイドに配置される前処理のモデル・アルゴリズムに大きく影響します。逆に言うとハードウェアの制約を超えた要素技術をクライアントサイドで使用することは出来ません。また、クライアントが複数存在し、同時に複数のユーザーが使用する可能性がある場合には、サーバーサイドも同時並列で複数の処理を実行する必要があることにも注意してください。ただしプロトタイピングフェーズではシステム構成を決定すれば十分であり、キャパシティプランニングは製品化フェーズの課題です。

3.2.5　システム設計のポイント

上記の要素は互いに関連しているため、各々を独自あるいは恣意的に選択することはできません。すなわちアルゴリズム・ソフトウェア・ハードウェアの 3 要素が一貫するように同時に設計を行う必要があります (図 4 参照)。また、これらの設計は技術的制約の制約の下で行わなければなりません。

アルゴリズム　各コンポーネントの要素技術に採用するモデル・アルゴリズム
ソフトウェア　使用するコンポーネントの選択とそれらの連携させるソフトウェアの構成
ハードウェア　ハードウェアの構成、特にマイク・スピーカーなどクライアントサイドのデバイスの構成

3.3　ユーザー評価試験

対話ルール作成、システム設計・実装の後にユーザー評価試験を行います。デザインフェーズにおける WOZ 方式によるユーザー評価試験では対話シナリオに沿って人間がシステム役を演じました。これに対してプロトタイピングフェーズでは、開発したプロトタイプ版の音声対話システムを使用します。ユーザーには、既定のシナリオに沿って対話を行うのではなく、既定の状況・目的の下で自由に対話を行ってもらいます。試験環境について、デザインフェーズのユーザー評価試験で使用したものと同じ環境、あるいはより製品版に近づけた環境を使用します。また評

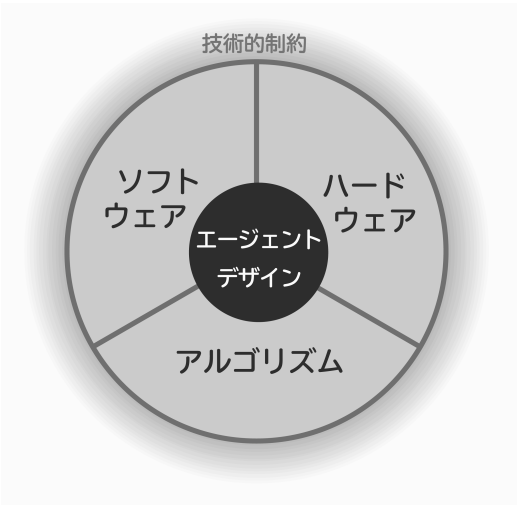

図 4　システム設計の 3 要素

価指標について、デザインフェーズのユーザー評価試験と同じ指標を用いてください。

プロトタイピングフェーズの試験の目的は以下の通りです。

- 音声対話システムの目標が程度の水準で達成しているかを測定
- 対話ルールの問題点を発見
- システム全体（ハードウェアおよびソフトウェア）の動作テスト

プロトタイプ版のシステムが目標とした性能指標を達成できているかどうかを検証するために、ここでプロトタイプと WOZ のシステムとの比較を行う場合もあります [34]。

ユーザー試験を実施することにより、システム設計・実装、対話ルールの問題が発見されるはずです。ユーザーに対するアンケート調査だけではなく、対話ログの分析から問題点を発見することも重要です。新たに大きな問題点が発見されなくなるまで、プロトタイピングフェーズ（システムの改善と評価）を繰り返してください。（ただし一切問題が発生しない水準に達することは難しいかもしれません。稀にしか発生しない、発生したとしても大きな障害を引き起こさない問題が残ってしまうことは、ある程度は許容する必要があるでしょう。）

3.4　次フェーズへ進むべきかの判断

プロトタイピングフェーズ終了時には、以下のことが実現されているはずです。

プロトタイピングフェーズで達成されること

- エージェントデザインに基づいたプロトタイプ版のシステムが完成している
- ユーザー評価試験の結果が得られている

この時点で、プロトタイプ版のシステムが期待した水準において音声対話システムの目的を達成しているか検証を行ってください。次フェーズではプロトタイプ版に基づいて製品版を開発します。製品版は安定性、セキュリティ、スケール性、メンテナンス性などを考慮して再実装されますが、エージェントデザイン・システム設計はそのまま変更なしに引き継がれます。したがってユーザーに与える印象の観点からは、プロトタイプ版の評価結果が最終的な製品の品質であると捉えるべきでしょう。プロトタイピングフェーズでユーザー評価試験を、重大な問題を見過ごしたまま製品版が完成してしまうリスクがあります。製品化フェーズに進んでから重大な問題が発見されても、修正でき

ない、あるいは修正に多大なコストがかかってしまう可能性あります。プロトタイピングフェーズから製品化フェーズに進むべきかの判断は慎重に行ってください。

また、以下のような判断に至った場合には、音声対話システムの目的自体を見直す必要があります。

- 技術的制約を考慮すると期待した水準で目的を達成できない
- 期待した水準で目的を達成するためには技術的制約を満たせない

場合によっては、現在の技術水準や想定するプロジェクト規模に対して開発プロジェクト自体の目標が過大である可能性もあります。このような場合には開発プロジェクト自体を見直す必要があります。

4　製品化フェーズ

プロトタイピングフェーズで決定した仕様に基づいて、製品版を開発します。プロトタイプ版とは異なり、製品版においては安定性、セキュリティ、スケール性、メンテナンス性などの要件について高い水準が要求されます。必要に応じて、これらの要求を満たすように設計の修正を行ってください。プロトタイプ版は参照実装にすぎないのため、プロトタイプ版のコードをベースとして製品版を実装することは多くの場合には適切ではありません。したがって製品版はスクラッチで実装することが望ましいでしょう。ただし各要素技術のコア部分（モデル・アルゴリズムに相当する部分）は、そのまま利用してください。要素技術のコンポーネントを入れ替えてしまうと、ユーザー評価試験により検証済みのプロトタイプ版とは異なるシステムになってしまいます。

既にシステム設計は決まっており、要素技術のコンポーネントも揃っている状態であるため、本フェーズの内容は通常のシステム開発と同様です。これは製品化フェーズに至るまでに、音声対話システムに特有のタスクをすべて決着させ、通常のシステム開発に落とし込んでおかなければいけない事を意味します。

5　まとめ

本稿ではビジネス目的で製品化に向けて音声対話システムを開発する場合の、典型的な開発プロセスを紹介しました。特にエージェントデザイン、システム設計、WOZ 方式およびプロトタイプ版を用いたユーザー評価試験について議論を行いました。音声対話システムの目的や想定する製品の特性ごとに、適切にデザインや設計を行う必要があります。これには音声言語処理・音声対話システムの技術的な専門知識、およびシステム設計・実装の経験が必要となります。

参考文献

[1] 長尾確, 大沢英一, 伊藤孝行, “エージェント・マルチエージェントの過去と現在,” 人工知能学会誌, Vol.35, No.3, pp.430-443, 2020.

[2] Lukas Mathis, Designed for Use: Create Usable Interfaces for Applications and the Web, 1st ed. (Pragmatic Bookshelf, 2011)（武舎広幸, 武舎るみ訳, インタフェースデザインの実践教室 —優れたユーザビリティを実現するアイデアとテクニック（第 1 版）（オライリージャパン, 東京, 2013)）.

[3] Byron Reeves, Clifford Nass, “The Media Equation," Cambridge University Press, 1996.

[4] Donald A. Norman, “Living with Complexity,” MIT Pless, 2010.

[5] Bruce Balentine, David P. Morgan, “How to Build a Speech Recognition Application: Second Edition: A Style Guide for Telephony Dialogues,” EIG Press, 2001.

[6] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus

Bühner, Heinrich Hussmann, “Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach,” Proceedings of the CHI Conference on Human Factors in Computing Systems, 2020.

[7] David C. Funder, “Personality,” Annual Review of Psychology, Vol.52, pp.197-221, 2001.

[8] Katherine Isbister, Clifford Nass, “Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics,” International Journal of Human-Computer Studies, Vol.53, No.2, pp.251-267, 2000.

[9] Ramiro H. Gálvez, Lara Gauder, Jordi Luque, Agustín Gravano, “A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora,” in Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2020.

[10] Amol Deshmukh, Bart Craenen, Alessandro Vinciarelli, Mary Ellen Foster, “Shaping Robot Gestures to Shape Users’ Perception: The Effect of Amplitude and Speed on Godspeed Ratings,” in Proceedings of the International Conference on Human-Agent Interaction, 2018.

[11] Seiji Yamada, Takanori Komatsu, “Designing Simple and Effective Expression of Robot’s Primitive Minds to a Human,” IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006.

[12] Roger K. Moore, “Appropriate Voices for Artefacts: Some Key Insights,” in 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR), 2017.

[13] Bruce Balentine, “It’s Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age,” ICMI Press, 2007.

[14] Byron Reeves, Cliffor Nass, “The media equartion: How people treat computers, television, and new media like real people and places,” Cambridge University Press, 1996.

[15] 山田誠二, 小野哲雄, 寺田和憲, 小松孝徳, 角所考, “HAI の方法論,” 人間とロボットの＜間＞をデザインする, 山田誠二編（東京電機大学出版局, 東京, 2007）.

[16] Ingo Siegert, Julia Kruger, Olga Egorow, Jannik Nietzold, Ralph Heinemann, Alicia Lotz, "Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon’s ALEXA", Proceedings of the Language Resources and Evaluation Conference, 2018.

[17] Ingo Siegert, “Alexa in the wild - Collecting unconstrained conversations with a modern voice assistant in a public environment,” in Proceedings of the Language Resources and Evaluation Conference (LREC), 2020.

[18] 原梓織, 深澤伸一, 赤津裕子, “音声入力と感情推定情報を活用する対話インタラクション方式：感情適応提示技術の提案と検証,” 情報処理学会インタラクション 2019 論文集, pp.852-857, 2019.

[19] 加藤淳, “ヒューマンインタフェース研究における再現性向上に向けた取り組み,” ヒューマンインタフェース学会誌, Vol.20, No.1, pp.23-28, 2018.

[20] Maria Schmidt, Wolfgang Minker, and Steffen Werner, “How Users React to Proactive Voice Assistant Behavior While Driving,” in Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, pp.485 – 490, 2020.

[21] 高橋舞羽, 小松孝徳, “ロボットへの認識における「外見」の影響をテセウスの船パラドクスから考察する,” HAI シンポジウム 2021 論文集, pp.G-5, 2021.

[22] Sarah Wilson, Roger K. Moore, “Robot, Alien and Cartoon Voices: Implications for Speech-Enabled Systems”, International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots, 2017.

[23] 大澤博隆, 今井倫太, “人と違う自由度のロボットと人はどう対話するか： 人-ロボット間のインタラクションデザイン再考,” 情報処理学会シンポジウム 2012 論文集, pp.137-144, 2012.

[24] Kazunori Terada, Atsushi Yamauchi, Akira Ito, “Artificial Emotion Expression for a Robot by Dynamic Color Change,” The IEEE International Symposium on Robot and Human Interactive Communication, 2012.

[25] Patrik Jonell, Mattias Bystedt, Per Fallgren, Dimosthenis Kontogiorgos, José Lopes, “FARMI: A FrAmework for Recording Multi-Modal Interactions,” Proceedings of the Language Resources and Evaluation Conference, 2018.

[26] Steve Young, Milica Gasi, Blaise Thomson, Jason D Williams, “POMDP-based Statistical Spoken Dialogue Systems: a Review,” Proceedings of the IEEE, Vol.101, No.5, pp.1160-1179, 2013.

[27] Ashish Mittal, Samarth Bharadwaj, Shreya Khare, Saneem Chemmengath, Karthik Sankaranarayanan, Brian Kingsbury, “Representation based meta-learning for few-shot spoken intent recognition,” INTERSPEECH, 2020.

[28] Jonathan Gratch, David DeVault, and Gale Lucas, “The Benefits of Virtual Humans for Teaching Negotiation,” in Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA), Los Angeles, 2016.

[29] Christoph Bartneck, Dana Kulić, Elizabeth Croft, Susana Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” International Journal of Social Robotics, Vol.1, No.1, pp.71 – 81, 2009.

[30] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, Kensuke Kato, “Prediction of human behavior in human-robot interaction using psychological scales for anxiety and negative attitudes toward robots”, IEEE Transactions on Robotics, Vol.24, No.2, pp.442-451, 2008.

[31] Aleksandra Cerekovic, Oya Aran, Daniel Gatica-Perez, “Rapport with Virtual Agents: What Do Human Social Cues and Personality Explain?,” IEEE Transactions on Affective Computing, Vol.8, No.8, pp.382-395, 2017.

[32] Cathy Pearl, Designing Voice User Interfaces: Principles of Conversational Experiences, 1st ed. (O’Reilly Media, 2017)（川本大功監訳, 高橋信夫訳, デザイニング・ボイスユーザーインターフェース――音声で対話するサービスのためのデザイン原則（第 1 版）（オライリージャパン, 東京, 2018））.

[33] Haojun Li, Dilara Soylu, and Christopher Manning, “Large-Scale Quantitative Evaluation of Dialogue Agents’ Response Strategies against Offensive Users,” in Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Singapore, 2021, pp.556 – 561.

[34] 井上昂治, ララ ディベッシュ, 山本賢太, 中村静, 高梨克也, 河原達也, “アンドロイド ERICA の傾聴対話システムにおける WOZ との比較評価,” 人工知能学会第 90 回 言語・音声理解と対話処理研究会, pp.85-90, 2020.