オープンソースの音声言語インタラクションの社会実験基盤を提供するMMIプラットフォームの開発

Development of Open-source Multi-modal Interaction Platform for Social Experiment of Conversational User Interface

李 晃伸
Akinobu Lee

名古屋工業大学大学院工学研究科
Nagoya Institute of Technology, Japan

A development of a multi-modal interaction platform for Social experiment of conversational user interface is proposed. In order to go over the simple spoken language interaction systems such as voice assistants, it is necessary to elucidate various factors of rich interactions quantitatively via thousands of wide variety of actual interaction data from users. The proposed system is based on a voice interaction building toolkit MMDAgent, adding some features to promote a testbed for social experiment and data collection of speech interaction system on cloud environment. It includes facilities for system distribution and management, collection of interaction log and speech data, and easy connection with cloud-based chat system. The beta version of the software is available, and it will be released as open-source software to promote wider use for various speech-based conversational user interface.

1. はじめに

近年，音声対話システムあるいは対話的な音声言語インタフェースは，情報機器・サービスと人をつなぐ次世代のモーダルとして期待が高まっている．特にここ数年，知的情報処理技術の深化と汎化に伴って，クラウドベースの知的情報サービスと人間を繋ぐスマートなコミュニケーションインタフェースとしての音声対話技術が注目されている．スマートスピーカーやWeb検索等において簡潔な発話による情報の授受やタスク達成はかなり実現しており，今後はより自然な会話スタイルでやりとりできることが求められている．

一方で，インタフェースとしての音声モーダルの一般的な設計論や方法論はまだ確立されているとはいえない．音声対話システムの研究事例や応用システム例はこれまでに多くあり，音声認識や音声合成の基盤技術も近年飛躍的進歩を得たが，それでもなお，現時点では音声は様々な機器やサービスに実際にアクセスするための主要なモーダルとなるに至っていない．また音声対話システムの構築には必要とされる技術が多く，CGエージェント等のインタフェースまで含めた設計・実装は多方面の多大な労力を必要とする．

音声対話を含む音声言語インタラクションが一般的なモーダルとして広く用いられるようになるためには，その設計や適用範囲，他のUIとの関連性を含め，統合的に実証・実験する枠組みが必要である．音声言語の領域だけでなく，自然言語処理や対話，ユーザインタフェースやヒューマンエージェントインタラクション，あるいはデザインやコンテンツの領域を検証可能な融合的な基盤を共有することで，多様な実際的・現実的なタスクのオープンなシステム運用から多くの実インタラクションデータを収集し，データドリブンなアプローチで様々な要素を量的に解明していくことが可能となる．様々な目的のための多様なシステムが大量に作成され実際に使われることで，多くの事例やデータを集める環境がボトムアップに形成されると期待される．

連絡先: 李晃伸，名古屋工業大学，愛知県名古屋市昭和区御器所町，052-735-7550，ri@nitech.ac.jp

本研究では，音声対話を含む音声言語インタラクションの多様な社会実験の実践のための基盤となるプラットフォームの実現を目標に開発された拡張版MMDAgentについて述べる．本アプリケーションは，MMDAgentで作成した音声対話・音声インタラクションシステムのネットワーク配信，ならびに利用ユーザのインタラクションログや音声データの収集を行える．これにより，音声対話システムを公開することで誰でもクラウドベースの広範囲な音声言語インタラクション実験およびデータ収集を行える．本アプリケーションはマルチプラットフォームで動作し，Android，iOSおよびデスクトップOS（Win/Mac/Linux）で同一の動作を行う．現在ベータ版が公開されており無償でダウンロード・試用が可能である．以下，ベースシステムであるMMDAgentについて概説したのち，提案システムについて述べる．

2. MMDAgent

まずベースとなるツールであるMMDAgentについて述べる．MMDAgentは音声ベースのインタラクションシステムのための構築ツールキットである[1]．多様な音声対話システムおよび音声インタラクションのための実験用プラットフォームとして開発が行われている．現在，ソフトウェア全体およびサンプルの音声対話システムがオープンソース（BSDライセンス）で公開されている．2010年に1.0が公開された．ソフトウェアの最新版は2016年に公開されたバージョン1.7である．公開以来14万件以上ダウンロードされている[2]．

MMDAgent自身は音声対話システムではなく，音声対話システムを動作させるソフトウェア（ブラウザ）である．音声対話システムの構成要素（辞書・ボイス・対話シナリオ・エージェントモデル・モーション等）は外部で定義する．MMDAgent用のサンプルシステム（サンプルコンテンツ）が同じWebで公開されており，2018年12月には従来の女性モデル・女性ボイスに加えて男性モデル・男性ボイスを追加したバージョン1.8が公開された．ドキュメントはJST CRESTのuDialogueプロジェクト(2011—2017)の成果物としてuDialogueサイトの「MMDAgentエンサイクロペディア」[3]に集約されている．

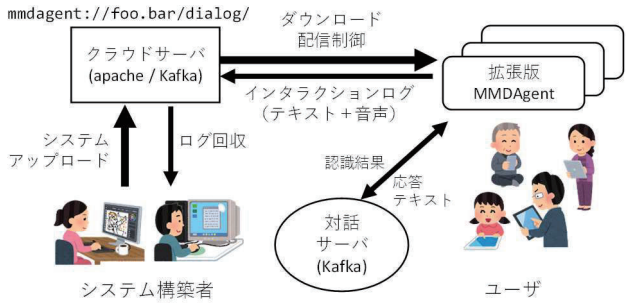

図1:システム全体像

MMDAgent では音声認識(Julius), 音声合成(Open JTalk(日本語), Flite+HTS Engine(英語)), エージェント表示がオールインワンで提供されている. 各部がモジュール化されていて拡張が容易であり, 言語モデルや音響モデルの拡張, ボイスモデルの入れ替え等が容易に行える. 画面表示は 3D 空間で, 物体とモーションの分離性や十分な表現力と自由度から, MikuMikuDance の互換プラットフォームとしており, プリミティブな 3D オブジェクトからヒューマノイドエージェントによるしぐさ・反応・ジェスチャの表出まで多様なインタラクション表現が可能である. 対話管理部はメッセージテキストを入出力とする状態遷移機械(FST)で実装されている.

3. 提案システムの概要

本研究ではスマートフォンから等身大サイネージまで多様なデバイスにおいて, 音声を主体とする会話的な UI を備えた音声インタラクション・対話システムのテストベッドとして双方向の実験および大規模データ収集を行えるよう MMDAgent を拡張した新たなプラットフォームを構築する[4]. 本システムはスマートフォンマルチ OS 環境で動作し, 双方向の大規模な社会実験環境の基盤を提供する. 以下, 実装された新たな機構を述べる. 全体像を図 1 に示す.

3.1 システムのクラウド配信および管理機構

音声対話システムを任意の URL から任意のユーザ端末へ直接配信可能にする. Web 上の URL で示されるディレクトリ以下にシステムのファイル一式(対話スクリプト, 3D モデル, モーション, 背景画像, ボイス定義ファイルなど)を置いておき, アプリケーションからその URL を開くことでシステムを自動取得できる. URL スキームに対応しており, ユーザが “mmdagent://…” のリンクを開くだけでシステムを自動ダウンロード・実行できる.

また, システム構築者からダウンロードしたユーザへの制御・連絡を行う仕組みを実装している. ダウンロードしたシステムに対する更新や削除, お知らせの表示を, 端末からサーバへ定期的な更新チェックを行うことで自動的に行う仕組みになっており, 利用ユーザに対する様々な配信制御が可能である.

3.2 ログ・フィードバック収集機構

システム構築者が自システム利用時の動作ログを収集するための機構を実装する. 収集する動作ログは, MMDAgent のメッセージキューを流れたすべてのメッセージおよびシステムログであり, 音声認識結果, 対話管理 FST の状態遷移などのインタラクション情報がミリ秒単位のタイムスタンプとともに保存されたものである. また, 認識対象となった生の音声波形データも収集できる. 端末の識別ごとにユニーク ID が自動生成され, 識別子としてログに付与される.

ログの収集スキームとして, Apache Kafka を用いたリアルタイムログ収集と, ローカルにファイル保存してから Web アップロードする方法の 2 種類を実装する. 前者は 2011 年にオープンソース化され, LinkedIn や Twitter でも用いられている分散ストリーミングのプラットフォームであり, 数十万以上の大量の端末に対して高性能なリアルタイムデータストリーミングを行うことができる. 即時性が高く, コンテンツのエラー検出やロギングをリアルタイムに行うことができる. 後者の Web アップロードは, 端末のローカル上にログを記録・保存し, 一定のタイミングでそれらを POST メソッドで Web 上のサーバへアップロードする. 音声波形データの収集は後者の場合のみ対応する. これにより, ユーザからコンテンツ公開者へのフィードバックを実現する.

3.3 クラウド型対話システムへの対応

システムが指定する Apache Kafka サーバと consumer モードで接続することで, 個々の端末が Kafka サーバを通じて認識結果や応答テキストなどのメッセージをリアルタイムにやりとりできる. これを用いることで, 対話サーバを接続して当アプリケーションをフロントエンドとしたクラウド型音声対話システムを容易に実現可能である. また接続先はシステム単位で変更できるため, あるシステムを利用中の端末全てに発話命令を一斉送信するような利用方法も可能である.

3.4 オープンソース開発体制

多様な音声インタラクションを対象とするためには, 様々なセンサーあるいは IFTTT のようなバックエンド接続サービス等との接続のための拡張を行いやすい開発環境である必要がある. 本アプリケーションの開発においては GitHub を活用した実践的な体制を構築し, ソースコード共有によるコード主体の多様かつ迅速な開発を行う予定である.

4. 現行システム

本稿で提案した拡張の多くを施したアプリケーションは, 既に「Pocket MMDAgent」という名称で試験公開中である[5]. サイトからは Android, iOS を含む各種 OS 用のベータ版アプリが無償で入手可能であり, 仕様もサイト上で公開されている.

5. まとめ

本システムの開発進捗は現在 80%程度であり, 最終的にはオープンソースで公開する予定である. マルチプラットフォームの UI 基盤としては Unity が著名であるが, 本ソフトウェアはオープンソースであり, 音声言語インタラクションのデータ収集基盤としての拡張性と可搬性を重視して作成されている.

本システムの開発と公開が, 音声を含めた知的インタラクションにかかる研究テーマの統合的でデータドリブンなアプローチの土台となり, 音声言語を含めた次世代インタフェースの幅広い試行錯誤と検討の一助となれば幸いである.

参考文献

[1] A. Lee, K. Oura, K. Tokuda: MMDAgent - A fully open-source toolkit for voice interaction systems, IEEE ICASSP, pp. 8382-8385, 2013.
[2] http://www.mmdagent.jp/
[3] http://www.udialogue.org/ja/encyclopedia-ja
[4] 李晃伸:音声対話コンテンツのネットワーク配信および大規模ログ収集を可能にするスマートフォン版 MMDAgent の開発, 日本音響学会秋季講演論文集, 2-2-8, 2019.
[5] Pocket MMDAgent (beta): https://mmdagent.lee-lab.org/