4ZE-02

複数人を対象としたマルチモーダル対話システムの開発

黒野明日希† 高橋伸弥‡
福岡大学大学院工学研究科† 福岡大学工学部‡

1. はじめに

近年，音声認識・合成システムの高精度化・高機能化に伴い，音声ユーザインタフェースを用いた製品が身近になってきている．しかし，人間同士のような自然な会話・対話を実現することのできる音声対話システムは未だ登場していない．特に，システムに対して複数の利用者が話しかけるような１対多の対話システムは，WOZ 実験によりその特性を検討した例はある[1]が，未だ実現はされていない．１対多の対話システムを実現する場合，誰が誰に話しているかを正しく判断するためには，音声情報だけではなく画像情報も必要となるが，こちらもまだ検討が不十分である．そこで本研究では，簡易的な音声対話が実現可能なソフトウェアである MMDAgent[2]と，人物検出や顔検出，音源方向検出が可能な Microsoft 社の Kinect とを連携させたシステムを実現し，これを用いて複数人対話で自然な振る舞いができるような音声対話システムを開発することを目指す．本稿では，話者数を二人として，その場合に考慮すべき対話状態を定義し，動作確認を行った結果を報告する．

2. 話者状態検出処理を組み込んだ音声対話システム

2.1 Kinect を用いた話者状態検出結果送信プログラム

KinectはMicrosoftが開発した入力デバイスであり，RGB カメラや超音波深度センサ，マイクロフォンアレイなどのセンサ類で構成されている．SDK をインストールすることで音声認識や音源方向の検出，骨格の表示，顔検出などが可能になる．本研究では，マーカーなしで顔を追跡することができるサンプルプログラム Face Tracking SDK に音源方向検出のための audio クラスを組み込んで音声検出部を thread として実装し，二人の話者それぞれの顔方向検出の結果（Up, Down, Left, Right, Front の５パターンの何れか）と発話者音源方向（Left, Right, None）を MMDAgent に送るような話者状態検出結果通信プログラムを作成した．

2.2 MMDAgent への機能追加と対話状態の推定

MMDAgent は名古屋工業大学によって開発された音声対話システム構築ツールキットであり，音声認識や音声合成，３D モデル制御などの機能がある．それぞれの機能はプラグインとして組み込まれており，自作のプラグインを新たに作成することも可能となっている．本研究では，Kinect を用いた話者状態検出プログラムとの間で socket 通信を行うプラグインを作成した．なお今回はユーザ１が左，ユーザ２が右に位置しているものとする．このとき，どちらのユーザが発話したのかという情報とそれぞれのユーザの顔の向きから，対話相手を判定することを考える．

表 1 音源方向 Left（ユーザ１の発話）の場合

User1 \ User2	Left	Front	Right	Up/Down
Left	Other	Other	Other	Other
Front	System	System	System	System
Right	User2	User2	User2	User2
Up/Down	Other	Other	Other	Other

表 2 音源方向 Right（ユーザ２の発話）の場合

User1 \ User2	Left	Front	Right	Up/Down
Left	User1	System	Other	Other
Front	User1	System	Other	Other
Right	User1	System	Other	Other
Up/Down	User1	System	Other	Other

上記の２つの表は，ユーザ１および２の顔の向きの各状態と話者情報の組み合わせから，誰に

Development of a multimodal dialogue system for multi-speakers
†Asuki KURONO, Graduate School of Eng. Fukuoka Univ.
‡Shinya TAKAHASHI, Faculty of Eng., Fukuoka Univ.

対して話しかけているかを判定した結果を示したものである．User1 または User2 とある欄は，もう一人の相手に対して話しかけている状況であり，System とある欄は，対話システムに対して話しかけている状況となる．Other はそれ以外（別の方向を向いている）と判断する．

3．動作確認実験

3.1　対話シナリオの設定

MMDAgent では，対話シナリオを FST （ Finite State Transducer)ファイルに記述することで対話を管理することが可能である．今回は発話者数を二人に限定して，動作確認のために以下のような対話シナリオを設定した．

① ユーザ１またはユーザ２が正面を向いていて「こんにちは」と発話した場合（すなわち対話相手がSystemの場合），「こんにちは，○○さん（または××さん）」と応答する．（○○はユーザ１，××はユーザ２の名前）

② 話者が正面ではなくもう一人のユーザの方を向いて「こんにちは」と発話した場合（すなわち対話相手が User1 または User2 の場合）で，かつお互いに向き合っている場合（つまり，ユーザ１の顔の向きが Right で ユーザ２の顔の向きが Left の場合）は，「こんにちは，○○さん，××さん」と応答する．

③ 話者が正面ではなくもう一人のユーザの方を向いて「こんにちは」と発話した場合（すなわち対話相手が User1 または User2 の場合）で，かつ相手が違う方向を向いている場合は，「○○さん，××さんが挨拶していますよ」（またはその逆）と応答する．

3.2 動作確認

実験時の状況を図 1 に示す．システムの前１m 離れて２名の話者が立っている状況で，上記①～③の動作確認を行った．いずれのパターンにおいても，想定通りの応答を返すことが確認できた．

4．まとめ

本研究では，複数人を対象とした音声対話システムの実現に向けて，音声対話システム MMDAgent と Kinect センサとを連携する仕組みについて検討し，顔の向きと音源方向を推定して，その情報を MMDAgent に送る仕組みを実装した．動作確認を行った結果，二人の話者の顔の向きと音源方向に応じて MMDAgent の反応が変わることが確認できた．

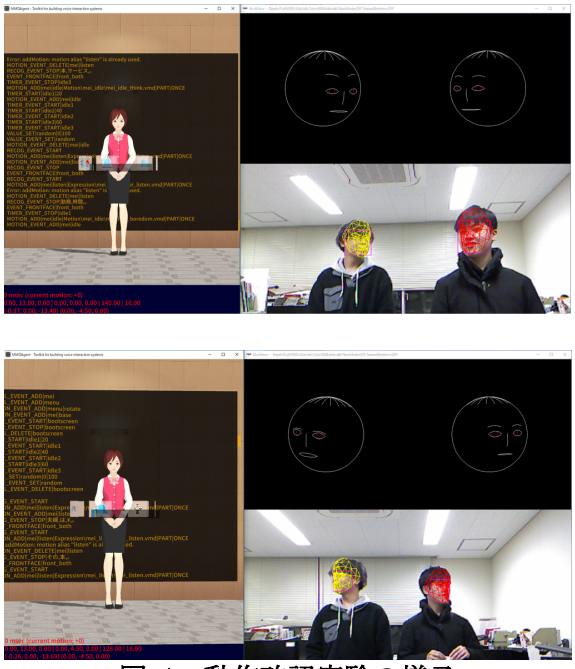

図 1　動作確認実験の様子

現在は話者が２名のケースを想定しているが，3 名以上の場合には誰に話しかけているのかの推定は困難になるものと予測されるため，基本的にはシステムに対しての発話なのかグループ内での会話なのかを判断することで対応する．また今回は Kinect による画像処理として顔の向きの検出のみを用いているが，今後は簡易的なジェスチャ認識を組み込むことで，さらに自然な対話を実現したいと考えている．Kinect 以外のデバイスや，OpenPose[3]やOpenFace[4]といった機械学習を利用したソフトウェアベースのジェスチャ検出・顔検出処理も検討する予定である．

参考文献

[1] 黄他，“多人数会話において積極的に情報提示ができるガイドエージェントの実現に向けての介入場面の検討”，人工知能学会論文誌，pp. DSF-514，2016.

[2] 李他,”魅力ある音声インタラクションシステムを構築するためのオープンソースツールキットMMDAgent”，信学技報，Vol.111，No.365，pp.159-164，2011.

[3] Z. Cao, *et. al.*, “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields,”. arXiv:1611.08050, 2016.

[4] B. Amos, *et. al.*, “Openface: A general-purpose face recognition library with mobile applications,” CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.