5S-1

環境音に頑健な同時合図を識別するクイズ司会者の構築

西牟田 勇哉　平山 直樹　大塚 琢馬　杉山 治　糸山 克寿　奥乃 博
京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

近年，実環境で複数人を相手にする対話システムの研究が進められている．従来の研究 [1] では様々なアプローチを用いてシステムが構築されているが，発話者は一人に絞られていた．しかし，実環境でロボットが複数人を相手にコミュニケーションを行うためには，ロボット自身に搭載されたマイクロホンで複数人が同時に話しかける状況を適切に処理し，対話を可能とする機能が必要となる．このような同時発話処理に必要となるのは，同時発話を分離することやそれぞれの発話がいつどこで行われたのかを正しく理解することである．我々は，そのような対話ロボットを実現するための第一歩として，テレビ番組「パネルクイズ アタック 25」をケーススタディとした複数人の同時発話が起こる「早言い」のクイズ “HATTACK25” (図 1) のロボット司会者を構築した．本稿では，ロボット聴覚ソフトウェア HARK [2] を用いてロボット司会者を構築する手法について述べ，構築したロボットの同時回答合図の処理性能について評価する．

2. 同時合図を識別するロボット司会者

2.1 HATTACK25 概要

多人数を相手に司会を務めるタスクのケーススタディとして，日本で代表的なクイズ番組である「パネルクイズ アタック 25」(朝日放送) を採用し，アタック 25 をモデルとして音声ベースで再現した HATTACK25 を実装した．HATTACK25 のルールはアタック 25 と同様に設定し，クイズゲーム参加者 (プレイヤー) はクイズによって 25 枚のパネルを取り合う．ただし，ロボット操作の制約から，以下の点について変更を加えた．

- 問題は読み上げによる一問一答クイズのみを扱い，読み上げはロボットが行う．
- 回答の合図は発話によって行い，早押しボタンは用いない．
- ロボットが問題を読み上げている途中でも回答の合図を行ってもよい．(バージイン発話を許容する．)

HATTACK25 を通して，同時発話を認識して処理する際の問題点を挙げ，その解決を図る．

2.2 問題設定

構築した司会者 (以下，ロボットと略す) は，自身に搭載されたマイクロホンによってプレイヤーの発話を認識し，HATTACK25 の司会者を務める．このようなロボットを構築するには，どの発話がどのプレイヤーによるものなのかをロボットが識別する必要がある．またロボットが自身の発話をいずれかのプレイヤーの発話として認識してしまわないように，(1) 発話からプレイヤーを同定することが必要となる．またマイクロホンをロボット自身に搭載するため，発話者とマイクロホンの距離が長くなり，雑音や環境音の影響を受けて音声認識精度が低下する．よって雑音や環境音に頑健な (2) 実環境での高い音声認識精度が必要となる．これらの問題を解決するうえで，システムにはプレイヤーはゲームが終了するまでその立ち位置を変えることがないと前提をおく．

Construction of robot identfying simultaneous sign under real environment noise: *Izaya Nishimuta, Naoki Hirayama, Takuma Otsuka, Osamu Sugiyama, Katsutoshi Itoyama, and Hiroshi G. Okuno (Kyoto Univ.)*

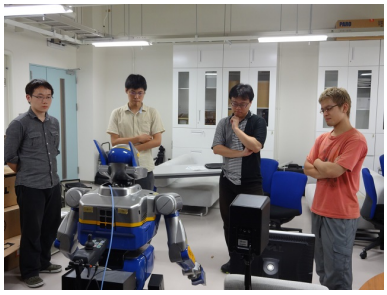

図 1: 多人数インタラクション “HATTACK25” の様子

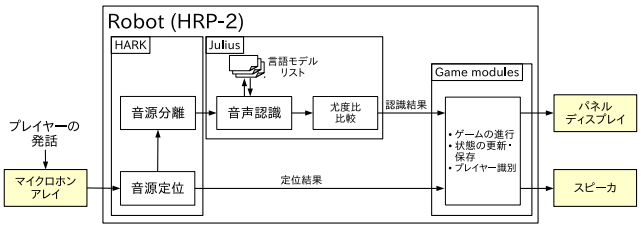

図 2: システムの構成

2.3 システム設計

ロボットはヒューマノイドロボット HRP-2 [3] を用いて構築した．システムの構成を図 2 に示す．HRP-2 の頭部には 8ch のマイクロホンアレイが搭載されており，外部デバイスとして合成音声を出力するスピーカ，パネル表示のためのディスプレイが接続されている．ロボットはプレイヤーの発話をマイクロホンアレイによって受け付け，HARK[1] を用いてその音響の定位・分離をする．分離した音声の認識には音声認識システム Julius[2] を用いた．HARK による定位結果，Julius による分離音の認識結果はゲームの管理に用いられ，状況に応じて合成音声の出力，パネルの更新を行う．

3. 問題点とその解決手法

2. 章で述べたロボットを構築する上での主な問題は，発話からのプレイヤー同定と実環境での高い音声認識精度である．プレイヤーの同定は HARK の音源定位，分離機能を用いて実現する．また音声認識精度の向上のために言語モデルの切り替え，音韻タイプライタを用いた尤度比較による雑音棄却を行った．

3.1 プレイヤーの同定

ロボットはゲーム中の発話がどのプレイヤーによるものなのかを同定する必要がある．本稿では，HARK の音源定位機能を用いたプレイヤー同定手法について述べる．

初期化

ゲーム開始前に位置同定に必要な初期化を行う．プレイヤーはロボットの前方に 45° 程度の間隔を開けて並ぶ．続いてロボットの位置確認に対して返事をし，その定位結果の水平角成分をプレイヤーの位置情報 $\theta_i\,(1 \leq i \leq 4)$ を登録する．

HARK を用いた話者位置同定

発話の定位結果 ϕ が θ_i と式 (1) の関係を満たすとき，

[1] http://www.hark.jp/
[2] http://julius.sourceforge.jp/

表 1: 発話内容

発話者数	2–people	3–people	4–people
発話スピーカ数	4 台中 2 台 (6 通り)	4 台中 3 台 (4 通り)	4 台全て (1 通り)
ディレイを与えるスピーカ	いずれか (2 通り)	3 台中 2 台 (3 通り)	4 台中 3 台 (4 通り)
繰り返し回数	5 回	5 回	15 回

プレイヤー i が発話したとみなす．HATTACK25 では，許容範囲が被らないよう $\varepsilon = 15°$ と設定した．

$$|\phi - \theta_i| \leq \varepsilon \qquad \varepsilon : \text{許容誤差} \tag{1}$$

3.2 言語モデルの切り替え

HATTACK25 ではゲームの進行状況に応じて音声認識用の言語モデルを切り替える．HATTACK25 はクイズ形式の対話であり，その進行状況によって求められる発話は異なるため，必要な情報のみを記述した文法を状況に応じて切り替えながら認識に用いることでルールセットにない状況が現れないようにしている．また言語モデルを切り替えることで語彙サイズを抑えることが可能なため，ルールに従った発話が誤認識されることを抑制することができる．HATTACK25 は出題，回答者決定，問題への回答，パネル選択が繰り返される．そこで，回答者の決定，問題への回答，パネルの選択それぞれに対し 3 つのモデルを用意し，切り替えながら音声認識を行う．

3.3 音韻タイプライタを用いた尤度比較

音韻タイプライタ [4] とは音節の構造のみを反映した文法であり，あらゆる入力の音響に対して認識結果の候補仮説の尤度の上限を求める．一方，記述文法を用いた認識では文法に記述されていない内容の入力に対しては尤度が小さくなる．そのため，音韻タイプライタを目的の文法に並行させて音声認識を行い，音韻タイプライタに対する目的の文法の尤度比が一定のしきい値より小さい入力を棄却した．これにより，周囲の環境音やロボットのモータ音などの自己雑音を棄却し，誤動作を防ぐことできる．ここで 3.2 節の言語モデルの切り替えには，ルール外の発話をルールに従う発話だと認識してしまう問題がある．しかし，ルール外の発話における尤度比は小さくなるため，尤度比較を用いることで問題となる発話を棄却することも可能となる．音声認識の尤度は発話時間が長くなるほど小さくなるため，この手法では短い雑音が十分棄却されない場合がある．よって，認識結果の尤度を直接用いるのではなく，単位時間あたりの尤度比を求めるなどの工夫が必要である．

4. 話者位置同定評価

提案した話者位置同定手法の精度について評価実験を行った．本稿では，同時発話が行われたときの最速発話者の検出と，その位置同定成功率について検証した．

4.1 実験環境

様々な条件下で繰り返し検証を行うため，人の代わりにスピーカを用いて環境を構築した．スピーカの間隔は，ロボット前方 120° の範囲に 40° 間隔で設置した．クイズゲームの司会者と回答者の関係が社会的距離に相当することからスピーカとロボットの距離は 1.5m，人の口の高さに揃えるために，スピーカの地上からの高さは 1.5m とした．スピーカから発生させた音声は，いずれも 20 歳代前半の男性の回答合図である「はい」の音声である．

4.2 実験内容

クイズにおける同時発話の状況を再現し，最速発話者の検出と位置同定の精度を調べるため表 1 に従って発話させるスピーカを選択する．先に 1 台で回答合図を再生し，残りのスピーカは 20-200msec のディレイを与えて同時に発話させる．複数の合図から最も早かった合図の定位結果から同定されたプレイヤーと正解のプレイヤーを比較する．それによって得られた話者位置同定の成功回数 $N_{success}$ と総発話回数 N_{all} から，式 (2) によって話者位置の同定成功率 N_{SR} を求める．

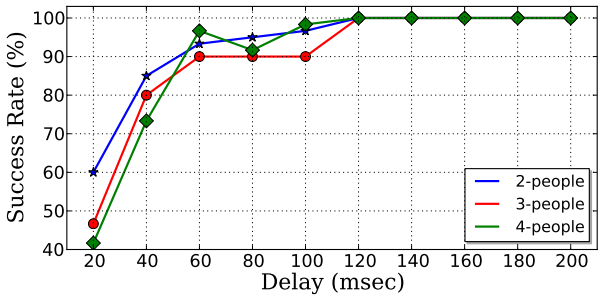

図 3: 実験結果

$$N_{SR} = \frac{N_{success}}{N_{all}} \tag{2}$$

4.3 実験結果と考察

図 3 にディレイと同定成功率の関係を示す．全ての発話者数において，60msec 以上のディレイであれば 90.0%以上の同定成功率を得た．この結果はクイズ司会者としては十分であると考えられるが，人で同様の実験をした結果と今回の結果と比較することで，ロボットと人の聞き分け能力の違いを考察する必要があると考えている．また本実験ではスピーカの配置や高さについて前述の 1 環境しか試していなかったため，それらを変化させた場合の実験結果との比較が必要である．

5. おわりに

本稿では，クイズゲーム “HATTACK25” の司会を行うロボットを構築した．同時発話の聞き分けやプレイヤーの識別はロボット聴覚ソフトウェア HARK の音源定位，分離結果を用いることで実現し，実環境における音声認識の精度向上のために，言語モデルの切り替えによる誤認識の抑制と音韻タイプライタを用いた雑音棄却を行った．今後の課題として，4.3 節で述べた更なる比較実験の追加や，音源定位・分離を用いた更なるインタラクションの考案などが挙げられる．本研究は科研費 基盤研究 (S) No.24220006 の支援を受けた．

参考文献

[1] S. Young, M. Gašić, B. Thomtson, and J. D Williams, “POMDP-based statistical spoken dialog systems: A review”, In Proc. of the IEEE, pp. 1160–1179, 2013.
[2] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, “Design and implementation of robot audition system ‘HARK’ – open source software for listening to three simultaneous speakers”, Advanced Robotics, vol.24(5-6), pp. 739–761, 2010.
[3] K. Kaneko, F. Kanehiro, S. Kajita, H. Hirukawa, T. Kawasaki, M. Hirata, K. Akachi, and T. Isozumi, “Humanoid robot HRP-2”, In Proc. of the ICRA, vol.2, pp. 1083–1090, 2004.
[4] 伊藤克亘, 速水悟, 田中穂積. “音声対話システムにおける未知語の扱い”, 人工知能学会研究会資料, SIGSLUD–9201, pp. 1–9, 1992.