「早言い」合図を識別しインタラクションに活用するロボットクイズ司会者

○西牟田 勇哉 † 吉井 和佳 † 西出 俊 † 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 ‡ 早稲田大学 実体情報学博士プログラム

1. はじめに

近年，実環境における Multi–Party Human Robot Interaction (HRI) の研究において，クイズゲームを題材としたインタラクションが着目されている．例えば，藤江ら [1] の研究では人同士の「難読ゲーム」を取り上げ，そこにロボットを介在させることでコミュニケーションの活性化を図っている．

本研究では複数人が参加するクイズゲームのロボット司会者を構築する．一般的にクイズ司会者に求められるタスクは以下の 2 つが考えられる．

(1) クイズゲームの進行管理
(2) 参加者・観衆を盛り上げる言動

(1) では出題時の問題読み上げ，回答合図の受理，正解判定時のプレイヤーの回答の正誤判断，(2) ではクイズゲームやプレイヤーの状況推定結果を踏まえた適切なコメントが必要となるため，これらのタスクの達成には音声による HRI が不可欠である．

我々はまず (1) に着目し，実環境において，複数人の音声を適切に処理しながらゲームの進行管理を行うロボットの開発に取り組む．従来の対話システムでは，対話参加者全員にマイクロフォンを持たせることで対話の管理を行っていたが，参加者がマイクに向かって話すスタイルは自然なインタラクションとは言えなかった．一方，複数人が対話に参加する環境では複数人の同時発話や，ロボットの発話中に人が割り込んで発話するバージイン発話が起こりうる．これらの問題に対処する有効な手段として，ロボット自身に搭載したマイクロフォンアレイを用いて同時発話に対して音源定位・分離するロボット聴覚機能が必要となる．このような音環境理解の有用性はロボット聴覚ソフトウェア HARK [2] を用いた水本らの Telepresence system [3] の開発において指摘されている．

本稿の構成は以下の通りである．2 章で，クイズゲーム *HATTACK25* の概要，またそのロボット司会者に必要な課題を述べる．3 章では，本ロボット司会者のシステムの設計を述べる．4 章では，実際に行われたインタラクション例を示す．5 章では実験設定，結果，考察について述べ，6 章で結論と今後の展望とする．

2. *HATTACK25* とロボット司会者

我々は「パネルクイズ アタック 25」(朝日放送) [1] をケーススタディとする音声ベースのクイズゲーム “*HATTACK25*” と，HARK による音源定位，分離を活用したロボット司会者をを考案し，開発を行っている [4]．ロボットは音源定位，分離によって発話とプレイヤーを位置同定し，「早言い」クイズの回答合図から回答者の決定を行う．これまで，音声認識結果の出力順で回答者を決定していた．そのため，僅かな時間差によって行われた複数人の同時回答合図に対しては音声認識処理時間の影響によって，正しく回答者を決定することができないといった課題があった．そこで，本研究では音源の定位・分離結果だけでなく音源の出現時刻 (音源のオンセット) に着目し，その時刻の比較を用いてクイズの回答者の決定を行う．これにより，複数人の同時合図から正しく発話順を検出できるようになり，その結果を用いたインタラクションが可能になった．

本節では多人数参加型の「早言い」クイズゲーム *HATTACK25* の概要，またそのロボット司会者に必要な課題について述べる．

図 1 「早言い」クイズゲーム *HATTACK25* におけるロボット司会者 (HRP–2) と 4 人のプレイヤー

2.1 *HATTACK25*

HATTACK25 は，4 人対戦形式の音声ベースのクイズゲームである．25 枚のパネルが 5×5 の正方形状に並べられており，プレイヤーはクイズによってこのパネルを取り合う．最終的にパネルを最も多く獲得したプレイヤーが勝者となる．ゲームの基本的な流れは，1) ロボット司会者による出題，2) プレイヤーによる合図，回答，3) ロボット司会者による正解判定，4) プレイヤーによるパネル選択であり，この一連の流れがゲームの終了条件を満たすまで繰り返される．

HATTACK25 が音声ベースのクイズゲームであることを踏まえ，ゲームには以下の仕様が存在する．

- 問題は読み上げによる一問一答クイズのみを取り扱い，問題の読み上げはロボットが行う．
- 回答者は「はい」という合図の「早言い」によって行い，ボタンなどの外部デバイスは使用しない．
- 複数人が合図をし，最も早かったプレイヤーが誤回答した場合は，次に早かったプレイヤーに回答権を与える．
- 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．)

[1] http://asahi.co.jp/attack25/index.html

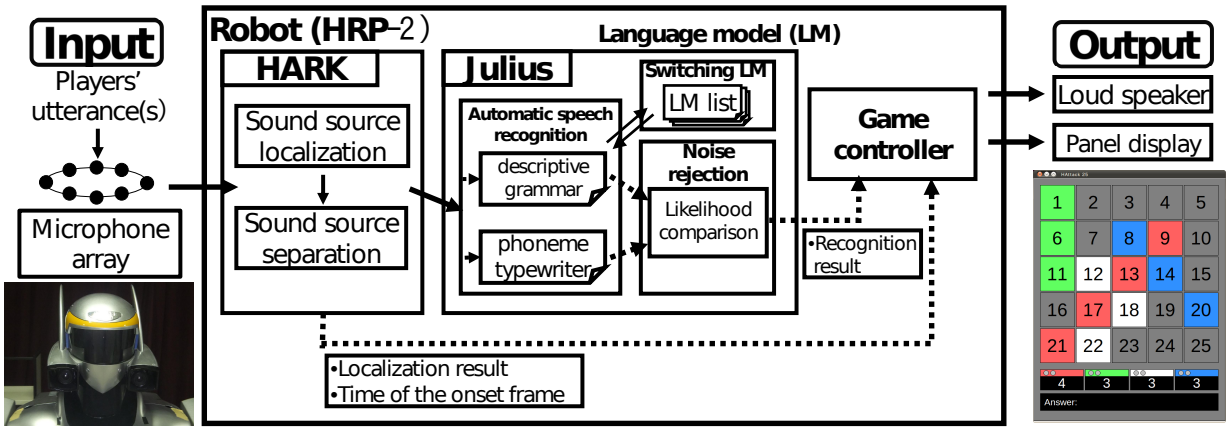

図 2*HATTACK25* システム構成

- プレイヤーはゲームが終了するまで位置を変えることがない

本研究では，*HATTACK25* の開発を通して，クイズゲームにおける同時発話やバージイン発話をインタラクションに組み込む際の課題を挙げ，その解決を図る．

2.2 ロボットのタスクと問題

構築したロボット司会者は，自身に搭載されたマイクロフォ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．ロボットは常時マイクロフォンを有効にし，いかなるときも入力を受け付ける．例えば回答合図を受け付けるときは，ロボットが出題中に合図を受け付けなければならない．そのため，出題中はロボット自身の発話を受理しないようにし，合図がどのプレイヤーから行われたかを識別する必要がある．また回答の正解判定では，回答権を持たないプレイヤーが回答者よりも先に行った発話を棄却する必要がある．よってロボットは，(1) 発話からプレイヤーを同定することが求められる．

また，(2) 環境音や雑音に頑健な，実環境での高い音声認識精度が必要となる．実環境でロボットは自身に装着したマイクロフォンでプレイヤーの発話を受け付ける．そのため，発話者とマイクロフォンの距離が長くなり，環境音の影響を受ける．更に，ロボット自身から発生するモータ音などの自己雑音にも影響されるため，それらに頑健であることも求められる．

3. システム設計

前節で述べたロボットのタスクと課題に基づき，ロボット司会者を設計，実装した．本節ではロボットの構成をハードウェア，ソフトウェアの両面から述べ，問題とその解決手法について述べる．

3.1 ハードウェア

ロボット司会者はヒューマノイドロボット HRP–2 [5] を用いて構築した．頭部には 8–ch のマイクロフォンアレイが搭載されており，外部には合成音声出力のためのスピーカー，パネルを表示するためのディスプレイが接続されている．

3.2 ソフトウェア

図 2 に本研究で設計したシステムの構成を示す．ロボットはプレイヤーの発話を頭部のマイクロフォンアレイで受け付け，HARK を用いてその音源の定位・分離をする．分離した音声の認識には大語彙連続音声認識システム Julius[2] を用いた．HARK による定位，分離結果や Julius による分離音の音声認識結果はゲームの管理に用いられ，必要に応じてパネルを更新し，合成音声の出力を行う．

3.3 問題点と解決手法

ロボット司会者構築の主な課題は 2. 章で述べた通りプレイヤーの同定と実環境での高い音声認識精度である．本研究では，プレイヤーの同定は HARK の音源定位・分離機能を用いた位置同定により実現する．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を行い，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．

3.3.1 プレイヤー位置同定

2.2 節で述べたように，ロボットはゲーム中の発話がどのプレイヤーのものであるのかを同定する必要がある．本稿では，HARK の音源定位，分離機能を用いた位置によるプレイヤー同定手法について述べる．

初期化: ゲーム開始前に必要な初期化を行う．プレイヤーはロボット前方に 40° 程度の間隔を開けて並ぶ．ロボットの位置確認の問いかけに対して返事をし，その定位結果の水平角成分をプレイヤーの位置情報 $\theta_i (1 \leq i \leq 4)$ として登録する．

定位結果による話者位置同定: 発話の定位結果 ϕ と θ_i が式 1 の関係を満たすとき，プレイヤー i が発話したとみなす．*HATTACK25* では，プレイヤー 4 人の許容誤差が被らないよう $\varepsilon = 15°$ と設定した．

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

ここで，本位置同定手法を用いた回答者 (回答順) 決定の流れを以下に示す．回答権は最初の合図の音源の

[2] http://julius.sourceforge.jp/

発話時刻から 100 [msec] の間に現れた合図の発生時刻を比較し，早い順に与えられる．

Algorithm 1 回答者 (回答順) 決定フロー

ロボットによる出題, 回答者受付
if プレイヤー i が合図 **then**
  $TIMELIST$ に t_i を追加
  t_i: プレイヤー i の合図のオンセット時刻
else
  合図なし，回答者なし
  return
end if
if $t_i \leq t \leq t_i + 100[msec]$ に音源 **then**
  $TIMELIST$ にそのオンセット時刻を追加
end if
$TIMELIST$ に記録されている時刻を比較
早い順に回答権を与える

3.3.2 言語モデルの切り替え

HATTACK25 では，誤認識によるロボットの誤動作を抑制するため，ゲームの進行状況に応じて言語モデルの切り替え [6] を行った．クイズの正解判定など，音声認識がロボットの行動決定のトリガとなる際，プレイヤーの発話の誤認識はロボットの誤動作を引き起こす原因となる．*HATTACK25* はクイズゲームであり，インタラクションの進行状況によって求められる発話は異なる．そのため，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することで，ロボットの行動決定のルールにない状況が現れないようにした．

HATTACK25 はクイズゲームであり出題，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，それぞれに対し「合図受付」,「(各問題に対応する) 回答」,「パネル選択」の 3 つのモデルを用意し，切り替えながら音声認識を行った．

3.3.3 音韻タイプライタを用いた尤度比較

雑音や環境音を棄却するため，目的の記述文法と音韻タイプライタを並列させて認識して尤度比較を行った．音節タイプライタに対する目的の記述文法の尤度比が一定の閾値より小さい場合にその入力を棄却した．音節タイプライタは音節構造のみを記述した文法であり，最も尤度が高くなる音節のつながりを認識結果として出力する．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．よって，その尤度比を求めて閾値で棄却処理することで，環境音やロボットのモータ音等の自己雑音を棄却することが出来る．

4. インタラクション例

HATTACK25 をプレイした際に，プレイヤーとロボットの間で行われたインタラクションの一例をに示す．**ロボット**は司会者ロボット，**赤**，**緑**はプレイヤーのうち 2 人，**システム**はシステム内部の処理を表す．以下のインタラクションは，ロボットが出題した問題に対して回答権を得た赤が誤回答して，次に合図をした緑に回答権が移る．そして緑が正解してパネルをを選択するといった，本システムの機能を一通り用いた一連の流れが示されている．

インタラクション例

ロボット: 「それでは次の問題」
ロボット: 「サッカー，野球，バスケットボール，プレイ人数が最も少ないのはどれでしょう」
システム: 「合図受付」言語モデルに切り替え
赤，緑: 「はい」,「はい」
システム: 発話時刻比較より赤に回答権を与える
ロボット: 「赤」
システム: 「問題」言語モデルに切り替え
赤: 「野球」**ロボット**: 「そっちにいっちゃったか」
ロボット: 「では次に早かった緑の方」
システム: 2 番目の緑に回答権を移す
緑: 「バスケットボール」
ロボット: 「その通り，バスケットボールだ」
ロボット: 「さぁ，緑の方，何番」
システム: 「パネル選択」言語モデルに切り替え
緑: 「16 番」
ロボット: 「16，12 と緑に変わった」
システム: パネル 16, 12 を緑に変化させる

5. 話者位置同定評価

4. 節で述べたインタラクションを実環境で行うには，僅かな時間差で行われる複数プレイヤーの合図の同時発話から正確に最速発話者を検出する必要がある．そのため，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，複数プレイヤーの合図の同時発話が行われたときの最速発話者の決定と，その位置同定成功率について検証した．

5.1 実験環境

様々な条件下で繰り返し検証を行うため，図 3 のように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，人の両眼視野が 120° であることを考慮して，ロボット前方 120° の範囲に 40° 間隔 (ロボット正面を 0° として，$-60°, -20°, 20°, 60°$) で設置した．クイズゲームの司会者と回答者の関係がホールの対人距離 [7] の定義において，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 に設定した．また，人の口の高さに相当させるため，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．

5.2 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した．全ての発話スピーカ数の設定において総発話回数が 600 回になるよう繰り返し実験を行った．本実験では，先に 1 台のスピーカか

表 1 発話内容

発話者数	2 players	3 players	4 players
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	20-200 ms (10 通り)	20-200 ms (10 通り)	20-200 ms (10 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

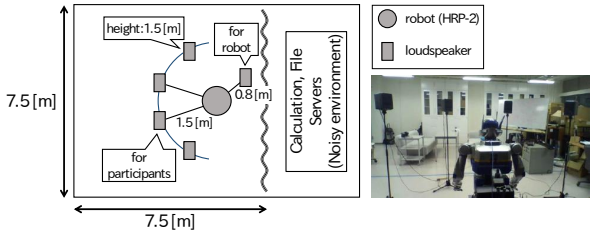

図 3 実験環境設定

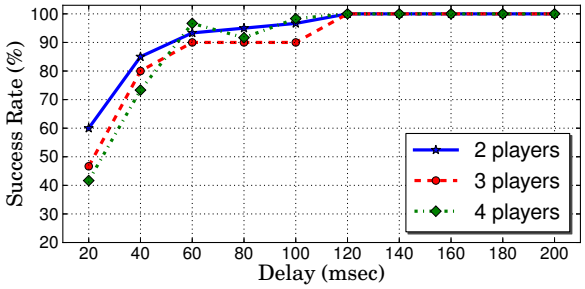

図 4 話者位置同定精度 (バージインなし)

ら回答合図を再生し，残りのスピーカは 20–200[msec] のディレイを与えて同時に再生させた．複数の合図から最速発話者の検出を行うため，同時に行われた合図の発話時刻を比較し，最も早かった合図の定位結果から 3.3.1 節の手法で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最速発話者の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 SR を求める．

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

5.3 実験結果と考察

図 4, 5 にディレイと同定成功率の関係を示す．バージインがあった場合は，ない場合に比べて少し成功率が低下したが，全ての発話スピーカ数において，60[msec] 以上のディレイであれば，バージインの有無に関わらず 90.0%の同定成功率を得た．この結果はクイズゲームの司会者としては十分であると考えられる．今後は，人で同様の実験を行った場合の結果と比較することで，人とロボットの聞き分け能力を考察したり，最速発話者だけでなく，同時発話合図の発話順についても同定成功率を求める必要があると考えている．

6. おわりに

本研究では，クイズゲーム *HATTACK25* の司会者を務めるロボットを構築した．ロボット聴覚ソフトウェア

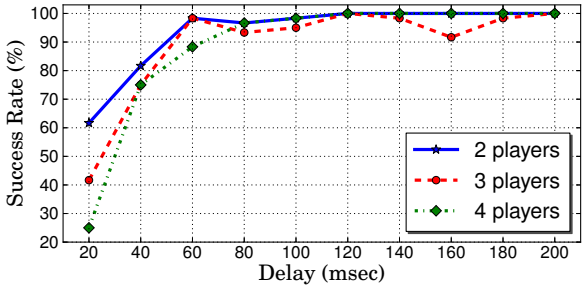

図 5 話者位置同定精度 (バージインあり)

HARK の音源定位，分離結果を活用することでプレイヤーの同定や同時回答合図の処理を行った．また，音源のオンセット時刻比較を実装することで，同時回答合図の処理をより正確に行うことが可能となった．加えて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，音韻タイプライタを用いた雑音棄却を行った．今後の展望として，音源定位．分離を活用した更なるインタラクションの設計，1. 節で述べたクイズ司会者のタスク (2) の場を盛り上げるためのコメントの生成などが挙げられる．

謝辞

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] 藤江真也 et al. 人同士のコミュニケーションに参加し活性化する会話ロボット．**電子情報通信学会論文誌**. *A*, **基礎・境界**, (1):37–45, 2012.

[2] Kazuhiro Nakadai et al. Design and implementation of robot audition system 'HARK' – open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[3] Takeshi Mizumoto et al. Design and implementation of selectable sound separation on the texai telepresence system using HARK. In *Proceedings of IEEE-RAS ICRA*, 2130–2137, 2011.

[4] 西牟田 勇哉 et al. HARK を用いたロボットクイズ司会者 HATTACK25 の開発. **日本ロボット学会第** *31* **回学術講演会**, 2013.

[5] Kenji Kaneko et al. Humanoid robot HRP-2. In *Proceedings of IEEE-RAS 2004*, volume 2, 1083–1090, 2004.

[6] Ian R Lane et al. Dialogue speech recognition by combining hierarchical topic classification and language model switching. *IEICE transactions on information and systems*, 88(3):446–454, 2005.

[7] Edward Twitchell Hall. *The hidden dimension*. Doubleday, 1966.