人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-09

ロボット聴覚ソフトウェア HARK を用いたクイズの同時回答を識別するロボット司会者の設計と実装

Design and implementation of emcee robot of the quiz distinguishing simultaneous answer using Robot Audition Software HARK

西牟田 勇哉 † **Izaya Nishimuta**
平山 直樹 † **Naoki Hirayama**
大塚 琢馬 † **Takuma Otsuka**
杉山 治 † **Osamu Sugiyama**
糸山 克寿 † **Katsutoshi Itoyama**
奥乃 博 † **Hiroshi G. Okuno**

† **京都大学 大学院情報学研究科** Graduate School of Informatics, Kyoto University

{nisimuta, hirayama, ohtsuka, sugiyama, itoyama, okuno}@kuis.kyoto-u.ac.jp

Abstract

実環境で複数の人とコミュニケーションを行うロボットの開発では，話者を一人に絞るカクテルパーティ効果ではなく，雑環境音下で複数人が同時に話しかける状況でも対話が可能な機能が必要である．本稿ではその第一歩として,「早言い」クイズのロボット司会者の設計と実装について報告する．本ロボット司会者は，ロボット聴覚ソフトウェア HARK を用いて発話者の位置を同定し，その発話を分離，音声認識することでクイズの同時回答を識別する．音声認識を頑健にするために，言語モデルの切り替えにより誤認識を抑制し，音韻タイプライタを用いた雑音棄却によって環境音の影響を抑制し，TV 番組「アタック 25」と類似したクイズを対象とした対話システムを開発した．また,「早言い」者の同定精度について評価を行い，60msec の発話のタイミングのずれでは，正しく発話者を同定できることを確認した．

1 はじめに

近年の音声認識技術の発展は著しく，その応用として音声対話システムが数多く登場している[Young *et al.*, 2013]．音声対話システムは Apple 社の Siri，NTT ドコモ社のしゃべってコンシェルといった携帯デバイスにおけるソフトウェアに始まり，PaPeRo [藤田善弘, 2003]や PALRO [富士ソフト株式会社, 2010] といったコミュニケーションロボットにも応用され，人とロボットのインタラクションに貢献している．既存のコミュニケーションロボットは，主に 1 対 1 で直接的にインタラクションを行っていた．一方で，実環境におけるインタラクションでは，多人数を相手にすることが想定される．そのため，参加人数を拡大し，多人数でインタラクションを行うことが可能なロボットが期待されている．ここで，従来のロボットが多人数インタラクションを行う研究では，同時発話の聞き分けや対話参加者の識別を行っていなかった．

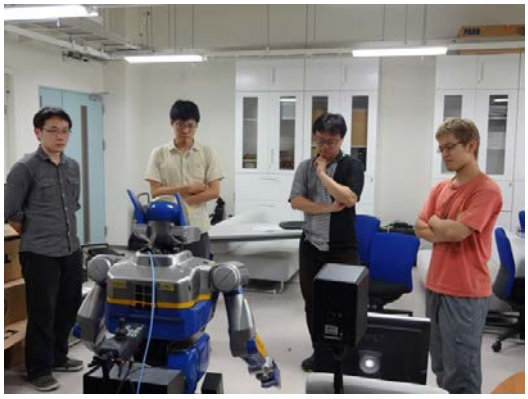

Figure 1: 実装した多人数インタラクション “HATTACK25” の様子．4 人の人がロボット司会者の質問を聞いている．

本研究では，ロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto Univercity) [Nakadai *et al.*, 2010]を用いることで前述の同時発話の聞き分け，対話参加者の識別を行うクイズ司会者を設計し実装する．そのために多人数が対戦形式で行うクイズゲーム「パネルクイズ アタック 25」(朝日放送) をケーススタディに採用した音声ベースのクイズゲーム “HATTACK25” (HARK を用いた ATTACK25) を設定した．また，実環境ではロボットの自己雑音，環境音が存在するため，音声認識精度が劣化する．そこで，言語モデルの切り替えによる誤認識の抑制や音韻タイプライタを用いた雑音の棄却によって，実環境に頑健な音声認識を行うことができるようロボットを実装した (Figure 1).

本稿の構成は次の通りである．2 章で関連研究を紹介する，3 章で本ロボットのタスクと課題を定義する．4 章でシステムを設計し，5 章で実装した HATTACK25 の実行例を示す．6 章で話者同定の性能評価の結果を示し，7 章でまとめとする．

2 関連研究

実環境において多人数で行う人・ロボットインタラクションの研究として，Matsusaka ら [Matsusaka *et al.*, 2003]，藤江ら [藤江真也 *et al.*, 2012]の研究が挙げられる．Matusaka らの研究では二人の人とロボットの一問一答形式の質疑応答を実環境で行うことを試みているが，発話の音声認識はロボット聴覚ではなく各対話参加者にマイクロフォンをもたせることで行なっている．藤江らの研究では人同士のクイズコミュニケーションにロボットを介在させることで，そのコミュニケーションを活性化させることを試みている．しかしこの研究では，多人数インタラクションの重要な要素である対話参加者の識別は行なっていない．また，どちらの研究も複数話者が同時に発話することは考慮されていない．

同時発話を処理するロボットの例として，1 章で述べたロボット聴覚ソフトウェア HARK を用いた口じゃんけんの審判を務めるロボット [Nakadai *et al.*, 2008]がある．この研究では同時発話の聞き分けの枠組みを述べているが，インタラクションへの応用は行なっていない．

対話システム構築の手法は，[河原達也 and 荒木雅弘, 2006]，[MacTear, 2004]で紹介されている．しかしこれらは音声認識に重点をおいており，音源の定位・分離結果を用いたシステムについては考えられていない．本研究では，同時発話処理によって得た情報を音声認識結果に加えてインタラクションに利用することで，従来研究では実現出来なかったインタラクションを取り扱う．

3 同時回答を識別するロボット司会者

多人数を相手にインタラクションを行う場合，ロボットはそれぞれの発話が自分に必要な音，聞きたい音だけを取捨選択する必要がある．例えば，ロボットが対話状況を管理する必要があるような役割を持つ場合，複数人が同時に発話したとき，だれに発話権を与えるかの決定をしなければならない．また，発話権を持っている人以外の発話を受理してしまうことがないようにしなければならない．本研究では，発話権の管理が重要となる例として多人数クイズの司会者を取り上げる．

ロボット司会者を実装するための課題は次の通りである．

- 対戦形式であるため対話者の識別が必要
- 音声の早言い (音声ベースであるため，早押しではなく早言いとなる) の合図を適切に処理するために発話混合音の分離が必要
- 回答権を持たない人の発話の棄却が必要

この章では本研究で実装したクイズゲーム”HATTACK25”について述べる．なお，以下ではクイズゲームに参加する人間を「プレイヤー」，司会者を務めるロボットを「ロボット」と表記する．

3.1 概要

本研究ではクイズゲームのケーススタディとして，日本の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) を採用した．アタック 25 は日本で最長寿のクイズ番組である．このクイズ番組の司会進行を参考にすることで，ロボットがクイズ司会者を務めるために必要な課題，要素技術について分析した．

本研究ではアタック 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 は基本的にアタック 25 と同じであるが，次のように音声ベースへの変更を施した．

- 問題は読み上げによる一問一答のクイズのみを取り扱う．映像，音楽を用いたクイズは用いない．
- 問題の読み上げはロボット司会者が行う．
- 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．
- ロボット司会者が問題を読み上げている最中でも回答の合図を行なってもよい．(バージイン発話を許容する)

ゲームは 4 人でプレイする．ディスプレイ上に 1 から 25 の数字が格子状に並んだパネルがあり，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獲得したプレイヤーが勝利となる．ゲームは Figure 2 のフローチャートに従って行われ，基本的に出題，回答，パネル選択が繰り返し行われる．またゲーム開始に先立って，ロボットがプレイヤーを識別するために必要な位置情報を取得するための初期化を行う．

3.2 ロボットのタスクと課題

上記の HATTACK25 の司会者をロボットで構築するにあたり，ロボットがなすべきタスクと，そこで発生する課題について明らかにする．HATTACK25 におけるロボットのタスクは，(1) 複数のプレイヤーの回答合図を処理し適切な回答者を決定する，(2) 発話とプレイヤーを対応付ける，(3) クイズの正解・不正解を判定，選択されたパネルを受け付ける，の 3 つである．

それぞれのタスクを達成するためには，(1)，(2) については同時発話の聞き分けやどのプレイヤーが発話した

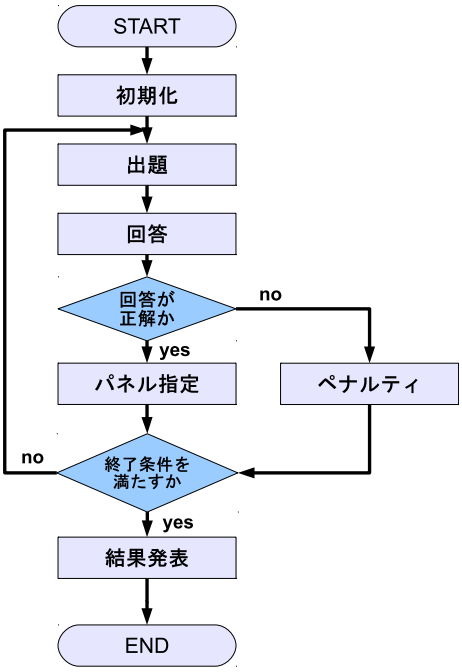

Figure 2: HATTACK25 フローチャート

のかの識別が，(3) については実環境における高い音声認識精度が課題となる．

4 章でシステムの構成と前述の各課題の解決方法について述べる．

4 システムの設計

3 章のロボットのタスク，解決すべき課題の分析結果に基づいて，HATTACK25 の司会を務めるロボットを設計し実装した．はじめに，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詳細に述べる．

4.1 ハードウェア構成

本研究ではロボットを HRP-2 [Kaneko *et al.*, 2004]を用いて実装した．HRP-2 は人の上半身を模したヒューマノイドロボットであり，頭部には 8ch のマイクロフォンアレイを搭載している．外部には合成音声を出力するためのスピーカが接続されており，パネルを表示するためのディスプレイが設置されている．

4.2 ソフトウェア構成

Figure 3 に本研究で設計したシステムの構成を示す．プレイヤーはマイクロフォンアレイを通してロボットへ音声を入力する．そして入力された音響を HARK を用いて定位・分離する．HARK で分離された音声の認識には大語彙連続音声認識システム Julius[1] を用いている．HARK によって得られた定位結果，Julius によって得られた認識結果は状況に応じてゲームの管理に用いられ，必要に応じてパネルディスプレイを変化，合成音声の出力を行う．

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

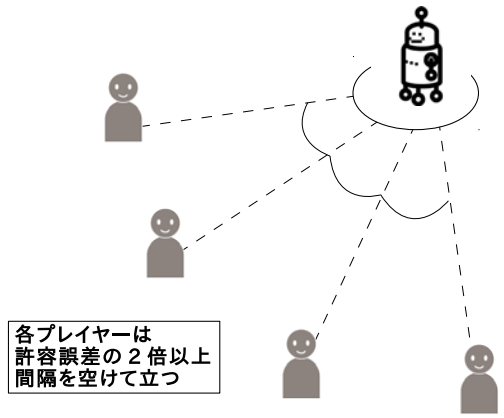

Figure 4: プレイヤーとロボットの位置関係

Figure 3 における Game Management Module とはゲームの管理モジュールの集合であり，この部分を変化させることで様々なインタラクションに応用が可能である．

4.3 課題と解決手法

ロボットを実装する上での主な課題は，3 章で述べたプレイヤーの識別と実環境の中での高い音声認識精度の 2 点である．本研究では HARK による音源定位・分離を用いてプレイヤーの識別を行い，音声認識精度の向上のために雑音・環境音の棄却，誤認識を抑制するための手法を実装した．以下にそれぞれの詳細について述べる．

4.3.1 プレイヤーの識別

HATTACK25 では，回答の合図を発話によって行うため，ロボットは同時に行われる合図の混合音を聞き分ける必要がある．また，どの発話がどのプレイヤーによるものなのかを識別する必要がある．本研究ではこのプレイヤーの識別を，HARK を用いた話者位置同定によって実現した．その手法を以下に示す．

初期化

まず，ゲームを開始する前に位置同定を行うために必要な初期化を行う．プレイヤーはロボットの前方に Figure 4 のように間隔を空けて立つ．続いてロボットの位置確認に対して返事をし，その返事の定位結果をプレイヤーの位置情報として登録する．

HARK を用いた話者位置同定

話者の位置同定は次のように行う．

1. 先に説明した初期化による各プレイヤーの登録位置を θ_i $(1 \leq i \leq 4)$ とする．
2. 発話の定位結果 ϕ が θ_i と式 1 の関係を満たすとき，プレイヤー i が発話したものとみなす．なお式 1 における ε は許容する定位結果の誤差を示す．HATTACK25 では，HARK の定位分解能が 5° 間隔であることと，各プレイヤーの許容誤差範囲が被らない限界を考慮して，$\varepsilon = 15°$ と設定した．

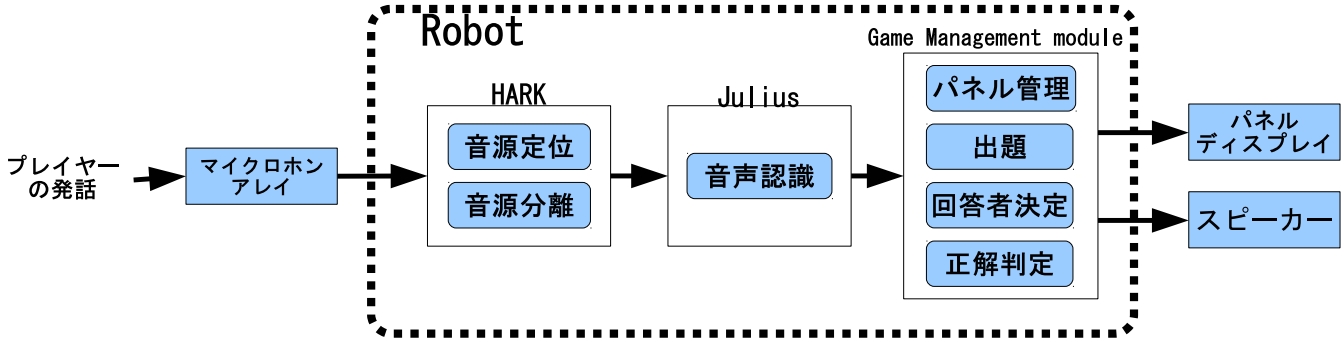

Figure 3: システム構成

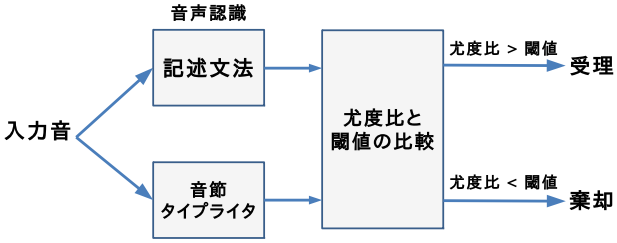

Figure 5: 雑音棄却方法の概念図

$$|\phi - \theta_i| \leq \varepsilon \tag{1}$$

4.3.2 雑音・環境音棄却

実環境でロボットが自身に搭載されたマイクロフォンで音声を認識する場合，周りの環境音やロボットのモータ音などの自己雑音，笑い声・独り言などを何らかの単語として認識し，誤作動を引き起こさないようそれらを棄却する必要がある．本システムでは音韻タイプライタ[伊藤克亘, 1992]を利用することで，そのような雑音や環境音，本来の目的ではない発話の認識結果を棄却する．音韻タイプライタとは，音韻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その認識結果の候補仮説の尤度の上限を求める．その音韻タイプライタと目的の文法を並行させて認識を行い，その際の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閾値より小さいとき発話を雑音とみなし棄却する．Figure 5 に音韻タイプライタを用いた雑音棄却方法の概要を示す．

4.3.3 誤認識の抑制

ロボットとのインタラクションにおいて，誤認識は誤動作を引き起こす原因となる．よって本システムでは誤認識を抑制するために，音声認識の際に言語モデルの切り替えを行った．今回の音声認識は，自分で記述した文法モデルを言語モデルとして使用している．ゲーム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した記述文法を複数用意し，状況に応じて切り替えながら用いることで想定外の発話が認識されないようにしている．例えば，HATTACK25 は基本的に，回答者の決定，問題への回答，パネルの選択が繰り返し行われるが，回答者を受け付けたり，パネルを選択する際に問題の回答がされることはなく，問題回答時に合図がなされたり，パネルが選択されることもない．そのため HATTACK25 では回答者決定における合図のみを受け付ける，問題ごとの回答候補を認識する，パネルの番号を受け付けるといった 3 つのモデルを用意し，切り替えながら音声認識を行っている．

5 実行例

本クイズゲームを実際にプレイした際に，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例を示す．これはロボットの出題からプレイヤーが回答し，パネルを選択するまでの一連の流れを示す．以下では Robot，Player がロボット，プレイヤーの発話，*はシステム内部の処理を示す．

インタラクション例

Robot: 次の問題，4 人です.
Robot: ブラジルの首都はどこでしょう.
* 言語モデル: 「はい」モデルへ変更
Player: はい.
Robot: 赤.
* 言語モデル: 「問題」モデルへ変更
Player: ブラジリア
Robot: 正解，ブラジリアだ.
Robot: さぁ，赤の方，何番.
* 言語モデル: 「番号」モデルへ変更
Player: 15 番.
Robot: 15, 14, 13 と赤に変わった.
* パネルディスプレイ: 15, 14, 13 番を赤に更新

6 性能確認

本研究で提案するロボット対話システムの動作確認を行った．動作確認では，4 人の被験者がいることを想定した実験環境を作り，その環境でシステムが設計通りに動くことを確認する．様々な確認項目のなかで今回は，同時発話が

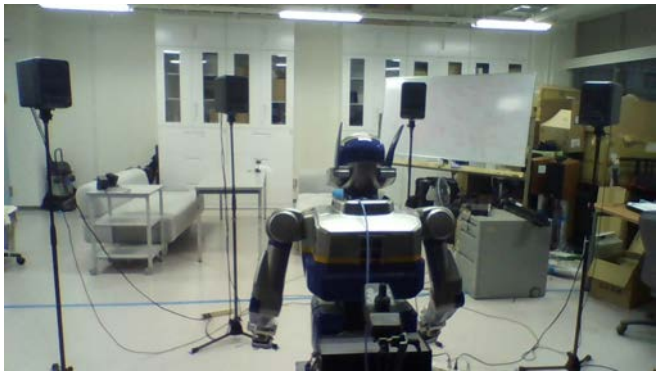

Figure 6: 実験環境

Table 1: 発話内容

発話スピーカ数	4 台中 2 台 (6 通り)
ディレイ	20-200 ms (10 通り)
ディレイを加えるスピーカ	いずれか (2 通り)
繰り返し回数	5 回
総発話回数	600 発話

行われたときの最速発話者の検出と，その位置同定精度の検証を行った．

6.1 環境設定

本実験では人の代わりにスピーカーを使用し，以下の設定に従い実験環境を Figure 6 のように構築した．プレイヤーの間隔は，人の両眼視野が 120° であることから，その視野内にスピーカが配置されるように 40 度間隔で設置した．今回の多人数インタラクションにおける司会者とプレイヤーの関係は Hall の対人距離の定義 [Hall, 1966, pp. 113–125] において，社会的距離に相当すると考えられる．そのためスピーカーはロボット頭部のマイクロフォンアレイの中心から 1.5m の位置に設置した．スピーカの高さは，人間の口の高さに近づけるために地上から 1.5m とした．また実験に先立ち，合図である「はい」という音声を研究室の学生 (いずれも 20 代男性)4 名に発話してもらい，スピーカから再生する音声を録音した．Figure 6 のロボット後方には多数の計算サーバ，ファイルサーバが稼働し，定常的にノイズが発生している．その実測値はロボットのマイクロフォンアレイ周辺において，A 特性音圧レベルの測定平均で求めたところ 61.2 [dB] であった．

6.2 実験内容

Table 1 の発話内容に従ってスピーカーを選択し，ディレイを与えて発話させる．複数の発話情報について，発話音源の最初のフレームの時刻 (Figure 7 では丸で囲った部分の時刻に相当) を比較し，最も早かった発話情報から最速発話者を決定する．そして，その発話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

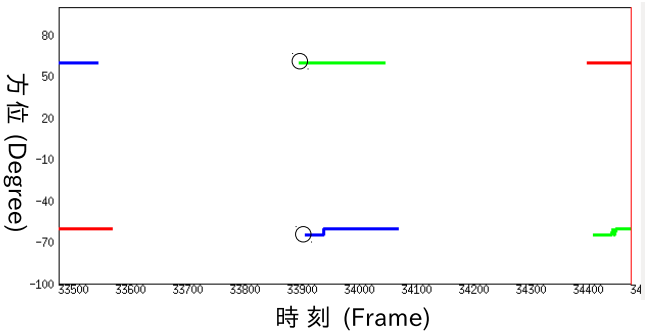

Figure 7: 定位結果

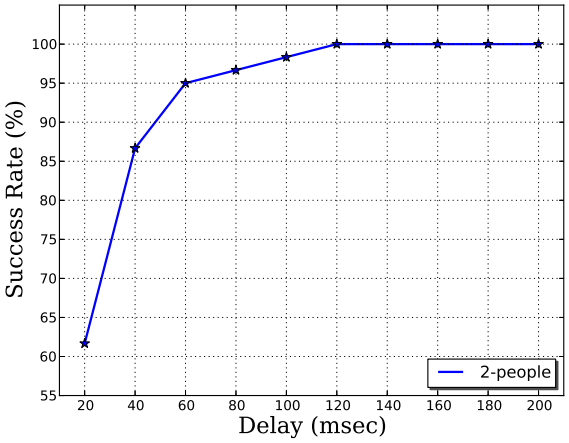

Figure 8: 話者位置同定成功率

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精度が求められる．

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

6.3 実験結果・考察

Figure 8 に，ディレイと同定成功率の関係を示す．同定成功率はディレイが大きくなるほど 100 %に近づき，60 msec で 90 % 以上，140 msec で 100 %の値を得た．この同定成功率は，2 話者の同定精度という点では十分であると考える．ただし，今回は 4 話者がそれぞれ 1 音声ずつ録音した 4 音声から 2 音声を選び出力する限られた条件での実験であり，結果が話者に依存している可能性もありうる．また，発話人数が 2 人であり，スピーカの間隔や範囲誤差を十分にとった理想的な条件で行なっていた．そのため，発話人数や間隔，誤差おいてより難しい環境を設定した場合，その同定成功率は低下する．よって今後は 3 話者以上が同時に発話した場合や，スピーカの間隔，範囲誤差を変更した場合の実験を行い，今回の実験結果と比較することで現状のシステムの問題発見と解決のために役立てたいと考えている．

7 まとめ

本稿では，同時回答を識別して多人数で対戦を行うクイズゲーム “HATTACK25” の司会を行うロボットを設計し実装した．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，性能評価の充実や音声認識の精度向上のために実装した技術の有効性を示すための実験を行うこと，同時発話の聞き分けについての情報をインタラクション部分に組み込むことが挙げられる．今回提案した聞き分けを用いることで，例えば，複数のプレイヤーが同時に反応した，あるプレイヤーが別のプレイヤーにわずかに遅れて反応したといったインタラクションも可能になるのではと考える．

謝辞

本研究は科研費 基盤研究 (S) No.24220006 の補助を受けた．

参考文献

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

[Kaneko *et al.*, 2004] Kenji. Kaneko, Fumio. Kanehiro, Shuuji. Kajita, Hiroshita. Hirukawa, Toshikazu. , Kawasaki, Masaru. Hirata, Kazuhiro. Akachi, and Takakatsu. Isozumi. Humanoid robot hrp-2. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 2, pages 1083–1090, 2004.

[MacTear, 2004] Michael F MacTear. *Spoken Dialogue Technology - Toward the Conversational User Interface*. Springer, 2004.

[Matsusaka *et al.*, 2003] Yosuke Matsusaka, TOJO Tsuyoshi, and Tetsunori Kobayashi. Conversation robot participating in group conversation. *IEICE transactions on information and systems*, 86(1):26–36, 2003.

[Nakadai *et al.*, 2008] Kazuhiro Nakadai, Shunichi Yamamoto, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. A robot referee for rock-paper-scissors sound games. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3469–3474, 2008.

[Nakadai *et al.*, 2010] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. 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.

[Young *et al.*, 2013] Steve Young, Milica Gašić, Blaise Thomthon, and Jason D Williams. Pomdpbased statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, pages 1160–1179, 2013.

[伊藤克亘, 1992] 田中穂積 伊藤克亘, 速水悟. 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-9201:1–9, 1992.

[河原達也 and 荒木雅弘, 2006] 河原達也 and 荒木雅弘. **知の科学 音声対話システム**. オーム社, 2006.

[藤江真也 *et al.*, 2012] 藤江真也, 松山洋一, 谷山輝, and 小林哲則. 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 (対話生成,< 特集 > 人とエージェントのインタラクション論文). **電子情報通信学会論文誌**. *A,* **基礎・境界**, (1):37–45, 2012.

[藤田善弘, 2003] 藤田善弘. 人工知能の現在と将来 パーソナルロボット PaPeRo の開発. **計測と制御**, 42(6), 2003.

[富士ソフト株式会社, 2010] 富士ソフト株式会社. 小型ヒューマノイド・ロボット PALRO, 2010. `http://www.fsi.co.jp/company/news/100201.html`.