人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216

ＡＩチャレンジ研究会 (第１６回)

Proceedings of the 16th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2002 年 11 月 22 日　　場　所　早稲田大学理工学部　62W 号館 1 階 大会議室
Waseda University, Nov. 22, 2002

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-1

ヒューマノイド・ロボットの技術とその展望

高西淳夫
早稲田大学理工学部・ヒューマノイド研究所

概要：
今，日本の製造業の中心的存在としてその牽引役を果たしてきた大企業が，一般家庭への普及をも視野に入れた人間あるいは生物の形をした非製造業用のロボットを続々と開発し，国内だけでなく世界的にも大きな社会的インパクトを与えている．産業用ロボットの市場規模に比べるとまだまだ小さいものの，ロボット産業全体のマーケットの方向にも徐々に変化が起こりつつある兆ではないかと思われる．このような時代背景のなかで，筆者らは人間型ロボットを用いた人間のメカニズム解明とその応用機器開発，ならびに将来のパーソナル・ロボットとしての人間型ロボットの基礎研究という視点でロボットの開発を行っている．本稿では，筆者らが現在開発中の様々なロボットを中心に紹介しながら，ロボット技術の現状および将来について展望を試みる．

１．はじめに

最近，多くの大学や研究機関でヒューマノイド・ロボット（以下，単に「ヒューマノイド」と呼ぶ）の研究が近年盛んに行われるようになってきた．企業でも，ソニーのペット・ロボットの AIBO とヒューマノイドの SDR-3X・4X，テムザックの遠隔操縦式人間型ロボットの T-4 や三洋電機と共同開発の警備用４足ロボットの番竜，ホンダの ASIMO，オムロンの猫型ペット・ロボットのネコロなどが発売・発表され，工業用一辺倒だった日本のロボット業界に大きな異変が起こっている．これらはマスコミで大々的に報じられ，国内だけでなく世界的にも大きな社会的インパクトを与えている．

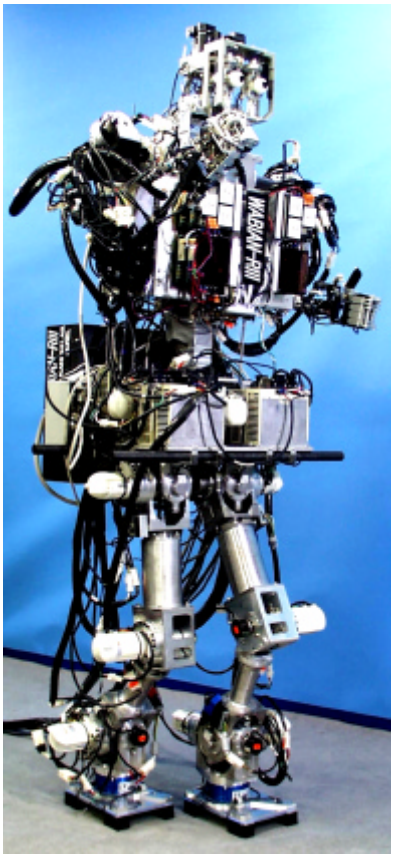
一方，早稲田大学では，故加藤一郎教授（1994 年他界）が 1960 年代前半より世界でも最初のヒューマノイドの研究に着手して以来，40 年近くに渡って研究が続けられている．現在は，早稲田大学ヒューマノイド研究所が設立され，８名の専任教員と国内外の６名の客員研究者を擁し，１０数社の企業とともにヒューマノイド・コンソーシアムを組織・運営し，ヒューマノイドの基盤・周辺技術に関する研究が続けられている．

今や過熱気味とも思われるブームになっているヒューマノイドだが，ロボットが人間社会の中で充分に役に立つほどのレベルに達するまでには，まだまだ，多くの時間と経験を要するというのが筆者の実感である．特に人間に対する安全性という観点からは，乗り越えなければならないハードルは何段もある．では何故，ヒューマノイド研究なのか，筆者は以下の２つの視点から研究を行っている．ひとつは，人間の形態と機能を模したロボットを設計･製作し，これを用いて人間の行動や機能を再現することで，構成論的に人間の工学モデルを構築するという視点である．すなわちヒューマノイドを道具として用いて人間を科学する「ロボット工学的人間科学(Robotic Human Science)」とでも呼べるものである．もうひとつは，そうして得られた人間の工学モデルを従来の基盤工学における各種モデル群と統合し，これと並行して人間のための様々なロボットや機器・装置の開発の実践を通して，人間に関わるロボットやシステムに関する演繹的設計論の構築を目指す「人間モデル規範型ロボット工学」とでも呼べるものである．以下，筆者らの研究を中心にその具体例を紹介する．

２．２足歩行ロボット

現在, WABIAN と呼ぶ２足人間型ロボットと WL-15 と呼ぶ２足歩行ロコモータの開発を行っている.

２．１ WABIAN

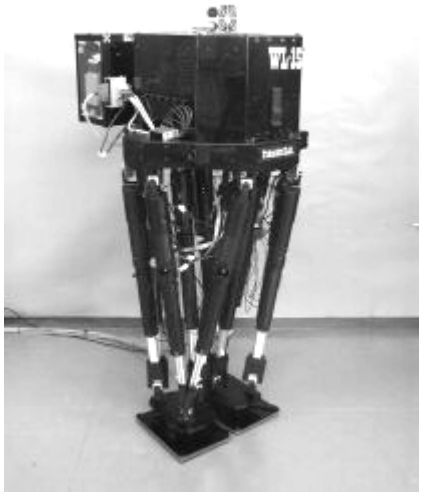


まず WABIAN(WAseda BIpedal humANoid)の全体写真を右図に示す．その身長は直立静止状態で 1.9[m]総重量は約 130[kg]ある．アクチュエータは下肢では片脚に 6 軸，体幹に 3 軸，上肢では片腕に 7 軸，手にも各 3 軸，首部に４軸あり，全身合計で 43 の自由度を有している．このロボットでは，M.Vukobratovich によって提案された ZMP(Zero Moment Point)の概念を用いて，２足歩行ならびに全身運動における軌道計画と安定化制御の問題に取り組んでいる．

この研究で構築された歩行に関する工学モデルは，ソニーのヒューマノイドの SDR シリーズだけでなく, 最近始まった日立との歩行障害者用支援機の性能評価用ヒューマノイドの開発やテムザックとの（車椅子では不可能な）階段昇降も可能な歩行障害者乗用２足歩行ロボットの開発にも応用されている．

２．２ WL-15

近年，多くの企業や研究機関でヒューマノイドが開発されているが，その移動機構には車輪やクローラを用いているものが，まだ圧倒的多数を占めている．その理由は，まずロボットによる２足歩行技術が未だ発展途上である上に，従来の２足歩行ロボット自体が，車輪やクローラのように，その上部に負荷や搬送物を搭載して移動するための独立した移動機構として設計・開発されていないところにある．

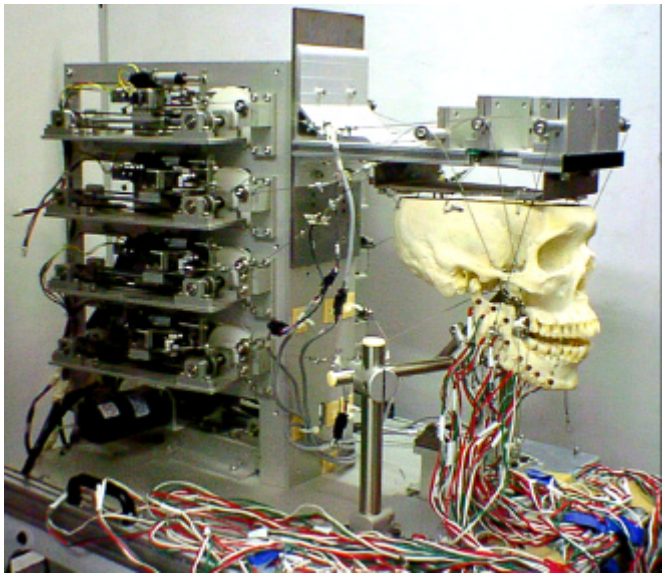


しかし一方で，車輪やクローラを人間の住環境で使うには，そのための様々なインフラの整備を必要とし，結果的に普及の弊害となる．そこで，現環境の変更・修正を必要としない移動機構が必要とされ，筆者らは「ロボットの移動用モジュールとしての２足ロコモータ」という発想から研究を開始した．これは下半身のみで自立歩行が可能な２足ロコモータを開発し，階段や斜面を昇降できる歩行障害者用の移動椅子等を開発できるようにするというものである．その結果，剛性の高いパラレルリンク機構を用いた脚と腰部のみで構成され，樹脂材料を多用することにより小型・軽量で，しかも電池駆動により自立性の高い歩行あるいは移動が可能な WL-15（Waseda Leg No.15）を開発した．（左図）

３．デンタル・ロボット

筆者らが「デンタル・ロボティクス」と呼び，「ヒトの顎構造を模したロボットを開発し，これを用いて咀嚼運動を再現することで，ヒトの咀嚼に関する工学モデルを構築すること」を目的とした咀嚼ロボットに関する研究，さらに，そこで得られたヒトの咀嚼モデルをもとに設計・製作された顎運動障害患者用の顎開閉口治療ロボットなどの開発を行っている．

３．１　咀嚼ロボット

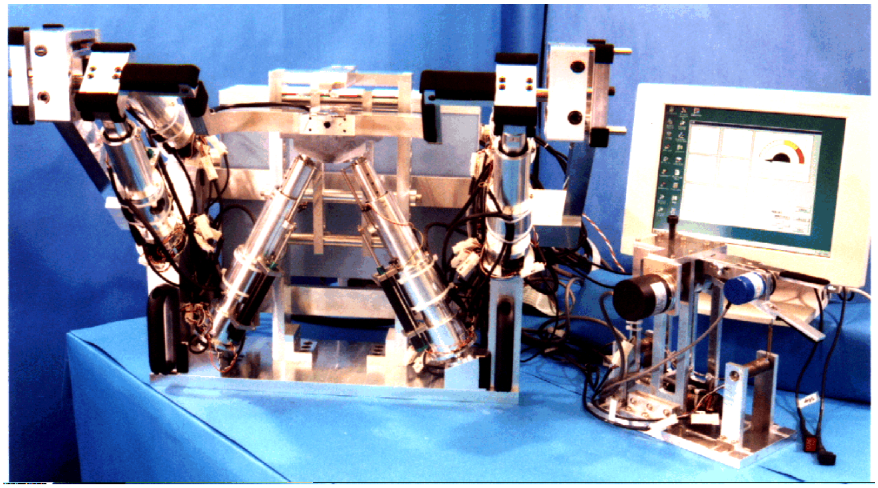


咀嚼ロボットを用いてヒトの咀嚼メカニズムに関する工学モデルの構築を目指している．右図に示す最新のロボットはオキノ工業と共同開発したWOJ-1R　(Waseda-Okino Jaw No.1 Refined)である．ヒトの筋特性をシミュレートするため，非線形粘弾性機構およびDCサーボ・モータとプーリ巻き上げ式のワイヤ機構を用いて下顎を駆動する合計 11 の自由度を有する．咀嚼生理学的視点から１咀嚼サイクルを４つの相に分け，それぞれの相の特徴と目的に合った顎運動パターンと制御パラメータを切り替える状態遷移型の制御を行うことで，ヒトの咀嚼運動を再現する．この研究で得られた顎運動の工学モデルは，山梨医大と共同で開発中の顎運動障害者用の開閉口訓練ロボットの臨床治療，ならびに和洋女子大と共同で開発中の食物物性測定用ロボットなどに応用されている．

３．２　顎開閉口訓練ロボット

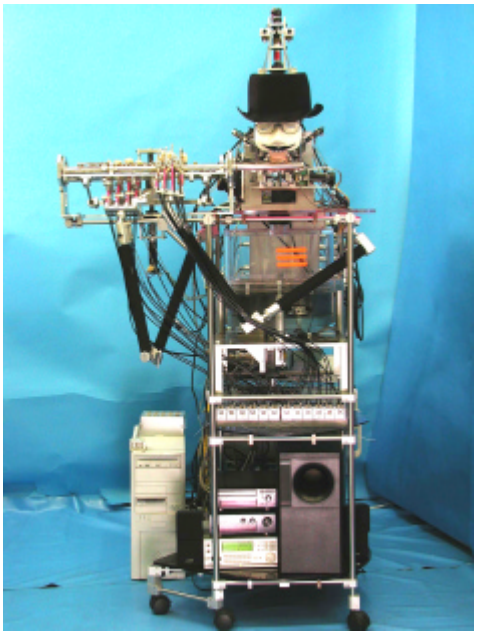
山梨医科大学および工学院大学と共同で，マスター・スレーブ方式と呼ぶロボットの制御・構成技術を用いることで，従来の手動式開口器に比べて多角的で高精度かつ定量性の高い顎運動障害患者の開閉口訓練・治療ができるロボットの開発を行っている．

現在は下図に示すように，直動型アクチュエータによる６自由度の並列駆動機構を有するWY(Waseda Yamanashi)-5R(Refined)の開発と，山梨医大に設置し患者への適用を行う実用プロトタイプWY-6の製作を完了している．WY-5Rでは仮想コンプライアンス制御と呼ぶ制御方式を導入している．各アクチュエータの位置ならびに６軸力センサにより患者の顎に加わる３軸力と３軸モーメントを計測，これと座標変換マトリックスを用いたリアルタイム演算により６個のアクチュエータの速度目標値をフィードバック制御する．これにより，あたかも全運動軸を回転中心とした開閉口方向にのみ強制力を発生し，それ以外の方向へは外力に応じて浮遊する仮想的な機構が存在しているかのようにロボットが動く．したがって，患者個々に異なる様々な顎運動経路へのロボットの自動適応動作が実現し，顎関節に余分な力をかけることなく安全な開閉口訓練が行える．また，マスター（術者）側とスレーブ（患者）側のロボットをISDN回線により接続することで，両者がお互い遠隔地にいても訓練・治療が可能となっている．さらに，熟練した術者の治療中における操作量（マスター側のレバーの変位と力）を記憶し，後でそれら（レバーの変位と力）を同時に入門者が感じることができる治



療術教育用のマスター・ロボットも開発している．

４．フルート演奏ロボット



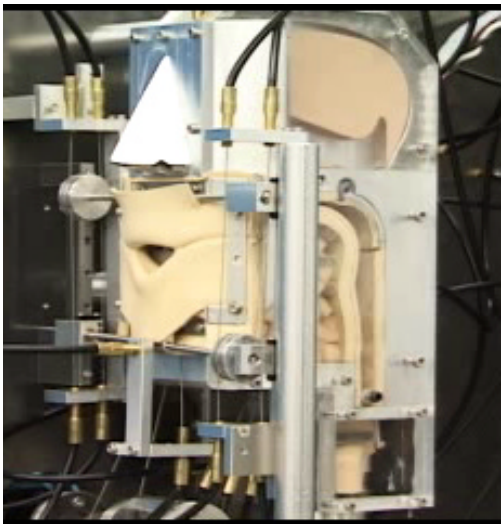
人間のフルート演奏は人体各部の巧みな協調動作により実現している．筆者らは，楽器演奏時の人体各器官の働きを機械モデルにより再現し，フルート演奏のメカニズムを工学的視点から解明することを目指し，現在は右図に示す人間形フルート演奏ロボット WF-3RIX（Waseda Flutist No.3 Refined version IX）の開発を行っている．ハードウェアとしては，呼気吸気を行う肺部，フルート歌口との相対位置決め用姿勢制御装置，トリル演奏のできる指部，口唇部，ダブルタンギング機構，ヴィブラート発生装置，MIDI 伴奏同期システムにより構成されている．ソフトウェアとしてはヒューマン・インタフェース機能の向上を図ったロボット制御用ソフトウェア，フルートの安定した吹鳴と楽曲の演奏，また音質評価パラメータを用いた無音状態からの吹鳴音探索，口唇部を固定してすべての音を吹鳴できる「ジェネラルポジション」の探索，さらに FFT を用いたリアルタイム吹鳴音評価システムを有している．それらを統合したロボット・システムにより，ジェネラルポジション決定後の口唇部パラメータの自律探索を実現し，さらに現在は人間とロボットの合奏する際，ロボットの演奏にリアルタイムなインタラクション性を持たせる研究を行っている．

５．表情ロボット

筆者らは，数年前から早稲田大学の心理学教室と連携し，ヒトの体幹と頭部を模した構造で，感情を表出することが可能なロボット WE シリーズを開発し，ヒトの心理に関する工学モデルの構築を試みている．

左図に示す最新の WE-4（Waseda Eye No.4）は合計で 29 自由度を持ち，感覚としては人間の五感のうち視覚，聴覚，嗅覚および皮膚感覚（触覚と温覚）の４つを有している．WE-4 は覚醒・不覚醒，快・不快，確信・不確信の３軸で構成される３次元の心理ベクトル空間を持ち，この中に，外部刺激ならびに内部状態を強制項として持つ２階の線形および非線形の常微分方程式群（情動方程式）と，その解としての情動ベクトルとムード・ベクトルが定義されている．この心理空間はまた，「幸せ」や「驚き」などの心理用語に対応する７つの領域に区分され，刻々と動き回る情動ベクトルが通過中の区分領域に対応した感情が，そのときのロボットの感情として一意に決定される．

６．発話ロボット



人が社会生活を営むとき，社会を構成している個人間には頻繁な意思の交流が必要であり，これがなくては社会を維持していくことは困難である．社会における意思の伝達には通常，音声言語 (言葉)が用いられ，それ以外に身振り・文字・絵などによる伝達方法があるが，伝達効率から考えると音声言語が最も優れたものであるといえる．音声言語の生成，つまり発声に関してはこれまでに多くの研究がなされているが，未だ脳における発声の運動計画処理機構から運動器官における音声生成の運動までを包括的に研究された例はなく，また人の発声動作は十分に解明されていないのが現状である．

筆者らは，文科省 CREST プロジェクトの一環として NTT を中心としたグループとの共同研究において，発声器官 (肺，声帯) と調音器官 (舌，唇，歯，鼻腔) を持つ発話ロボットを開発し，これを用いてヒトの発声メカニズムの解明を目指している．また，これによって発声の訓練や外国語習得の有効な手段を提案できるものと考えられる．上図に示すものが，最新の機構を有する発話ロボット WT-2(Waseda Talker No.2)である．WT-2 は全体で 15 の自由度有し，声帯から口唇までの声道長さは 175[mm]と，成人男性と同程度である．まだ不自然さは残るものの，WT-2 は日本語の 50 音全て，濁音および半濁音の子音までの発声が可能となっている．

７．まとめ

本稿では，筆者らの研究室で開発しているロボットの具体例を紹介しながら，人間の形態と機能を模したロボットを設計・製作し，これを用いて人間の行動や機能を再現することで，構成論的に人間の工学モデルを構築する「ロボット工学的人間科学(Robotic Human Science)」と，そうして得られた人間の工学モデルを従来の基盤工学における各種モデル群と統合し，これと並行して人間のための様々なロボットや機器・装置の開発の実践を通して，人間に関わるロボットやシステムに関する演繹的設計論の構築を目指す「人間モデル規範型ロボット工学」について述べた．

本稿が同分野に関する理解の一助になれば，存外の喜びである．

参考文献：

(1) 早稲田大学ヒューマノイドプロジェクト編著，“人間型ロボットのはなし，”日刊工業新聞社，東京，1999 年

(2) 高西淳夫，“デンタル・ロボティクス，”日本歯科医師会雑誌，vol.53，no.3，June 2000.

(3) 三輪洋靖，高西淳夫，他，“ヒューマノイドロボット用心理モデルの構築 -学習システム・気分ベクトル・2 次情動方程式の導入-，”第 20 回日本ロボット学会学術講演会予稿集，Sep. 2002.

(4) 西川員史，高西淳夫，他，“人間に近い発声を目的とした新形発話ロボットの開発，”第 20 回日本ロボット学会学術講演会予稿集，Sep. 2002.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-2

身体動作に基づく人ロボット対話の解析
－主観的評価・性格との関係－
An analysis of body movement on human-robot interaction

神田崇行　石黒浩　今井倫太　小野哲雄
Takayuki Kanda　Hiroshi Ishiguro　Michita Imai　Tetsuo Ono
ATR 知能ロボティクス研究所
ATR Intelligent Robotics and Communication Laboratories
kanda@atr.co.jp

Abstract

Toward an ideal human-robot communication, we created an interactive humanoid robot that complexly behaves and autonomously interacts with humans by speaking and taking gesture. This is a testbed for embodied communication. Our approach is to precisely measure the interaction represented as body movements between the robot and humans to identify the essential embodiment and behavior. We performed an experiment about the interaction. The body movements is measured by a motion capture system, and then compared with subjective evaluation about the robot and personality of the subjects. It reveals the important role of the body movements. As the result, we have verified effect of the well-coordinated behaviors on the subjective evaluation, and found how important to measure the body movements for evaluating human-robot interaction.

1　はじめに

我々は、ロボットが人間似の身体を持つことの意味はコミュニケーションにあると考えている。つまり、将来的にヒューマノイドロボット([1][2]など)は自らの身体性を活用して、コミュニケーションのための新しいメディアとして、人間社会と情報社会のインターフェース的役割を果たすことが期待されている。

これまでにも、身体を用いたロボットと人間のコミュニケーションに関する研究が行われてきた。身体の中でも非常に効果があると考えられているのが視線である。たとえば、視線を対話相手に向けてアイコンタクトを行うことで、ロボットは自らのコミュニケーション意図を人に伝え、人間と自然に対話することが可能になる。これまでにも音声と視覚による人物追従[3]など多くの研究が行われた。このほか、身振りにおける腕や身体の向き[4]、顔の表情[5]などの研究が行われている。

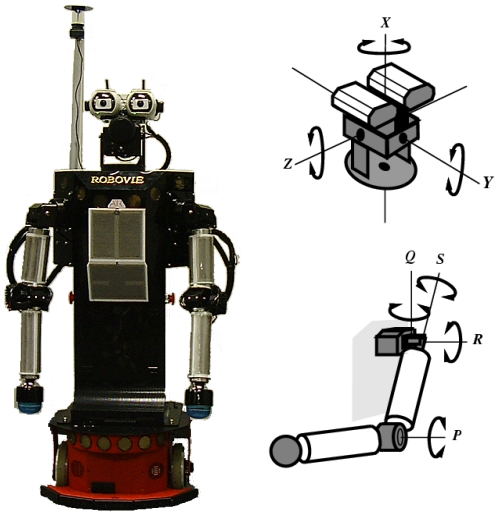


Figure 1： Robovie

このように、これまでにロボットの身体の個別の機能（視線や腕のゼスチャ）の有効性は確かめられている。我々はむしろ身体機能を組み合わせた総合的なロボットにおける個別の機能の役割に関心を持つ。このために、日常的会話を行うインタラクティブロボットを作り、相互作用を解析する中で、人とロボットの身体や身体動作の相互の関わり合いを見いだす。特に、ロボットが人間に何かする一方的な関係ではなく、ロボットと人が協調的に動作するような、コミュニケーションにおける協創的関係に関して調査する。

2　コミュニケーションロボットの開発

2.1　ハードウェア構成

図 1,2 に我々の開発したロボット ”Robovie ”を示す. Robovie は人間とコミュニケーションするために人間に類似した上半身を持つヒューマノイドロボットである。人間が視覚・触覚・聴覚をもつようにカメラ、マイク、接触センサなどの様々なセンサを持つ. このような人間に類似した身体とセンサを

Figure 2: interactive behaviors

用いて，Robovie は人間とのコミュニケーションに必要な様々な音声とゼスチャを交えた対話的行動を生成することができる．また、Robovie はすべての必要な制御機器を内蔵している．本体下部に搭載されている Pentium III 850MHz の PC を用いて，音声認識や画像処理を行うとともに，すべてのモータおよびセンサを制御している．

2.2 自律対話行動のためのソフトウェア

ロボットの身体を活用するためにこれまでに行われた認知科学的実験に基づき、我々は自律的に人間の行動に反応して動作するコミュニケーションロボットのソフトウェアを考案し、Robovie 上に実装した。このソフトウェアは、状況に応じて動作するシンプルな個々の行動モジュール(状況依存モジュールと名付けられる) を大量に用意し、このモジュール間の関係をシンプルなルールによって記述することからなる。このようにシンプルな構造から複雑な自律システムの挙動が生じることがこのソフトウェアの特徴である。[6]

2.2.1 コミュニカティブユニット

これまでにヒューマノイドロボットの視線や腕の動作に関する研究が行われてきた。コミュニカティブユニット(Communicative unit)は，このようなコミュニケーションにおける身体の活用に関する知見に基づき、身体を利用したコミュニケーションのために必要な基本要素行動である. 具体的には，視線を合わせる，物の方を見る，物を指さすといったコミュニケーションの基礎となる要素行動である（図 3 Indication 内の Communicative Unit）.

2.2.2 状況依存モジュール

状況依存モジュール(Situated module)はこのソフトウェアの基礎となる行動モジュールである．これは，

特定の限られた状況で，ロボットにある特定の動作をさせる行動モジュール

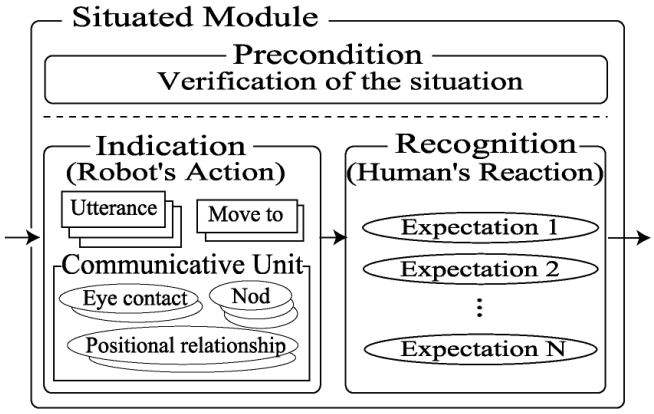


Figure 3 : Situated module

と定義することが出来る. 状況依存モジュールは前提条件部（precondition）、提示部(indication)、認識部(recognition)からなる。前提条件部は、現在の状況がこの状況依存モジュールが実行可能な状況かどうかを判断する。たとえば、インターネットに接続して明日の天気について話すモジュールは、インターネットに接続できない場合には実行できない。また、握手を求めるモジュールはロボットの前方に人(と想定される近距離の物体）がない場合には実行されない。

提示部は人間に働きかけを行う。これはコミュニケーティブユニットを組み合わせ，また不足する行動（特定の発話，ある地点に移動する，等)を直接実装することで実現される. たとえば握手モジュールは、アイコンタクトを行い、適切な位置関係をとり、そして「握手してね」といって人間に手を差し出す。このロボットの行動は、アイコンタクトや位置関係に関するコミュニカティブユニットに、特定の発話を追加することで実現される。

認識部は提示部でロボットが行った行動に対する様々な人間の反応的な行動を認識するように設計される。これは、人間の行動の予期を行うことを意味する。状況依存モジュールは実行する状況を限定するのみでなく、モジュール自身が特定の状況を作り出し、そしてこの特定の状況下において人間の複雑な行動を認識する。たとえば「握手しよう」といって手を差し出した時に、手先が触れられれば、それは人間の握手行動である。また、「どこから来たの」とのロボットの問いかけに対しては、地名の返答を期待して音声認識を行うことが出来る。

ロボットシステムは逐次的に常に１つの状況依存モジュールロボット実行することにより自律行動を実現する. このモジュールの実行が終了すると、状態遷移モデルと同様に、状況依存モジュールの実行結果に応じて、あらかじめ決められた遷移を行い、次に実行する状況依存モジュールが決まる。

2.2.3 自律行動の動作例

このようなアーキテクチャに基づいて, 人間と日常的なコミュニケーション行動をする自律行動を実装した. このようなインタラクション機能は将来日常生活の場で活動するロボットには欠かせないものである. 自律インタラクション機能の実験のために，「抱擁」「握手」「簡単な会話」「物の指

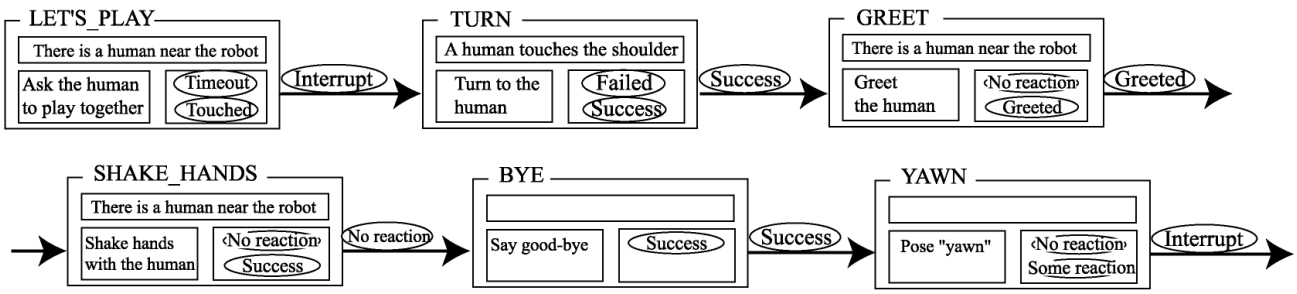


Figure 4: example of situated module transition

さし」といった 80 種類程度の人間との遊び行動および，「頭をかく」「腕組みする」といった 20 種類程度の待機行動，パトロールの真似をする等の環境を移動する 10 程度の行動を実装した．ロボットシステムは、人間からの働きかけがないときは待機行動や移動行動を実行し、人間からの働きかけがあった場合には、働きかけがある限り遊び行動を続けるように設計されている（図４）。

3　相互作用の数値解析実験

3.1　実験設定

本実験では、個々の被験者は前章で述べた自律的に動作するロボットを 10 分間観察した。被験者は 26 名であり、平均年齢 19.9 歳の大学生である。観察前に、被験者はあらかじめロボットとのコミュニケーションの見本を示された。実験は 7.5 m x 10 m の部屋で行われた。

また、被験者とロボットはモーションキャプチャシステムの動作計測用のマーカーを取り付けて実験を行った。この身体的動作の数値的結果と、被験者のロボットに関する主観的印象（実験に用いた形容詞対を表 1 に示す）、被験者の性格テストの結果を分析することで、人ロボット相互作用を数値的に解析することを試みる。

3.2　モーションキャプチャシステムを用いた身体動作の数値化

実験に用いたモーションキャプチャシステム[7]は、部屋の外周に沿って取り付けられた１２台の赤外線照射装置付きの赤外線カメラと、赤外線を反射するマーカーから構成される。モーションキャプチャシステムはすべてのカメラ画像上での各マーカーの２次元位置をもとに各マーカーの３次元位置を計算する。用いたシステムの時間分解能は 120Hz, 空間分解能は実験環境において約 1mm である。

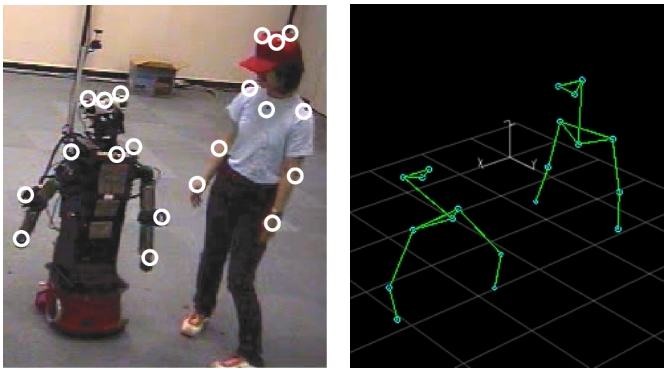


Figure 5: モーションキャプチャシステムによる相互作用解析

Adjective-pairs		Mean	Std. Dev.
Good	Bad	4.88	0.95
Kind	Cruel	4.85	1.29
Pretty	Ugly	5.08	0.93
Exciting	Dull	4.46	1.61
Likable	Dislikeable	4.77	1.03
Evaluation score		4.81	0.92

Table 1: The used adjective-pairs for subjective evaluation, and the mean and standard deviation as the result

	Mean	Std. Dev.
Distance (m)	0.547	0.103
Eye contact (s)	328	61.8
Eye height (m)	1.55	0.124
Moved distance (m)	35.2	17.0
Moved distance of hands (m)	108	29.5
Synchronized movements (s)	7.95	6.58
Touch (num. of times)	54.9	20.8

Table 2: The result about the body movement

図５に示すように、このマーカーを取り付けた。取り付け位置は、頭部(人間はマーカーが取り付けられた帽子をかぶる)、肩、首の付け根、腕の各関節である。このようにロボットと人の双方に類似の位置に取り付けることで、人とロボットの身体動作の相互作用を分析する。頭部の３点のマーカーにより視線の高さや向き（アイコンタクト）を検出する。また、肩と首の付け根のマーカーにより、対ロボット距離および移動距離を測定する。また、腕のマーカーにより、手先の運動量（体に対する手先の相対位置の移動量）や、腕の同調的動作(ここでは、人とロボットの間で、体に対する手先の相対位置の３秒間の相関値が高い時間領域、と定義する)の検出を試みる。なお、接触行動の分析はロボットの内部ログを用いた。

3.3　実験結果

相互作用の際の身体的動作の数値的解析の実験結果を示す。実験の結果、主観的評価が被験者の身体動作と相関を示

	Evaluation	Dist.	E. C.	E. H.	M. D.	M.D.H.	S. M.	Touch
Dist.	-0.04	1.00						
E. C.	**0.57**	**-0.47**	1.00					
E.H.	0.08	**-0.39**	0.29	1.00				
M. D.	-0.32	0.20	**-0.43**	0.02	1.00			
M.D.H.	0.01	-0.04	-0.21	-0.09	**0.49**	1.00		
S. M.	**0.54**	-0.05	0.28	-0.05	0.15	**0.61**	1.00	
Touch	0.21	-0.45	**0.49**	-0.07	-0.15	**0.35**	**0.41**	1.00

Table 3: The correlation between body movements and subjective evaluation

（E.C. : eye contact, E.H. : eye height, M. D. : moved distance, M.D.H.: moved distance of hands, S.M.: synchronized movements）

	Eval.	Dist.	E. C.	M. D.	M.D.H.	E. H.	S. M.	Touch
Depression	**-0.35**	-0.25	-0.29	0.33	0.02	-0.07	**-0.36**	-0.08
Cyclic Tendency	0.02	-0.03	-0.12	0.21	-0.03	-0.21	-0.10	0.24
Inferiority	0.08	-0.22	0.10	0.18	0.09	-0.19	-0.10	0.20
Nervousness	-0.13	-0.24	-0.03	0.15	-0.07	-0.10	-0.25	0.23
Lack of Objectivity	0.21	0.15	0.01	0.21	0.18	**-0.39**	0.17	0.17
Lack of Cooperativeness	-0.27	-0.08	-0.13	0.20	-0.19	0.23	**-0.41**	-0.10
Lack of Agreeableness	0.15	-0.07	0.19	-0.04	-0.27	**0.47**	0.00	-0.06
General Activity	0.13	-0.25	0.06	0.01	0.11	**0.42**	0.20	0.17
Rhathymia	0.23	-0.21	0.15	0.28	0.04	-0.03	0.14	0.21
Thinking Extroversion	0.23	0.26	0.18	-0.08	-0.20	**-0.33**	0.12	0.14
Ascendance	0.28	0.03	**0.34**	-0.19	-0.11	**0.38**	0.25	0.12
Social Extroversion	**0.51**	-0.01	0.32	**-0.43**	-0.05	0.25	0.26	0.09

Table 4: The subjects' personality and their correlation with the subjective evaluation and body movements

（E.C. : eye contact, E.H. : eye height, M. D. : moved distance, M.D.H.: moved distance of hands, S.M.: synchronized movements）

した。つまり、コミュニケーションの際にコミュニケーションをどのように感じているかが身体動作に表出する、と考えられる。また、主観的評価と強い相関を示したのは人とロボットが協調的に振る舞った結果生じる動作であったことはたいへん興味深い。

3.3.1　SD 法に基づく印象の主観的評価

７段階の SD 法を用いてロボットの印象の主観的評価を行った。表 1 に用いた形容詞対と、２６人の実験結果の平均値と標準偏差を示す。実験に用いた形容詞対は、従来研究[8]における因子分析で見いだされた第一因子である評価性因子に負荷を持つ形容詞を選択した。また、これらの５つの形容詞対への評価を平均することにより、ロボットの印象に関する評価性得点を計算した。

3.3.2　主観的評価と身体的動作の相関

表 2 に身体動作に関する数値的解析結果を示す。アイコンタクトの平均時間は 328(sec.)であり、実験時間の半数を上回った。ロボットの視線の高さは 1.13(m)であることから、一部の被験者は若干かがんで時にロボットと視線の高さを合わせたことが分かる。また、一部の被験者に、ロボットの体操などの腕の動きを真似する同調行動が見られた。

さらに、表 1 に示した形容詞対を基に計算された評価性得点と、表 2 の身体動作との相関を計算した。この結果を表 3 にしめす。被験者数が 26 であるので、相関値の絶対値が 0.3297 以上が有意な相関である（表中に太字で示す）。分析の結果、アイコンタクトと同調が主観的評価と比較的強い相関を示した。一方で、身体動作の間での相関をみると、アイコンタクト-距離、アイコンタクト-移動量、同調-指先移動量、同調-接触といった項目で相関がみられるものの、距離・移動量・指先移動量・接触は主観的評価とあまり相関が見られなかった。つまり、単にロボットに近づき、指先を活発に動かしたり、ロボットに触れたり、という活発なインタラクション行動が良い印象につながるわけではなく、むしろロボットとの間にアイコンタクトや同調的動作という協調的な関係を築き上げた被験者がロボットに良い印象を持ったことがわかる。

3.3.3　性格の影響

さらに、我々はこのような身体動作と被験者の personality の関連を見るために被験者の性格テストを行った。テストには YG 法検査（矢田部-ギルフォード法）を用いた。YG 法は 1949 年に開発された Guilford-Zimmerman Temperament Survey（10 factor scale が用いられる）を元に、1958 年に矢

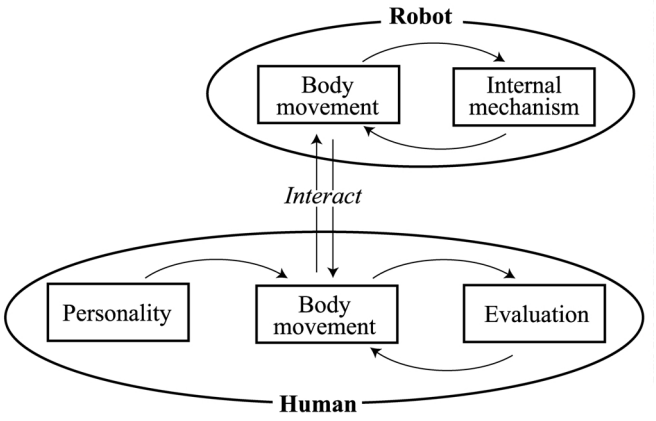


Figure 6: A model of forming evaluation through the interaction of the body movements

田部が日本人向けに翻訳したものである。被験者は 120 の質問(3-point rating scale)に回答することにより、12 factor に関する性格・気質の傾向が明らかになる[9]。

この性格テストの結果と印象、身体動作の相関を調べた（表 4）。この結果、評価には抑鬱性と社会的外向が相関を示し、身体動作には表中の太字に示すように多くの項目が相関を示した。たとえば、協調性の欠ける(Cooperativeness)被験者は同調的動作と負の相関を示した。つまり、協調的でない被験者ほど、同調的動作をしなかった。また、抑鬱性（Depression:厭世的気分を表している病的な精神状態）の高い被験者も同調的動作をせず、かつロボットを高く評価しなかった。支配的(Ascendance)な被験者はアイコンタクトをしがちであった。社会的外向(Social Extroversion)の大きい被験者は対話の際にあまり歩き回らず、またロボットを高く評価した。

3.4 重回帰分析による身体動作間の関係の検証

実験結果から、主観的印象と身体動作の間には相関関係が見られた。そこで、重回帰分析により印象に関する評価性得点をこのような身体的動作から推定することを試みる。このような分析から身体動作間の主観的印象における関係を見いだす。

このため、身体動作と評価性得点に関して重回帰分析を行い、表 5 に示す標準化偏回帰係数(standardized partial regression coefficient)を得た。得られた重回帰式を(1)に示す(DIST, EC, EH, MD, MDH, SM, TOUCH はそれぞれ各身体動作の測定値を正規化したものである)。なお、E は 7 段階評定のスコアの平均値であることから値域は 1 から 7 である。この回帰式の重相関係数は 0.77 であることから、評価性得点の 59%がこの式から説明されることが分かる。なお、この重回帰式に関する有意性を分散分析により検証したところ、F(7,18)=3.71, P<0.05 で有意であった。

$$E = \alpha_{dist} \bullet DIST + \alpha_{ec} \bullet EC + \alpha_{eh} \bullet EH + \alpha_{md} \bullet MD + \alpha_{mdh} \bullet MDH + \alpha_{sm} \bullet SM + \alpha_{touch} \bullet TOUCH + \alpha_{const}$$

分析の結果、偏回帰係数が示すように、ロボットの評価性得点について、アイコンタクトと同調的動作が多い被験者ほどロボットへの評価が高くなったことが分かる。一方、単に距離が近く、移動距離が大きく、または接触行動が多い被験者はロボットへの評価が低くなることが分かる。

	Coefficient	Value
Distance	α_{dist}	0.173
Eye contact	α_{ec}	0.476
Eye height	α_{eh}	0.019
Moved distance	α_{md}	-0.228
Moved distance of hands	α_{mdh}	-0.029
Synchronized movements	α_{sm}	0.535
Touch	α_{touch}	-0.186

Table 5: the standardized partial regression coefficients obtained by the multiple linear regression analysis

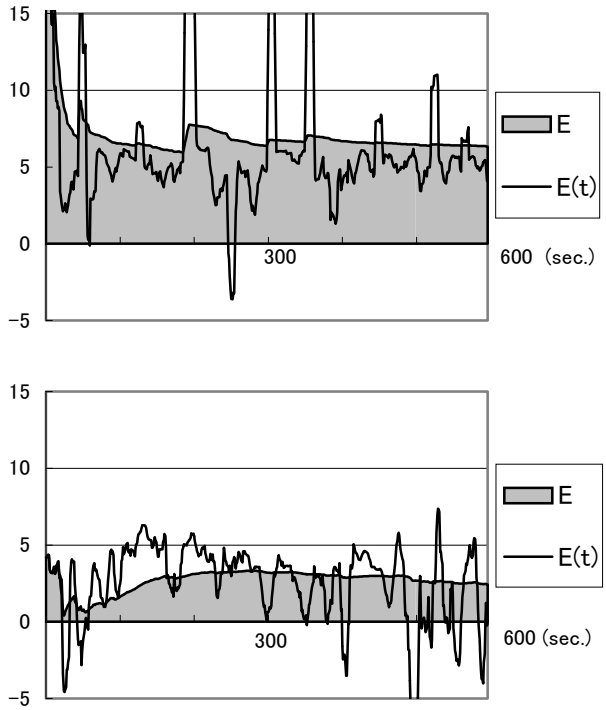


Figure 7: illustration of the entrainment score (upper: about the subject who treated the robot as if it was humans child, lower : about the subject who embarrassed with interacting with it)

ここで、この回帰式の各項目は瞬間毎に測定可能な値であることから、式(2)のように瞬間毎の評価の推定値を計算することができる（DIST(t)などは時刻 t における身体動作の測定値を意味する）。このような瞬間毎のインタラクションの評価値を引き込み感得点と名付ける。

$$E(t) = \alpha_{dist} \bullet DIST(t) + \alpha_{ec} \bullet EC(t) + \alpha_{eh} \bullet EH(t) + \alpha_{md} \bullet MD(t) + \alpha_{mdh} \bullet MDH(t) + \alpha_{sm} \bullet SM(t) + \alpha_{touch} \bullet TOUCH(t) + \alpha_{const} \quad (2)$$

また、当然 E と E(t)には次の関係がなりたつ。

$$E = \int E(t) \quad (3)$$

実際にこの瞬間値に基づき実験データをグラフ化した。被験者１は実験後に「目の動きが本当に子供に見上げられて見つめられている気分になった。本当に人間の子で無邪気な人格

ID	Contents	Evaluation
TICKLE	Tickle	-2.09
APOLOGIZE	Apologize	-1.96
NOT_TURN	Say, “I’m busy”, and refuse to play together	-0.51
SLEEP_POSE	A pose of sleeping	-0.42
FULLY_FED	A pose of fully fed	0.32

Table 6: the worst 5 *situated modules* based on the average of the entrainment scores

ID	Contents	Evaluation
EXERCISE	Exercise	5.75
ASK_SING	Ask humans, “May I sing a song?”	5.59
CONDUCTOR	A imitating pose of conductor	4.85
WHERE_FROM	Ask humans, “Where are you from?”	4.55
LET’S_PLAY	Say, “Let’s play, touch me”	4.24

Table 7: the best 5 *situated modules* based on the average of the entrainment scores

であるかのように錯覚しそうになった。」とコメントし、ロボットと上手く遊んだ被験者である。図６にこの被験者の引き込み感の時間遷移を示す。グラフ中の実線は引き込み感を示し、塗りつぶされた領域は時刻 t までの E(t)の積分値である。また、被験者２は対照的に、ロボットと上手く相互作用が出来なかった被験者の例である。両者のグラフを比較すると、被験者１のグラフは引き込み感が 5 付近で推移し、時に大きな値を示していることが分かる。被験者 1 はロボットの近くでロボットと目を合わせながら子供に話しかけるように対話を続けた。時刻 200[s]付近での大きな値は、被験者がロボットの体操する腕の動きを真似たことから生じた。これに対して、被験者２のグラフはしばしば 0 以下の値を示し、特に実験の終盤では非常に不安定に低い値を示している。実際に、被験者２は実験の終盤にロボットの目を隠し、苛立つようにロボットの接触センサを触り、ロボットから遠ざかるような動きを見せた。

3.5 引き込み感に基づくインタラクティブロボットの行動開発

ここまでの引き込み感評価は、ロボットの開発アプローチとは独立であり、どのような相互作用型ロボットにも適用可能である。ここでは、さらに我々の状況依存の行動モジュールと単純なルールに基づく開発手法の上で、この評価方法を利用することを考える。これは、同時に引き込み感がインタラクションの瞬間的評価を表していることの検証にもつながる。

我々は、すべての被験者の実験データに関して、ロボットが実行したモジュール毎に、そのモジュールが実行されていた間の瞬間毎の評価値を計算し、引き込み感（entrainment score）と名付けた。表７、表８はこの評価値が最も低かった５モジュールと、最も高かった５モジュールを示す。評価値が低かったモジュールはあまりインタラクティブなモジュールではなかった。例えば、SLEEP_POSE, FULLY_FED は人間の行動に反応せず“ 寝たふり ”などを行うモジュールであり、NOT_TURN は、このような人間の行動に反応しないモジュールを実行している際に人間がロボットの肩をたたいて遊ぶように呼びかけた際に、人の手を振り払って“ 忙しい ”と言うモジュールである。これに対して評価が高かったのは体操や指揮の真似といった、人間がロボットの身体動作を真似しがちなモジュールや、問いかけや遊びの呼びかけと言ったインタラクティブ性が高く、人間をロボットとの対話に引き込むようなモジュールであった。

このように、引き込み感を利用することが、ロボットのインタラクティブな行動の設計に利用できることが示された。同時に、この方法による引き込み感推定の正しさを示すものである。このように身体動作から推定される引き込み感を利用することで、例えば相互作用行動の学習や内界センサを用いた引き込み感センシングなど、同様のインタラクティブなロボットに一般的に応用可能であると考える。

4 考察

4.1 身体動作に基づく協創対話

分析の結果、インタラクションの評価には、人の協調的行動（アイコンタクトや同調行動）が深く関わることが見いだされた。つまり、ロボットの行動に対して協調的に振る舞う人ほどロボットのことを高く評価する。一方で、ロボットに対する協調的な振る舞いと、人の性格との間にはそれほど相関関係が見られない。ここから考察されるのは、人はロボットとの相互作用の間に、ロボットとの協調的な関係を築き上げ、このような関係を上手く構築できた人はロボットと上手くインタラクションし、ロボットを高く評価する、という仮説である。近年、相互作用における引き込み(Entrainment)現象に関する研究が行われるようになってきたが([5][10]など)、このような引き込み現象の概念に基づくと、ロボットが多様な身体的動作を行うことで、人間の身体動作の引き込み現象を生じさせて、ロボットへの協調的な関係を生じさせていると言える。

4.2 主観的評価と性格、身体動作の関係

性格テストと身体動作の関係に関する実験結果は、我々には理解がしやすい。たとえば協調的でない被験者は同調的な動作を行わなかった。つまり、被験者の身体動作は性格により異なったものとなると考えられる。一方で、主観的評価は被験者の性格だけでなく、むしろ協創的な身体動作からも形成される。このような主観的評価形成のモデルを図６に示す。

このモデルはまた、身体動作の観察から主観的評価を推定できる可能性を示唆する。実験結果に関して、重回帰分析により 60%の主観的評価が身体動作から説明された。これは言語的内容にかかわらずに行われたにも関わらず大きな値

である。ロボットは、音声認識により小さい子供程度の能力ではあるが人間と会話することができる。実験においても、被験者は時にロボットに話しかけた。発話の内容は、主にロボットに対する要求的な内容であった（特にロボットが前に実行した行動を再び要求するものが多い）。このような呼びかけに対して、ロボットは時に正しく反応し、あるいは誤って反応した。このような発話内容の分析は会話分析などにより可能ではあるが、どちらかと言えばこれらの手法は主観的であり、また主に探索的に事象を見いだすことに用いられることが多い。むしろ、本稿にて報告した身体動作による評価は、客観的な尺度“文脈と独立に数値的に得られた身体動作”により行われたものであり、この客観性から様々な応用可能性があると考えられる。

本研究では被験者の性格が相互作用におよぼす影響を分析した。一方、これまでにも[11]のようにロボットの性格を変化させる研究が行われている。両者のアプローチを組み合わせ、同時にロボットの性格も考慮することで、性格に関する相互作用が生じることが考えられる。このような性格レベルのインタラクションに関して本手法を適用することは興味深い将来課題のひとつである。

5 おわりに

本稿では、モーションキャプチャシステムを利用した相互作用の数値解析について報告した。我々の開発したロボットは自律的に人間とコミュニケーションを行う。特に、アイコンタクトや同調的行動といった人間とロボットが協調的にふるまう身体行動が主観的評価に大きく貢献した。この結果は、人間とロボットとのコミュニケーションは単にロボットが人間に何かを提示するのみでなく、人間とロボットが協調的関係を築き上げることの重要性を示していると考えられる。また、我々の相互作用の数値解析アプローチの可能性は大きい。今後、本手法を人-ロボットのみでなく、人同士の相互作用にも適用することで、コミュニケーションにおける身体動作の役割の解明を進めたい。

参考文献

[1] K. Hirai, M. Hirose, Y. Haikawa, and T. Takenaka. The development of honda humanoid robot. IEEE Int. Conference on Robotics and Automation, 1998.

[2] 橋本周司, 成田誠之助, 白井克彦, 小林哲則, 高西淳夫, 菅野重樹, 笠原博徳, ``ヒューマノイド -人間型高度情報処理ロボット-,"情報処理, Vol.38, No.11, pp.959-969, 1997.

[3] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano: “Real-Time Auditory and Visual Multiple-Object Tracking for Robots,” Proc. Int. Joint Conf. on Artificial Intelligence, pp.1425-1432, 2001.

[4] C. Breazeal, B. Scassellati: ``A context-dependent attention system for a social robot.” Proc. Int. Joint Conf. on Artificial Intelligence, pp.1146-1151. 1999.

[5] 小野哲雄, 今井倫太, 石黒浩, 中津良平：身体表現を用いた人とロボットの共創対話, *情報処理学会論文誌*, Vol.42, No.6, pp.1348-1358, (2001)

[6] T. Kanda, H. Ishiguro, M. Imai, T. Ono and K. Mase, An approach for developing interactive humanoid robots, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2002.

[7]Vicon Motion Capture System. http://www.vicon.com/

[8] 神田崇行,石黒浩,小野哲雄,今井倫太,中津良平, ``人間と相互作用する自律型ロボット Robovie の評価," 日本ロボット学会誌, Vol.20, No.3, pp.315-323, 2002.

[9]竹井機器工業, ``YG 性格検査''
http://www03.u-page.so-net.ne.jp/ta2/tkk/YG/YG.htm

[10]渡辺富夫, 大久保雅史, 小川浩基: ” 発話音声に基づく身体的インタラクションロボットシステム, ” 日本機械学会論文集(C編), 66 巻 648 号, pp.251-258, 2000.

[11] H. G. Okuno, K. Nakadai, and H. Kitano: ”Realizing Audio-Visually Triggered ELIZA-Like Non-verbal Behaviors,” PRICAI2002, LNAI 2417, Lecture Notes in Artificial Intelligence, Springer-Verlag, pp.552-562, 2002.

[12]今井倫太, 小野哲雄, 石黒浩:``身体表現を用いたロボットの発話生成,'' 第 13 回人工知能学会 AI チャレンジ研究会資料(JSAI SIG-Challenge-0113), pp.9-16, 2001.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-3

ロボット対話における自然な新規語彙の獲得

New word acquisition in the dialogue with Robot

小川 浩明
Hiroaki OGAWA
ソニー（株）デジタルクリーチャーズラボラトリ
Digital creatures laboratory, SONY Corporation
ogawa@pdp.crl.sony.co.jp

Abstract

新規語彙の獲得は、ロボットとの対話における重要な機能の一つである。従来の対話システムでは、新規語彙の獲得のためのモードが用意されていた。これに対し自律型 2 足歩行ロボット SDR-4X では、対話文中からの自然な新規語彙の獲得が可能となった。本報告では、対話文中に埋め込まれた新規語彙の発見とその発音の獲得の実装方法と評価結果について考察する。

1 はじめに

自律型ロボットにおいて、音声認識は最も重要なユーザインタフェースの一つである。

現在の一般的な音声認識では、辞書に定義された語彙以外の入力を受け付けない。また、連続単語認識では文法や N-gram などの語の連接関係を用いているため、受け付けなかった単語のみならずその周辺単語の認識結果にも悪影響を及ぼす。このためユーザが発声する辞書に含まれない語彙（Out-of-vocaburaly: OOV）は認識率に大きなダメージを与える [1]。しかし人名、地名などの固有名詞のすべてを辞書に含めることは非常に困難であり、なんらかの対策が必要となる。

この問題を解決する方法として、発話中に含まれる OOV の区間を推定し無視する方法が提案されている [2, 3, 4]。しかし、OOV は間投詞などのような不要な言葉[1] である場合もあるが、固有名詞など文章の重要な意味を示す場合も多い。

したがって OOV を単に無視するのではなく、新規語彙として取り込むことが出来れば、ロボットとのコミュニケーションが大きく広がることが期待される [5, 6]。例えばロボット自身の名前やロボットのユーザの名前を音声

[1] 感情理解などでは重要な役割を果たす可能性はある

図 1: SDR-4X

合成用および、音声認識用の語彙として獲得することで、それらを用いた呼びかけや、対話が可能となる。また、視覚系を用いた人物同定と共に人物の名前を認識語彙として獲得できればユーザは、

“小川 さんに会ったらメッセージを伝えて”

などの発話が可能となる。視覚系を用いた物体の認識と組み合わせれば、

“モモちゃん人形 を持ち上げて”

などの物体に付けられた固有名詞を用いた指示なども可能になる。さらに、ロボットにの動作の教示が可能であれば、

“このポーズは ガッツポーズ をだよ”

などとロボットの動作自体に名前付けを行うことも出来るなど、応用範囲が大きく広がる。

OOV を獲得する方法として、OOV を獲得するための特別なモードを設け、ユーザに 1 語ずつ登録してもらう方法がある。これは単純な方法であるが、ロボットの名称やユーザの名前など利用頻度が高く重要な語を間違いなく入力する手段として有効である。エンターテイメントロボット AIBO ERS-210 とそれ以降の同シリーズでは、ロボットの名前の登録のためのモードを用意した。登録後、ユーザは自分の好きな名前で AIBO を呼ぶことができるようになる。

しかし、特別なモードによる語彙の登録は確実である反面、特別な操作をユーザに要求する。また、登録される語彙の種類を増化させようとすると操作はさらに複雑にならざるをえない。

これらの問題を解決する方法として、音声認識の対象文中に埋め込まれた OOV の獲得が考えられる。例えば、

“私の名前は <OOV> です。”

という発話で <OOV> 部分の発音を獲得出来れば、私の名前というカテゴリと共に OOV を獲得することができる。同様に、

“お前の名前は <OOV> だよ。”
“このポーズは <OOV> だよ。”

など各種のカテゴリに属する OOV を自然に獲得することが出来る。

連続音声中の OOV を新規語彙として獲得するためには、

- OOV 区間の同定
- 認識の為の OOV の発音の獲得

が必要となる。獲得された OOV を表示あるいは合成音声などで利用するためには、

- 合成のための OOV の発音の獲得

がさらに必要となる。

音声認識のための発音の獲得では、OOV 区間に対応する認識単位の列、例えば音素列を求めれば、認識が可能となる。

ところが、合成音声のための発音の獲得では、認識用に獲得した音素系列を利用することも出来るが、汎用の Text-to-Speech などを利用するためには仮名レベルでの文字列があることが望ましい。また、汎用の Text-to-Speech との接続により、獲得した語へのアクセントの自動付与などの機能も期待できる。

SDR-4X では、上記 3 つの OOV 獲得機能を実装した。以下の節では、まず連続音声中の OOV の検出方法、および認識・合成のための OOV の発音の獲得について説明する。続いて、SDR-4X で実装された OOV の獲得性能についての実験結果とその考察を示し、問題点と今後の取り組みについて述べる。

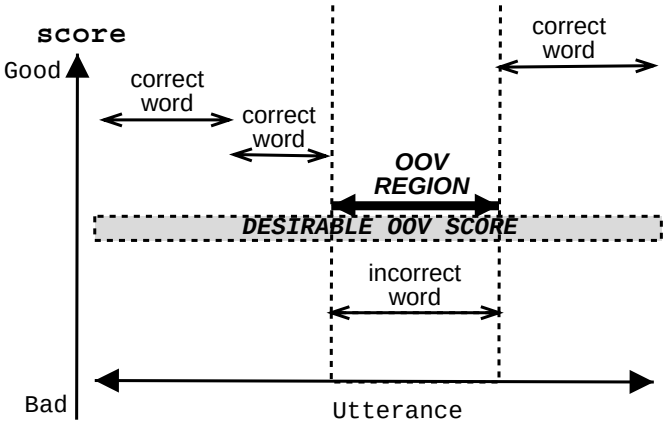


図 2: OOV 区間発見方法の模式図

2 連続音声中の OOV 区間の検出

2.1 OOV 区間のスコアリング

OOV の検出は正解単語に付与されるスコアと、不正解単語に付与されるスコアの差を利用して行なわれる。OOV 区間候補のスコアを正解単語のスコアより悪く不正解単語のスコアより良いスコアに調整することで、認識語彙に含まれず不正解となる区間を OOV 区間として認識処理の中で発見することができる。模式図を図 2 に示す。

音声認識中に OOV 区間の候補に対して、図 2 に示されるような適切なスコアを与えることが出来れば、不正解単語のかわりに OOV が単語候補として残る。

この OOV 区間のスコアを求める方法として、いわゆる garbage model を用いる方法 [2, 3]、サブワード系列を用いる方法 [4, 7, 8, 9] などが提案されている。

Garbage model を用いる方法では、音声全般に対して適度に良いスコアを出力する音響モデル (Garbage model) を用意し、この音響モデル の出力するスコアを OOV のスコアとして利用する。また、さらに簡単な実装では全音響モデルの出力するスコアの平均値などを利用することも出来る。

サブワード系列を用いる方法では、通常の音声認識以外に、音声にマッチする最適なサブワードの系列を求め、サブワードそれぞれに与えられたスコアから OOV 区間のスコアを求める。通常、サブワード系列から得られたスコアは連続単語認識から得られるスコアよりも良い傾向がある[2] 。このため、サブワード系列から得られたスコアにはヒューリスティックなペナルティーが付与される。

Garbage model を用いる方法は実装が容易で計算量も小さくてすむ反面、OOV 区間以外の情報が得られず、発音の獲得を別途行なう必要がある。一方サブワード系列を用いる方法は連続単語認識の枠組が必要であり計算量も

[2] サブワード系列の方が、語彙的制約が小さいためより良い音響的マッチングが行なえるため

大きくなるが、OOV 区間の同定と同時に対応する区間の発音の認識も行なうことが出来る。また、筆者らが予備的に行なった実験では、Garbage model を用いる OOV 区間推定よりも、サブワード系列を用いる OOV 区間推定の方が性能が高かった。このため、SDR-4X ではサブワード系列を用いる OOV のスコアリングを採用した。

2.2 システム構成

OOV 区間のスコアを効率良く求めるために [4] では、通常の連続単語認識と並列に音声に対応するサブワード系列の推定を行ない、その結果より得られるスコアを通常の音声認識で利用している。この方法には、通常の音声認識部分のサーチスペースの広がりとは独立に、一定の計算コストでサブワード系列とスコアが得られる利点がある。

[9] では文法の階層化とデコーダの見直しにより、通常の音声認識に連動して OOV 区間に対応する部分のサブワード系列の認識も行ない、上記のヒューリスティックに依存しない認識を実現しているが、実装の簡易さと、実行効率の点から筆者らは [4] と同様にサブワード系列のサーチが連続単語認識と並列する構成をとった。

図 3 に音声認識システムの概要を示す。連続単語認識のためのスタックサーチ型のデコーダと、OOV 区間のスコアと発音を推定するための時間同期型のサブワードデコーダーが並列で動作している。入力音声から求められた音響特徴量と音響モデルは共通して両デコーダーに供給される。言語モデル (LM) は異なる物を用いる。

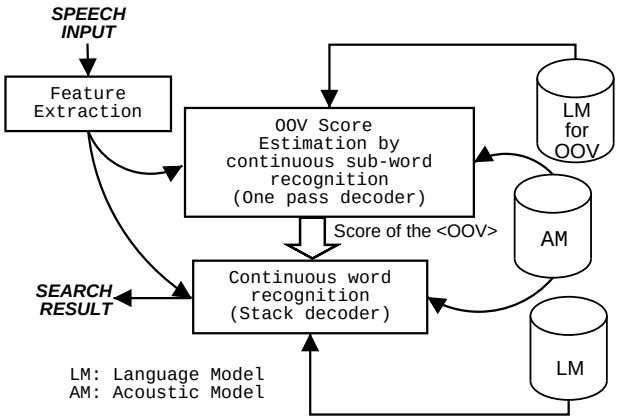


図 3: 音声認識システムの概要

2.3 OOV を含むサーチ

図 4 は認識時のサーチの進み方とスコアの供給関係を示している。サブワードのサーチは連続単語認識よりも時間的に先だって逐次進められ、連続単語認識のためのスタックサーチ部分では仮説を展開する際に通常の単語に対しては辞書から得られた音素系列を用いて Viterbi サーチを行ない、<OOV> に対してはサブワードのサーチの

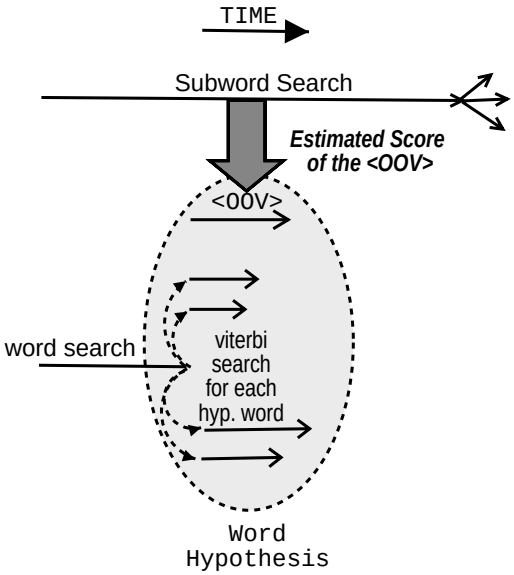


図 4: 二つのサーチの時間的関係と、スコアの関係

(途中) 結果から得られるスコアを用いる。

単語列のサーチの結果得られる単語列の単語境界の時刻と、サブワード列のサーチの結果得られるサブワード列の時間的境界は必ずしも一致しない。例えば認識結果が

```
word1 <OOV> word2
```

のとき、<OOV> と周辺単語の境界では図 5 に示すような単語境界とサブワード境界の不一致が発声する。この時、図の例では、$syl4, syl7$ は OOV に含まれるべきかどうか判断する必要がある。そこで、認識終了時に図 6 に示されるようなネットワークを作成して、入力音声に対して尤度が最大となる最適な単語とサブワードの系列を求めた。

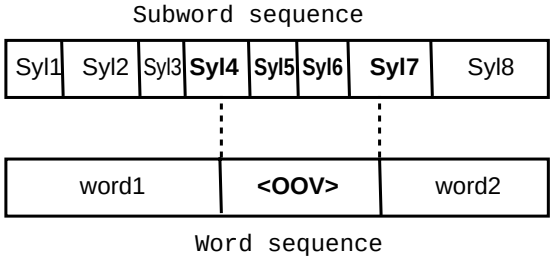


図 5: 単語境界とサブワード境界

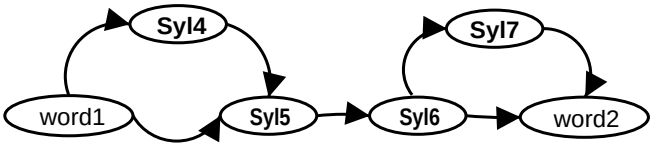


図 6: サブワード境界決定のためのネットワーク

表 1: tab:exp-condition

特徴量	・16bit 16KHz サンプリング ・フレーム周期 10msec ・フレーム長 25msec ・12 次の MFCC および、0 次から 12 次までの MFCC の 1 次回帰係数 (25 次元)
音響モデル	16mixture, 1000 tied-state HMM
言語モデル	サブワードトライグラム Cut-off trigram 5, bigram 5 Trigram を 25% に削減

3 認識・合成のための発音の獲得

サブワードの単位としては、音節 [10, 4, 11] や、モーラ (およびモーラ連鎖)[9]、などが用いられている。さらに、最適化されたサブワードユニットによる性能の向上も報告されている [7, 12]。今回はサブワード単位に、Text-to-Speech モジュールが必要とする仮名文字列へのマッピングが簡単でサブワードの語彙数が比較的少なくなる音節および音節の連鎖[3] を用いた。

また、音節系列の言語的制約として、音節トライグラムを用いた。音節単位のトライグラムはタスクに依存したトレーニングを行なうと高い性能を示すという報告 [11] があり、我々も予備実験で確認している[4] が、今回の目的は汎用の利用を念頭においたものであるので、大量の新聞コーパスで音節トライグラムの学習を行なった。

4 実験

4.1 サブワード系列のサーチ性能

12 名 (男女各 6 名) の旅行ドメイン[5] の 752 発話 (1 人あたり 63 または 62 発話) を用いて、サブワード系列のサーチ部分の性能を評価した。性能はサブワード認識精度

$$Acc = \frac{N_C - N_I}{N}$$

で示す。ここで、N, N_C, N_I はそれぞれ総サブワード数、正解の数、挿入誤りの数を示す。

実験条件を表 1 に示す。314 種類の音節および音節連鎖をサブワードとして用いた。言語モデルは日経新聞 6 年分のコーパスで学習したサブワードのトライグラムを用いた。また、トライグラムで消費されるメモリを削減するために、エントロピーに基づく逐次削減手法 [13] を用いてトライグラムサイズを 25% に削減した。デコーダには今回開発した時間同期型ワンパスビームサーチ [14] を用いた。

SDR-4X 上での CPU およびメモリの消費量を出来るだけ押えつつ、各ビームパラメータを調整した結果、サブワード認識精度は 72.1% を得た。この時 SDR-4X 上での Real-time-factor (RTF) は 0.2 であった。

[3] 簡単のために以降では単に音節と記述する
[4] 旅行会話のドメインで 95% を越える音節認識精度を得た
[5] ホテルのチェックインやレストランでの注文など

表 2: OOV を含む/含まない音声の単語認識精度

		文法	
		OOV なし	OOV あり
テスト音声	OOV なし	89.4	89.4
	OOV あり	88.5	91.5

4.2 OOV の獲得

前節で評価したサブワード系列のサーチを用いて OOV の検出実験を行なった。評価データとして 8 名 (男女各 4 名) が発話する二通りのコーパス (OOV を含まない文章１５文、OOV を含む文章 10 文) を用いた。OOV を含む文章では OOV として人名 (名字) を用い、同一の名字を含んだ文章を 5 文ずつ 10 文発話した。

発話音声の収録には作動中の SDR-4X(静かに立った状態) を用いた。このため、SDR-4X のオーディオ系の特性や、立った状態を維持するためのアクチュエーターのノイズも音声に含まれている。

言語モデルとして、語彙数 165 パープレキシティー 3.9 の有限状態文法を用いて認識した。その他の実験条件は表 1 に示した。前述したサブワード認識精度と同様な単語認識精度を算出して表 2 す。ここで、「OOV あり」の文法では、

(私 | ぼく | おれ)[の (名前 | 名)] は <OOV> (です | だよ | といいます)

などのルールを含むのに対し、「OOV なし」の文法では <OOV> の部分を、

(正解単語 1| 正解単語 2| 正解単語 3| 正解単語 4…)

で置換した。

「OOV あり」の文法では、OOV 部分に認識結果として「<OOV>」を出力し、単語認識精度計算の際には 1 単語としてカウントした。

「コーパスに OOV あり&文法に OOV あり」での OOV の検出率は 90% だったが、10% の誤り 8 文のうち「コーパスに OOV あり&文法に OOV なし」で正しく認識出来たにも関わらず OOV の検出を失敗していた文章は 3 文だけであった。この意味で正しく認識できる発話を OOV 検出すると検出率は 96% となる。また、false alarm は 0% だった。

OOV として検出された区間の平均サブワード認識精度は 48.5% だった。話者毎のサブワード認識精度を表 3 に示す。

図 7 に OOV 部分の認識結果の例を示した。

表 3: 話者毎の OOV 区間のサブワード認識精度

話者	f01	f02	f03	f04
	74.3	48.2	62.9	60.0
話者	m01	m02	m03	m04
	40.7	62.9	16.7	22.2

話者 f01		話者 m01	
正解	認識結果	正解	認識結果
クロサキ	クロタチ	ヒロエ	リノイ
クロサキ	オロサ	ヒロエ	リヲイ
クロサキ	ロサキ	ヒロエ	リヲエ
クロサキ	ロサキ	ヒロエ	ヒロイ
クロサキ	クロサキ	ヒロエ	イロイ
カズミ	カズミ	ミナミノ	イニナリノ
カズミ	カツニ	ミナミノ	イミナミノ
カズミ	カズミ	ミナミノ	ニナミノ
カズミ	カツミ	<OOV> 検出失敗	
カズミ	カスミ	<OOV> 検出失敗	

図 7: <OOV> 部分のサブワード列の認識結果と正解列

5 考察

false alarm は 0%、つまり OOV なしのテスト音声で、文法に “<OOV>” が含まれる・含まれないに関わらず認識率に変化がなかった。OOV の検出率と false alarm は、サーチ中に OOV の仮説のスコアに与えるペナルティーによってトレードオフされる関係にあるが、今回用いたペナルティーの値は、OOV を含む音声だけで最適化されているので、この false alarm は実験に用いた認識タスクでの実際の性能といえる。

この結果は、今回用いた辞書・文法が小規模なものであり、発話の前後関係から <OOV> の出現が強く予測できてしまうためと考えられる。OOV の検出率が 90% と非常に高いのも、文法の制約が大きく貢献していると考えられる。

文法的制約の緩和と OOV 検出率の劣化に関して定量的な検討は行っていないが、<OOV> を含む文法ルールや語彙のバリエーションを多くしていくと OOV の検出率が急速に劣化していく傾向がみらた。このため、現段階では文法的制約を大幅に緩めるのは困難であると予想される。

したがって、現段階でのロボットとの対話アプリケーションでは、文法の動的切替え等を用いて、語彙を獲得するべき状況で比較的小規模な OOV 獲得用文法が使われるように制御するなどの配慮が必要となる。

音節の認識精度 72.1% に対して、OOV 部分の音節の認識精度は 48.5% とかなり劣化している。認識結果を詳細に見ると、OOV 区間では文章全体に対して置換誤りと、挿入誤りの大幅な増化があった。置換誤りに関しては、音節トライグラムの学習に新聞コーパスを用いた点が妥当であったか、再検討する必要がある。挿入誤りに関しては、特に今回実験に用いた OOV は 3〜4 音節程度の短い単語であるので、OOV 区間の境界での 1 音節の挿入・削除誤りが認識に大きな影響を与えたと考えられる。また、表 3 に示された通り、OOV 部分の音節の認識精度は話者によって大きな違いが出た。話者毎に OOV として発話している語彙が異なるので語彙の影響とも考えられる。

獲得された発音の精度が平均的に低い語彙でも正解に近い語彙が得られる場合もある。例えば図 7 の「クロサキ」に対する「クロサキ」や、「ミナミノ」に対する「ニナミノ」などである。この事象を的確に捉えることが出来れば、OOV の獲得という目的を達成することができる。この問題を一種の学習問題として捉えれば、学習方法は大きく二つに分けることが出来る。

一方は「教師あり」学習である。音声対話のインターフェイスを用いてユーザに直接確認を行うことにより、悪い発音を排除して、ユーザの気に入った発音を直接獲得するすことができる。

もう一方の「教師なし」では、ユーザの教示なしに正解データを獲得する。[15] ではその方法の一つが示されている。ただし、OOV のトランスクリプションの平均的性能が非常に低い場合や、OOV の区間検出の誤りが発音の誤りとなって現れている場合など、注意して検討すべきであろう。

本稿では OOV の発音の獲得精度を、正解音節列との比較から定義した。しかし、自律ロボットが例えばユーザの名前を覚えて発音するようなアプリケーションを想定すると、必ずしも正解音節列は一つとは限らない。ロボットの持つテキスト音声合成から発せられた音が、ユーザにとって満足のいくものならば OOV の獲得は成功したといえる。この意味で、主観評価などを用いた性能の測定も重要と考えられる。

6 まとめと今後の課題

OOV の検出方法と発音の獲得について、SDR-4X での実現方法を中心に解説し、簡単な OOV 獲得タスクでの性能を調べた。語彙数 165 パープレキシティー 3.9 の文法を用いた場合 false alarm 0%、OOV の検出率は 90% であった。また、この時の OOV の音節認識精度は良い話者で 74.3% 悪い話者で 16.7%、平均 48.5% だった。この実験結果と、実際の利用のためのユーザーインターフェイスについて考察した。

更なる OOV 区間の検出精度向上のために認識単語の確信度の利用 [16, 17] や、OOV のトランスクリプションの精度向上のためにサブワードユニットの最適化 [7, 12] などに関して、我々のシステムでの有効性の確認を行ない

たい。

また、より汎用的な文法や N-gram 言語モデルでの獲得された OOV の利用および、OOV 獲得の性能に関してさらに検討を進めたい。

謝辞

本稿の執筆にあたって、貴重なコメントを数多く頂いた、ソニーデジタルクリーチャーズラボラトリの本田等氏に感謝します。

参考文献

[1] Lin Lawrance Chase. *Error-responsive Feedback mechanisms for speech recognizers.* PhD thesis, Carnegie Melon University, April 1997.

[2] J.G.Wilpon, L.R.Rabiner, C-H. Lee, and E.R. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. ASSP*, 38(11):1870–1878, 1990.

[3] Ayman Asadi, Richard Schwartz, and John Makhoul. Automatic detection of new words in a large vocabulary continuous speech recognition system. In *Proc. of ICASSP*, pages 125–128, Albuquerque, 1990.

[4] Atsuhiko Kai and Seiichi Nakagawa. 冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価. 電子情報通信学会論文誌, J80-D-II(10):2615–2625, Octber 1997.

[5] F. Kaplan and P-Y. Oudeyer. A method for teaching actions to an autonomous robot. In *Proc. of Sony Research Forum 2001*, December 2001.

[6] F. Kaplan. Talking AIBO: First experimentation of verbal interactions with an autonomous for-legged robot. In *Learning to Behave: Interacting agents CELE-TWENTE Workshop on Language Technology*, pages 57–63, October 2000.

[7] Dietrich Klakow, Greg Rose, and Xavier Aubert. OOV-detection in large vocabulary system using automatically defined word-fragments as fillers. In *Proc. of EUROSPEECH1999*, volume 1, pages 49–52, Budapest, September 1999.

[8] Issam Bazzi and James Glass. A multi-class approach for modelling out-of-vocaburary words. In *Proc. of ICSLP2002*, pages 1613–1616, Denver, September 2002.

[9] 小窪 浩明, 大西 茂彦, 山本 博史, and 菊井 玄一郎. サブワードモデルを用いた未登録語認識の効率的探索手法. 情報処理学会論文誌, 43(7):2082–2090, July 2002.

[10] 北 研二, 江原 暉将, and 森元 逞. 連続音声認識における未知語処理. In 日本音響学会講演論文集, pages 93–94, March 1991.

[11] Issam Bazzi and James R. Glass. Modeling out-of-vocabulary words for robust speech recognition. In *Proc. of ICSLP2000*, pages 433–436, Beijing, October 2000.

[12] Issam Bazzi and James Glass. Learning units for domain-independent out-of-vocabulary word modelling. In *Proc. of EUROSPEECH2001*, pages 61–64, Aalborg, September 2001.

[13] 踊堂憲道. 形態素単位の n-gram モデル の構築と圧縮に関する研究. Master's thesis, 奈良先端科学技術大学院大学, March 1998.

[14] Herman Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999.

[15] Helmut Lucke and Masanori Omote. Automatic word acquisition from continuous speech. In *Proc. of EUROSPEECH2001*, pages 2667–2670, Aalborg, September 2001.

[16] Aruna Bayya. Rejection in speech recognition systems with limited training. In *Proc. of ICLSP1998*, pages 572–575, Sydney, December 1998.

[17] Timothy J. Hazen and Issam Bazzi. A comparison and combination of methods for OOV word detection and word confidence scoring. In *Proc. of ICASSP2001*, pages 397–340, Salt Lake City, May 2001.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-4

A technology of intelligence on ASIMO and a system introduction

Yoshiaki Sakagami

Honda R&D Co.Ltd.
1-4-1 Chuo Wako-shi Saitama, Japan
sakagami@f.rd.honda.co.jp

Abstract

We present the intelligent function on ASIMO and integration of the system for autonomous robot.

In this paper, we introduce a technology of intelligence of ASIMO that used sub system of vision, auditory and external. The vision system and the auditory system are useful for interaction with human to obtain a timing of attention and lead visual intention. The planning agent handle with several type of data from sub systems concurently. We also discuss the external online database system that can be accessed using internet to retrieve desired information. We show an experimental results using intelligent function.

1.Introduction

Humanoid robots are especially desirable in human society as they can work well in indoor environments that have been designed for humans. Having a human form makes it easier for people to identify as compared to other forms. Humanoid robots like ASIMO can potentially assist humans in their daily tasks, bringing additional value to the human society.

Currently, however, humanoid robots are not skilled enough to perform many tasks that humans routinely do. This is because their sensory system is poor compared to humans. Robots have to process several types of raw sensory data such as orientation obtained by a gyro sensor, forces obtained by force sensors, image pixel data obtained by cameras and sound signals obtained by microphones. Then there is filtered intermediate level data such as velocity, acceleration, optical flow, edges and color. Finally there is high-level identifiable data like human face, gesture, posture, staircase, door, and natural language sentences.

We have developed a sensory system that processes multiple types and levels of data on ASIMO. We have also developed functions to understand human intentions to interact with them and to perform various tasks.

2.Related work

We have developed biped walking robots, humanoid robots like P1, P2, P3 and ASIMO over the past 16 years. Our initial goal was to realize a biped walk control to walk and turn in any direction, as well as for going up or down stairs in ordinary indoor human environments [1].

In this paper, we focus on the vision and auditory system for human interaction on ASIMO.

Various vision and auditory systems have been developed for robot intelligence. Applications vary from robust navigation in uncertain environments to identification of humans and interaction with them using gestures and voice. The humanoid robot Cog [2] realized humanlike intelligence by distributed computer systems outside the robot body. SDR-4X [3] is a small stand-alone humanoid entertainment robot. This robot has a vision system so as to recognize a human face, avoid obstacle and keep balance on slanted boad. Flo [4] is wheel type robot to serve elderly people, providing healthcare and other information related to activities of daily living. In this system, map based navigation as well as human interaction using voice, face detection and tracking were used. Robovie [5] is a wheel type robot for interacting with people by generating speech based on joint attention. This system is able to draw the person's attention to the same sensor information as the robot and omit words that are clear from the context.

There are various approaches for robot intelligence. Artificial intelligence techniques for simulating human thinking include symbolic processing, cognitive modeling based on brain mechanism, and emerging new logics such as in artificial life. Our approach is to model the relationship between sensory information and behavior directly.

In section 3, we describe the system hardware and software structure. In section 4, we describe the vision and auditory sensing system, navigation and human interaction. In section 5, we explain the planning behavior based architecture. Sections 6, 7, and 8 contain discussions on external database system, demonstrations and conclusions, respectively.

3.System structure

The current experimental model of ASIMO is a highly autonomous system compared to our first version that used external computation for planning and action selection by GUI. A vision and auditory system has been installed on ASIMO for navigation in ordinary environments and for

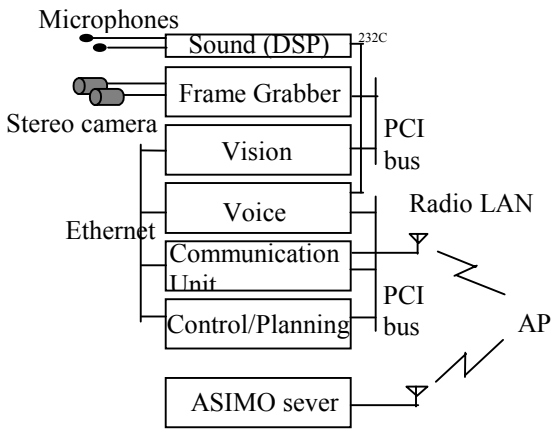


Figure 1: ASIMO computational structure

human interaction. The intelligence system consists of a frame grabber, a PC for image processing, a PC for speech recognition and synthesis, a processor for control and planning, a radio communication network controller unit for communication with the external system and DSP board for detecting sound sources (Figure 1).

Two board color cameras are installed on head unit to obtain stereo images which are processed to compute depth. A frame grabber connects the vision computer with the PCI bus for high speed data transfer. The vision system is separated from the processor group of control and planning. Two microphones for sound detection system are installed on the front side of head.

In the external system of ASIMO, we have developed a map management system for navigation and specification of tasks in man-made environments such as offices, museums, and hospitals. This system is able to send commands to the robot and allows selection of tasks for execution at specified locations (such as recognition and speech dialogue). There is also a database of face images of different people used to identify people and recognize them at subsequent meetings.

The experimental model of ASIMO uses several different operating systems, and a message board for asynchronous internal communication to achieve tasks and motions. The control system needs fast processing to perform actions. Vision processing is slower compared to control processing. The planning system has to handle both fast (e.g. obstacle avoidance) and slow situations (e.g. route trace moving).

A communication server manages socket port numbers for communication among vision and planning processes. The planning system is an agent-based distributed architecture system. The planning system is event driven with no central control to handle unforeseen situations. The construction of various functions and software of ASIMO is shown in figure 2.

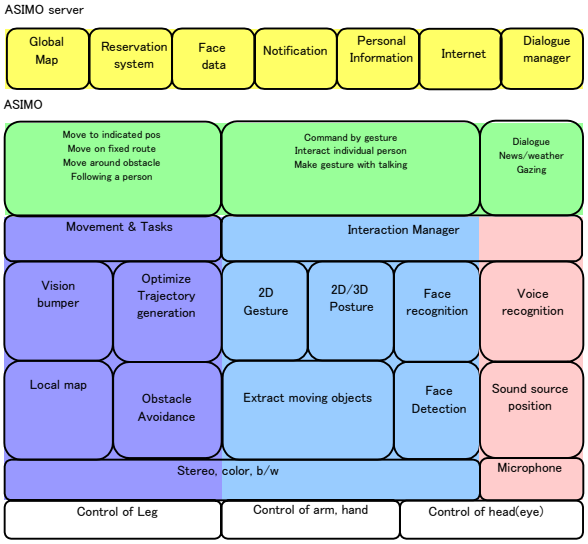


Figure 2: Function and software

4.Sensory system

4.1 Camera system and image capture

The vision system of ASIMO runs on a PC. The stereo method is based on SAD (Sum of Absolute values of Differences). It computes a depth map from two CCD cameras with b/w images, and does calibration for lens distortion and rectification (Table 2). The cameras in the head are located on top of the robot body with many degrees of freedom. To calculate camera pose, image capture is synchronized with all joint angles. The frame grabber has several types of I/O signals and captured images are synchronized with body motion (Table 3).

The vision system for navigation and interaction takes images from the frame grabber and processes them to extract 3D objects and moving objects (Figure 3).

Table 2: Specification of stereo system

Base line length	74mm
Imaging sensor	1/3"Color CCD x 2
Picture elements	768(H)x480(V)
Focal length	4mm
CPU	Mobile Pentium III-M 1.2GHz
Disparity image size	320(H) x 240(V)
Disparity range	32 pixels
Stereo frame rate	20fps
Software processing	Correction of lens distortion and rectification

Table 3: Specification of frame grabber

bus interface	CompactPCI
input signal	S-Video/RS-170A
input channel	2ch(S-VIDEO), 2ch(b/w)
Sampling	NTSC 4fsc(14.31818MHz)
sync.	VS, HD, VD
LUT	Y 8bit→16bit, CrCb16bit→16bit
frame memory	256Mbyte(90frame ring buffer)
sync. Pulse	0～16msec/field. Delay;5VTTL
async. Pulse	ON/OFF range;0～16msec
time code	sync pulse stamp on frame
Shutter	SONY XC-ES alternate
Transfer	64bit DMA
host OS	WindowsNT 4.0

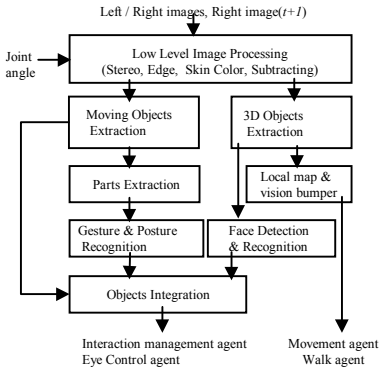


Figure 3: Vision system software configuration

4.2 Speech system and sound localization

For speech recognition and synthesis, we use a commercial voice recognition engine and speech synthesis product. However, the audio quality and intonation of voice need more work and they are not yet satisfactory for use on the robot. We control sentence and words tags for smooth speech. The sound and speech are important to get human intention under an ordinary environment. Sound source detection is able to identify human voice tones and step sounds. The shape of envelope of sound signal and the direction of the sound is computed from the volume and time difference of the signals at two microphones. The sound detection resolution is able to detect 1 degree by estimation (Figure 4). When someone calls ASIMO, it turns its head to face the person. When something falls on floor, it turns its head to gaze what happened.

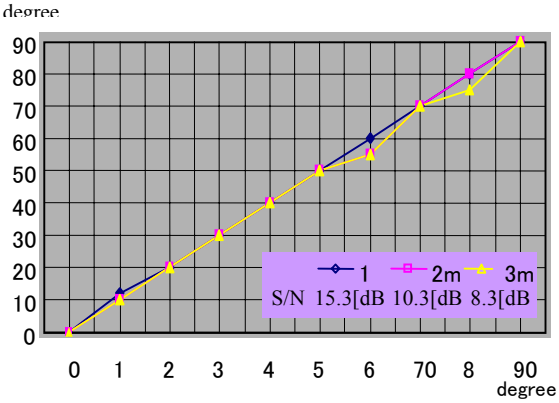


Figure 4: Sound direction
Horizontal axis: Real sound direction, Vertical axis: Estimated sound direction

4.3. Obstacle detection for navigation

Many autonomous robots perform navigation under an ordinary environment using vision, sonar, infrared sensor and range finders. ASIMO makes an in memory local map using vision. It reconstructs a map of the neighborhood surrounding the robot and uses it to move to the point of interest on a pre-defined route. The vision bumper handles obstacle surrounding the robot with simple 8 bit pattern from the local map data (Figure 5). Obstacles detected from stereo depth map data are updated frame by frame in

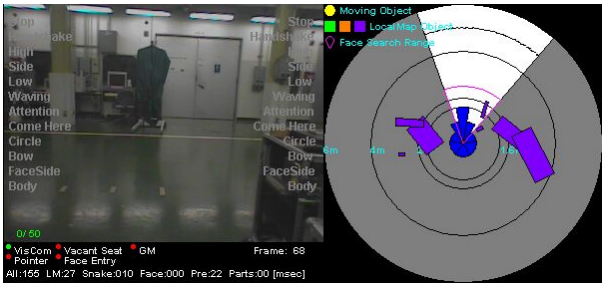


Figure 5: Image and top view of surroundings.
White region is robot's view. Purple rectangles are obstacle. Blue cone patterns are range of vision bumper.

time sequence. Obstacles are modeled using their bounding box. ASIMO can turn its head horizontally from +/-83 degrees, and can therefore make a wider local map covering +/-113 degrees.

4.4.Human and gesture recognition for interaction

Interaction is important for any kind of robot to perform tasks in human society like carrying luggage, pushing a cart, serving drinks, taking tools from the table and so on. The robot has to understand human's high-level task requirements by using its low-level sensory modules such as eyes, ears and tactile sensors .

Human tracking algorithm is able to track humans and their actions. An optical flow based algorithm extracts the foreground from the image even when robot head and body are in motion. Snake algorithm extracts contours of human shapes and can separate multiple people in the scene. Human head position is estimated at the top of the contour.

Face detection makes use of a model of skin color to extract face contours. Face recognition is based on the Eigenvector Method [6] (Figure 6).

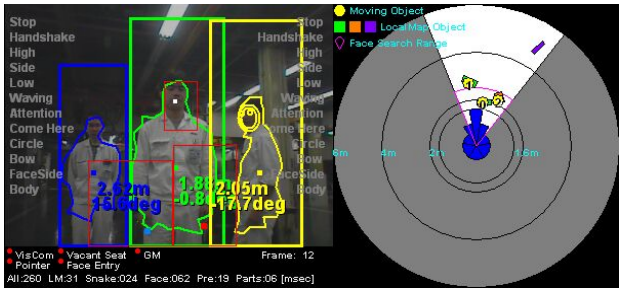


Figure 6: Extraction of humans in the scene.
Each color region is human position, point in contour is the human contour center. Numbers describe the direction and length of human from robot. Oval on face signifies recognized person.

Our 2D gesture recognition algorithm detects the position of the hand and estimates the action using a Bayes statistical model. It is able to identify hand, face and side and front profiles of body. Recognized gestures include handshake, hand circling, bye-bye, hand swing, high-hand and come here call. Human gesture computation is done using the robot front view image (Figure 7).

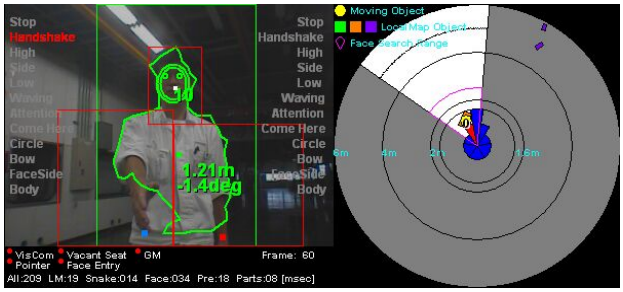


Figure 7: Handshake gesture recognition.

Our 3D gesture recognition algorithm can recognize pointed hand gesture based on the head and hand position relationship using depth map data from stereo (Figure 8).

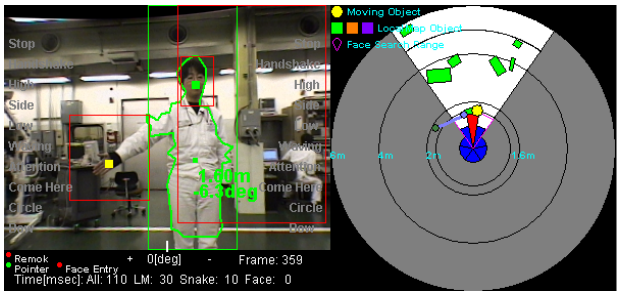


Figure 8: Pointing gesture recognition (blue bar).

5.Planning system

5.1 Software architecture

Many models of autonomous robot architecture for navigation and interaction have been proposed. RHINO [7] gave tours in museum for visitors. The system was organized in hierarchical modules to control a robot.

The ASIMO planning architecture is behavior-based and has deliberative and reactive layers. Behavior-based architecture [8] combines deliberative planning like high-level human commands and reactive behavior control for navigation in a rapidly changing environment. Behavior agents use distributed processing, are event driven, use asynchronous communication, and have no supervisor in each layer. The behavior agents in our system are listed in table 4.

Each agent handles data from the vision system, auditory system and other agents. The cycle time of deliberative agents is larger than that of reactive agents. For example, *Movement agent* makes a route every 500 ms. to avoid static obstacles. Static obstacles are detected by a potential method using a local map data that is updated by the vision system every 120 ms. Object avoidance in *Walk agent* has a 33 ms. cycle. *Walk agent* can stop and turn the robot when an obstacle suddenly appears. It uses the sensory behavior space generated by reward and penalty based learning on a simulator to decide motion direction. It is updated under real environment, if a robot was able to avoid obstacles or not(Figures 9,10,11).

Table 4: Functions of planning agents

1) Agents in Deliberative Layer

Movement Agent	Getting route data from Global map. Calculating direction towards sub goal and approach there. Obstacle avoidance by potential method. Issue command for task.
Interaction management agent	Switching scenario of task. Selecting a dialogue depending on task. Selecting an action using posture/gesture recognition result and voice recognition.
Dialogue Agent	Switching dialogue based on voice recognition. Switching a dialogue to speech with action Retrieving and making dialogue form Internet.

2) Agents in Reactive Layer

Walk Agent	Making a step command for walk depends on surrounding objects. Obstacle avoidance by vision bumper.
Eye control Agent	Attention a moving object. Attention and gazing sound. Gaze by command. Sleeping./ look around as no any input.
Sound source detection Agent	Detecting a position of sound source. Evaluation of human voice tone.
Robot control interface agent	Walk command, action command accepted from other agent. Broadcast a latest state of robot.

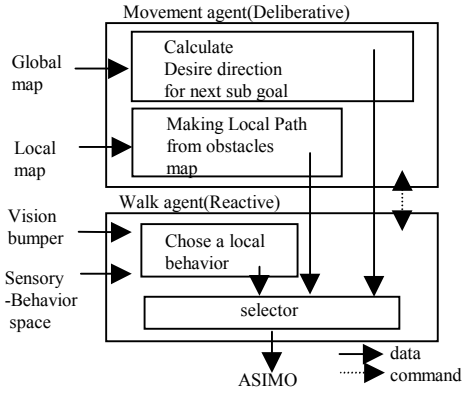


Figure 9: Obstacle avoidance mechanism

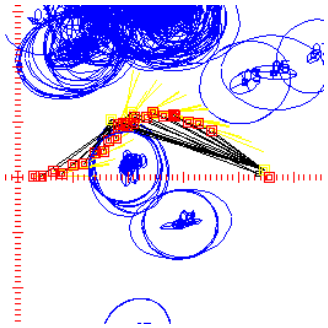


Figure 10: New route for obstacle avoidance. Red rectangles show generated local path. Blue oval are obsatcles.

Figure 11: Sensory-behavior space. Vertical dots; a case of vision bumper pattern (256). Horizontal dot;safe direction to move (5 degrees/dot). Black; not safe to move, White; safe to move. Gray; no data.

5.2 Integration

All agents run simultaneously and handle minimally necessary input data. Once the condition for agent invocation is satisfied, the agents become active and work on achieving their tasks. In case of conflicting behaviors, agents negotiate with each other by sending commands.

Agents receiving commands inhibit or modify their behaviors by updating parameters for acting behavior. For example, when *Interaction management agent* gets a command from a known person, it forwards the same command, frame number of image, and other information to *Dialogue agent*, *Move agent* and *Eye control agent*. *Dialogue agent* selects and executes appropriate speech synthesis. *Move agent* inhibits motion behavior toward sub goal. *Eye control agent* receives data of a moving object with frame number at every cycle and selects the object to gaze. The default eyes (head pan/tilt) moving are based on stimuli of processing of vision and auditory. When some stimuli received a human position data from a vision system, then gaze human face. When some stimuli received a human voice from an auditory system, then gaze a direction. A time of gazing depends on volume of sound and a kind of sound. When a pan/tilt command received simply from deliberative agents, then gaze that position.

For multi processor communication, we use a message board based communication in real time operating system and a socket based communication between real time system and non real time system. Vision and auditory systems are part of non real time system. A name server mechanism is used to manage port numbers to establish communication among agents automatically. Agents can start or terminate a process any time.

6.External system

6.1 Database of individual information

To recognize known people, the face recognition algorithm uses eigenvector algorithms on face and eye image data. Known people's face image database is kept on the ASIMO server. When the robot system initializes, the algorithm takes the stored eigenvector data from server. Other information like names, company names, address and hobbies is also available.

Processes in vision and planning system can refer to the external database via the network.

6.2 Reservation system

The example of task of ASIMO receptionist is to guide a visitor to a meeting room or a waiting place and to call a company representative to notify him about the visitor. The reservation system allows company people expecting visitors to choose meeting place and time and obtain visitor details in a web GUI. When a visitor approaches ASIMO, the moving human shape is extracted, and face is detected and recognized. Then the interaction manager confirms the schedule of meeting with the reservation system and sends a notification command to client program on user PC. Once the company user replies to the notification, ASIMO guides the visitor to the meeting place specified in the web GUI.

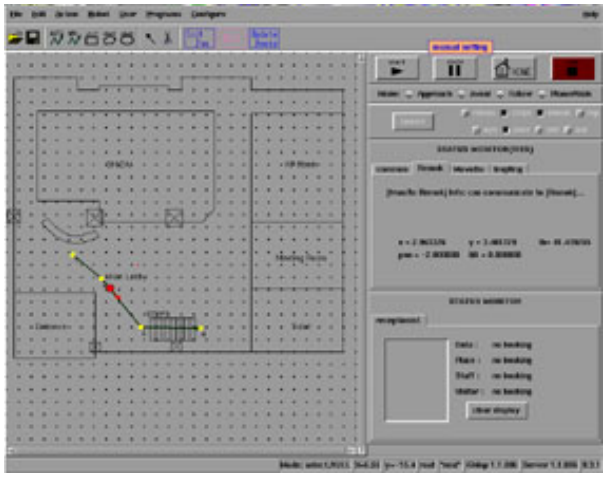


Figure 12: A GUI of Global Map system

6.3 Global map system

The Global map system GUI is shown in figure 12. It supports the following functions:

1) Loading, saving, updating and scaling a map data.
2) Playing an object at any position on map, changing its characteristics and manipulating it.
3) Making a route and sub-goal.
4) Selecting a route with minimum cost.
5) Switching a route from planning agents.
6) Setting a task (e.g. talking, action, image processing) at sub-goal.
7) Deciding a region for walking.
8) Sending a command (start, stop, home, pause) to planning.
9) Monitoring robot behavior state.

7. Demonstrations

The intelligent ASIMO can perform reception work autonomously in the following basic order (Video 1)

1) Detect an approaching visitor.
2) Find and track a visitor's face.
3) Verbally greet the visitor with a gesture.
4) Recognize the individual and confirm.
5) Check appointment meeting time, place and people.
6) Notify the visitor's arrival.
7) Guide visitor to meeting place.

The functions for navigation are as follows (Video 2):

1) Walking on a route specified in the GUI of Global map system, and performing pre-specified tasks at sub-goal.
2) Detecting obstacles.
3) Stopping if an obstacle is encountered.

The other interaction functions are as follows:

1) It is possible to call the robot. This is useful when calling person is far, and/or the voice command has too much noise. The robot first finds the direction of calling person by sound source detection and tries to find a call sign gesture. If a call sign is identified, the robot approaches the calling

person and follows him/her. The robot can also be stopped by gesture. (Video 3)
2) The robot knows handshake gesture of people. This is a good and basic action at first meeting.
3) Sometime, we want ASIMO to wait at the some location. 3D hand gesture can be used to point a location to the robot where it should go and wait. (Video 4)
ASIMO can inform a latest news source from pre-loaded file. It is also easy to retrieve from internet directory. (Figure 13, Video 5)

8. Conclusion

We have integrated autonomous functions for navigation and interaction in ASIMO. A stereo camera, frame grabber and a computer image processing have been added to the current rental model. Equipped with the vision system, the robot can not only navigate in an ordinary environment, but also understand human requirements. Our auditory system is useful for interacting with people and understanding commands from a distant location. The deliberative and reactive architecture can perform high level planning and rapid response. The vision, auditory, and planning systems also use information from the external database system.
In the future we plan to develop a more robust sensory and planning system to realize highly autonomous functions on ASIMO. Our goal is to demonstrate a personal robot that can perform helpful tasks to support human daily activities.

Figure 13: Inform a News/Topics

References

[1] Hirai, K.,Hirose, M., Haikawa, Y., Takenaka, T., ìThe Development of Honda Humanoid Robotî, Proc. of the 1998 IEEE International Conference on Robotics & Automation, pp.1321-1326, 1998.

[2] Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., Williamson, M., ìThe Cog Project: Building a Humanoid Robot, , Computation for Metaphors, Analogy and Agents, Vol. 1562 of Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998.

[3] Kuroki, Y., ìA Small Biped Entertainment Robot and its Attractive Applicationsî, Proc. of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, Plenary talk, 2002.

[4] Galtus, G., Fox, D., Gemperle, F., Goetz, J., Hirsch, T., Magaritis, D., Montemerolo, M., Pineau, J., Roy, N., Schulte, J., Thrun, S., ìTowards Personal Service Robots for the Elderly.î, Computer Science and Robotics, Carnegie Mellon University, 1998.

[5] Imai, M., Ono, T., Ishiguro, H., îPhysical Relation and Expression: Joint Attention for Human-Robot Interactionî, Proc. of 10th IEEE International Workshop on Robot and Human Communication, 2001.

[6] Turk, M., Pentland, A., ìEigenfaces for Recognitionî, Journal of Cognitive Neuroscience, Vol.3, No.1, pp.71-86, 1991.

[7] Burgard, W., Cremers, B. A., Fox, D., Hahnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S., ìExperiences with an interactive museum tour-guide robot.î, Artificial Intelligence 114, pp.3-55, 1999.

[8] Arkin, R.C., Behavior-based Robotics, MIT press, 1998.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-5

アクティブオーディションによる複数音源の定位・分離・認識

Speech Localization, Separation and Recognition by Active Audition for Humanoid

中臺 一博[†], 奥乃 博[*,†], 北野 宏明[†]
Kazuhiro Nakadai[†], Hiroshi G. Okuno[*,†] and Hiroaki Kitano[†]
[†] 科学技術振興事業団 ERATO 北野共生システムプロジェクト
[*] 京都大学大学院情報学研究科
[†]Kitano Symbiotic Systems Project, ERATO, JST, [*]Graduate School of Kyoto University
{nakadai, okuno, kitano}@symbio.jst.go.jp

Abstract

This paper presents active audition based real-time robot audition for speech localization, separation and recognition. The robot audition system consists of three sub-systems; real-time multiple human tracking system, sound source separation by active direction-pass filter (ADPF) and speech recognition for separated speech. The real-time multiple human tracking system localizes and tracks multiple people by stream based integration of auditory and visual processing such as sound localization by a pair of microphones, face recognition and localization and object localization by stereo vision. The ADPF separates sounds originating from directions obtained from the real-time multiple human tracking system by hypothetical reasoning on interaural phase difference and interaural intensity difference for each sub-band. The speech recognition uses multiple acoustic models trained by speaker and direction and integrates their results to recognize separated speeches. The whole system is implemented on a upper-torso humanoid and runs in real-time with 5 PCs distributed processing. Our experimental results show that the robot localizes, separates and recognizes simultaneous three voices properly and the efficiency of active audition such as turning to sound source and the active control of the filter sensitivity.

1 はじめに

近年, ヒューマノイドを代表とするロボットは AI のテストベッドなど研究目的としての利用にとどまらず, 人間と知的なソーシャルインタラクションを行い, 将来ロボットが “人間のパートナー” となることが期待されている. このような人間とのソーシャルインタラクションでは, 人間のコミュニケーションの多くが音声に依存していることから明らかなように, 聴覚は最も重要な機能の一つである. そうした状況では, ロボットは複数のイベントを同時に聞き分け, 他の音声や自分自身が作り出すモータノイズを抑制して, 混合音をうまく扱えるように音源分離機能を備える必要がある. この機能は, 音源分離をフロントエンド処理とした音声認識にも有効である. 音源分離問題については音環境理解 (*Computational Auditory Scene Analysis*, CASA) の分野で, 様々なアプローチが取り組まれてきた. しかし, 心理学的知見を利用した音源分離[11, 17], マイクロホンアレーを用いたビームフォーミング[3, 18], 独立成分分析 (*Independent Component Analysis*, ICA) [19, 9], 視聴覚統合による音源分離[16] など, そのほとんどはシミュレーション環境で行われており, 実環境・実時間処理への配慮があまりされていない. 実時間処理に関しては, 2 本のマイク間の強度差と位相差を利用した音声強調[2]が報告されているが, 環境が既知であること, マイクや音源が静止していることが前提であり, ロボットへの適用は難しい. また, ロボットでの音声認識については, ロボット自身が動作時に発するノイズ問題から, 動作中の音声認識は難しく, Sony AIBO のように “*stop-perceive-act*” 原理に従わざるを得ないか, 部屋やロボット自身の影響による音声の伝達歪みや他の音源からのノイズの影響で, ロボットではなく話者の口元にマイクを設置している.

アクティブオーディションはこれらの問題に対する解決の糸口を与える[14]. これは, 人間や動物を見習って, アクティブな動作を積極的に利用してロボット聴覚を向上させるもので, ロボット動作時に問題となるノイズをキャンセルすることにより, これまで, 実時間・実環境での音源定位・追跡、分離を報告してきた[14, 13, 15].

本稿では, 音源定位・分離だけでなく, 分離音の認識までを対象として, アクティブオーディションを利用したロボット聴覚システムを提案する. システムは, 視聴覚を統合し正確に複数人物を定位・追跡する実時間人物追跡システム, フィルタの通過帯域を音源方向に応じてアクティブに制御し, 分離精度を向上できるアク ティブ方向通過型フィルタ (*Active Direction-Pass Filter*, ADPF), 分離

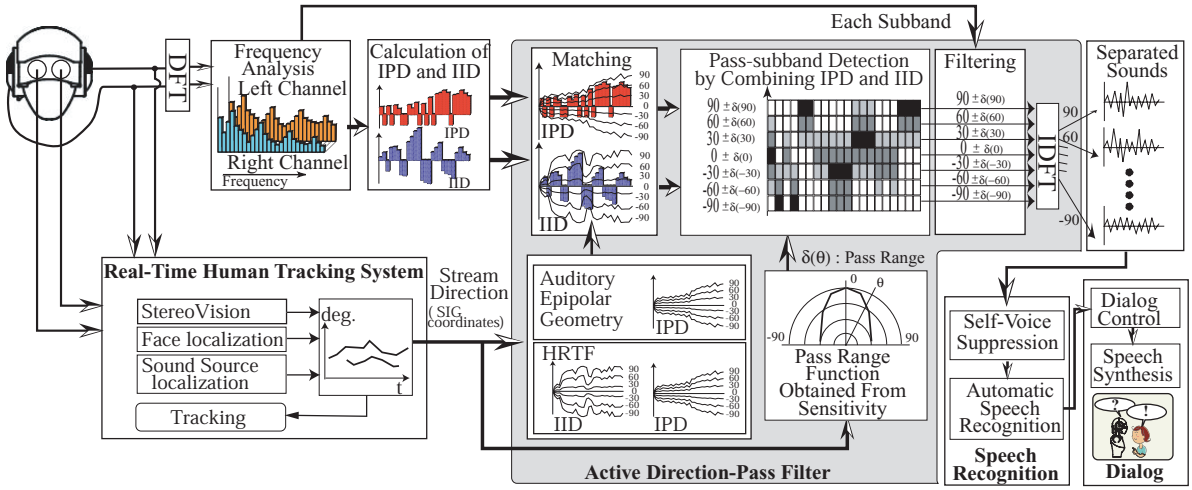


Figure 1: The Architecture of The Robot Audition System

音声に対し, 複数の音響モデルを用いた音声認識結果を統合して認識率を向上できる音声認識システムからなっている. また, 実際にシステムを 2 本のマイクを搭載したロボットに実装し, その評価を行う.

以降, 2 章ではロボット聴覚システムについて述べる. 3, 4, 5 章では, ロボット聴覚システムを構成する実時間複数人物追跡システムによる音源定位・追跡, アクティブ方向通過型フィルタによる音源分離, および複数の音響モデルを利用した 音声認識をそれぞれ説明する. 6 章でシステムを評価し, 7 章でまとめる.

2 ロボット聴覚システム

アクティブオーディションを利用したロボット聴覚システムを Fig. 1 に示す. システムは, ロボット (*SIG*) のカメラ, マイク入力から, 音源が複数ありかつ動作している場合でも, ロボット自身のアクティブな動作, 視聴覚の統合により, これらを定位・分離・認識することが可能である. システムは大きく 3 つのサブシステム「視聴覚を統合した実時間複数人物追跡」, 「アクティブ方向通過型フィルタ (以後, ADPF) による音源分離」, 「複数の音響モデルを使用した音声認識」からなる. 以下にヒューマノイド *SIG* を紹介し, 次章以降で各サブシステムを説明する.

2.1 ヒューマノイドプラットホーム: SIG

研究のテストベッドとして, 上半身のヒューマノイド *SIG* を使用している. *SIG* は 4 自由度を有し, 各モータには, ポテンショメータによって, 位置制御, 速度制御が可能な DC モータを用いている. また, 音響的にロボットの内外を区別できるよう設計された FRP 製の外装を備えている. カメラには, 左右の目の位置に一組の CCD カメラ (Sony EVI-G20) を, マイクには, 計 4 本の無指向性マイク (Sony ECM-77S) を使用している. 4 本のマイクは, 外装を挟んで一組ずつ取り付けられており, 内部に設置されている一組は, 主にロボット自身のモータによって発生する内部ノイズをキャンセルするために使用している [14]. 音源定位・分離には, 外界からの音響信号を収音するよう *SIG* の左右の耳部に設置されているもう一組のマイクを利用している.

3 視聴覚統合による実時間複数人物追跡

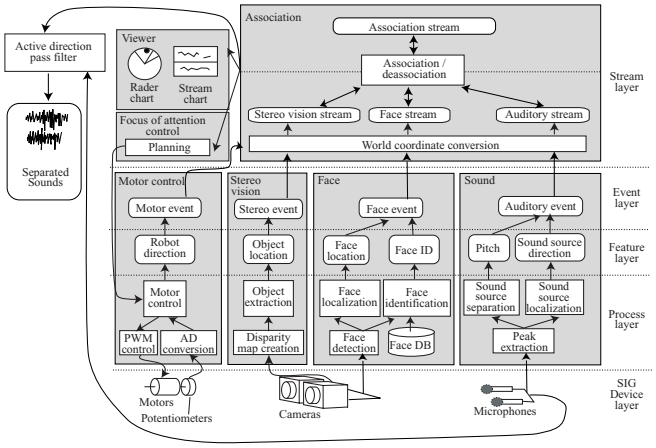


Figure 2: Architecture of Real-Time Human Tracking System

Fig.1 の実時間複数人物追跡システムの詳細な構成を Fig.2 に示す. このシステムは, *SIG* のカメラやマイクなどから得られるセンサ情報を統合して, 複数の人物の位置を把握し, これを追跡することができる. 本稿では, このシステムで認識される正確でロバストな音源方向情報をアクティブ方向通過型フィルタへの入力としている. システムは, **音源定位**, **顔認識・定位**, **ステレオビジョン**, **アソシエーション**, **アテンション制御**, **モータ制御**, **ビューワ** の 7 モジュールから構成されている. なお, 以降ではモジュール名はボールド体で表す.

モジュール内のサブモジュールや情報は, 5 階層に分けられる. *SIG* のカメラ, マイク, モータシステムなどのセンサデバイスは *SIG* デバイス層に属している.

プロセス層, 特徴層, イベント層では, *SIG* デバイス層から得られるセンサ情報から, 位置, 名前情報といった特徴を抽出し, これらに時間情報を付加して, 特徴の種類に拠らない透過的な表現であるイベントに変換し, ストリーム層に出力する. 各センサ情報の観測タイミングは非同期であるため, イベントも非同期に発生する. これらの処理は, **音源定位**, **顔認識・定位**, **ステレオビジョン**, **モータ制御**といったイベント抽出モジュールで行われる. 具体的には, **音源定位**では, マイク入力から, ピッチ (音高) を抽出し, 倍音の調波構造を利用したグルーピングを行い, 複数の音源が同時に存在する場合でも, それらを定位し, 音イベントを生成する. **顔認識・定位**では, 顔画像の認識と定位を行い, 顔の ID と位置情報を含んだ顔イベントを生成する. 顔認識では, まず, 肌色抽出と相関演算に基づくパターンマッチングにより, 顔領域抽出を行う[7]. 次に, 抽出顔領域に対し オンデマンド更新が可能で最適な判別空間を生成できるオンライン判別分析 (LDA)[8] を適用して顔認識を行う. 顔定位は, 抽出した顔の大きさに一定の仮定をおくことで, 顔の 3 次元定位を行う. **ステレオビジョン**は, 高速な視差マップ生成法[10]を用いて, 人物のように縦に長い物体を抽出・定位し, 人物位置情報を含んだステレオイベントを生成する. 横を向くなど顔が見えない場合にも人物の位置情報を得ることができるため, システムのロバスト性を向上することができる. **モータ制御**では, モータのポテンショメータから得られるロボットの姿勢情報をもとに, モータイベントを生成する.

ストリーム層では, イベントを種類ごとに時間方向に接続し, ストリームを形成する. ストリームとイベントの接続には, Kalman フィルタを用いて, 観測誤差・処理誤差にロバストな処理を実現している[15]. このようにして生成されたストリームうち, 複数のストリームが同じ人物に由来すると判断すると, これらを一つに束ねアソシエーションストリームを生成することにより, センサ情報の統合を行っている. これらの処理は, **アソシエーション**で行われる. アクティブ方向通過型フィルタの入力は, 音ストリーム, および音情報を含んでいるアソシエーションストリームの方向情報である. 送出されるストリーム方向情報には, 時間情報およびストリーム ID が含まれているため, データの同期および複数音源分離が可能である.

また, **注意制御**はストリームの状態に応じて *SIG* の動作を決定し, **ビューワ**はストリームの状態を, レーダチャート, ストリームチャートとして可視化するためのモジュールである. **モータ制御**は, **注意制御**からの信号をもとに, PWM(Pulse Width Modulation) 信号を生成し, DC モータを駆動するためにも使われている.

実装上は, これらのモジュールをギガビットイーサ, ファストイーサの 2 つのインタフェースを備えた 5 台の Linux ノード (Pentium III 1GHz) に分散させている. ギガビットイーサは, トラフィックが多く, 通信量も大きいモジュール間通信用に, ファストイーサは, 同期信号などの通信用に使い分けている. 結果として, 200 ms のレイテンシ, および 100 μs 以下の精度のノード間同期を実現している. レイテンシについては, ストリーム生成で用いた Kalman フィルタを予測に利用することで補っており, これによりリアルタイム動作を可能としている.

4 アクティブ方向通過型フィルタによる音源分離

Fig. 1 の網掛け部分がアクティブ方向通過型フィルタの構成に対応する. アクティブ方向通過型フィルタへの入力は 4 つあり, 入力のスペクトル, 入力スペクトルから計算される IPD と IID, および, 実時間人物追跡システムから得られる音源方向情報である. 出力は, 入力方向に対する分離音響信号である.

アクティブ方向通過型フィルタでは, 方向通過型フィルタに対し, 聴覚中心窩に基づくアクティブな通過帯域制御とロボットの伝達関数を利用した仮説生成により, 実環境での高速な音源抽出を可能にしている. ここで, ロボットの伝達関数は, 部屋の伝達関数, ロボット頭部による音の歪みなどを考慮して, 特定方向の IPD および IID を推定するための関数である. 以下では, アクティブ方向通過型フィルタのアルゴリズムの詳細について説明する.

4.1 アクティブ方向通過型フィルタのアルゴリズム

アクティブ方向通過型フィルタのアルゴリズムは以下の 6 ステップで構成される.

1. 入力音のスペクトルから, 各サブバンドの IPD $\Delta\varphi'$ と IID $\Delta\rho'$ を計算する. ここで, Sp_l, Sp_r は, それぞれある時刻に左右のマイク入力信号から得られたスペクトルである.

$$\Delta\varphi' = \arctan\left(\frac{\Im[Sp_l]}{\Re[Sp_l]}\right) - \arctan\left(\frac{\Im[Sp_r]}{\Re[Sp_r]}\right) \quad (1)$$

$$\Delta\rho' = 20\log_{10}\left(\frac{|Sp_l|}{|Sp_r|}\right) \quad (2)$$

2. 抽出すべき音源の方向を θ_s とする. θ_s は 3 節で述べる実時間人物追跡システムから, ロボット座標系での水平角として得られる.

3. 通過帯域関数に従って, θ_s に対応するアクティブ方向通過型フィルタの通過帯域 $\delta(\theta_s)$ が選択される. 通過帯域関数は, 聴覚中心窩に基づき, ロボットの正面方向で最小となり, 周辺部で大きな値をとる関数である. 詳細は 4.2.2 節で述べる. 選択された通過帯域

$\delta(\theta_s)$ を用いて, $\theta_l = \theta_s - \delta(\theta_s)$, $\theta_h = \theta_s + \delta(\theta_s)$ と定義すると, θ_l から θ_h の範囲にある音響信号を抽出するのがアクティブ方向通過型フィルタの基本的な動作である.

4. θ_l と θ_h に対する IPD, IID を推定する. これらの推定には, ロボットの伝達関数を利用する.

5. 音源方向 θ に対して, ロボットの伝達関数を利用して, 入力スペクトルから以下の条件を満たすサブバンドを選択する.

$$f < f_{th}: \quad \Delta\varphi_E(\theta_l) \le \Delta\varphi' \le \Delta\varphi_E(\theta_h),$$
$$f \ge f_{th}: \quad \Delta\rho_H(\theta_l) \le \Delta\rho' \le \Delta\rho_H(\theta_h)$$

$\Delta\varphi_E(\theta)$, $\Delta\rho_H(\theta)$ は, それぞれロボットの伝達関数から推定される IPD, IID である. f_{th} は, フィルタリングの判断基準に IPD と IID のどちらを用いるかを決める閾値である. 一般に, 低周波数域では IPD, 高周波数域では IID が大きく影響し, この閾値はマイク間距離に依存する. 我々のロボットでは, 理論的にも, 実験的にも f_{th} として 1500 Hz が妥当であることが報告されている[15].

6. 選択されたサブバンドから, 音響信号を再合成し, 該当範囲にある音響信号を抽出する.

実際には, 音源方向 θ_s は時間 t の関数であるため, 特定音源を抽出し続ける際には, 時間方向の連続性を考慮する必要がある. 本稿では, 3 節に述べる実時間人物追跡システムから音源方向を得ることでこれを解決している. 実時間人物追跡システムでは, すべての情報をストリームという時間的な流れを考慮した表現を用いて表しているので, 同時に複数の音源が存在したり, 音源や自分自身が移動する場合でも, 一つのストリームに注目することによって, 特定音源からの方向情報を連続的に得ることができる. また, ストリームは視聴覚情報を統合するためにも使用しており, これにより, 視覚情報による音源定位精度向上を実現している.

4.2 聴覚中心窩による通過帯域制御

4.2.1 聴覚中心窩とは

霊長類の視覚は, 中心窩と呼ばれる解像度が高い部分が中心部に存在し, 周辺部では解像度が低くなる代わりに, 広範囲な視野を得ている. このような構造を用いれば, 対象物を中心窩で捕らえることにより, 高解像度の情報を取得することができる. つまり, 広い視野と高い解像度を併せ持ち, かつ脳の情報処理量を劇的に削減できる効率的な構造を有している. ロボットでも, 同様の構造により計算量を削減できることから, 中心窩を利用した視覚処理はアクティブビジョン (*Active Vision*) [1]の典型的な例として, しばしば利用されている[12, 22].

人間の聴覚においても, 水平方向の音源定位の精度は正面方向で最も高く, 周辺部に行くに従い低くなることは, 古くから知られている[5]. 耳に 2 つのマイクを備えたロボットによる音源定位でも, 人間と同様の傾向が見られる. Fig. 3 は, 3 節で説明した 実時間人物追跡システム[13]における 3 つの定位モジュール**音源定位**, **顔定位**, **ステレオ物体定位**による定位結果の平均値, Fig. 4 は, **音源定位**による定位結果の分布を表している.

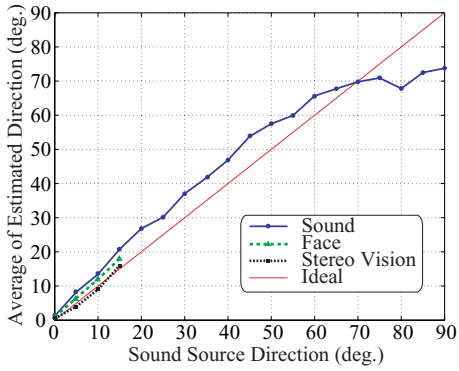


Figure 3: Localization by Face, Stereo Vision and Sound

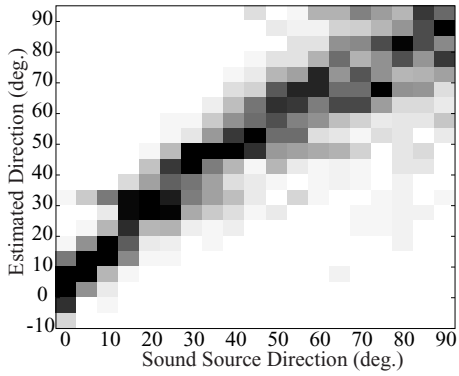


Figure 4: Distribution of Sound localization

Fig. 3 から, **音源定位**による定位誤差は, 正面方向から 20° 付近まで増加した後, 70° 付近までは 6° 程度で一定だが, それ以降は大きく悪化し, 90° では, 15° 以上になる. また, Fig. 4 から, 正面方向のばらつきは少なく, 正面から離れるにつれ, ばらつきが目立ち, 分散が大きくなる. このように定位結果の平均, 分散は, ともに正面方向で音源定位の精度が高くなることを示しているので, 本現象をロボットにおける聴覚中心窩と呼ぶ.

なお, 神経行動学 (neuroethology) では, ドップラー効果によるエコー音の周波数変化を抽出するため, キクガシラコウモリの蝸牛殻で特定の周波数に対する感度が高くなっている部分を聴覚中心窩と呼んでいる [23]. 選択的注意という広義の意味では, 両者は似ているが, 本稿では, ロボット頭部の正面方向で感度が高いという意味で聴覚中心窩という言葉を使用する.

Fig. 3 では, ステレオビジョンによる定位誤差は 1°, 顔

定位による誤差は 2° 程度と, 聴覚処理よりも正確であることがわかる. これは, 音源方向が正面に近く, 視覚情報が利用できる場合では, 高精度の視覚情報によって, 聴覚の精度不足を補うことが可能であることを示している.

これらから, 音源定位では, 視覚の中心窩と同様に, 聴覚中心窩を利用して音源に正対するようなアクティブな動作を行えばシステムの精度の向上が期待できる. さらに正面方向で視覚情報が利用できれば, 視聴覚統合によりシステムのロバスト性を向上できると考えられる.

4.2.2 通過帯域制御

方向情報を利用した音源の分離抽出を考えた場合, 正面方向の音源であれば, 正確な音源方向を利用することができるが, 音源方向が正面から離れるにつれ, 方向情報に精度を期待できなくなるため, 音源方向によってフィルタの通過帯域を制御する必要がある.

従来の方向通過型フィルタ[20]はスペクトルの各サブバンドで, 両耳間位相差 (*Interaural Phase Difference*, IPD) と 両耳間強度差 (*Interaural Intensity Difference*, IID) に対する仮説推論を行うことによって特定方向の音を抽出するものであるが, フィルタの通過帯域が音源方向によらず一定であることが, 十分な精度が得られない一因であった.

そこで, アクティブ方向通過型フィルタでは最適な通過帯域を求めるために, 音源数 1 の場合に音源方向や通過帯域を様々に変化させて, 抽出精度の違いを調べた. 音源には, スピーカから出力される音声信号を用いた. スピーカとロボットの距離 は 1m とし, スピーカの水平方向を, ロボットの正面から, 0°〜90° まで 10° おきに変化させた. また, 音源を抽出する際には, スピーカ方向は既知であるものとし, 方向通過フィルタの通過帯域を ±5°〜±90° まで ±5° 単位で変化させて音源を抽出し, S/N 比による比較を行った.

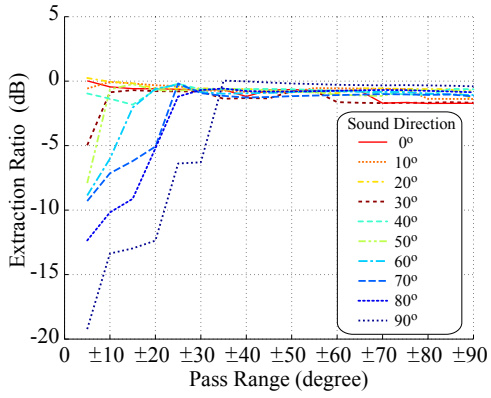


Figure 5: Extraciton of Single Sound Source

Fig. 5 に結果を示す. 実験では, 背景雑音は無視できる程度に小さかったため, 音源数が 1 の場合は, S/N 比 が $0\,dB$ となった時に, 元波形が完全に抽出できたと解釈する. 音源方向が 0°〜30° と正面方向に近い場合には, 通過帯域が ±10° 程度で元波形を抽出できているが, 音源方向が正面から離れるに従い, 元の波形に含まれるパワーを抽出するために, 広い通過帯域を必要とし, 音源方向が 90° の場合には, 最低でも ±35° 程度の通過帯域が必要である.

音源数が 1 の場合には, 通過帯域が広ければ広いほど, S/N 比の高い信号を抽出することができるが, 実環境では, 背景雑音を含め, 複数の音源を考慮する必要があるため, なるべく通過帯域を狭くとることが望ましい. そこで, Fig. 5 から, ほぼ元波形が抽出でき, かつ極力狭い通過帯域を音源方向ごとに抽出し, Fig. 6 のように通過帯域関数を導出した. 通過帯域は正面方向では狭く, 周辺部では広くなっていることがわかる. これは, 音源定位と同様に, 音源分離でも聴覚中心窩を利用することが可能であることを示している. アクティブ方向通過型フィルタでは, このような通過帯域制御を行って, 正面方向では S/N 比の高い音響信号を抽出し, 正面方向から離れた音源に対しては帯域を広く取り, 背景雑音の混入により S/N 比は多少落ちるものの, 必要な情報をできるだけ抑制せずに, 特定の音源の強調を行う. 正面方向から離れた音源を精度よく抽出する必要がある場合は, 聴覚中心窩を利用できるように, 音源方向を向くような制御を行う.

実際の利用では, 他の音源の音を極力抽出したくない場合, 単なる音響信号の強調として利用したい場合など, 状況に応じたチューニングが必要な場合もあると考えられるが, 以後の実験では, Fig. 6 に示された通過帯域関数を利用するものとする.

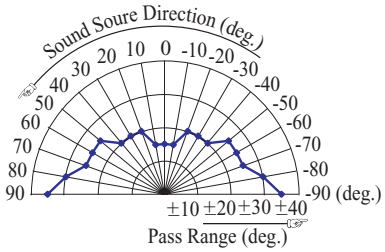


Figure 6: Pass Range Function

4.3 ロボットの伝達関数

一般に, ロボットの伝達関数は計測によって求める. 本稿では, 無響室で水平方向について 5° 刻みでインパルス応答を計測して得られた計測伝達関数を用いている.

しかし, 部屋の音響環境の変化に動的に対応するためには, 音響環境が変わるたびに部屋の伝達関数, もしくはロボットの伝達関数の再計測が必要となるなど残響や動的な音響環境の変化に追従させることが難しい. また, 各方向からの測定が必要であるため, 測定にも時間がかかるといった欠点を抱えている.

そこで, IPD については, 水平角から計算的に IPD を推定する手法である聴覚用エピポーラ幾何 (*Auditory Epipo-*

lar Geometry)[14] を利用している．これは, ステレオビジョンで利用されるエピポーラ幾何[6]と同様の概念を 2 本のマイクによる定位に当てはめたものである．音源とロボット間の距離が 50 cm 以上では無限遠の音源を仮定できるので[15], 頭部形状による影響を考慮すると, 式 (3) として表すことができる．

$$\Delta\varphi = \frac{2\pi f}{v} \times r\,(\theta + \sin\theta) \qquad (3)$$

ここで, $\Delta\varphi$, θ, f, v は, それぞれ IPD, 音源方向, 周波数, 音速を示す．また, r はロボット頭部を球形とみなした場合の半径である．

最終的に, ロボットの伝達関数として IID については, 計測した伝達関数, IPD については, 式 (3) を用いている．IID について計測によらない手法が望まれるが, これは今後の課題である．

5 分離音の音声認識

音声認識の分野では, マルチコンディショニングやミッシングデータなどノイズにロバストな音声認識へのアプローチが行われている[4, 21]．しかし, これらは S/N 比が小さい場合は有効ではない．このような場合には音声認識のフロントエンドとして音源分離が必要である．また, S/N 比が大きい場合も有効である．フロントエンドとして ADPF を使用し, 複数の音響モデルを使った音声認識を提案する．

5.1 音響モデル

音声認識エンジンには, 京大で開発された “Julian”[24]を利用している．本稿では, 音声データは, 男性 2 名, 女性 1 名の計 3 名 の発話による 色, 数字, 食べ物といった 150 語を使用している．

音響モデル用の音声データとして, まず, 3m×3m の部屋で, *SIG* から 1m の距離にスピーカを置き, その音を *SIG* のマイクで録音した．スピーカは, *SIG* から $0, \pm 60°$ の位置におき, それぞれの方向について, すべてのデータを録音した．また, $0, \pm 60°$ の 2 箇所から同時に音声を出力する場合, 3 箇所から同時に音声を出力する場合についてもすべての組合せについて録音を行った．次に, 音源方向を既知として ADPF による音声抽出を行った．抽出した音声を話者, 発話方向ごとに整理し, 音響モデルのトレーニングセットとした．音響モデルにはトライフォンを用い, 各トレーニングセットごとに, Hidden Markov Model Toolkit (HTK) を用いて作成した．したがって, 本稿では, 3 話者, 3 方向の組合せで 9 種類の音響モデルを使用している．

5.2 複数の音響モデルを利用した音声認識

音声認識では, 並列に 9 つの音声認識プロセスが実行される．各音声認識は, Fig. 7 に示されるようにそれぞれ異なる音響モデルを用いる．セレクタはすべての音声認識結果を統合し, 最も信頼性が高いと判断される結果を出力する．

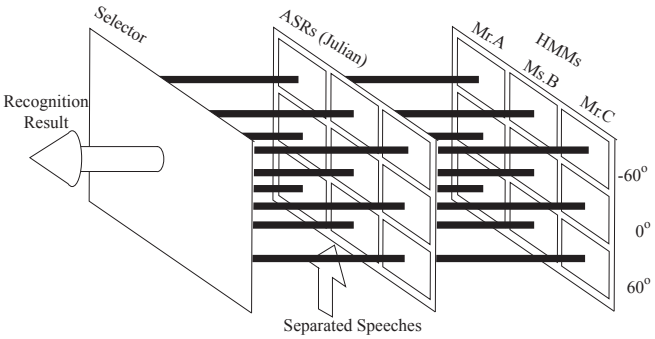


Figure 7: Speech Recognition by Multiple Acoustic Models

統合のアルゴリズムを定義するために, 特定話者の音響モデルに対する単語認識率を調べた．Fig. 8 に示した結果から, 話者よりも方向の違いによる認識率の低下が少ないことがわかる．また, 話者も方向もあっている場合は 80% 以上の認識率であることがわかる．この結果を踏まえ, 音声認識の際には, 音源方向は既知であることを利用し, セレクタは 式 (4) に示すコスト関数を統合のために使用している．

$$V(p_e) = \left(\sum_d r(p_e, d) \cdot v(p_e, d) + \sum_p r(p, d_e) \cdot v(p, d_e) - r(p_e, d_e) \right) \cdot P_v(p_e). \qquad (4)$$

$$v(p, d) = \begin{cases} 1 & if \quad Res(p, d) = Res(p_e, d_e), \\ 0 & if \quad Res(p, d) \neq Res(p_e, d_e). \end{cases}$$

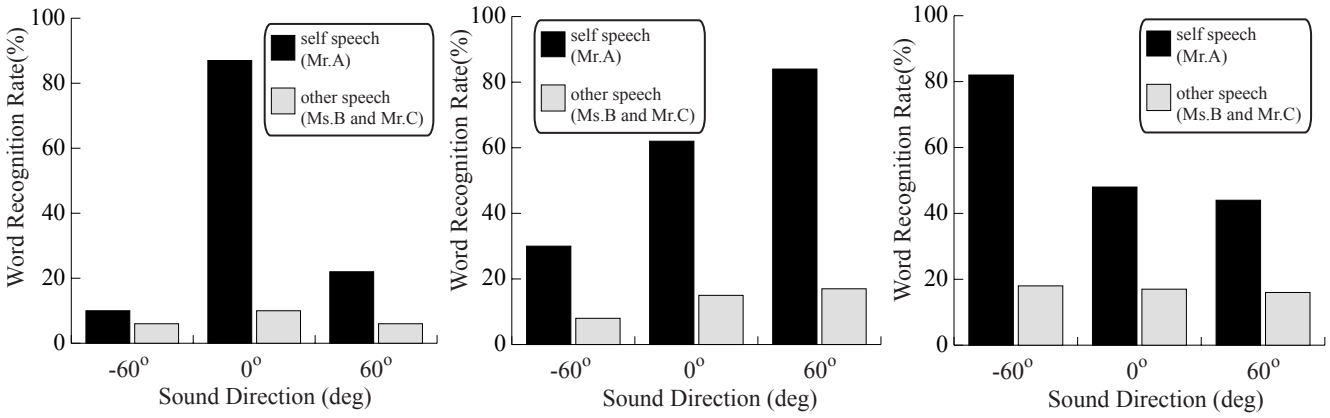
ここで $r(p, d)$, $Res(p, d)$ は, 話者 p, 方向 d の音響モデルを使用した場合の 単語認識率と入力音声に対する認識結果を示している．また, d_e は実時間人物追跡システムから得られた音源方向であり, p_e は, 評価対象の人物である．$P_v(p_e)$ は顔認識モジュールで生成される確率であり, 顔認識ができない場合は, 常に 1.0 となる．最終的に, セレクタは最も大きな $V(p_e)$ を持つ人物 p_e と認識結果 $Res(p_e, d_e)$ を出力する．

$V(p_e)$ の最大値が 1.0 以下もしくは, 2 番目に大きい値と近い場合は, *SIG* は認識が失敗もしくは, 一つの候補に絞りきれなかったと判断して, 音源方向を向き該当の人物に再度尋ねなおす．このように, 複数の音響モデルを利用して, 分離音と話者の認識を行う．また, 顔認識が利用可能であれば, 人物名がわかるためロバスト性を向上できる．

6 実験と評価

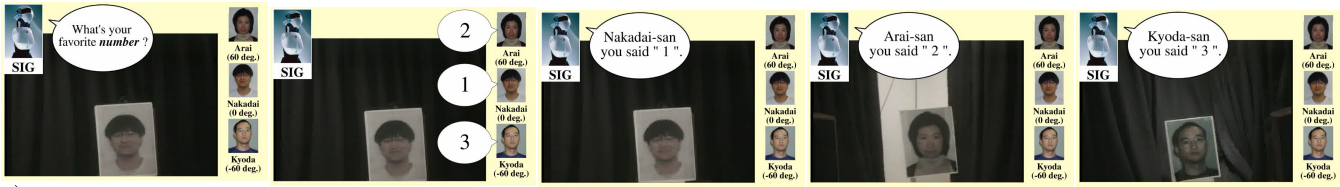
同時 3 話者発話のシナリオを通じてロボット聴覚システムを評価した．シナリオの内容を以下に示す．

1. ロボットから 1m の距離に 60 度間隔 (*SIG* から見て, $0°, \pm 60°$) で 3 名の人間が並んでいる．



a) acoustic model of Mr.*A* and 0°　b) acoustic model of Mr.*A* and 60°　c) acoustic model of Mr.*A* and -60°

Figure 8: Recognition Results by Acoustic Models of Mr.*A*



a) SIG asks a question.　b) Everyone replies simultaneously.　c) SIG answers what the front man said.　d) SIG answers what the 60° woman said.　e) SIG answers what the -60° man said.

Figure 9: Snapshots of Three Simultaneous Speech Recognition

2. *SIG* は,3 名に質問をする.
3. 各話者は 3 人同時に質問に対する回答を行う.
4. *SIG* は 3 話者の混合音声の定位・分離・認識を行う.
5. 最終的に,*SIG* は各話者に向きながら, 向いた方向の人が誰で何を言ったかを答えていく.
6. 音声認識に失敗したと判断した場合は, 該当話者の方向を向いた時に再び尋ねなおす.

本稿では, 実際の人間の代わりにスピーカとその前面に貼られた写真を用いている. スピーカは音響モデル作成時に使用したスピーカと同じものである. 各スピーカから流れる音声は, そのスピーカに貼られた写真と同じ人物のものである. 以下にこのシナリオの典型的な結果を 2 例示す.

1. *SIG* が好きな数字に関する質問をする (Fig. 9a)).
2. 各スピーカから 1 から 10 までの互いに異なる数字が同時に流れる. ただし, 数字の組合せはトレーニングセットに含まれる組合せと同じものである (Fig. 9b)).
3. *SIG* は各音声を実時間人物追跡システムを利用して定位する定位情報を利用して ADPF がその方向の音声を抽出する. 各分離音に対し 9 つの音声認識プロセスが同時に実行され, 結果を統合し, 最も適合のよい話者名, 認識結果を求める.
4. *SIG* は各話者に向きながら, 求めた話者名, 認識結果を答える (Fig. 9c)-e)).

結果では, 3 話者の認識がすべて成功しており, 同時発話の場合でもロボット自身のマイクを使った音声の定位・分離・認識を行うロボット聴覚システムの有効性を示すことができた. しかし, Fig. 8 に見られるように, 各分離音声の認識率は高々80%程度である. 音声認識に失敗する場合は, 対象音源の方向を向き, 聴覚中心窩をうまく利用し, 曖昧性を解消するように訊き返すようなアクティブオーディションを用いて解決することができる. また, 事前に顔認識によって顔の名前がわかっているときには, 音声認識で使用する音響モデルの数を削減することができるので, 高速で正確な認識が可能であるという結果も得られている.

7　結論

本稿では, ロボットへ搭載してアクティブオーディションに基づき音源の定位・分離・認識を行うロボット聴覚システムを提案・評価し, その有効性を示した. システムは, 聴覚中心窩に基づき音源方向に応じたアクティブな通過帯域制御と音源方向を向くというアクティブな動作を行い, 高速で高精度な音源定位・分離・認識を実現した. これは, アクティブオーディションの有効性を示しており, ロボット聴覚では, アクティブな動作による知覚向上が本質的であることを示している. 音源方向を向くという動作は, ロボット聴覚の向上だけではなく, 人間とのフレンドリーなインタラクションを実現したり, テレイグジスタンスによる会議では, 相手の注意を向けさせるという

意味でも重要であろう．しかし，より制約の少ない環境での利用に耐えうる音声分離・認識には，多くの課題がある．これには，よりロバストで高精度な音源分離も必要であろうが，音声認識エンジン にも，missing data や missing feature など分離データの性質を考慮した改良[4, 21]が必要であろう．

謝辞

豊橋科学技術大学の中川聖一教授と京都大学の河原達也助教授の助言に感謝する．また，北野共生システムプロジェクトのメンバに感謝する．

参考文献

[1] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1987.

[2] M. Aoki, M. Okamoto, S. Aoki, H. Matusi, T. Sakurai, and Y. Kaneda. Sound source segregation based on estimationg incident angle of each frequency component of input signals acquired by multiple microphones. *Acoust. Sci. and Tech*, 22(2):149–157, 2001.

[3] F. Asano, M. Goto, K. Itou, and H. Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Proceedings of International Conference on Speech Processing (Eurospeech 2001)*, pages 1013–1016. ESCA, Sep. 2001.

[4] J. Barker, M.Cooke, and P.Green. Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proc. of 7th European Conference on Speech Communication Technology (EUROSPEECH-01)*, volume 1, pages 213–216. ESCA, 2001.

[5] J. Blauert. *Spatial Hearing*. The MIT Press, 1999.

[6] O. D. Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint*. The MIT Press, MA., 1993.

[7] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima. Robust face detection against brightness fluctuation and size variation. In *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-2000)*, pages 1397–1384. IEEE, 2000.

[8] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi, and T. Mishima. Convergence analysis of online linear discriminant analysis. In *Proc. of IEEE/INNS/ENNS Int. Joint Conference on Neural Networks*, pages III–387–391. IEEE, 2000.

[9] M. Z. Ikram and D. R. Morgan. A multiresolution approach to blind separation of speech signals in a reverberant environment. In *Proceedings of 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*, pages 2757–2760. IEEE, 2001.

[10] S. Kagami, K. Okada, M. Inaba, and H. Inoue. Real-time 3d optical flow generation system. In *Proc. of Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI'99)*, pages 237–242, 1999.

[11] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of bayesian probability network to music scene analysis. In *Working Notes of the IJCAI-95 Computational Auditory Scene Analysis Workshop*, pages 52–59. AAAI, 1995.

[12] W.N. Klarquist and A.C. Bovik. Fovea: A foveated vergent active stereo vision system for dynamic 3-dimensional scene recovery. *RA*, 14(5):755–770, October 1998.

[13] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano. Real-time auditory and visual multiple-object tracking for robots. In *Proc. of the 17th Int. Joint Conf. on Atificial Intelligence (IJCAI-01)*, pages 1424–1432. MIT Press, 2001.

[14] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pages 832–839. AAAI, 2000.

[15] K. Nakadai, H. G. Okuno, and H. Kitano. Exploiting auditory fovea in humanoid-human interaction. In *Proceedings of 18th National Conference on Artificial Intelligence (AAAI-2002)*, pages 431–438. AAAI, 2002.

[16] Y. Nakagawa, H. G. Okuno, and H. Kitano. Using vision to improve sound source separation. In *Proc. of 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 768–775. AAAI, 1999.

[17] T. Nakatani and H. G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27(3-4):209–222, 1999.

[18] T. Nishiura, M. Nakamura, A. Lee, H. Saruwatari, and K. Shikano. Talker tracking display on autonomous mobile robot with a moving microphone array. In *Proceedings of the 2002 International Conference on Auditory Display (ICAD 2002)*, 2002.

[19] H.G. Okuno, S. Ikeda, and T. Nakatani. Combining independent component analysis and sound stream segregation. In *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pages 92–98. IJCAI, 1999.

[20] H.G. Okuno, K. Nakadai, T. Lourens, and H. Kitano. Separating three simultaneous speeches with two microphones by integrating auditory and visual processing. In *Proc. of European Conf. on Speech Processing(Eurospeech 2001)*. ESCA, 2001.

[21] Philippe Renevey, Rolf Vetter, and Jens Kraus. Robust speech recognition using missing feature theory and vector quantization. In *Proc. of 7th European Conference on Speech Communication Technology (EUROSPEECH-01)*, volume 2, pages 1107–1110. ESCA, 2001.

[22] S. Rougeaux and Y. Kuniyoshi. Robust real-time tracking on an active vision head. In *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS-97)*, pages 873–879. IEEE, 1997.

[23] G. Schuller and G. Pollak. Disproportionate frequency representation in the inferior colliculus of horsehoe bats: evidence for an "acoustic fovea". In *J. Comp. Physiol. A*, volume 132, pages 47–54, 1979.

[24] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, and 山本 幹雄. **音声認識システム**. オーム社, 2001.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-6

状況検知を利用したロボット用音声認識インタフェースの一手法とその評価

A Speech Recognition Interface for Robots using Notification of Ill-Suited Conditions.

岩沢 透 Toru IWASAWA　**大中 慎一** Shin'ichi OHNAKA　**藤田 善弘** Yoshihiro FUJITA

NEC マルチメディア研究所

Multimedia Res. Labs., NEC Laboratories

t-iwasawa@bp.jp.nec.com

Abstract

This paper describes a speech recognition interface for robots which uses notification of ill-suited condition for speech recognition. Because our hope of speech recognition interface for robot is hands-free and flexible for distance, it is sensitive for the condition such as noise and utterance power. To reduce the influence of ill-suited conditions for speech recognition, we tried to detect these conditions and made them notified by robot. There are two ways of notification method from robot, one is active method that the robot actively improve the condition, another is passive method that the robot asks a person to improve. We implemented the passive method on mobile robot PaPeRo, and evaluated the effect.

1 はじめに

ロボットの研究開発は，主に工場を対象とした産業用ロボットの分野から人とのインタラクションをテーマとしたパートナー型ロボットの分野へと徐々にシフトしてきている．人とロボットのインタラクション手段としては，人がロボットを撫でたり叩いたりする身体的インタラクションの他に音声を利用したコミュニケーションが重要な役割を果たすものと考えられる．このようなコミュニケーションの手段として，音声認識を利用し人と対話することを目的としたロボットの研究開発が実用化を目指し進められている [1]-[6]．筆者らは人と対話する自律移動型パーソナルロボット PaPeRo の研究開発を行っており，家庭環境での利用を想定した離散単語音声認識インタフェースの研究開発を行ってきた [7]．PaPeRo の目指す音声認識インタフェースは，家庭環境において利用可能でかつ利用者の身体的自由度の高いインタフェースである．利用者の身体的自由度という観点では，ハンズフリーでかつ発話距離に融通性を持つディスタンスフリーな音声認識インタフェースの構築を目標としている．

このような実環境での使用を前提としたハンズフリー，ディスタンスフリーな音声認識インタフェースの構築には 2 つの大きな問題がある．一つは周囲雑音を音声認識語彙と誤認識してしまうことによる誤動作の問題，もう一つは利用者の発話に何らかの問題が混入することによる音声認識精度劣化の問題である．このうち前者の誤動作の問題に対し，PaPeRo では棄却辞書を利用した不要音声棄却を行っている [7]．本稿では，後者の音声認識精度劣化の改善策について述べる．

音声認識精度劣化の問題は，利用者の発話に関する問題と環境の問題に大別できる．利用者の発話の問題としては，発声のパワー，明瞭さ，タイミングといった発声方法の問題とシステム側で受理不能な未知語発話の問題がある．これに対し環境の問題は，物音などの環境音，周囲会話，システム内部雑音と言った利用者とは別の因子により音声認識を阻害する雑音が混入する問題である．家庭環境での使用を想定した音声認識に対しては，学習データを実環境の音響空間に適応させるよう変換させ実環境での認識精度を向上させる方法などが提案されている [8]．しかしながら，先に述べた利用者側の発話の問題や雑音が含まれる音声を正確に認識させることは難しい．また環境面の問題点に対しては，マイクロフォンアレイを利用し雑音を除去する技術が研究されているが [5],[9]，実環境において発話音声に重畳される未知で非定常な雑音を除去することは極めて難しい．

そこで本研究では，このような発話面や環境面の問題が

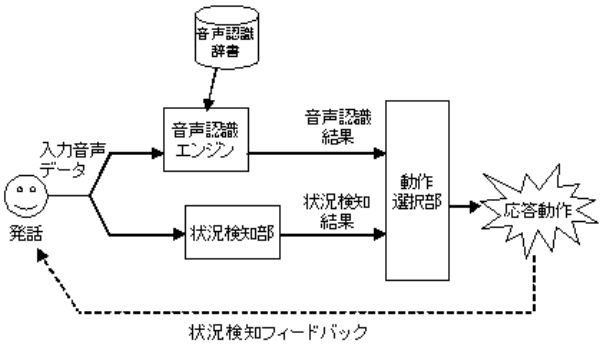


Figure 1: 状況検知を利用した音声認識システムの動作図

発生する状況を抑える方法として，利用者とシステム双方が音声認識エンジンの弱点をカバーするよう適応することを考えた．そして，問題状況の発生を抑える改善策として，音声認識を行う上で問題となる要因やシステムの利用状況（音声認識エンジンが雑音の影響を受けているかどうかなど）を推定する状況検知の手法を検討した．本稿では，状況検知を利用した音声認識インタフェースと状況検知の利用方法について述べる．状況検知のフィードバックに対するロボット側の対処行動としては，能動的に対処する Active な対処行動と，利用者に対処を依頼する Passive な対処行動が考えられるが，今回は主に Passive な対処行動を PaPeRo に実装し評価したのでその結果についても述べる．

2　状況検知を利用した音声認識インタフェース

始めに，状況検知を利用した音声認識インタフェースを組み込んだ音声認識システムの動作について説明する．状況検知は，音声認識時にマイクから入力される音声を解析し音声認識精度の劣化につながる問題状況を抽出することにより行われる．図 1 に状況検知を利用した音声認識システムの動作の流れを示す．利用者が発話した音声は，音声認識エンジンで認識され音声認識結果を返すと同時に状況検知部において問題状況が推定され状況検知結果として動作選択部へ出力される．動作選択部では，音声認識結果と状況検知結果をもとにシステム側の応答動作が選択される．応答動作の選択は，音声認識結果が音声認識語彙かリジェクトかといった情報と状況検知結果の種類などに応じなされ，状況検知動作が選択された場合に状況検知結果が利用者にフィードバックされる．

次に状況検知の対象となる問題状況の例を波形データを利用し説明する．波形データは，家庭環境で PaPeRo に入力された音声に対し音声認識エンジンで音声区間を切り出した音声（以後，切り出し音声と呼ぶ）である．まず，理想的な音声データを図 2 に示す．図 2 の音声データは，音

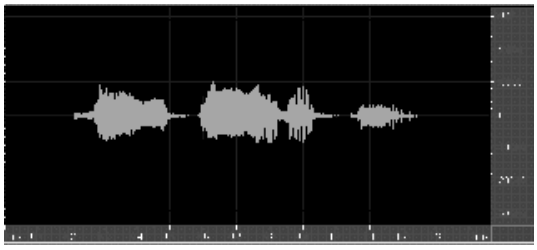


Figure 2: 理想的な音声の例

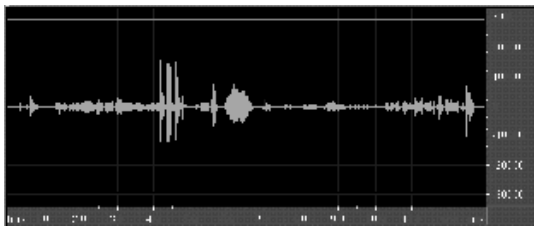


Figure 3: 周囲雑音の影響を受けた音声の例

声の前後に 200〜300ms 程度の無音区間が付与されており発話全体が網羅されるよう理想的に切り出されている．次に問題状況における音声の例として継続的な周囲雑音が重畳される例を図 3 に示す．この例の音声は，子供の足音や笑い声に続く形で発話がなされ，さらにその後に周囲会話が続くなど切り出し音声に様々な音声が 10 秒以上に渡り含まれている状況である．次に発話タイミングの問題に伴う頭切れ音声の例を図 4 に示す．この例では，音声認識エンジンが動作開始する前に発話が始まっているため，音声データの先頭部分から発話音声が出現している．次に声量（パワー）に問題があり誤認識したとみられる音声の例を図 5 に示す．上が声量不足で下が声量過多である．パワー不足の音声は，殆ど静環境における無音区間と変わらないような波形データである．このような音声は，利用者とマイクの距離が離れすぎている場合や発話自体が怠け発話の場合によく発生する．一方で，パワー過多の波形データは音声がクリッピングされ音声がひずんでいる．このようなパワー過多によるひずみ音声は，マイクの近傍で怒鳴り気味に発声することにより多く発生する．また，図 5 中の 2 つの音声比較から利用者，使用環境，マイク距離の違いにより入力される音声に大きな差が出ることが分かる．

このような問題状況における音声は一概に全て誤認識するという訳ではないが，誤認識となる場合が多い．

状況検知結果に対する応答動作としては，次に示す Active な応答動作と Passive な応答動作の 2 つが考えられる．

Active な応答動作：　問題状況に対しシステム側が能動的に対処行動を行う動作方法．例としては，入力音声が小さい場合にマイクを人間に近づける（人に近づく）方法，周囲雑音の影響を検知した場合にマイクの感度

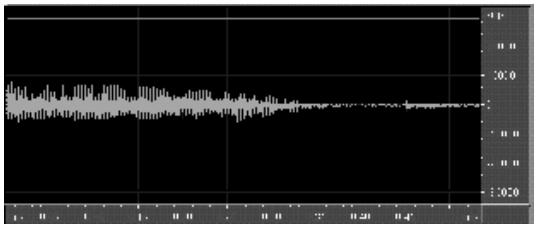


Figure 4: 発声タイミング不具合のある音声の例

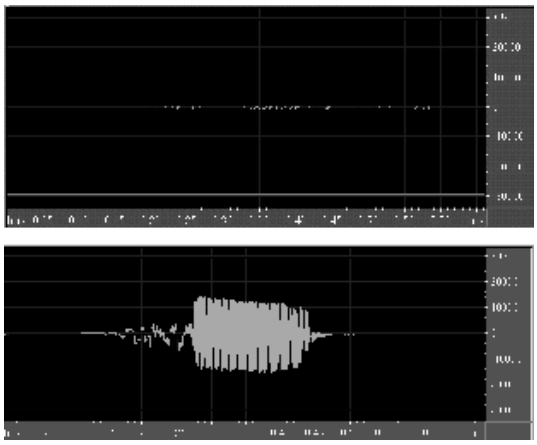


Figure 5: 声量の不具合がある音声の例（上がパワー不足，下がパワー過多）

を調整したり利用者の方向へビームフォーミングなどを行い音声入力のフォーカスをあてる方法などが考えられる．

Passive な応答動作： 利用者に対しシステム側から問題状況の存在を警告し対処行動を依頼する動作方法．利用者に対し大きい声で明瞭に発話するよう警告したり周囲雑音を抑制するように警告する方法が考えられる．

今回は，状況検知結果に対し Passive な応答動作を行う音声認識インタフェースを PaPeRo に実装した．次章において，PaPeRo に実装した状況検知と動作選択の方法及び応答動作について述べる．

3 パーソナルロボット PaPeRo への実装

PaPeRo の音声認識は，不特定話者の離散単語音声認識であり，対話メインモードにおける音声認識語彙数は約 650 である．音声認識インタフェースは，ハンズフリーでマイク間距離 0.5m〜2m の発話音声の認識を想定しマイクの入力を調整していることを特徴とする．マイクの指向性については，対話中の移動動作に伴う床のすべりなどの影響を考慮し側面から背面にかけての入力レベルが低下する程度の単一指向性マイクを利用している．

PaPeRo への実装にあたっては，状況検知部に入力させる音声データに音声認識エンジンで使用した切り出し音声

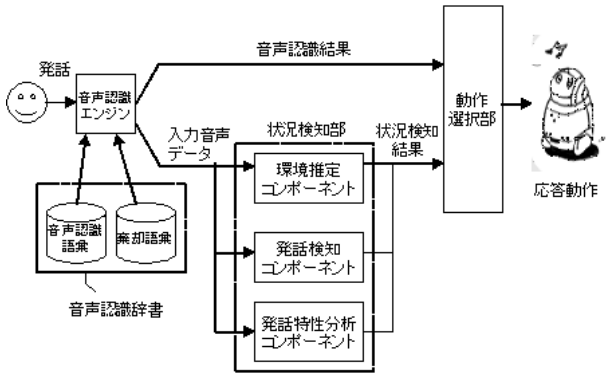


Figure 6: PaPeRo に実装した状況検知の動作

を利用した．音声認識辞書には，PaPeRo の認識語彙に加え，文献 [7] に記載の棄却辞書を搭載し棄却辞書へのヒットをリジェクトと判定した．PaPeRo に実装した状況検知の動作図を図 6 に示す．

状況検知部には以下の 3 つのコンポーネントを実装した．

環境推定コンポーネント： 主に周囲雑音の影響を検知するコンポーネント．定常もしくは非定常な雑音が継続的に発生する状況を検知する周囲雑音検知を実装した．

発話検知コンポーネント： 入力音声中に発話音声の有無を検出するコンポーネント．発話音声の検出には入力音声の調波構造を示す特徴量を利用した．

発話特性分析コンポーネント： 発話音声に含まれる問題状況を検出するコンポーネント．入力音声の始端に発話らしい音声の有無を検出する頭切れ検知と入力音声のパワーが特定の閾値条件を満たす状況を検知するパワー不足検知とパワー過多検知を実装した．なお，声量の不具合に関しては，問題あり，やや問題ありの 2 段階の閾値を与えそれぞれパワー不足（過多），パワー不足（過多）気味のように分類し検知するようにした．

なお, 状況検知の際に用いる諸々の閾値や検知条件のチューニング，状況検知結果を利用した動作選択方針の決定には，実環境（家庭環境）において PaPeRo を動作させ収集した音声データを学習データとして利用した．

各状況検知コンポーネントが返す検知結果は，表 1 に示す状況検知結果選択表に基づき評価され最終的な状況検知結果が決定される．表中の「−」は結果に非依存であることを意味する．表 1 において，環境推定コンポーネントの検知結果が継続雑音検知を返した場合は，最優先で環境推定コンポーネントの結果が状況検知結果として返される．環境推定コンポーネントの検知結果が「なし」の場合は，次に発話検知の有無を調べ，発話ありの場合は発話特性分析コンポーネントの返す結果を，発話なしの場合は「非発話」

Table 1: 状況検知結果選択表

環境推定	発話検知	発話特性分析	検知結果
周囲雑音	ー	ー	周囲雑音検知
なし	なし	ー	非発話
	あり	なし	通常発話
		頭切れ	頭切れ
		パワー不足気味	パワー不足気味
		パワー不足	パワー不足
		パワー過多気味	パワー過多気味
		パワー過多	パワー過多

Table 2: 音声認識結果と状況検知結果を統合した動作選択表

音声認識結果	状況検知結果	対応動作
ー	周囲雑音検知	周囲雑音警告
	非発話	無視
	頭切れ	頭切れ警告
	パワー不足	無視
	パワー過多	パワー過多警告
認識語彙	通常発話	通常動作
	パワー不足気味	通常動作
	パワー過多気味	通常動作
リジェクト	通常発話	通常棄却動作
	パワー不足気味	パワー不足警告
	パワー過多気味	パワー過多警告

を検知結果として返す．発話ありで発話特性分析コンポーネントの検知結果がない場合は「通常発話」を検知結果として返す．

次に音声認識結果と状況検知結果を利用した動作選択について表 2 を利用し説明する．状況検知結果が「周囲雑音検知」「非発話」「頭切れ」「パワー不足」「パワー過多」の場合は，状況検知結果が音声認識結果より優先され警告や無視といった動作選択がなされる．なお，極端なパワー不足音声に対しては，周囲雑音と混同しやすいためパワーが一定の閾値（＝パワー不足とパワー不足気味の境界の閾値）を下回る場合は無視することとした．本来話しかけに対する無視は好ましい反応ではないが，パワー不足音声に限っては周囲会話との混同が多い点と無視された後の声量が必然的に増加し問題状況が回避されやすいとの予測に基づき無視することとした．また，状況検知結果が「パワー不足気味」「パワー過多気味」「通常発話」の場合は，音声認識結果の種別に応じ対応動作が異なる．すなわち，音声認識結果が認識語彙がである場合は認識語彙に対応した動作を行い，リジェクトである場合には問題状況の警告ガイダンスもしくは通常棄却動作を行う．

最後に，表 2 の対応動作の各項目に対する具体的な反応動作を表 3 を利用し説明する．まず対応動作が「無視」の場合は何もせず再び音声認識待ち状態になる．「通常動作」

Table 3: 具体的な反応動作

対応動作	動作内容
無視	何もしない
通常動作	認識語彙に対応した動作（あいさつする，ダンスするなど）
通常棄却動作	通常棄却反応（「うんうん」とうなづく,「えっ何？」と聞き返すなど）
周囲雑音警告	「周りがうるさい」「みんなで同時にしゃべらないで」「あまりいろいろなことといわないで」などの警告
頭切れ警告	発話タイミング教示（「しゃべるのは耳が光ってからにして」）
パワー不足警告	声量不足警告（「もう少し大きい声でしゃべってみて」「もう少し近くでしゃべってみて」）
パワー過多警告	声量過多警告（「もう少し小さい声でしゃべってみて」「もう少し離れてしゃべってみて」）

の場合は音声認識語彙に関連付けられた動作を，「通常棄却動作」の場合は，「うんうん」とうなづいたり「えっ何」と聞き返すといった棄却反応動作を行う．周囲雑音警告に関しては，周囲が騒がしい旨警告ガイダンスを行う．頭切れとパワー不足とパワー過多に関しては，各々発話方法の改善を促す警告ガイダンスを行う．頭切れ警告の「しゃべるのは耳が光ってからにして」という警告は，PaPeRo が耳が光っている時にのみ音声認識をしていることの教示である．

4 評価

PaPeRo の実環境評価の一環として，実際の家庭環境において PaPeRo を動作させ得られた入力音声の収集とその分析を行っている．ここでは，実環境評価の音声を利用した状況検知の評価について述べる．状況検知の評価対象には，周囲雑音，頭切れ，パワー不足，パワー過多といった特定の警告ガイダンスを行う場合（以後特定状況検知と呼ぶ）を利用する．なお，パワー不足警告に関しては，警告を行わない「パワー不足」検知は対象から除き，状況検知結果が「パワー不足気味」の時のみを対象とした．

4.1 家庭環境評価の概要

家庭環境での評価は，PaPeRo を家庭内で 2 週間程度継続使用してもらい収集された音声データを利用している．評価の前には事前に代表者に簡単な使用方法の説明を行っており受理可能な認識語彙や発話のコツをガイダンスしている．

このようにして収集された音声データを利用し次の 2 項目の評価を行った．

経過日数に対する状況検知への適応度合： 特定状況検知

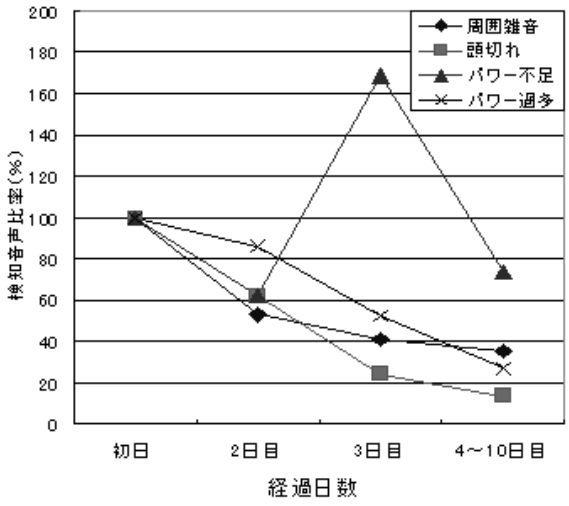


Figure 7: 使用日数の経過に伴う特定状況検知の発生頻度（初日を 100 とした相対数値）

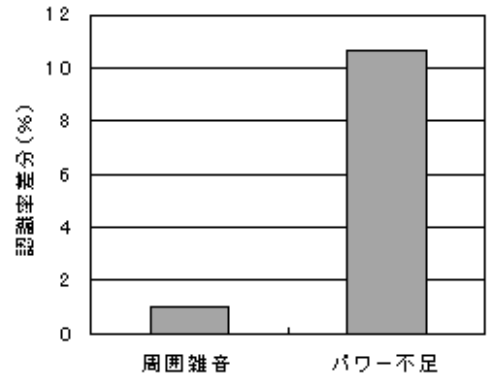


Figure 8: 状況検知フィードバックを返した場合と返さなかった場合の直後発話に対する音声認識率（状況検知フィードバックを返した場合の認識率から返さなかった場合の認識率を引いた差分）

の発生頻度の使用日数の経過に対する変化を評価した．発生頻度は，使用日ごと話しかけ総数に対する状況検知の発生比率として計算した．

状況検知フィードバックの音声認識への寄与： 特定状況検知のフィードバックが利用者の次発話に与える音声認識率への寄与を評価した．評価方法は，状況検知フィードバックを的確に返した場合と返さない場合の直後発話に対する音声認識率差分で比較した．

なお，一日あたりの起動時間には制約を設けていないため，家庭ごとに慣れの進行度合いにはばらつきがある．これを考慮し，状況検知の評価には使用開始から 3 日目までの使用時間にばらつきの少ない 5 家庭を抽出し利用した．

4.2 結果

使用日数の経過に伴う状況検知の発生頻度の変化に関して，初日の発生頻度を 100 として正規化したグラフを図 7 に示す．図 7 より，周囲雑音，頭切れ，パワー過多検知は使用日数の経過と共に特定状況検知の頻度が減少するのに対し，パワー不足は減少しないことが分かる．検知頻度が使用日数の経過に伴い減少したものに関しては，頭切れ警告とパワー過多警告が日数の経過と共に順調に減少している．特に頭切れ警告は頻度が大きく減少し，使用 3 日目の段階で初日に比べ頻度が 20 %程度まで減少している．一方で，周囲雑音警告は 2 日目以降の減少が小さく 3 日目以降も初日の 40 %程度の頻度で発生していることが分かる．なお，検知頻度の減少がみられないパワー不足検知の音声を試聴してみた所，遠距離発声と思われる反響音の強い発話音声や怠け気味の発話音声が多かった．

次に，特定状況検知フィードバック直後の話しかけ音声（未知語発話は除く）に対する認識率改善について説明する．ここでは，周囲雑音とパワー不足に対する状況検知フィードバックのケースを対象とし，状況検知フィードバックを返した場合と返さなかった場合の次発話の音声認識率を評価した．周囲雑音とパワー不足両者の状況検知フィードバックは，前者が利用者が周囲環境に働きかけ適応する必要があるものであり，後者が利用者の矯正で適応可能なものである点で質的に異なっている．

評価は，同じ家庭の被験者に対し状況検知フィードバックを返した場合と返さない場合を一定の条件のもとに制御し両者の比較を行った．図 8 に状況検知フィードバックを返した場合と返さなかった場合の直後発話認識率の差分を示す．パワー不足検知はフィードバックを行う方が次発話の音声認識率が 10 %以上高いのに対し，周囲雑音検知はフィードバックを行うことによる音声認識率の改善がわずかであった．

4.3 考察と検討

図 7 に示した経過日数と状況検知の発生頻度の関係から，パワー不足以外の特定状況検知に対しては，問題状況のフィードバック効果により利用者が問題状況を発生させないよう適応していると考えられる．一方で，パワー不足検知だけは検知頻度が使用日数の経過と共に減少せず利用者が適応していないという結果が得られた．この結果には利用者の慣れが関係していると考えられる．実際に，PaPeRo の使用開始時には利用者は PaPeRo の近傍かつ正面に座りはっきりとした声で話しかけるが，時間の経過と共に離れた位置からもしくは怠け気味に発話するように変化していく使用例が確認されており，慣れと発話の因果関係は存在しているもの思われる．一方で，パワー不足警告は，状況

検知フィードバックを行うことにより次発話の音声認識率が大きく上昇している．このデータは慣れに伴い発生する問題が状況検知フィードバックによりその都度矯正され，次発話の正確な認識につながっていることを裏付けるものである．このことから，パワー不足警告は利用者の慣れに起因する発話の問題を定期的に矯正する効果を持っているものと考えられる．

特定状況検知の音声認識への寄与に関しては，上記の通りパワー不足警告に主な改善が見られた．一方で，周囲雑音検知に関しては，経過日数に伴う発生頻度が 2 日目以降ほぼ横ばいであまり減少していない結果と検知フィードバックの効果がわずかであることから，利用者が周囲環境へ改善を働きかけることの難しさが現れているものと考えられる．周囲会話などの雑音に対しては，PaPeRo 側で話者を識別し近づく，話者の方向にビームフォーミングするなどの Active な応答動作を行い問題状況に適応させることが望まれる．

最後に，インタフェースの観点から状況検知を利用した音声認識インタフェースを考察してみる．一般的に音声認識の性能は人間の聴覚には及ばないため，利用者の側から見るといつ，どういう環境で，どのような声量で話しかければ正しく認識できるのか分からないという問題がある．今回実装した利用者に問題状況をフィードバックする状況検知は，利用者に対しシステム状態の透過性を高める役割を持つものと考えられる．一方で，利用者に対しロボット側から適応を求めることは，利用者にとっては負担となる．実際，今回の家庭環境評価の被験者からは「ロボットに命令されているようで違和感がある」との声が聞かれた．この点については，今回実装した Passive な応答動作に加え Active な応答動作を取り入れ，利用者の負担を軽減するような音声認識インタフェースを構築していく必要があるものと考えられる．このように，Active な応答動作と Passive な応答動作を状況に応じ使い分けていくことで，人間とロボットが共同で適応する双方向適応型の音声認識インタフェースを構築していくことが望ましいと考えられる．

また，ガイダンスの的確性という点に関しても改善の余地がある．例えば，パワー不足の警告ガイダンスを行う場合に「もう少し大きい声でしゃべって」と「もう少し近くでしゃべって」の 2 つの質が異なるガイダンスがある．これら 2 つのガイダンスは，話者との距離を画像やセンサの情報を利用し取得することにより状況依存で使い分けることが可能であると考えられる．

5 おわりに

本稿では，状況検知を利用した音声認識インタフェースの一手法と状況検知に対する Passive な対応動作をパーソナルロボット PaPeRo へ実装し評価した結果について述べた．問題状況の警告ガイダンスを利用者に対して行うことにより，使用時間の経過に伴う問題状況検知の頻度減少やパワー不足警告に対する検知後発話の音声認識精度改善の効果が見られた．一方で，周囲雑音検知のような利用者が周囲環境に働きかけ問題を改善させる必要があるものに関しては，利用者側で適応することの難しさが見受けられた．

今後は，状況検知の精度向上を図ると共に，画像やセンサ情報を統合し状況検知ガイダンスの適切化やロボット側で問題状況の改善を図る Active な適応についてさらに検討を進める予定である．

参考文献

[1] 藤田. , “パーソナルロボット R100”, 日本ロボット学会誌 Vol.18 No.2, pp.40-41, (2000).

[2] 藤田. , “NEC におけるパーソナルロボットの開発”, 日本ロボット学会誌 Vol.20 No.7, pp.676-679, (2002).

[3] Perzanowski D, Schultz A C, Adams W, Marsh E, Bugajska M., “Building a Multimodal Human-Robot Interface”, IEEE Intelligent Systems, pp. 16-21, (2001).

[4] Chong S, Kuno Y, Shimada N, Shirai Y., “Human-Robot Interface Based on Speech Understanding Assisted by Vision.”, Lecture Notes in Computer Science, pp.16-23, (2000).

[5] 松井, 麻生, J.Fly, 浅野, 本村, 原, 栗田, 速水, 山崎., “オフィス移動ロボット Jijo-2 の音声対話システム”, 日本ロボット学会誌 Vol.18 No.2, pp.300-307, (2000).

[6] 松坂, 東条, 小林., “グループ会話に参与する対話ロボットの構築”, 電子情報通信学会論文誌 Vol.J84-D-II，No.6，pp.898-908 (2001).

[7] 岩沢., “パーソナルロボット PaPeRo の音声認識インタフェース”, 人工知能学会 AI チャレンジ研究会論文集, Vol.13, pp.17-23, (2001).

[8] Stahl V, Fischer A, Bippus R., “Acoustic Synthesis of Training Data for Speech Recognition in Living Room Environment.”, Proc ICASSP, pp.21-24, (2001).

[9] 中村.,“外乱に強い音声認識を目指して”, 日本音響学会誌, Vol.57, No.10, pp.662-667, (2001).

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-7

視聴覚定位能力を同時に獲得するロボットヘッドの構築

Construction of a Robot Head which Acquires Audiovisual Localization Ability Simultaneously

中島 弘道
Hiromichi NAKASHIMA
理化学研究所 BMC
RIKEN BMC
nakas@bmc.riken.go.jp

大西 昇
Noboru OHNISHI
名古屋大学
Nagoya University
ohnishi@ohnishi.nuie.nagoya-u.ac.jp

向井 利春
Toshiharu MUKAI
理化学研究所 BMC
RIKEN BMC
tosh@bmc.riken.go.jp

Abstract

A robot system which acquires sound source localization ability in self-organization through repetition of movement and perception was built. A robot consists of one set of two microphones (hearing : ear) and a video camera (vision : only an eye and a right eye use) , and a rotation stand (movement : head). Moreover, a learning model consists of two modules by neural network. One is a visual module and another is a auditory module. The system can be learned without the external explicit supervision. As a result, it becomes possible after learning to catch a sound source in the center of a view by rotating the head in the direction of a sound source.

1 はじめに

生体の持つ優れた感覚情報処理能力と運動制御能力を模倣し，柔軟かつ統合的な機能を実現するための多くの研究がなされている．知覚及び運動能力獲得の基本となるのが，知覚と運動の繰り返し（知覚循環）[Neisser, 1976]による学習である．目や耳など感覚器による知覚と，それに基づく環境への働きかけ，つまり身体の運動を通して情報を収集し，その情報を用いて知覚および運動を学習する．つまり，生体は環境からの明示的な教師なしで，運動制御や感覚能力を獲得していると考えられる．

音源定位もこのような学習によって獲得される能力の一つであると考えられる．音源定位とは，音源が発する音からその音源物体の位置を判定することである．もし，音の早さと左右の耳の距離が与えられれば，我々は時間差から音源の方向を簡単に計算することが出来る．しかし，生体はそのようなパラメータは知らないし，時間差と角度の

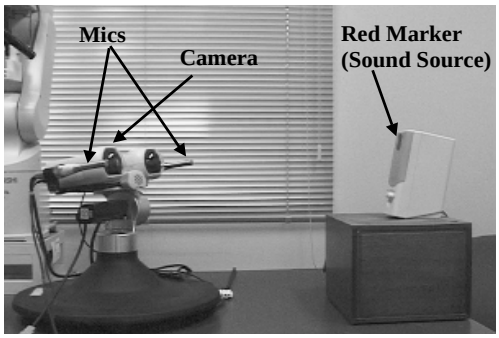


Figure 1: 音源定位ロボット

幾何的関係も知らない． また生体は成長等により身体パラメータが変化する．それにもかかわらず，生体は左右の耳での音の違い（到達時間差や音圧差など）を基にして，音源の方向を推定することが出来る[寺西 *et al.*, 1969][吉田 and 亀田, 1980]．メンフクロウを用いた実験から，この音源定位能力は，視覚情報を手がかりにして獲得されていると考えられる[Knudsen and Knudsen, 1985]．このことから，繰り返しの知覚と運動によって得られた視聴覚情報等を用いて，音源定位能力は獲得されていると考えられる．

知覚と運動の繰り返しによる学習の研究には，視覚と眼筋の情報から腕の姿勢を学習するモデル[Kuperstein, 1988]や，首と両眼を持つ冗長なシステムでの物体追跡学習[Kuniyoshi and Berthouze, 1998]，音源定位学習モデル[Irie, 1990] [浅井 *et al.*, 2000]などがある．

本論文では，図 1 に示されるような 2 本のマイク（聴覚：耳），ビデオカメラ（視覚：目），ロボット（運動：首）からなる人間の頭部に似せたロボットが，音源定位能力を自己組織的に獲得するシステムを構築し，ロボットと実音源を用いた実験によってその有効性を示す．さらに，精度及び汎用性を向上させる為の拡張について述べ，シミュレーションによってその有効性を検証する．

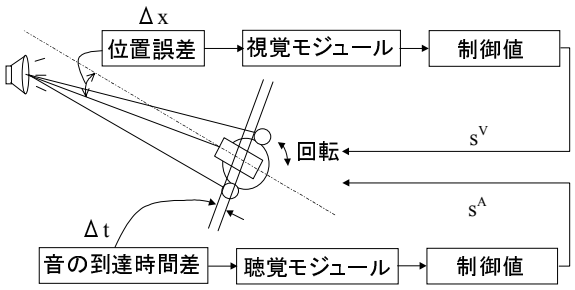


Figure 2: 音源定位モデル

本システムは，教師なしで学習が可能で，学習後には音のする方向に首を回転させて，視野の中央に音源を捕らえる事が可能となる．また，外部環境やマイクやカメラ等の配置が変化しても，自己組織的に再学習され音源定位が可能である．このような柔軟性のある能力をもつ学習モデルは，音源定位以外の学習機能を持った工学的システム (ロボット) に応用できると考えられる．

2 音源定位学習モデル

2.1 前提条件

定位を行う音源は静止しており，周囲の環境においてただ 1 つであるとする．また，それが目で見て，音源であるという判断はすでにできるものとする． 知覚される情報は，聴覚からは音の両耳間時間差（Δt），視覚からは視野中央から音源物体までの水平距離（Δx）である．制御対象は，制御値（s）によって首が回転する． さらに両耳間距離，耳，目，首の配置などの幾何学的情報及び，音速などの環境情報は未知とする．

2.2 音源定位モデル

図 2 に音源定位モデルを示す．このシステムは，視覚モジュール，聴覚モジュール及び制御対象 (首の回転) から構成される．システムへの入力は視覚からの位置誤差 Δx と聴覚からの時間差 Δt である．また，システムからの出力は，水平面上の頭の角度 θ である．システムは，入力の Δx と Δt をもとに頭を回転させて，音源を視野の中央にもってくる．

視覚モジュールは，その入力 Δx から制御値 s^V を計算する．また，聴覚モジュールは，入力 Δt から制御値 s^A を計算する．システムの出力である頭の角度 θ は，制御値 s^V または s^A によって決まる．システムは運動と感覚のインタラクションによって教師なしで，時間差 Δt (位置誤差 Δx) と制御値 s^A (s^V) の関係を獲得する．

2.3 学習アルゴリズム

視覚モジュール，聴覚モジュールは，自己組織特徴マップにより構成される．モジュールのネットワーク構成を，図 3 に示す．それぞれのモジュールは，入力空間 W と出力空間 S の 2 つの空間を持つ．入力空間の参照ベクトル w_k は，出力空間の制御値 s_k に一対一で写像される．時間差 Δt（又は位置誤差 Δx）が入力されると，その値と最も近い参照ベクトルが選択され，それに対応する制御値 s が最も大きく出力される．視覚モジュールは，直接逆モデリング[Miller *et al.*, 1990]によって入出力関係を学習し，また聴覚モジュールは，視覚モジュールからの出力値を用いて学習がなされる．以下に学習の流れを示す.（図 4）

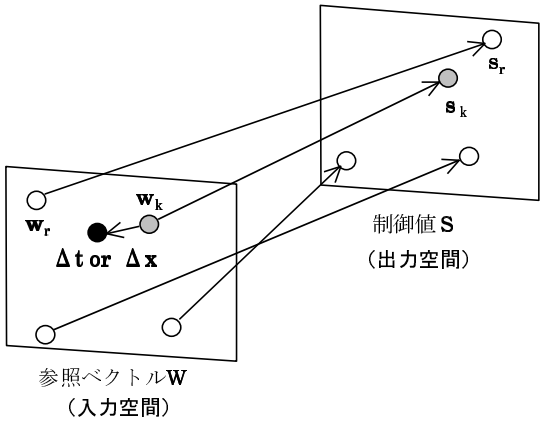


Figure 3: 視覚・聴覚モジュールのネットワーク構成

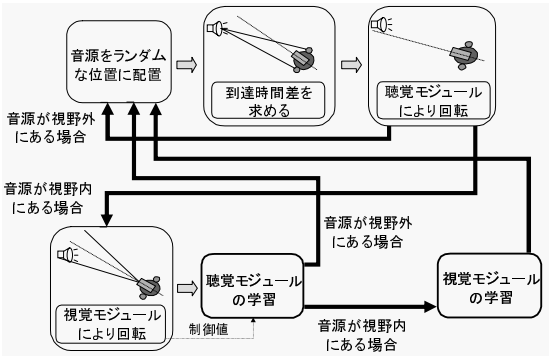


Figure 4: 学習の流れ

- 学習の流れ

Step 1 音源をランダムな位置に提示[1] し，その音のマイクへの到達時間差 Δt が聴覚モジュールに入力される．

Step 2 聴覚モジュールから制御値 s^A が式（3）を用いて出力され，制御対象の首を回転する．(図 3)

Step 3 もし音源が視野内にないときは，Step 1 にもどる．

Step 4 音源が視野内にある時，視覚モジュールに Δx が入力され，式（4）から得られた制御値 s^V で首を回転する．

[1] 実環境において，音源の位置と首の向きはさまざまに変化する．その状況を再現する為に音源をランダムな位置に繰り返し提示する．

Step 5 聴覚モジュールの参照ベクトル w^A と制御値 s^A を式（1）を用いて更新する．

Step 6 音源が視野内にある場合，ランダムな制御値 s_{rand} によって首を回転させ，その結果音源が視野内にある場合，網膜上での移動距離 Δx_{rand} を求める．

Step 7 視覚モジュールの参照ベクトル w_r^V と制御値 s_r^V を式（2）を用いて更新する．

Step 8 Step 1 にもどる．

聴覚モジュール，視覚モジュールそれぞれのパラメータである参照ベクトル及び制御値の更新則を以下に示す．

$$\begin{cases} w_r^A & \leftarrow & w_r^A + \epsilon_1^A \cdot g_r^A(k) \cdot (\Delta t - w_r^A) \\ s_r^A & \leftarrow & s_r^A + \epsilon_2^A \cdot g_r^A(k) \cdot s^V \\ g_r^A(k) & = & e^{-\sigma^A (r-k)^2} \\ k & = & \arg\min_{\forall r} \|w_r - \Delta t\| \end{cases} \quad (1)$$

$$\begin{cases} w_r^V & \leftarrow & w_r^V + \epsilon_1^V \cdot g_r^V(k) \cdot (\Delta x_{rand} - w_r^V) \\ s_r^V & \leftarrow & s_r^V + \epsilon_2^V \cdot g_r^V(k) \cdot (s_{rand} - s_r^V) \\ g_r^V(k) & = & e^{-\sigma^V (r-k)^2} \\ k & = & \arg\min_{\forall r} \|w_r - \Delta x\| \end{cases} \quad (2)$$

ここで ϵ は，学習速度係数であり，σ は学習の影響する範囲を表す係数である．

式 (1)(2) の 1 つ目の式は，Kohonen の自己組織特徴マップの学習則であり，学習が進むと w^A, w^V は，$\Delta t, \Delta x$ の特徴マップを形成する．聴覚モジュールの制御値 s^A は，式 (1) の 2 つ目の式によって視覚モジュールから出力される制御値 s^V の方向に修正がなされる．もし視覚モジュールが十分な学習がされていて，正しい出力値が出力されるとすると，聴覚モジュールの制御値は，徐々に音源を視野の中央に回転させることが出来る制御値に近づいてゆくことになる．よって学習は，視覚モジュールが先に行われ，その後，徐々に聴覚モジュールの学習がなされる．視覚モジュールの制御値の更新は式 (2) の 2 つめの式でなされ，制御値は値 s_{rand} に徐々に近づいてゆくことになる．これによって，視覚モジュールへの入力値 Δx_{rand} に対応した出力値 s_{rand} のマッピングが形成される．また式 (1)(2) の 3 つ目の式は，ベル型のガウス関数であり，4 つ目の式から求められる選択されたニューロン k の近傍領域にも影響を与える為に用いる．

各モジュールの出力値は，以下の式のように出力空間の制御値を”winning neuron” からの距離で重み付け平均したものを用いた．これは入力が連続値を取るので，出力値も離散値ではなく連続値とする為である．

$$s^A = \frac{\sum_{i=0}^{n-1} s_i^A * e^{-\xi^A (w_i^A - \Delta t)^2}}{\sum_{i=0}^{n-1} e^{-\xi^A (w_i^A - \Delta t)^2}} \quad (3)$$

$$s^V = \frac{\sum_{i=0}^{n-1} s_i^V * e^{-\xi^V (w_i^V - \Delta x)^2}}{\sum_{i=0}^{n-1} e^{-\xi^V (w_i^V - \Delta x)^2}} \quad (4)$$

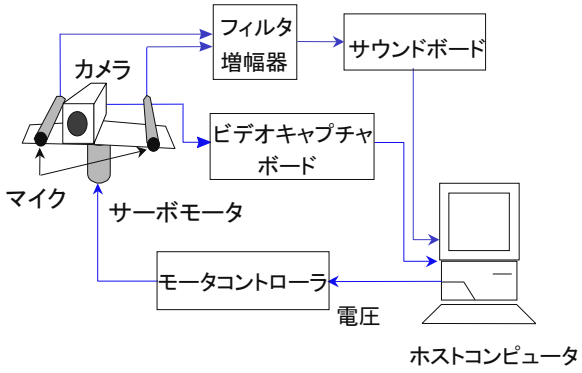


Figure 5: システム構成図

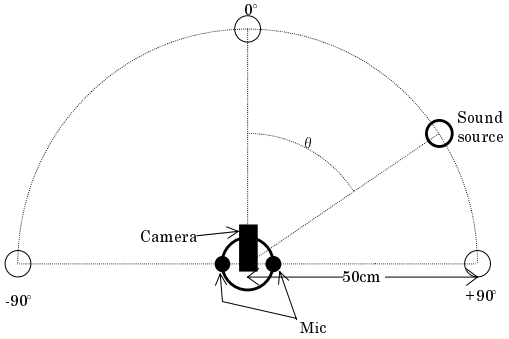


Figure 6: 実験環境

3 音源定位学習

3.1 システム構成

図 5 にシステム構成を示す．ロボット，コンピュータ，およびいくつかのサポート機器によって構成される．図 1 がロボット及び音源として用いたスピーカの写真である．ロボットには 2 本の無指向性マイクと CCD カメラ（視野角 43 °）が搭載されている．ロボットはサーボモータによって垂直軸回りに回転する．CCD カメラは，1 台のみを使用し水平方向に向けられている．また，左右のマイク間距離は 30cm で，その中点はほぼ回転軸上にある．

本研究の目的は画像処理ではないため，音源には赤のマーカーを付け，他に赤色物体はないものとして，視覚的位置 Δx の検出処理を簡略化した．時間差の計算方法は，左右の音データをシフトさせて相互相関係数値を求め，その値が最大となるポイント数を時間差として用いる．また音源位置の計算方法は，キャプチャされた RGB 画像をホストコンピュータ内で HSV 変換及び閾値処理によって赤色ピクセルだけを抽出した後，その画素の重心座標を音源位置として用いる．

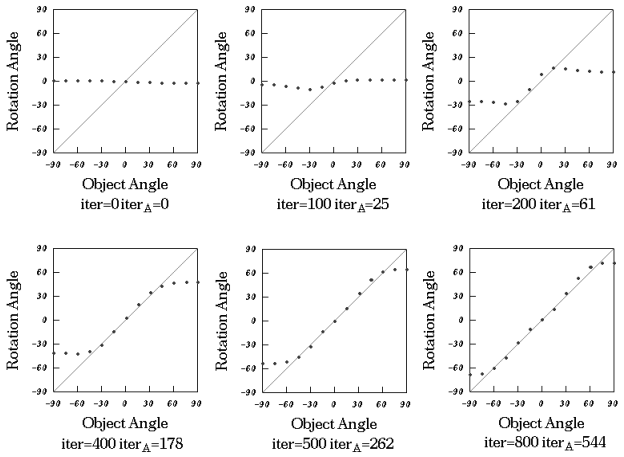


Figure 7: 聴覚モジュールの定位能力の推移 (横軸:音源の位置，縦軸:首の回転角，iter：音源提示回数，iter_A：聴覚モジュールの学習回数)．図中の右上に伸びる直線は，音源方向の角度と首の回転角が一致する理想値である．

3.2 実験環境

図 6 に実験環境を示す．音源として赤色のマーカーを付けたスピーカを用いて，拍手の音を録音したものを間欠的に発生させる．スピーカはビデオカメラを含む水平面上にある．ロボット上に 1 つのカメラ（右のカメラのみ使用）と 2 本のマイクがあり，水平面上で回転する．音源は半径 50cm の半円上にランダムに配置し，θ の範囲は ± 90 度である．また，視覚・聴覚モジュールに使用した特徴マップの数は，参照ベクトルと制御値それぞれ 40 個，参照ベクトル w の初期値にはランダムな値を，制御値 s の初期値にはランダムな小さな値を用いた．

3.3 結果と考察

学習過程における，音源の方向とシステムの回転角の関係の推移を図 7 に示す．800 回の音源提示後，この 2 つの角がほぼ等しい値となり，システムは音源を正確に視野の中央に位置させることが可能になっていることがわかる．また，図からシステムがはじめは視野内の音源の定位能力を獲得し，徐々に視野外の定位能力を獲得してゆくことがわかる．聴覚モジュールの学習回数 (図 7 内の iter_A) を見ても，音源提示回数 100 回での聴覚モジュールの学習回数は 25 回であり，音源提示回数の 25% しか学習がなされていない．これは，カメラの視野が約 43 °であり音源の提示範囲が 180 °である為，およそ 25%しか音源が視野に入らない事によるものである．しかし，音源提示回数 500 回では聴覚モジュールの学習回数は音源提示回数のおよそ半分程度になっており，学習が進むにつれて音源を視野内に捉えられる割合が増えている事が分かる．

また，図 7 の 800 回の音源提示後のグラフから，± 60 °を超えた部分において精度が悪いことがわかる．この原因は，時間差 Δt の分解能が粗いことによるものである．サンプリングポイント数で表される時間差 Δt は ± 42 の範囲の値を取るが，60 °では 39，75 °では 41，90 °では 42 となり，両端ではかなり粗いものとなってしまう．この為，割り当てられるニューロン数が少なくなってしまい，十分な精度が得られないと考えられる．

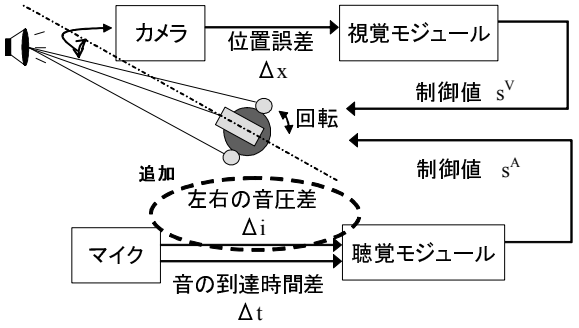


Figure 8: 変更された音源定位モデル

4 定位能力の向上

4.1 音圧差情報の利用

聴覚モジュールの入力に，時間差情報のみを用いた場合，分解能が低い為定位精度が悪い事が実験により明らかとなった．また両耳間到達時間差の情報のみでは，音源までの距離が変化する場合には，時間差と回転角度の関係が 1 対 1 とはならない為，学習が行えない．そこで，聴覚モジュールの入力として，左右の音圧差の情報を追加し，定位能力を向上させる改良を行った．

4.2 モデルの変更

聴覚モジュールの入力に音圧差を加えたモデルを図 8 に示す．マイクから得られた情報から音圧差を求めて，その値を聴覚モジュールへ入力する．入力として用いる音圧差 (Δi) の導出には，式 (5) を用いることとする．

$$\text{音圧差} = \log\left(\frac{\text{右の音の振幅の 2 乗和}}{\text{左の音の振幅の 2 乗和}}\right) \tag{5}$$

また，聴覚モジュールへの入力が 1 次元から 2 次元に増えた事に伴い，モジュールのネットーワーク構成及び学習則も変更する必要がある．変更されたネットーワークを図 9 に示す．また，学習則は式 (6) のように変更した．

$$\begin{cases} w_p^t & \leftarrow w_p^t + \epsilon^t \cdot e^{-\sigma^t (p-k)^2} \cdot (\Delta t - w_p^t) \\ w_q^i & \leftarrow w_q^i + \epsilon^i \cdot e^{-\sigma^i (q-l)^2} \cdot (\Delta i - w_q^i) \\ s_{pq}^A & \leftarrow s_{pq}^A + \epsilon^A \cdot e^{-\sigma^t (p-k)^2 - \sigma^i (q-l)^2} \cdot s^V \\ k & = \arg\min_{\forall p} \|w_p^t - \Delta t\| \\ l & = \arg\min_{\forall q} \|w_q^i - \Delta i\| \end{cases} \tag{6}$$

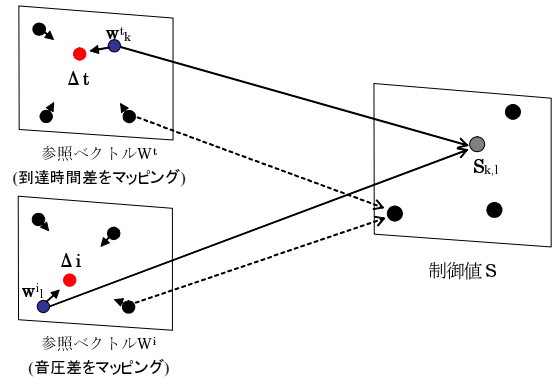


Figure 9: 聴覚モジュールのネットワーク構成

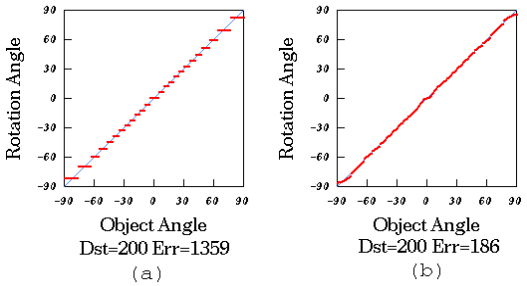


Figure 10: 聴覚モジュールによる定位結果 (a) 時間差のみ (b) 時間差＋音圧差（横軸:音源の位置，縦軸:首の回転角，Dst:音源からロボットの回転中心までの距離 (cm)，Err:音源方向と回転角度の誤差の 2 乗和）

4.3 マイク-音源間距離が一定の場合

まず，精度の向上を調べる為のシミュレーションを行った．図 10 に，音源からロボットまでの距離を 2m に固定して学習を行った結果を示す．図の左が聴覚モジュールの入力に時間差のみを用いた場合，右が時間差と音圧差２つの入力を用いた場合の定位結果を示している．入力が時間差のみの場合には，獲得される入力値の不連続性から定位結果も不連続となり誤差 (図中:Err) も大きいが，時間差と音圧差を用いた場合には，定位結果はほぼ理想値と一致し，誤差もかなり小さくなっていることがわかる．

4.4 マイク-音源間距離が可変の場合

次に，音圧差を加える事によって，音源までの距離の変化にロバストになる事を調べる為のシミュレーション実験を行った．実験は，入力が時間差のみの場合と，時間差＋音圧差の場合それぞれについて４回行った．また，音源までの距離は 50cm〜500cm までランダムに変化させた．

図 11 に学習後の誤差の分布を示す．入力が時間差のみの場合には，距離が 100cm 以内の近い所で誤差が大きくなっているが，音圧差を加えた場合には，誤差は距離に依らずほぼ一定であることが分かる．

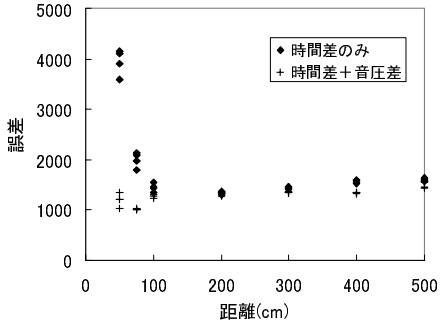


Figure 11: 誤差分布の比較（横軸:音源からロボットの回転中心までの距離 (cm)，縦軸:音源方向と回転角度の誤差の 2 乗和）

5 視聴覚情報の統合

5.1 感覚系と運動系の分離

現在のモデルでは，視聴覚情報を同時にシステムに入力する事が出来ない．さらに，視聴覚の感覚情報からダイレクトに制御値の計算を行っている．生体が音源定位を行う場合において，感覚系と運動系の問題を一度にまとめて解いているとは考えにくい．また工学的にも，感覚系と運動系を別々に分けたモデルの方が有利である．なぜなら，例えば左右のマイクの距離が変化した場合には，感覚系のみを再学習すればよいからである．

これらの理由から，音源方向を表すような内部表現[入江 and 川人, 1990]を用いて，モデルを感覚処理部と運動処理部に分割し，さらに視聴覚情報を統合するシステムに変更を行う．このような感覚情報の統合は，聴覚情報から視覚情報の推定や，さらにその他の感覚情報とのインタラクションにも有効であると考えられる．

5.2 学習モデルの変更

変更されたモデルのネットワーク構成を図 12 に示す．前のモデル (図 2) に図 3 の視聴覚モジュールと同様のネットワーク構成を持つ制御モジュールが追加されている．視聴覚モジュールからの出力を内部表現（中間コード）とし，その値を制御モジュールへの入力とする．制御モジュールからの出力は，聴覚モジュールからの中間コード (I_a) に対応した参照ベクトル (w_a^s) と視覚モジュールからの中間コード (I_v) に対応した参照ベクトル (w_v^s) の中央のニューロンに対応した制御値 s となる．

学習は，2.3 で述べたのと同様の方法を用いる．すなわち，視覚系からの情報 (I_v, s_v) を用いて聴覚系のパラメータ (I_a, s_a) の学習を行う．但し，I_v 及び I_a は教師として特定の値が与えられないので，それぞれの最大値・最小値に対して，教師をそれぞれ 5,-5 として学習を行って，値の範囲の拡大操作を行う．

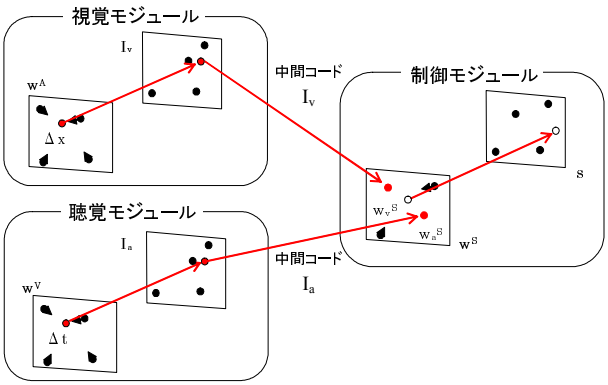


Figure 12: ネットワーク構成

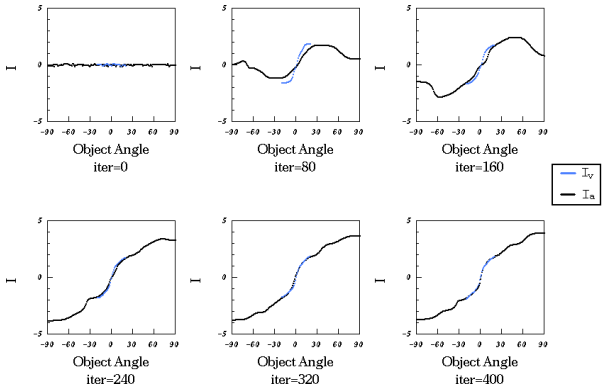


Figure 13: 中間コードの学習による推移

5.3 シミュレーション

変更されたモデルの検証を行う為に，シミュレーションを行った．実験環境は 3.2 と同様とし，視覚・聴覚・制御モジュールのニューロン数はそれぞれ 80 個とした．

シミュレーションの結果，前のモデルと同程度の音源定位能力は獲得された．視聴覚モジュールの出力である中間コードの学習課程を図 13 に示す．学習初期には，視覚モジュールの出力と聴覚モジュールの出力は異なっているが，学習が進むにつれて一致するようになっている．また，音源角度と 1 対 1 の関係が徐々に獲得されて行っていることがわかる．

6 まとめ

音源定位能力を学習するシステムを構築し，モデルの有効性を示す為の実験を行い，さらに能力を向上させる為のモデルの拡張についての検討を行った．実験結果から，システムはカメラやマイク，音源までの距離等の環境に関する情報及び，外部からの明示的な教師信号もなしで，自ら動作する事によって獲得した感覚情報を用いて学習を行い，音源定位能力及び視覚定位能力を獲得出来る事が示された．さらに，聴覚モジュールへの入力に音圧差を追加することにより，精度の向上及び音源距離の変化へ対応した．また，視聴覚情報の統合についても検討を行い，シミュレーションによって内部表現が獲得されることを示した．

現在のシステムでは，上下方向の音源定位は行う事が出来ないので，今後は，聴覚モジュールへの入力を検討し，上下左右の音源定位を可能とするようなシステムへの拡張を行いたい．さらに，本システムを移動ロボットに搭載し，音源を追跡ロボットの構築も行っていきたい．

参考文献

[Irie, 1990] Robert E. Irie. Multimodal Sensory Integration for Localization in a Humanoid Robot. In *IJCAI-97*, pages 55–59, 1990.

[Knudsen and Knudsen, 1985] Eric I. Knudsen and Phyllis F. Knudsen. Vision Guides the Adjustment of Auditory Localization in Young Barn Owls. *SCIENCE*, 230:545–548, 1985.

[Kuniyoshi and Berthouze, 1998] Yasuo Kuniyoshi and Luc Berthouze. Neural learning of embodied interaction dynamics. *Neural Networks*, 11:1261–1276, 1998.

[Kuperstein, 1988] Michael Kuperstein. Neural Model of Adaptive Hand-Eye Coodination for Single Postures. *SCIENCE*, 239:1308–1311, 1988.

[Miller *et al.*, 1990] W. Thomas Miller, Richard S. Sutton, and Paul J. Werbos. *Neural Networks for Control*. Mit Press, 1990.

[Neisser, 1976] Neisser. *Cognition and reality: Principles and implications of cognitive psychology*. W.H.Freeman, 1976.

[浅井 *et al.*, 2000] 浅井 恭行, 中島 弘道, 山村 毅, 黄 捷, and 大西 昇. 運動を介した視聴覚による物体定位能力の獲得. **電気情報通信学会論文誌**, J83-DII(7):1676–1684, 2000.

[吉田 and 亀田, 1980] 吉田 登美男 and 亀田 和夫. **新版聴覚と音声**, pages 73–240. 電子情報通信学会, 1980.

[入江 and 川人, 1990] 入江 文平 and 川人 光男. 多層パーセプトロンによる内部表現の獲得. **電子情報通信学会論文誌**, J73-DII(8):1173–1178, 1990.

[寺西 *et al.*, 1969] 寺西 立年, 竹川 忠男, and 中田 和男. **感覚・知覚心理学ハンドブック**, pages 684–724. 誠信書房, 1969.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-8

頭部の３次元運動に追従するダミーヘッドシステム － テレヘッド －
The *TeleHead* : A dummy head that tracks 3D head movement

平原達也、戸嶋巌樹、植松 尚 *
Tatsuya HIRAHARA, Iwaki TOSHIMA, Hisashi UEMATSU *

日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories,　NTT Corporation
hirahara@idea.brl.ntt.co.jp, toshima@avg.brl.ntt.co.jp, uematsu.hisashi@lab.ntt.co.jp

Abstract

Auditory functions are closely related to head size and head movement. To seek a comprehensive understanding of the nature of the auditory information processing mechanisms including head size and head movement, we are developing a remote dummy head which tracks 3- dimensional human head movement in real-time, named "*TeleHead*". This paper describes design concept and architecture of the *TeleHead*. The tracking and acoustic characteristics of the *TeleHead-I* are also described.

１．はじめに

感覚系の役割は、いわゆる五感センサーを通じて外界から取り込んだ信号の断片から、自分の周りの状況を把握したり相手が伝えようとしているメッセージを解読するための脳内情報を、その場の状況とつじつまが合うように頑健かつ迅速に創りだし、自分が次に取るべき行動を決める一助とすることにある。そして、この脳内情報は、自らが生活する世界の物理的な環境と自らの身体という物理的な形状とその運動と密接な関係を持つ。

聴覚系は、音を通じて身に迫る危険を察知して回避行動を起こしたり、餌を見つけたり、異性を誘引して繁殖のためのプロトコルを確認したりするために発達してきた遠方感覚系である。聴覚系の仕事としてよく知られているのは、音がどの方向から来たのかを判断する音源定位機能と、それが何の音であるかを判断する音源識別機能の二つである。しかし、聴覚系はこの他にも、コミュニケーション音を処理・制御する機能、新奇な物音に対して自動的に注意を向ける早期警報機能、そして情動系の賦活機能を持つ[1][2]。これら聴覚系の諸機能も、頭部という身体形状とその運動と無縁ではない。

聴覚系は、音の時間差と音圧差とスペクトル差を元にして、音の到来方向を計算している[3]。左右の耳が受ける音信号の差は両耳の間にある頭部と耳介の形状によって形成されるため、音源定位の処理系は、自分の頭部と耳介の形状にチューンされている。身体の形状は個体ごとに微妙に異なるために、他人の頭部と耳介を使うことを強いられると、しばらくの間は混乱をきたすであろう。また、自前のものと全く異なる形状の頭部と耳介、例えば象のような大きな頭部と耳介あるいはネズミのような小さな頭部と耳介を使うことを強いられると、音源定位はできなくなるであろう。このように、聴覚は身体の形状と密接な関係を持つ。

私たちは、実環境の中で音の到来方向を探る場合に頭部や体を動かすし、小型哺乳類の多くは頭部だけでなく耳介を動かして捕食者や餌が出す音の到来方向を探る。小型哺乳類の多くでは、聴覚系の２次ニューロンである蝸牛神経核背側核（DCN: Dosal Cochlear Nucleus）に耳介の体性感覚ニューロンが投射されていて、高域にあるスペクトルの谷を検出して耳介の向きを補正している[4][5]。耳介を随意に動かせないヒトではこのDCNの神経回路は退化している[6]。

頭部運動と音源定位機能の関係についてはH. Wallach (1939)[7]以来いくつかの報告があり、頭部回転は水平面内の音像定位精度を上げることが知られている[8]-[14]。我々も、仮想音源呈示システムを用いた音像定位実験を行い[15][16]、自発的な頭部運動が音像定位精度の向上に貢献することを確認するとともに、自発的な頭部運動を許すとHRTFのスペクトルを鈍らせても音像定位精度は低下しないが、強制的な頭部運動は音像定位精度を低下させることを明らかにした。さらに、我々は、聴覚系には音像の位置だけでなく音像の移動速度の知覚が順応の影響を受けること、すなわち速度残効が存在することを発見し、音像の移動速度を処理するモジュールが聴覚系に存在することも明らかにした[17]。このように、聴覚は身体の運動とも密接な関係を持つ。

そこで、人間と同等もしくは人間を超える聴覚機能を機械に与えたり、聴覚が本来持つ諸機能を有効に利用できる情報通信技術を創り出すためには、頭部形状と頭部運動を含めた聴覚系の総合的な理解が必要となってくる。本稿では、このような聴覚系の総合的な理解を深めるための道具として、現在作成しつつある、頭部の３次元運動に追従するダミーヘッドシステム－ テレヘッド (TeleHead) － について述べる。

2. 頭部形状と頭部運動を伝える方法

頭部形状を折り込み頭部運動を聴こえに反映させる方法としては、HRTF技術を利用するバーチャルリアリティ方式とバイノーラル技術を利用するテレロボティクス方式とが考えられる。本章では両方式の長所と短所について比較する。

2.1 バーチャルリアリティ方式

バーチャルリアリティ（ＶＲ）方式とは、所望する音場を頭部の周囲に物理的に再現してそこで頭部を自由に動かすことができるようにしたり、所望する音場を物理的に再現せずに頭部運動を反映させた音の聴こえを刻々と創りだす方式である。すなわち、信号処理技術を駆使して受聴者の頭部の動きに応じて音源の方向と頭部の位置関係を計算し、マイクロフォンで拾った音響信号に刻々と適切なHRTFを畳み込んで受聴者に呈示する方法である（図１）。E.Wenzelらは1988年代後半にＶＲシステム用の３次元音響ディスプレイとして、頭部運動に同期させてHRTFをリアルタイムで変化させるシステムを開発している[18]。その後、様々な方式の３次元音響ディスプレイや音場再生技術の研究が行われている[19][20][21]。

HRTF技術を利用したＶＲ方式は、受聴者側で全ての処理を電子的に行えるという利点があるが、HRTFにかかわる諸問題から逃れられない。第一に、あらかじめ音源の方向が分かっていないと畳み込むHRTFを選択することができない。そのため、複数の音源がありそれらが動いている状況では処理が複雑になる。この問題は、人工的な音環境を生成する場合はともかくとして、実環境を再現する場合には致命的である。第二に、受聴者の各方向のHRTFをあらかじめ計測しておかなければならない。例えば、方位・高さ方向共に5度間隔で有効な位置全てのHRTFを用意しようとすると、1945組ものHRTFが必要となる。第三に、受聴者ごとに頭部と耳介の形状は異なるために、各人のHRTFを用意しておかなければならない。他人のHRTFを用いると、音像の位置がずれて、定位精度が落ちる。第四に、音像の遠近は主として音圧の差でしか表せないために、近くで鳴る小さな音と遠くで鳴る大きな音の区別をつけにくく、距離の制御が困難である。HRTF をコンパクトに表現する方法[22]、HRTFの個人差を解消する方法[23]、HRTFを頭部形状から算出する方法[24]、距離感を制御する方法[25]などが検討されてはいるが、いずれも未解決問題として残されている。

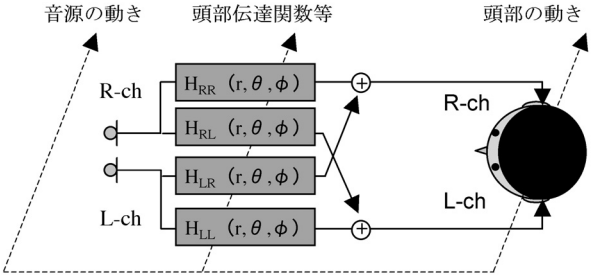


図１　ヴァーチャルリアリティ方式の一例

2.2 テレロボティクス方式

テレロボティクス方式とは、可動な頭部を含めた耳を所望する音場に置く方法である。即ち、受聴者の頭部運動に追従するダミーヘッドで収録した音響信号を、バイノーラル技術[19][26][27]を用いて受聴者の耳に呈示する方式である(図２)。この方法は、ダミーヘッドという実体を作成し遠隔操作で動かす必要があり、駆動に伴う騒音や駆動系の時間遅れなど物理的な機械系の問題がある。

しかし、ダミーヘッドという実体を用いることにより、上述したHRTFにかかわる諸問題から逃れることができる。つまり、ダミーヘッドの形状と駆動系そして音響信号伝達系を適切に設計することにより、ＶＲ方式より容易に、ダミーヘッドを置いた場所の音響空間を受聴者の耳に再現できる。また、１章で述べたように、頭部を自発的に動かせる場合には頭部伝達関数を精密に再現しなくとも音像定位精度が悪化しない。つまり、受聴者の頭部形状と一致していないダミーヘッドを用いても、受聴者の頭部運動に追従させられれば、ダミーヘッドが置かれた場所の音響空間を受聴者の耳に再現できる可能性もある。

以下の章では、このテレロボティクス方式を用いたテレヘッドの実現を目指していくつかの点を検討する。

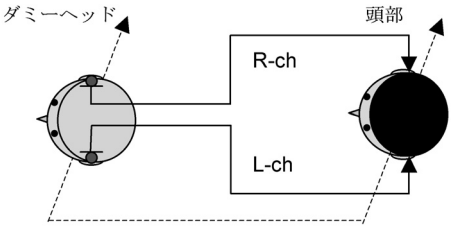


図２　テレロボティクス方式の一例

3. 人間の頭部運動

本章では、人間の頭部運動機構を概説するとともに、頭部の可動範囲と頭部の運動速度を計測した結果について述べる。

3.1 頭部の回旋運動機構

人間の首は上位頚椎と下位頚椎の２つから成っている。上位頚椎は 3 軸 3 自由度で下位頚椎の動きを補償し，目標となる運動を生成するための調整を行う。下位頚椎は屈曲・伸展運動及び，回旋運動を行うが、回旋によって側屈は一意に決定されるため屈曲・伸展とあわせて 2 自由度をもつ。

それぞれの頚椎には複数の筋肉が接続されている。椎椎の前部にある筋群は，様々な角度で斜めに頚椎の各骨を引っぱるように筋が配置されている。これらの筋が同時に収縮すると屈曲が起こり，片側だけ収縮するとそちらの方向に側屈が起こる。一方、頚椎の後部にある筋群は下内側後方へ伸びており、伸展・側屈・回旋の複合運動を行う。一部の筋は頚椎に巻きついており、これらの収縮によって回旋運動が起こる。[28]

3.2 頭部の可動範囲

頚椎の可動域については、屈曲・伸展(Pitch)方向は 130°、側屈方向(Roll)は片側 45°、回旋(Yaw)方向は片側 80°程度と言われている。

実際に、被験者４名（20-30 歳代の男女各2名）で頭部の最大回転角度をヘッドトラッカー(Polhemus FASTRAK)を用いて計測した結果、屈曲方向は47°（σ=8.7°）伸展方向は63°（σ=17°）合わせて110°、側屈方向は44°（σ=7.3°）回旋方向は片側69°（σ=12°）であった。

3.3 頭部運動速度

頭部の運動速度に関する既存のデータが見つからなかったために、ヘッドトラッカーを用いて頭部の水平回旋運動速度を計測した。まず、被験者は椅子に座り､正面に置かれた LED に顔を向ける。正面の LED が 0.5 s 点灯の後、被験者正面から右 60° あるいは左 60° に置かれた LED が点灯し、正面の LED は消灯する。被験者には、点灯した LED が顔の正面となるように、かつ、できるだけ早く頭部を回転するよう教示した。

20ー30歳代の男性3名、女性2名に対して左右方向それぞれについて10回計測した。その結果、正面から左右60° 方向へ水平回旋運動を行う際の最大回旋運動速度は382°/s（σ=72°/s)、加速度は5644°/s^2（σ=1180°/s^2)であった。別途、通常の音像定位を行う際の頭部運動速度も予備的に計測したが、その値は100°/sー200°/sであった。

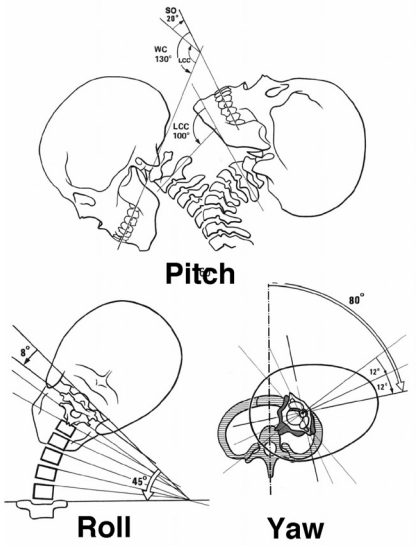


図３　人間の首の機構と頭部運動の範囲

3.4 頭部運動機構に関する考察

このように、人間の頭部運動を実現しているメカニズムは巧妙であり、動作音を発生せずに、重い頭部を高速に運動させることができる。特に、頭部回転運動に伴う動作音が発生しない点は、驚嘆に値する。

また、3.1 では触れなかったが、人間は頭部を回旋させずに前後・左右・上下に移動させることもできるが、多くの場合、肩や胴の運動を伴う。さらに、頭部回旋運動の中心は、運動に伴って移動すると思われる。

4. 頭部形状を再現するダミーヘッド

本章では、テレヘッドに不可欠な頭部形状を精密に再現するダミーヘッドについて述べる[29]。市販のダミーヘッドは音響計測やバイノーラル録音を目的として、多数の頭部形状の平均値に基づいて作成されているし、運動させるには重すぎる。テレヘッドに求められるダミーヘッドは受聴者の頭部形状の再現であり、可動性である。

4.1 ダミーヘッドの構造と作成方法

男性1名、女性1名について、型取りをすることによって頭部の形状を精密に再現するダミーヘッドを作成した。このダミーヘッドは3層構造になっている。すなわち、FRP樹脂で作成した中空の骨格相当部分に発砲ポリウレタンを約2cm緩衝材として付着させ、その上に厚さ1 mmの軟質ウレタンの表面形状を張り付けてある。この頭部の表面形状は、石膏を用いて取った型に軟質ウレタン樹脂を流し込んで再現した。また、凹凸の多い耳介部分については別途石膏による型取りを行い、塩化ビ

ニルゾル素材で成型し、頭部に装着した。ダミーヘッドの重量は約1kgである。

4.2 頭部形状の再現精度

作成したダミーヘッドの精度を評価するために、３次元形状計測用レンジファインダ（Danae-R, NEC）を用いて、実頭とダミーヘッドの形状比較を行った。３次元の位置計測における計測密度は1.3mm、計測精度は0.5mm、計測点数は15万点（300×500点）である。

図4に示す男性の実頭とダミーヘッドの形状の差を比較すると、目元付近に4mm程度、両耳前方のもみあげ付近で7mm程度の差違があった。また後頭部に大きな差異が認められたが、これは頭髪の影響である。耳介部分は概ね数mm以内の差異に収まっているが、部分的に5mmを越える差異がある箇所もあった。また、窪み部分は光計測ができないために比較もできなかった。さらに、ダミーヘッドの耳介は実頭の耳介よりも立っていて、その部分の差異は17mm以上もあった。女性のダミーヘッドについても同様の結果であった。

4.3 頭部形状の再現に関する考察

このように、作成したダミーヘッドの形状と実頭の形状とは最大で17mm相違した。型取りを行ったにも係わらず、このような形状の差異が生じた原因としては、型取り精度、ダミーヘッドの工作精度、頭髪の影響、などが考えられる。石膏や印象材を顔面に付着させると、重力の影響で柔らかな皮膚の部分が変形する。また、頭髪の部分はストッキングや薄いゴム膜で押さえつけて型取りをしなくてはならないので、その部分が盛り上がることは避けられない。光学的な３次元計測法を用いる場合でも、頭髪部分の計測は困難である。

頭部形状の計測方法としてはMRIを用いる方法もある。、ただし、多くのMRIでは横臥してヘッドバンドによって額部分を固定して撮像を行う。そのために、重力の影響が立位と異なり、測定した頭部形状は光学的な３次元計測結果と異なってくる。

なお、頭部形状を再現するダミーヘッドの作成方法としては、MRIや光計測をして得られた３次元形状データを元にして、光造形などの手法を用いて成型する方法もある。

このように、頭部の３次元形状を精度よく再現するためには更なるアイディアが必要である。また、ダミーヘッドにどの程度の再現精度が必要かということについては、知覚的な観点からの精査が必要である。

5. 実頭と疑似頭の頭部伝達関数

5.1 HRTFの測定方法

ダミーヘッドとそのモデルとなった被験者のHRTFを無響室（4.8 m×5.4 m×4.7 m）内で測定した。被験者は無響室内の椅子に座った状態で測定し、疑似頭を B&K のヘッドアンドトルソーの首から下の部分に装着して、衣服を着せ、椅子に置いた状態で測定した。

頭部の中心位置から音源までの距離は 1.5 m とした。測定点は、水平方向には全周を 10° 間隔、仰角方向には−40° から 80° までを 10° 間隔、合計 468 点である。音源は最適化時間引き延ばしパルス[30][31]を用い、各測定点について標本化周波数 48 kHz、512 点のインパルス応答を求めた。

ダミーヘッドの HRTF 測定に用いたマイクロホンは ECM-77B (SONY)で、実頭の HRTF 測定には UC-92H (RION)を用いた。それぞれの HRTF は、別途測定したマイクロホンの周波数特性の逆特性を用いて補正した。ダミーヘッドにおけるマイクロホンの設置位置は、外耳道入り口から 2 mm 奥で、外耳道はマイクロホンによって完全に塞がれている。一方、実頭におけるマイクロホンの設置位置は外耳道入り口から 2 mm 奥であるが、マイクロホンの外径が外耳道の外径より小さいため、外耳道は完全には塞がれていない。

図４　モデル（左）とその頭部形状を再現したダミーヘッド（中）の一例。右はダミーヘッドの内部構造。FRP樹脂の骨格部分と発砲ポリウレタンの緩衝材と軟質ウレタンの表面の三層構造を成す。下の方に見えるのは、駆動機構との接続治具である。

5.2 HRTF の比較結果

右耳の仰角0°で前方0°40°80°と後方180°，220°，260°の HRTF の周波数特性を図５に示す。左列は男性、右列は女性のデータで、濃い細線はダミーヘッド、薄い太線は実頭の HRTF を示す。

前方の HRTF については、いずれの被験者でも、概ね 1.5 kHz までは、伝達関数の相対的なレベル差は±5dB 以内で類似したものとなっていた。それ以上の帯域では、実頭の結果に認められる 6kHz のピークは再現されているが、さらに高域のピークやディップの位置と深さが異なっていた。一方、後方の HRTF では、3 kHz 以上の帯域のピーク・ディップの位置が異なり、伝達関数の相対レベルも大きく異なった。特に、女性のダミーヘッドでは実頭の HRTF に認められる 2kHz，3.8 kHz，5 kHz のディップがダミーヘッドでは再現されていない。

5.3 HRTF の差異の原因

実頭とその形状を模擬したダミーヘッドでこのような HRTF の差異が生じた原因としては、以下のようなことが考えられる。

まず、型取りしたにも係わらず、でき上がったダミーヘッドの形状が、被験者自身のものと多少異なっていたことがあげられる。特に、前述したように、ダミーヘッドの耳介が実頭よりも立っている。また、耳甲介腔（コンチャ）の形状もダミーヘッドと実耳とでは異なっている。

次に、HRTF 測定時のマイクロホンの設置条件の違いがあげられる。つまり、マイクロホンが外耳道を塞いでいるか否かの差であり、実頭ではマイクロフォンから後ろ側の音響インピーダンスが含まれている。

さらに、軟質ウレタンでできているダミーヘッド表面と皮膚の音響インピーダンスの違いや、首から下の胴の形状の差も可能性としてあげられる。これらが HRTF の周波数特性に及ぼす影響については、更なる検討が必要である。

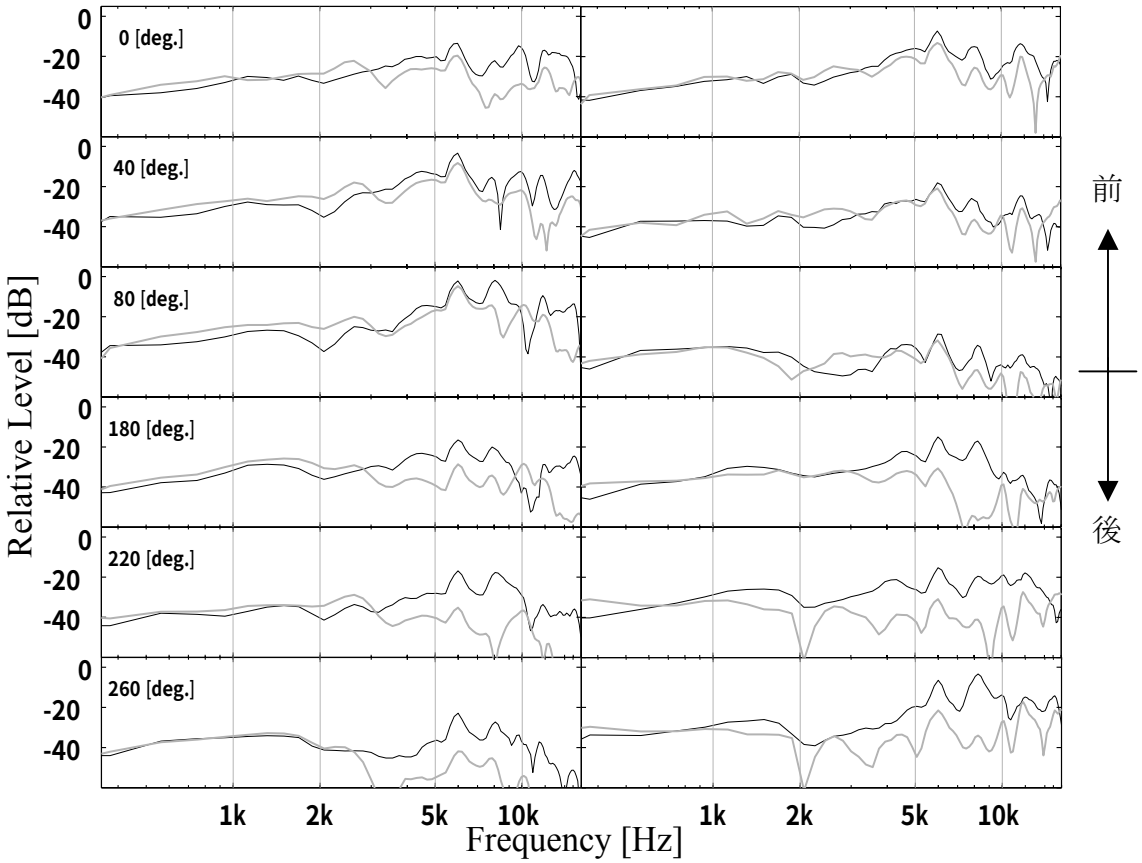


図５　ダミーヘッドと被験者の実頭で測定した HRTF の周波数特性。左列は男性、右列は女性のデータ。濃い細線はダミーヘッド、薄い太線は被験者の実頭の HRTF をそれぞれ示す。

6. テレヘッド１号機の駆動機構と諸特性

テレヘッドは、頭部姿勢検出部、ダミーヘッド部とその３次元運動駆動部、音響信号伝達部で構成される。その概要を図６に示す。本章では、前述したダミーヘッドを人間の頭部の３次元運動に追従させるための駆動機構とその追従特性、および動作騒音特性について述べる［32］。

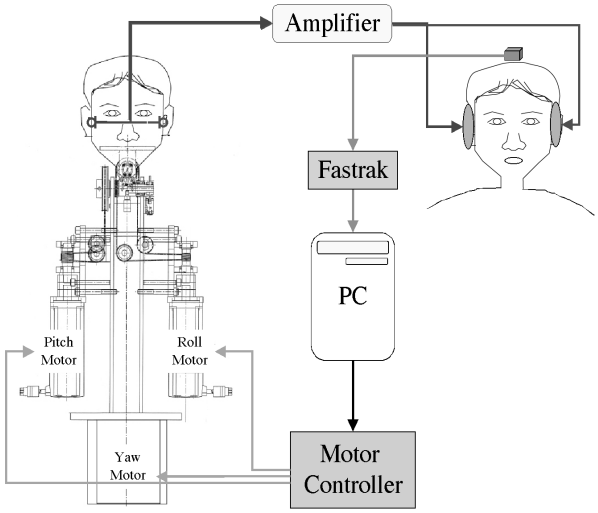


図６　テレヘッド１号機の概略

6.1　頭部姿勢検出部

頭部の姿勢は，頭頂に装着した 3 次元位置姿勢センサーFASTRAK（Polhmus）を用いてサンプリング周波数 120Hz で検出する。そして、この頭頂の姿勢（オイラー角）は RS232C を通じて PC に取り込み、駆動部へと送られる。

6.2　３次元運動駆動部

ダミーヘッドは球面関節に取り付け、屈曲・伸展(Pitch)方向と側屈(Roll)方向はそれぞれ AC サーボモータ（400W）を用いてワイヤーとプーリで駆動する。駆動に伴う機械的な雑音の発生を抑えるために、ギヤは使用していない。回旋(Yaw)方向については、ダミーヘッドと屈曲・伸展方向と側屈方向の駆動機構を含めた全体を、DD サーボモータ（トルク 2.1Nm，120rpm 時）で、上部の機構全体を直接駆動する。

ダミーヘッドの可動域は，屈曲方向 54°、伸展方向 26°、側屈方向片側 30°、回旋方向片側 90°とした。回旋方向を除いて人間の可動域よりも狭いが、これは機構上の制約による。また、人間の頭部運動の回転中心は運動によって移動するが、今回は、ダミーヘッド内部に固定した。なお、駆動機構部分は防音のため着脱式のウレタンフォームの胴体で覆ってある。

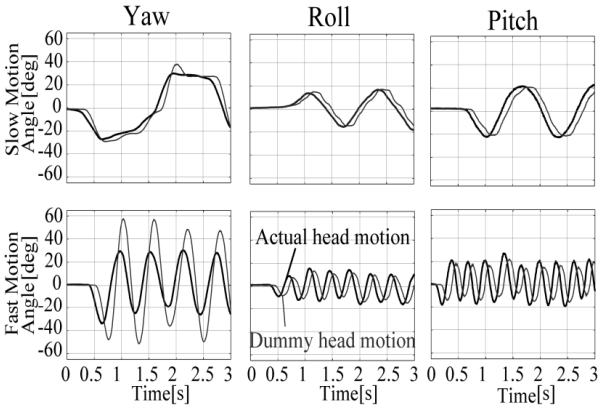


図７　テレヘッド１号機の運動追従特性

6.3 頭部運動追従特性

テレヘッド 1 号機を人間の頭部運動に追従させた場合の追従特性の一例を図７に示す。左列は回旋（Yaw）、中列は側屈（Roll）、右列は屈曲・伸展（Pitch）運動である。それぞれ、上段は最大速度が 100[deg/s]から 200[deg/s]程度の遅い運動、下段は同じく 300[deg/s]から 500[deg/s]の速い運動を示す。図中の太線は人間の動作、細線がダミーヘッドの追従動作を表している。

遅い運動に対しては 200ms 程度の遅延はあるものの、いずれの方向の運動も比較的良好な追従運動が実現されている。一方、速い運動に対しては、屈曲・伸展方向と側屈方向では遅延が大きく、回旋方向の運動では目標に対して最大 70%程度オーバーシュートする。これは屈曲・伸展・側屈方向と回旋方向とでは駆動系が異なるために生じた差異と考えられる。駆動系の遅延および運動の動特性については満足の行くものではなく、改善が必要である。

6.4 音響信号伝達部

ダミーヘッドには 2 つのマイクロフォン ECM77B（Sony）が内蔵されており、ダミーヘッドの外耳道入り口から 2mm 奥での音を拾う。音響信号はオーディオアンプを経てヘッドフォンを介して受聴者の耳に導かれる。このヘッドフォンは、何を使ってもいいわけではなく、受音点での音響条件を再現する必要がある。今回は、挿入型イヤフォン ER4（Etymotic Research）と密閉型の耳覆い型ヘッドホン HDA200（Sennheiser)[33]を使用したが、ヘッドホンを通じてバイノーラル信号を忠実に再生するためには、受音点からヘッドホンを見込んだ音響インピーダンスがヘッドホンを装着していない場合とどれだけ近いかが重要といわれている[34][35]。

6.5　動作騒音特性

図８にテレヘッド１号機の前方0.5mにおける外部放射騒音特性を示す。破線は防音室の暗騒音レベル、細線は静止時の騒音レベル、太線は動作時の騒音レベルである。同図に示されるように、動作時に 1kHz 近傍の騒音レベルが増加するが、最大でも 40dBSPL 程度であり、外部放射騒音レベルは低く抑えられている。なお，防音カバーを外すと 2-4kHz の騒音レベルは 10-15dB ほど増加する。

図９にダミーヘッドに装着したマイクロフォンへのライン混入騒音特性を示す。混入騒音レベルは、マイクロフォンからヘッドフォン HDA200（Sennheiser）[(5)]までの音響系のゲインを 1kHz の音圧で校正した後に、ヘッドフォンを IEC カップラに装着して測定した。ダミーヘッドの頭部運動追従動作に伴って 0.3 - 2kHz の帯域の音圧レベルが増加し，ラインに混入する騒音音圧は 0.2kHz で 70dBSPL にも達している。これは，AC サーボモータの振動とプーリの回転に伴う振動が、ダミーヘッドに装着したマイクロフォンのダイヤフラムを振動させているものと考えられる。機械的な振動の低減と、マイクロフォンの取り付け方などの根本的な振動と騒音の低減対策が必要である。

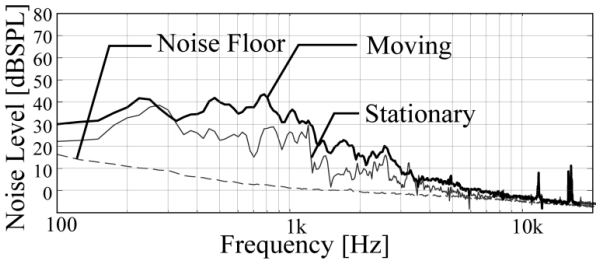


図８：外部放射騒音特性

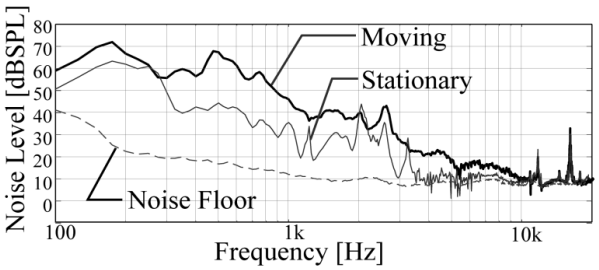


図９：ライン混入騒音特性

6.6　テレヘッドを通じた聴こえ

テレヘッドを用いた実環境での聴こえ方については予備実験的に確認したに留まっている。ダミーヘッドのモデルは、これまでにない音場のリアルさがあるとの内観報告している。また、モデル以外の人々でも、テレヘッドを頭部運動に追従させた場合には、静止時には聞き分けられなかった近接した２つの音源位置をはっきりと分離できるようになることを確認している。テレヘッドの知覚的な性能評価に関しては今後の課題である。

7. おわりに

本稿では、頭部形状と頭部運動を含めた聴覚系の総合的な理解を深めるための道具として作成しつつある、頭部３次元運動に追従するダミーヘッドシステム- テレヘッド- について述べてきた。今回紹介したテレヘッド 1 号機は、いわば問題を洗い出すためのテストベッドであり、各章で述べてきたように改善すべき点が多々ある。今後は、これらの解決に向けた諸検討を進め、次期テレヘッドの設計指針を明確にする予定である。

また、情報通信技術の進展により、高速・広帯域ネットワークが張り巡らされ、誰もが容易に利用できるようになりつつある。つまり、これまでのテレコミュニケーション技術を束縛していた帯域の制約は緩くなり、脳が産み出し処理することができる「情報」を制約せずに遠方とやり取りする環境が整ってきた。その場合、利用可能な帯域が広くなるからといって音や影像のベースバンドの伝送帯域を単純に拡張したり、複数のメディアを無節操に組み合わせて送り込もうとするのではなく、私たちの脳の情報処理の仕組みと処理特性のツボを押さえた帯域の利用を考える必要がある。たとえば、これまで見過ごされてきた頭部運動情報をどのように伝え合うかは重要な観点であろう。

参考文献

[1] 平原達也 (2002): 音を通じて世界を伺う聴覚の仕組み、日本バーチャルリアリティ学会誌 **7** (1), 23-29

[2] 平原達也 (1994): 聴覚のメカニズム, *視聴覚情報科学*, ATR視聴覚機構研究所 編, (オーム社, 東京, 1994) pp.145-200

[3] Yin,T.C.T. (2002), "Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem," in *Integrateve Functions in the Mammalian Auditory Pathway*, D.Oertel, R.R.Fay and A.N.Popper Eds., Springer-verlag New York, pp.99-159

[4] Young E.D. et al (1992)：Neural organization and responses to complex stimuli in the dosal cochlear nucleus, *Processing of Complex Sounds by the Auditory System*, Carlyon, R.P. and Darwin, C.J. & Russel, I.J. Eds, Clarendon Press, Oxford, pp.113-119

[5] Young E.D. and Davis K. (2002) Circuitry and

Function of the Dosal Cochlear Nucleus, in *Integrateve Functions in the Mammalian Auditory Pathway*, D.Oertel, R.R.Fay and A.N.Popper Eds., Springer-verlag New York, pp.160-206
[6] Moore J.K. (1987) : The human auditory brain stem: A comparative view, Hearing Research, 29, 1-32
[7] Wallach H. (1939) : "On sound localization," *J. Acoust. Soc. Am.* **10**, 270-274
[8] Wallach H. (1940): "The role of head movements and vestibular and visual cues in sound localization," J.Exp.Psychol., **27**, 339-368
[9] Thrlow W.R. and Runge P.S. (1967) : "Effect of induced head movement in localization of direction of sound", *J. Acoust. Soc. Am.*, **42**, 480-488
[10] Thrlow W.R. and Runge P.S. (1967): "Head movements during sound localization", *J. Acoust. Soc. Am.*, **42**, 489-
[11] Freedman S.J. and Fisher H.G. (1968): "The role of the pinna in auditory localization," in *Neuropsychology of Spatially Oriented Behavior* Freedman S.J.Ed. (Dorsey Press Illinios, 1968)
[12] Perrett S. and Noble W. (1997): "The effect of head rotations on vertical plan sound locali-zation" *J. Acoust. Soc. Am.* **102**, 2325-2332
[13] Perrett S. and Noble W. (1997): "The contri-bution of head motion cues to localization of low-pass noise". *Perception & Psychophysics,* **59** (7), 1018-
[14] Wightman F. and Kistler D.J. (1999): "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, **105**, 2841-2853
[15] 植松尚,柏野牧夫,平原達也, (2001.10): "頭外音像定位における自発的な頭部回転の影響," 日本音響学会講演論文集, 501-502.
[16] 加藤正晴,植松尚,柏野牧夫,平原達也, (2001.10) : "頭部回転運動の種類が音像定位の精度に及ぼす影響," 日本音響学会講演論文集, 505-506.
[17] 植松尚,柏野牧夫,平原達也, (2001.10) : "聴覚系における速度残効," 日本音響学会講演論文集, 503-504.
[18] Wenzel E. M. (1992): "Localization in virtual acoustic displays," Presence, **1**, 80-107.
[19]三好正人 (1996) : "音場を創る", 日本音響学会誌 **52**(6), 466-469
[20]伊勢史郎 (2001) : "音場制御におけるコンピュータシミュレーションの応用," 日本音響学会誌, **57**(7), 457-462
[21] 神沼 充伸, 伊勢 史郎, 鹿野 清宏 (2000): "受聴者の頭部の動きを考慮した多チャンネル音場再現システム", 日本バーチャルリアリティ学会論文誌, **5** (3), 957-964,
[22] 西野隆典、梶田将司、武田一哉、板倉文忠(2001): "水平方向及び仰角方向に関する頭部伝達関数の補完, " 日本音響学会誌 **57**, 685-692
[23] 西野隆典、梶田将司、武田一哉、板倉文忠(2001): "重回帰分析に基づく頭部伝達関数の推定, " 電子情報通信学会論文誌 **J84-A**, 260-268
[24] 大谷真、伊勢史郎 (2002.10): "境界要素法により求めた頭部伝達関数の妥当性及び計算の高速化に関する検討," 日本音響学会講演論文集, 585-586
[25] 金海永, 鈴木陽一, 高根昭一, 小澤賢司, 曽根敏夫 (1999): "絶対判断と相対判断による音像距離知覚の比較," 日本バーチャルリアリティ学会論文誌, **4** (2), 455-460.
[26] Møller H. (1992): "Fundamentals of binaural technology," Applied Acoustics, **36**, 171-218.
[27] 稲永潔文、山田祐司、小泉博司 (1995) "頭部運動による動的頭部伝達関数を模擬したヘッドホンシステム," 信学技報 EA-94-94
[28] I.A.Kapandji 著, 荻島監訳, (1986) "カパンディ関節の生理学Ⅲ体幹・脊柱", 医歯薬出版, 東京
[29] 植松尚,平原達也 (2002.04): "頭部形状を精密に模擬したダミーヘッドの頭部伝達関数", 日本音響学会講演論文集, 467-468
[30] 鈴木陽一、浅野太、曽根敏夫 (1989): 音響系の伝達関数の模擬を巡って (その２) ,日本音響学会誌, **45** (1), 44-50
[31] Suzuki Y., Asano F, Kim H-Y., and Sone T. (1995): "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.,* **97**, 1119-1123
[32] 戸嶋巌樹, 植松尚,平原達也 (2002.10): "頭部運動に追従するダミーヘッド", 日本音響学会講演論文集, 467-468
[33]平原達也(1997):"聴覚実験に用いられるヘッドホンの物理特性", 日本音響学会誌, **53** (10), 798-806
[34] Møller H., Hammershøi D., Jensen C.B. and Sørensen M. F. (1995): "Tranfer characte-ristics of headphones measured on human ears," J. Audio Eng. Soc., **43** (4), 203-217.
[35] Møller H., Hammershøi D., Jensen C.B. and Sørensen M.F. (1995), "Design criteria for headphones," J. Audio Eng. Soc., **43** (4), 218-232.

＊ 植松 尚の現所属は NTT サイバースペース研究所
＊ Hisashi UEMATSU is currently working for NTT Cyber Space Laboratories.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-9

パワーパターンとピーク時周波数パターンを利用した環境音認識

豊田 義之、黄 捷、Yong Liu
会津大学、コンピュータ理工学部
{m5051125, j-huang, yliu}@u-aizu.ac.jp

概 要

環境音認識は知能ロボットにとって環境を認識するための重要な機能の１つである。今まで行われた環境音認識の数少ない試みは音声認識の技術をそのまま環境音認識に適用したものと言える。しかし、環境音の特徴は音声に比べて、遥かに長い時間スパンで捕らえる必要がある一方、周波数の変動は音声と比べて緩やかなものである。我々は音の長時間パワーパターンとピーク時の周波数パターンを組み合わせて、一次元の特徴ベクトルを使うことによって、環境音の時間周波数特徴を捕らえることを試みた。なお、環境音の認識方法としてはニューラルネットワークによる学習方法を用いた。

1 はじめに

環境音認識は人が周囲状況を把握し、危険を回避する上で不可欠な機能である。例えば室内にいながら雨音で雨が降り出した事が分かるし、夜道で後ろから来る足音で危険を感じることができる。また、ドアをノックする音で誰かが訪ねて来たことがわかるなど視覚情報だけでは把握できない多くの有益な情報を環境音認識から得ることが出来る。ロボットにおいても同様で、パトロールロボットやお手伝いロボットの様に実環境で作動するロボットには必要な機能であると言える [1]。

環境音認識は近年その重要性が次第に認知され、音声認識で蓄積された技術を使った環境音認識の研究がいくつか試みられた。RWC プロジェクトで研究開発用の環境音データベースが作成され、瞬時スペクトルによる認識法が提案された [2, 3]。また、音声認識の技術の応用として HMM を使った認識法が提案され、RWCP 環境音データを使った実験で成果をおさめている [4]。

これらの方法で、瞬時スペクトルによる方法は時間のパターンが使われないので、繰り返し音など、音の特徴が長時間に渡る音は認識できない。また、HMM による方法は音声認識で養われた技術を使う利点はあるが、入力データとして音のスペクトログラムが必要であるから、処理量は増える。一方、環境音は、例えば物体の衝突音、ドアのノックの音などは物体の振動に起因するものが多く、その周波数パターンは時間と共に減衰はするが、大きな変動は見られないのが一般的である。このような特徴を利用して、計算量の少ない認識方法が実現できないかと考えたのが我々の研究動機である。

本研究では、音の特徴としてのパワーパターンと周波数パターンを別々に扱うことによって入力データの量を減らし、また、音の認識には多層ニューラルネットワーク（以降では、NN と略す）を利用する方法を試みる。実験ではニューラルネットワークの組み合わせにより２種類の認識実験を行い、環境音認識に対する本提案手法の有効性を示す。

2 環境音のデータ収集と前処理

2.1 環境音データベース

環境音は音声（広義的には音声も環境音の一種類であるが）とは違って、その種類が無限にある。前述 RWCP プロジェクトとして、環境音データベース [3] が作成されているが、今回は我々のロボットの活動範囲を想定して、既成のデータベースを使う代りに、我々のターゲットとする環境音を選定し、データベースを作成した。

作成したデータベースには、環境音認識システムを搭載予定のロボットの行動範囲から考えて、ドア

の開閉音、ノック、金属衝撃音、電話の着信音などの環境音 10 種類、日本語の母音５種類を各 30 個ずつ録音し、データベースとして収録した。録音環境は多少のノイズのある大学の実験室という一般環境で、収録装置はマイクから Fostex のデジタルマルチチャンネルレコーダー VF160（その 1 チャンネルを利用）に録音し、サンプリング周波数 44.1kHZ の WAV ファイルで CD1 枚に記録されている。

表 1: データベース収録音

音の種類
ボールと床の衝突音
金属音
キーボード（単音）
キーボード（繰り返し音）
ドアが閉まる音
ドアを開ける音
施錠
ノック
照明のスイッチ
電話の着信音
母音

2.2 前処理

一般的に HMM や時間遅延 NN（TDNN）[5] による音声認識は音のスペクトログラムを必要とする。しかし環境音の特徴は音声とは異なり、単発で指数減衰する音もあれば、単音が複数回繰り返される音もあり、認識には長い時間区間を必要とする。したがって音声認識のように入力にスペクトルグラムを用いるとデータサイズが大きくなってしまう。一方、環境音の周波数は音声に比べて変動が少ない。これらの環境音の特徴から、本研究では音の時系列変動と周波数特徴を別々に抽出して組み合わせ、一次元の特徴ベクトルを作成する手法を提案する。時系列パワー変動は短区間 LPC[6] によって、周波数特性は求められたパワーパターンからピークを検出し、ピーク時の周波数をＦＦＴにより抽出する（中間処理での音のサンプリング周波数は 8kHz までおとして処理した）。この時、認識の効率をあげるために、時間パターンに区間の長さが違う 2 種類のデータを用意し（図 1 と図 2 参照）、

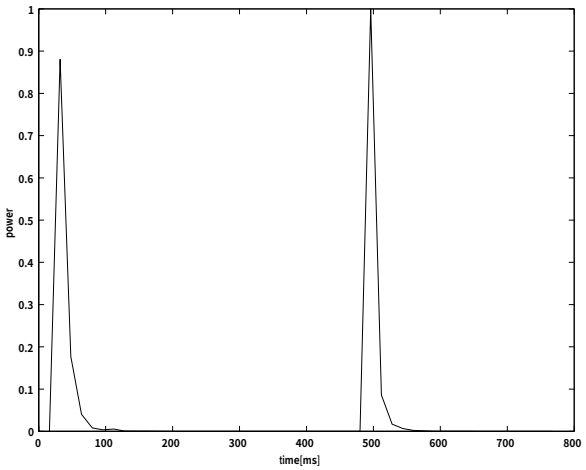


図 1: 環境音の長時間パワーパターン

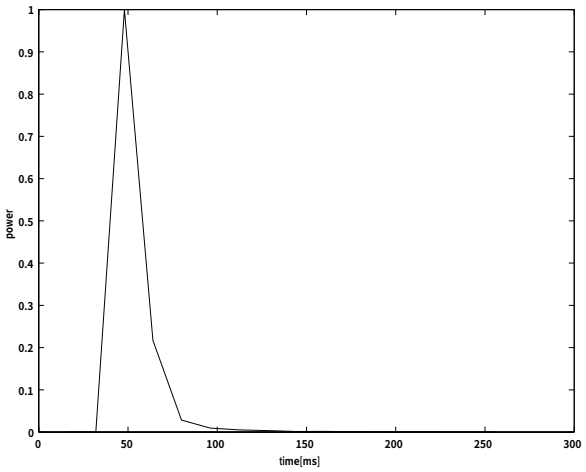


図 2: 環境音の短時間パワーパターン

階層 NN により、予め繰り返し音であるかどうかを判断し、そして次の層で前の層の結果に基づいて、音の種類を判断するという方法を採用した（図 3）。この場合、長時間範囲のパワーパターンからは音の種類を単発衝突音、繰り返し衝突音、連続音の３種類に分類する。また、例えばノックはドアを数回たたくので、短い区間だけで見るとドアに何かがぶつかった音と識別されてしまい、ノックという意味が失われてしまいます。逆にキーボードをタイプする音などは１回でも複数回でも意味は同じなので、こ

の場合はむしろまとめて単音として識別する方が無駄がないといえる。

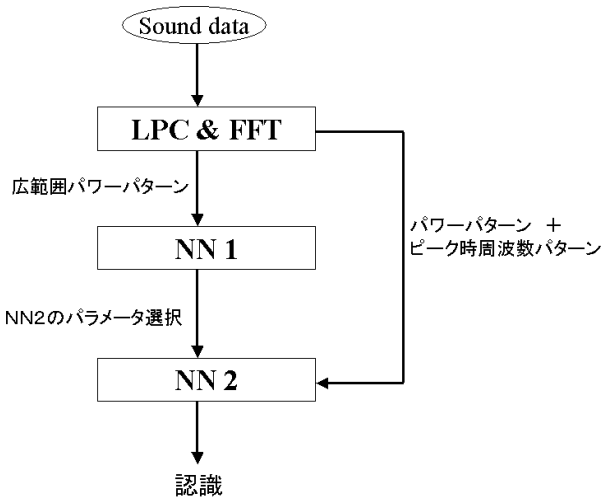


図 3: NN の階層モデル

3　ニューラルネットワークの構成

提案手法で環境音の識別を行うために、階層型ニューラルネットワーク（NN）を使用した。第一層の入力は長い時間範囲のパワーパターンを用いて入力信号を単発衝突音、繰り返し衝突音、連続音の３つのグループに分類する。第二層では分類された音のグループから音の種類識別を行う。認識方法としてパワーパターンとピーク時の周波数パターンを一度に入力できる NN を用いることにした。NN は３層パーセプトロンを用いた（図 4)。入力層の数は 48 で、その内 16 個にパワーパターンを、残りの 32 個には周波数パターンを入力として与える。隠れ層のニューロン数は 24 で、出力層はトレーニングに用いるデータの種類と同じ数のニューロンを用いた。

4　認識実験と結果

実験は第二層の NN の構造を変えて二種類の実験を行った。

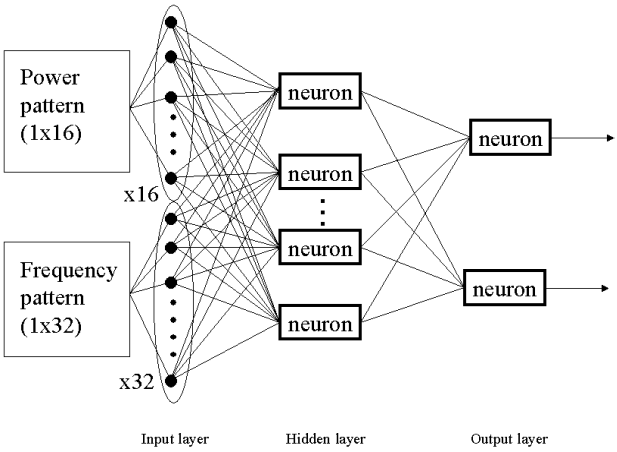


図 4: 第二層に用いる単体 NN モデル

4.1　実験１

実験１では単発衝突音の識別をすべて一つの NN で行う。入力は以下に示す 10 種類でトレーニングに 15 個、テストに 15 個のデータを用いて実験を行った。ただし、第二層での音声データの認識は外した。実験環境は CPU が AMD Athlone1900kHz、メモリが 512M の WindowsPC を使用しプログラミング言語は MATLAB を用い、トレーニングには約 1 時間を要した。表 2 に第一層の認識結果、表 3 に第二層の認識結果を示す。それぞれ平均で 97

表 2: 実験１結果 第一層認識率

入力	認識率	入力	認識率
ボール	100%	キーボード１	100%
金属音	100%	キーボード２	80%
ドア	100%	スイッチ	100%
施錠	100%	ノック	95%
音声	99%	着信音	100%

表 3: 実験１結果 第二層認識率

入力	認識率	入力	認識率
ボール	40%	キーボード１	55%
金属音	95%	キーボード２	100%
ドア	85%	スイッチ	25%
施錠	60%	ノック	95%
音声	—%	着信音	100%

4.2 実験２

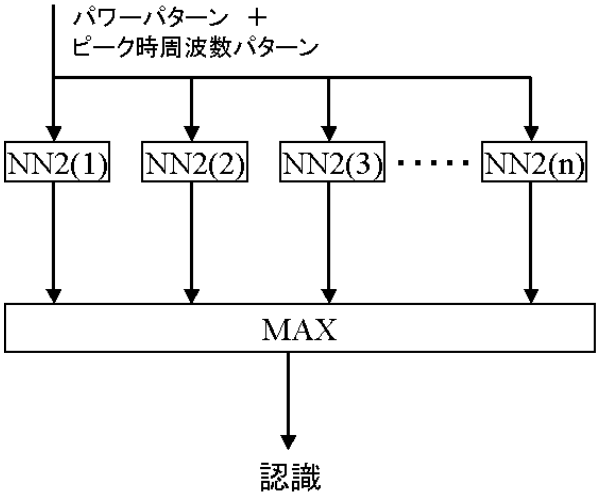


図 5: 第二層 NN の構成

実験２は実験１と同様に二層構造の NN を使用するが、それぞれの種類の音に対して１個ずつの専用 NN を用意する。例えば入力信号が金属衝撃音かどうかと、一種類の音に対する識別を行う NN を並列に組み合わせたものを使用する。結果は最も強い出力を返したものの出力とする。入力は以下に示す 10 種類でトレーニングに 15 個、テストに 15 個のデータを用いて実験を行った。実験２では音声 (母音５種類）が入力された場合は全て音声と識別させたためトレーニングに用いた入力は各母音ごとに 15 個ずつ 75 個用いた。実験環境は実験１と同様です。トレーニングには約 1 時間半を要した。

表 4: 実験結果２ 第二層

入力	認識率	入力	認識率
ボール	73%	キーボード１	60%
金属音	100%	キーボード２	100%
ドア	100%	スイッチ	65%
施錠	45%	ノック	95%
音声	88%	着信音	100

実験２は結果１よりも平均認識率が改善され、平均 83 そして一つの音の対して一つの NN を用いるために新たな音の追加でも他の音の認識率に影響することなく柔軟に対処出来るなどのメリットがあると言える。

5 まとめ

本研究では音の時間パターンと周波数特性合わせた 1 次元のデータを用いてニューラルネットワークによる環境音認識の手法の提案を行い、平均認識率は約 83 本手法は時間遅延ニューラルネットワークによる方法や HMM による方法よりも処理時間は速く、ロボット本体に搭載する目的には適していると言える。しかし、環境音認識は範囲が無限に広いことから、これからもデータベースの作成やそれぞれの方法の比較検討が必要と考える。

参考文献

[1] J. Huang, N. Ohnishi, and N. Sugie. Building ears for robots: Sound localization and separation. *Artificial Life and Robotics*, 1(4):157–163, 1997.

[2] S. Nakamura, K. Hiyane, F. Asano, and T. Endo. Sound scene data collection in real acoustical environments. *J. Acoust. Soc. Jpn*, 20(3):225–231, 1999.

[3] 比屋根一雄 and 飯尾淳. マイクロホンアレーを用いた非音声認識. RWC NEWS, 2000. vol.17.

[4] 三木 一浩, 西浦 敬信, 中村 哲, and 鹿野 清宏. Hmm による環境音識別の評価. 日本音響学会 2000 年春季大会論文集, 2000. 1-8-8.

[5] K. J. Lang, A. H. Waibel, and G. E. Hinton. A time-delay neural network architecture for isolated world recognition. *Neural Networks*, 3:23–43, 1990.

[6] 今井 聖. **音声認識**. 共立出版株式会社, 1995.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0216-10

ETSI AURORA プロジェクトの動向と雑音下音声認識評価ワーキンググループの活動報告

Progress Report for Current Status of ETSI AURORA Project and SLP Working Group for Noisy Speech Recognition

中村 哲 (ATR-SLT)[1]　西浦敬信 (和歌山大学)[2]　武田一哉 (名古屋大学)[3]　黒岩眞吾 (徳島大学)[4]
Satoshi NAKAMURA　Takanobu NISHIURA　Kazuya TAKEDA　Shingo KUROIWA

山田武志 (筑波大学)[5]　北岡教英 (豊橋技科大学)[6]　山本一公 (信州大学)[7]
Takeshi YAMADA　Norihide KITAOKA　Kazumasa YAMAMOTO

藤本雅清 (龍谷大学)[8]　水町光徳 (ATR-SLT)[1]
Masakiyo FUJIMOTO　Mitsunori MIZUMACHI

Abstract

This paper reports current status of the IPSJ-SLP working group (IPSJ: Information Processing Society of Japan, SLP: Spoken Language Processing) establised in October 2001 on the noisy speech recognition and current Europian ETSI AURORA evaluation projects. This working group aims to develop standards, common corpus, and noisy speech recognition system in conjunction with Europian ETSI AURORA evaluation projects.

1　はじめに

音声認識においては，隠れマルコフモデルと確率言語モデル，および大量の学習データの収集により，学習データと同一の性質のテストデータに対しては，非常に高い認識性能が得られるようになった．しかしながら，実際に認識装置を利用する状況での音声は，種々のバリエーションが生じて，学習データとは異なったものが生じてくる．このバリエーションとしては，大きなものとして加法性，乗法性雑音の混入，発話スタイルの変化がある．しかし，学習データとして集められるデータの量は，限られているためこのような場合，十分な性能を獲得することは難しい．特に，本稿では雑音の混入による音声の認識性能の劣化について述べる．

本稿では，2001 年 10 月に情報処理学会 音声言語情報処理研究会内に設立した雑音下音声認識の評価に関するワーキンググループの活動状況の報告を行う．このワーキンググループは，特に EUROSPEECH2001 における欧州の AURORA セッションに刺激を受けて発足したものである．この欧州の AURORA プロジェクトというのは，携帯電話などとネットワークを利用した場合の音声認識サービスを実現する際に，音声認識の前処理部を標準化しようという試みであり，標準化においては，十分利用環境において高い性能を期待できる音声認識前処理が必要ということである．このネットワークで利用する音声認識は分散型音声認識 (DSR:Distributed Speech Recognition) と呼ばれている．この標準化は実際には，ETSI の DSR 標準化に加わっている主として企業が具体的な要求基準，技術開発，評価を進めているが，それと平行して評価データを ELRA(European Language Resources Association) から一般の研究者に配布し，主に ISCA の EUROSPEECH,

[1] ATR 音声言語コミュニケーション研究所 (ATR Spoken Language Translation Research Laboratories, Kyoto), {satoshi.nakamura, mitsunori.mizumachi}@atr.co.jp
[2] 和歌山大学 (Wakayama University, Wakayama), nishiura@sys.wakayama-u.ac.jp
[3] 名古屋大学 (Nagoya University, Aichi), takeda@nuee.nagoya-u.ac.jp
[4] 徳島大学 (Tokushima University), kuroiwa@is.tokushima-u.ac.jp
[5] 筑波大学 (University of Tsukuba, Ibaraki), takeshi@is.tsukuba.ac.jp
[6] 豊橋技科大学 (Toyohashi University of Technology, Aichi), kitaoka@slp.ics.tut.ac.jp
[7] 信州大学 (Shinshu University, Nagano), kyama@sp.shinshu-u.ac.jp
[8] 龍谷大学 (Ryukoku University, Shiga), masa@arikilab.elec.ryukoku.ac.jp

ICSLP において評価結果を発表する AURORA スペシャルセッションが開催されている．本ワーキンググループでは，この AURORA スペシャルセッションに関連して，日本において，並列に同様の雑音下音声認識の評価のためのデータ収集，評価法の検討，性能評価のための仕組みの検討，AURORA スペシャルセッションへのコミットメント，日本からの参加者のプロモーションなどを活動の趣旨としている．

以下，第 2 章でワーキンググループ提案概要，第 3 章で AURORA スペシャルセッションの評価の方法，内容などを述べ，第 4 章でワーキンググループの活動経過とその内容を述べる．

2 ワーキンググループ提案概要

EUROSPEECH2001 において，AURORA2 という Special Session が ETSI，ISCA 共同で企画され，雑音下音声認識のための性能比較が行われた．比較対象となっているのは，音声認識の前処理の信号処理部分である．ある程度の種類の雑音と符号化方式に頑健な前処理方式が確立できれば，携帯電話などの分散型の音声認識の特徴抽出方式として標準化される可能性もある．実際，ETSI は ITU の配下の組織であり，学会での評価活動と平行して実際に標準化の活動も進められていると聞いている．本 WG 提案は，このような活動を日本でも並行して進め，問題を明確化しさらに必要な研究のフォーカスを定める，日本語での評価を行って言語依存性を調査する，研究レベルの技術の利用可能性を探る，日本の技術レベルを積極的にアピールすることを目的としている．

2.1 雑音下音声認識評価ワーキンググループ構成員

現在，本ワーキンググループは下記メンバーにより構成されている．今後，企業の研究者に対しても積極的に本ワーキンググループへの参加を呼びかける予定である．

- **主査：**
  中村　哲　（ATR 音声言語コミュニケーション研）

- **コアメンバー：**
  武田一哉　（名古屋大学）
  黒岩眞吾　（徳島大学）
  山田武志　（筑波大学）
  北岡教英　（豊橋技術科学大学）
  山本一公　（信州大学）
  西浦敬信　（和歌山大学）
  水町光徳　（ATR 音声言語コミュニケーション研）
  藤本雅清　（龍谷大学）

- **アドバイザリーメンバー：**
  古井貞熙　（東京工業大学）
  松本　弘　（信州大学）
  新田恒雄　（豊橋技術科学大学）
  鹿野清宏　（奈良先端科学技術大学院大学）
  有木康雄　（龍谷大学）
  中川聖一　（豊橋技術科学大学）
  小林哲則　（早稲田大学）

3 AURORA スペシャルセッション

この欧州 AURORA プロジェクト[1]は，ETSI の DSR (Distributed Speech Recognition) の標準化活動[2]に同期して進められているもので，雑音に強い音声認識の前処理を開発し，標準化することを目的としている．米国で行われている DARPA の SPINE プロジェクトなどに比べて，雑音処理のみに焦点を当てるため比較的小さな TI-DIGITS の数字認識をタスクとしている[3].

3.1 AURORA2 データベースの入手方法

AURORA2 の音声データが TIDIGITS であるため，TIDIGITS のライセンスを LDC から入手する．実際には，TIDIGITS を LDC から入手するのと同様である

1. 下記サイトを基に LDC ライセンスを入手.
   http://www.ldc.upenn.edu/Catalog/LDC93S10.html

   93 年度会員の方は，ライセンスの依頼のメールを LDC に送るだけ利用可能である．また会員でない方は，$250 を支払えば利用可能である．
2. 下記サイトを基に ELRA から配布されている AURORA2 のデータを入手.
   http://www.icp.grenet.fr/ELRA/aurora2.html

   250 ユーロを支払えば，入手できる．AURORA2 のデータ CD の中に音響モデルの学習等に関する HTK スクリプトが含まれている．

3.2 AURORA2 タスク

2001 年の EUROSPEECH で扱われたものが英語の TIMIT 連続数字データ (LDC から公開済み) を対象としたタスクである．

3.2.1 配布物

- Training Set
  - Clean-condition training:
    8440 発話の Clean（雑音重畳のない）TIMIT データ
  - Multi-condition training:
    上記データに種々の雑音を種々の SNR で混入させ

Multicondition training, multicondition testing

	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.59	98.67	98.57	98.83	98.67	98.59	98.67	98.57	98.83	98.67	98.62	98.67	98.65	**98.66**
20 dB	97.82	97.94	98.24	97.47	97.87	97.73	97.61	97.61	97.66	97.65	97.67	97.58	97.63	**97.73**
15 dB	96.65	97.43	97.70	96.88	97.17	96.16	96.67	96.60	96.27	96.43	96.50	96.31	96.41	**96.72**
10 dB	94.38	95.47	96.18	94.11	95.04	92.94	94.86	93.71	93.92	93.86	93.83	93.92	93.88	**94.33**
5 dB	89.01	88.21	87.53	87.60	88.09	85.05	86.58	87.53	85.16	86.08	83.11	84.16	83.64	**86.39**
0 dB	67.85	63.18	54.10	63.71	62.21	60.88	63.06	66.27	58.07	62.07	46.21	56.35	51.28	**59.97**
-5dB	26.56	27.33	20.22	23.63	24.44	27.11	27.66	29.91	21.75	26.61	19.22	24.73	21.98	**24.81**
Average	89.14	88.45	86.75	87.95	**88.07**	86.55	87.76	88.34	86.22	**87.22**	83.46	85.66	**84.56**	**87.03**

Clean training, multicondition testing

	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.89	99.03	99.05	99.26	99.06	98.89	99.03	99.05	99.26	99.06	99.17	99.09	99.13	99.07
20 dB	96.75	90.54	97.08	96.20	95.14	90.14	95.86	89.95	94.79	92.69	93.37	95.13	94.25	93.98
15 dB	91.53	72.19	88.55	90.03	85.58	74.52	88.15	73.84	81.24	79.44	86.03	89.09	87.56	83.52
10 dB	75.53	47.61	63.53	72.29	64.74	51.89	66.05	49.27	55.20	55.60	71.94	75.03	73.49	62.83
5 dB	47.34	22.91	30.75	39.08	35.02	26.80	36.28	24.60	24.96	28.16	50.63	50.57	50.60	35.39
0 dB	22.44	5.53	10.71	14.25	13.23	7.12	17.35	10.50	9.50	11.12	24.53	23.64	24.09	14.56
-5dB	10.65	0.12	6.83	6.85	6.11	0.95	8.62	5.28	6.14	5.25	12.90	11.19	12.05	6.95
Average	66.72	47.76	58.12	62.37	**58.74**	50.09	60.74	49.63	53.14	**53.40**	65.30	66.69	**66.00**	**58.06**

Figure 1: Baseline results for clean- and multi-condition training.

たもの.

* 雑音：
  subway, babble, car, exhibition
* SNR:
  5dB, 10dB, 15dB, 20dB, Clean

- Test Set
  - Set A
    学習データと同一種の雑音
    （subway, babble, car, exhibition）
  - Set B
    学習データと異なる雑音
    （restaurant, street, airport, station）
  - Set C
    音響特性が異なるもの
    （subway, street に MIRS 特性を付与したもの）
  - Test set SNR
    -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, Clean
- HTK スクリプト
  - 分析条件
    * Sampling frequency 8kHz
    * フレーム周期　10msec
    * フレーム長　　25msec
    * Pre-emphasis　0.97
    * Feature: 12MFCC+pow+Δ+(ΔΔ)
  - HMM
    * Unit:
      数字 Whole Word Model 16-states, 3-mixtures
    * 無音モデル (sil):
      3-states，6-mixtures
    * sp モデル:
      sil の第 2 状態と tied
  - 評価モード
    * Clean-condition test:
      Clean 学習データで学習し，テストセット A-C で評価
    * Multi-condition test:
      Multi-condition 学習データで学習し，テストセット A-C で評価
    * 評価は共通の Excel spread sheet で平均の性能，ベースラインからの改善率で評価.

Figure 1 に上記分析条件に於けるベースライン結果をあげておく．参加者はこのベースライン結果に対し前処理や適応化処理を行い性能改善率を競うことになる．全体の性能改善率は，改善率の平均として与えられる．

AURORA は，次に示すように，さらに自動車内で実収録した数字，コマンドデータの認識評価，さらには，雑音下における大語彙連続音声認識評価へと継続，発展していく予定である[4].

3.3 AURORA3 タスク

雑音を加算するのではなく実際に雑音下で収録した音声を用いて評価を行うことを目的としている．SpeechDat-Car プロジェクトで収録した自動車内発話音声を評価に用いる．対象言語もフィンランド語，イタリア語，スペイン語，ドイツ語，デンマーク語を対象にして，言語による差を確認することができる．また，実環境で遭遇する可能性

のある次の 3 つの状況を設定し評価項目としている．

1. Well matched training and testing:
   ハンズフリーマイクロホンで種々のスピードで走行した場合の音声で学習，テストを行う．
2. Moderate mismatch training and testing:
   中程度のミスマッチがある条件，例えば，ハンズフリーマイクロホンを用いて低速走行中音声で学習し，高速走行中音声でテストを行う．
3. High mismatch training and testing:
   非常に学習とテストが異なる条件，例えば，近接マイクロホンで収録した音声で学習し，種々の速度で走行中の音声をハンズフリーマイクロホンで収録したものでテストを行う．

3.4 AURORA4 タスク

Noisy WSJ-large vocabulary evaluation: Wall Street Journal データベースに雑音とフィルタ特性を付加した大語彙データベースを用いて評価を行う[5]．評価条件は，次の 4 つの平均性能である．

- 8kHz clean training
- 8kHz multi-condition training
- 16kHz clean training
- 16kHz multi-condition training

この評価のためのスクリプトはミシシッピ大学により準備されつつある[6]．

4 WG の活動内容

WG 設立以来，AURORA 性能評価に参加するグループの啓蒙，雑音下音声認識を評価するための評価法，データ設計，スクリプト，それらの短期，長期の計画などについて，下記に示す 7 回の会合を通して議論を進めてきた．

1. 2001 年 12 月 20 日　東京工業大学
2. 2002 年　3 月　8 日　名古屋大学
3. 2002 年　3 月 19 日　神奈川大学付近
4. 2002 年　4 月 26 日　機械振興会館
5. 2002 年　7 月 19 日　ATR 音声言語コミュニケーション研究所
6. 2002 年　9 月 26 日　秋田大学付近
7. 2002 年 11 月 19 日　名古屋大学

これまでの活動のまとめを以下に述べる．また最新情報などは[7]を参照のこと．

4.1 ICSLP2002 AURORA セッション

ICSLP の AURORA セッションについては，Organizing Committee にメンバーの一人が入り，情報の流れをよくしたほか，これまでの論文のサーベイなどもWGのメンバーで行い，調査活動も共同で行った．ICSLP2002 の AURORA セッションは，主として AURORA3 の自動車内音声データの評価を中心に行うこととなったが，WGとしては AURORA2 の評価結果を投稿することとした．AURORA2 関連で約 3 件[8, 9, 10]，AURORA3 に 1 件[11]の投稿を行い，すべて採録された．下記に，現在 AURORA プロジェクトに参加している海外の主な研究グループを列挙しておく．

1. AT&T
2. UCLA
3. Philips Phicos
4. Alcatel + France Telecom
5. Siemens
6. Maribor Univ.
7. Motorola
8. Nijmegen Univ. + FT + Alcat
9. Granada Univ.
10. Lucent Bell Labs
11. ICSI + OGI + Qualcomm
12. Sheffield Univ.
13. IBM
14. UPC
15. Columbia Univ.

4.2 評価データの設計について

現在の AURORA タスクは，日本語が含まれていない．前処理なので，言語依存性は低いと言われているが，AURORA3 の 4 言語でも相当な認識性能の隔たりがあるので，日本語の評価データを作る必要がある．現在，次のようなデータ収録を計画している．

[**AURORA2J**]
AURORA2 の日本語版である．TIDIGIT の数字を日本語にして，AURORA2 と同一の雑音，伝達特性を付与したものを作成する．

すでに WG では，AURORA2 に含まれるている雑音データを入手した．現在，TIDIGIT の数字を日本語にしたものを収録中である．なお話者数や語彙数も AURORA2 と同様にする予定である．収録が終了しだい雑音と伝達特性の付与作業に取り掛かる予定である．

[**AURORA2.5J**]
AURORA2J の単語を AURORA2 の雑音再生下で発話する．雑音の加算と実際の環境での発話の違いを考察できる．

現在，雑音の加算と実環境での発話の認識性能の違いなどについて調査を進めている状況である．

[**AURORA3J**]

AURORA3 と同様に自動車内発話の音声を収録する．発話内容は，孤立単語，バランス文を収録する．現在，自動車内発話の音声は名古屋大学にて収録中である．

[**AURORA4J**]

大語彙の雑音下音声認識を中心としたタスクとするか，よりアプリケーションに近い環境での発話にするか，単語でなく自由発話にするかなど現在検討中である．

[**雑音 DB**]

これまで，電子協の雑音 DB[12], Noisex92[13], RWC-DB[14]などが収録されてきたが，未だ不十分であり，種々多様な評価を行う際に基準となるような雑音データベースが必要である．そこで現在，本 WG では新たな雑音 DB の構築に向けて現在検討中である．

4.3 評価ツールの設計

本WGでは，AURORA と同様にデータの配布とともに評価のため，HTK のスクリプトと評価のための Spread Sheet を配布する．また，より一般性のある評価基準を確立することを目指して，種々の共通ツールの作成を目指している．ツールには，SNR 測定プログラムなどが含まれる．

4.4 アンケート

雑音 DB，AURORA4J に関連して，雑音環境として，どのような環境の雑音，残響を考慮すべきかは，非常に重要でかつ設定するのが困難な課題である．WG としては，実際の音声認識装置の開発者，利用者，利用想定者にアンケートを行い，雑音環境の洗い出しを行いたいと考えている．

5 今後の計画とまとめ

雑音下音声認識評価に関する SLP-WG の活動の内容と現状および ETSI AURORA プロジェクトについて報告を行った．音声認識の雑音環境に於ける頑健性の問題は，今日では非常に重要な課題である．WG 設立以来，AURORA の評価プロジェクトと平行して活動を進めてきた．今後，AURORA については，ICSLP のセッションに参加し情報交換をするとともに，日本語データの収録が終了した際には，AURORA の評価 DB に日本語 DB を公開することなどを計画している．また，上述の日本語のデータ収集計画の議論を継続するとともに，評価手法などの検討を進めていく．WG としては，本活動が雑音下音声認識に於ける評価の枠組みの確立の一助になればと考えている．

[謝辞] 本研究の一部は，通信・放送機構の研究委託により実施したものである．

参考文献

[1] http://eurospeech2001.org/ese/NoiseRobust/index.html, http://www.elda.fr/proj/aurora1.html, http://www.elda.fr/proj/aurora2.html

[2] ETSI standard document, “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.2 (2000-04), 2000

[3] H.G.Hirsh, D.Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions”, ISCA ITRW ASR2000, september, 2000

[4] D.Pearce, “Developing the ETSI AURORA advanced distributed speech recognition front-end & What next”, Proc. EUROSPEECH2001, 2001

[5] Aurora document no. AU/337/01, “Experimental framework for the performance evaluation of speech recognition fron-ends on a large vocabulary task: Version 1.0”, Ericsson, June 2001

[6] Aurora document no. AU/345/01, “Large vocabulary evaluation of front-ends- baseline recognition system description”, Mississippi State University, Aug 2001

[7] http://www.slt.atr.co.jp/〜nakamura/SLPWG2001/ Noise-WG.html

[8] Masaki Ida, and Satoshi Nakamura, “HMM Composition-Based Rapid Model Adaptation Using a Priori Noise GMM Adaptation Evaluation on Aurora2 Corpus,” Proc. ICSLP2002, pp. 437–440, Sept. 2002.

[9] Masakiyo Fujimoto, Yasuo Ariki, “Evaluation of Noisy Speech Recognition Based on Noise Reduction and Acoustic Model Adaptation on the Aurora2 Tasks,” Proc. ICSLP2002, pp. 465–468, Sept. 2002.

[10] Norihide Kitaoka, Seiichi Nakagawa, "Evaluation of Spectral Subtraction with Smoothing of Time Direction on the Aurora 2 Task," Proc. ICSLP2002, pp. 477–480, Sept. 2002.

[11] Kaisheng Yao, Dong-Lai Zhu, Satoshi Nakamura, "Evaluation of a Noise Adaptive Speech Recognition System on the Aurora 3 Database," Proc. ICSLP2002, pp. 457–460, Sept. 2002.

[12] http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html

[13] http://www.speech.cs.cmu.edu/comp.speech/ Section1/Data/noisex.html

[14] http://tosa.mri.co.jp/sounddb/index.htm

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

ＡＩチャレンジ研究会
主 査
奥乃 博
京都大学 大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
工学部 10 号館
075-753-5376 Fax: 075-753-5977
okuno@i.kyoto-u.ac.jp

幹 事
浅田 稔
大阪大学大学院 工学研究科
知能・機能創成工学専攻
asada@robotics.ccm.eng.osaka-u.ac.jp
武田 英明
国立情報学研究所 知能システム研究系
takeda@nii.ac.jp
樋口 哲也
独立行政法人 産業技術総合研究所
t.higuchi@aist.go.jp, suzuki.a@aist.go.jp
田所 諭
神戸大学 工学部 情報知能工学科
tadokoro@octopus.cs.kobe-u.ac.jp

Executive Committee
Chair
Hiroshi G. Okuno
Dept. of Intelligence Science and Technology,
Gradulate School of Informatics
Kyoto University
Sakyo, Kyoto 606-8501 JAPAN
Secretary
Minoru Asada
Dept. of Information and Intelligent Engineering
Graduate School of Engineering
Osaka University
Hideaki Takeda
National Institute of Informatics
Tetsuya Higuchi
National Institute of Advanced Industrial Science and Technology
Satoshi Tadokoro
Dept. of Information and Intelligent Engineering
Kobe University

SIG-AI-Challenges home page (WWW): http://www.symbio.jst.go.jp/SIG-Challenge/